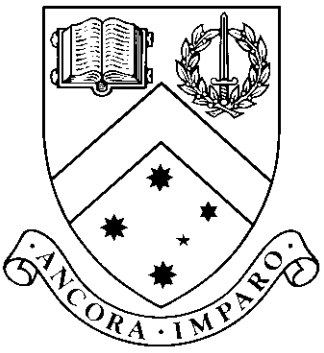
Bringing Order to User-Generated Content on the Web: Estimating User Expertise and Information Quality

by

Wern Han Lim



Thesis

Submitted by Wern Han Lim

for fulfillment of the Requirements for the Degree of

Doctor of Philosophy (0190)

Supervisor: Dr. Mark James Carman

Associate Supervisor: Dr. Jojo Sze Meng Wong

Associate Supervisor: Dr. Kok Sheik Wong

School of Information Technology
Monash University

May, 2018

Contents

List of Definitions

List of Tables

List of Figures

Bringing Order to User-Generated Content on the Web: Estimating User Expertise and Information Quality

Wern Han Lim
[redacted]
Monash University, 2018

Supervisor: Dr. Mark James Carman
[redacted]

Abstract

This research investigates the modern challenge in managing large amount of information on the web. Unlike the traditional World Wide Web (WWW), the current Web 2.0 ecosystem encourages users to not only consume information but to also create, organize and share information in the form of user generated content (UGC). This change creates the challenge in information management where large amounts of unmoderated data of varying structure, format and quality are being uploaded at an increasing rate. Processing such extensive information, especially with content-based approaches alone, is suboptimal. Thus, this research focuses on the content-agnostic approaches to information management on the WWW, and its applications such as content classification and retrieval.

This research uses graph-based algorithms and pairwise comparison approaches for the estimation of user expertise. This research observes that it is possible to estimate the expertise of users with the proposed algorithms in a content-agnostic manner. Moreover, the information quality of UGC can be inferred from the estimated user expertise. This resulted in various improvements, such as:- (1) better WWW resource classification and descriptions with user annotations on Collaborative Tagging (CT) platforms; (2) generation of new knowledge in the form of user answers on Community Question-Answering (CQA); and (3) supplementing shared content on Content Aggregation (CA) platforms through social curation with meaningful user discussions. The identified experts on various UGC platforms can then be motivated and rewarded for their contributions; whereas unwanted malicious users can be penalised and filtered from these platforms.

Bringing Order to User-Generated Content on the Web: Estimating User Expertise and Information Quality

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Wern Han Lim
May 19, 2018

Acknowledgments

The completion of this thesis would not be possible without the support of those around me. Thank you!

Firstly, I would like to express my sincerest gratitude to Dr Mark James Carman. You have been more than just my supervisor for this thesis – A great teacher; Mentor; Role Model. It is through such guidance that I can grow and produce research that I can be proud of today. I could not have hope for anything more. Thank you for all the engaging discussions. Thank you for highlighting my weaknesses and helping to overcome it. Thank you for your invaluable contributions to this thesis!

I need to extend my deepest appreciation to Dr Jojo Sze-Meng Wong who has been there as both my supervisor and as a friend. Thank you for believing in me; in providing support throughout this journey – Your kind words were the extra push needed. Your attention to detail had helped me realise so much throughout my studies. Thank you!

Also, I would need to thank Dr Kok Sheik Wong who had helped immensely in the time of need and in providing for the research. Thank you for the time and effort throughout in ensuring that this thesis can be completed without any hiccups.

Not to forget, I am grateful to the other supervisors who have imparted their wisdom throughout my journey. Thank you Dr Saadat M. Alhashmi, Dr Eu-Gene Siew and Dr Eng Keong Lua.

A shoutout to the Monash University School of Information Technology staff. Thank you for providing the resources and facilities to make this journey possible. A special thank you to Dr Anuja Dharmaratne, Miss Misha Supremaniam, Miss Wan Nurul and the late Miss Siriyaten Md Ali for all the contributions throughout. Besides that, thank you to my colleagues who have helped with their input, especially Kokum Weeratunga and Kuan Yew Leong.

I cannot end this without expressing the utmost gratitude to my parents (Mr Teong Seng Lim and Madam Ai Wah Tee) who have given me so much throughout this journey from the start. No value can be placed on all the moral and financial support given. My sincerest thank you to both of you for always being there.

Finally, a special thank you to Germaine Ooi for being understanding and caring throughout this journey. The cat, dog and bear videos helped.

Wern Han Lim

Monash University
May 2018

Chapter 1

Introduction

The World Wide Web (WWW) is a large information space contributed by content creators for the consumption of users. It provides a crucial source for everyday information [98]. The content creators on the WWW are often regarded as expert authors or editors in their respective domains. Users navigate through the WWW to retrieve and consume information, often assisted by Information Retrieval (IR) tools such as web search.

Since its introduction however, the WWW has evolved throughout the years with the modern WWW seeing a great change in its users. The users of the modern WWW are no longer just consumers of content. Instead, the very same users are also now the creators, contributors, organisers and sharers of content on the WWW. This blurring of user roles between content consumers and expert contributors [4] can be attributed to the growth of User-Generated Content (UGC) platforms.

Definition 1 (User-Generated Content, UGC)**.** Content of any form created by the users of a service, made public for the consumption of other users on that same service platform.

The influx of contributors from the popularity of UGC platforms complicates the information management and content curation on the WWW. Contributors on UGC platforms are diverse with varying levels of expertise and reliability [155]. Thus, the generated content is of varying information quality [111]. While it is possible to discard this user generated information, leaving only author or editorial content on the web; it is hard to ignore the information potential of UGC [132]. In fact, it is a dominant source of content on the WWW today [50].

In this thesis, we explore several UGC platforms within the following theme:- What are the features of UGC platforms that can be leveraged for the estimation of user expertise? The estimated user expertise is then used to infer information quality of generated content in order to improve content curation on these platforms, knowledge creation and retrieval. Findings from this research contribute towards the thesis goal of bringing order to the user generated content on the WWW.

1.0.1 Chapter Organisation

The rest of the chapter is organised as follows. First, we briefly introduce and discuss the categorisation of UGC platforms on the WWW in Section 1.1; selecting three categories for exploration. In Section 1.2, we present an overview for the information potential of UGC that motivates our research and highlight the challenges faced in Section 1.3. Section 1.4 outlines the research questions to address the challenges and the resulting research contribution. Finally, we present the overall structure for the thesis in Section 1.5 .

1.1 UGC Platforms

The emphasis of UGC in the modern WWW has resulted in the WWW being known as the Web 2.0. Today, UGC platforms are aplenty on the WWW and can be categorised into (but not limited to) the following categories:

- **Social Networking Services (SNS).**
  An SNS on the WWW acts as a social platform for users to connect with other users – often of similar interests, of the same demography or who have real-life connections. SNSs enable users to create a profile and then share content or interact with the other connected users [19]. Example of popular SNS of today include Facebook[1] and Twitter[2].

- **Weblog (Blog).**
  Weblogs are channels for users to communicate their generated content with their intended audience [100] without the hassle of managing their own website. These platforms include Blogger[3] and Wordpress[4].

- **Collaborative Tagging (CT).**
  CT provides a platform for the users to store and organise WWW resources by attaching short text or annotations. This collaborative organisation process results in a categorisation system known as Folksonomies [133] with websites such as Delicious[5], Pinterest[6] and Flickr[7].

- **Community Question-Answering (CQA).**
  CQA websites provide a platform for users to describe their information needs with natural language questions instead of keyword-based queries. Community members of the platform can then try to meet the information needs of the questioner by contributing answers [95]. Some of the largest CQA platforms are Quora[8] and StackOverflow[9].

- **Discussion Boards and Forums.**
  Discussion boards and forums are mediums on the WWW for the users to express their opinions and hold discussions. The messages are posted publicly on the conversation threads [141]. An example of a popular discussion board is 4chan[10].

- **Content Aggregation (CA).**
  CA platforms provide a one-stop-fits-all destination where content from the WWW is identified, shared and curated for the consumption of users [89, 134]. The largest content aggregation website of today is Reddit[11].

This thesis investigates three UGC platform categories:- (1) collaborative tagging in Part I; (2) community question-answering in Part II; and (3) content aggregation in Part III. The selected platforms are diverse and distinct, providing us with a good coverage of UGC platforms as illustrated in Figure 1.1. The research aims to better understand the

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://www.blogger.com/
[4] https://wordpress.com/
[5] https://del.icio.us/
[6] https://www.pinterest.com/
[7] https://www.flickr.com/
[8] https://www.quora.com/
[9] https://stackoverflow.com/
[10] http://www.4chan.org/
[11] https://www.reddit.com/

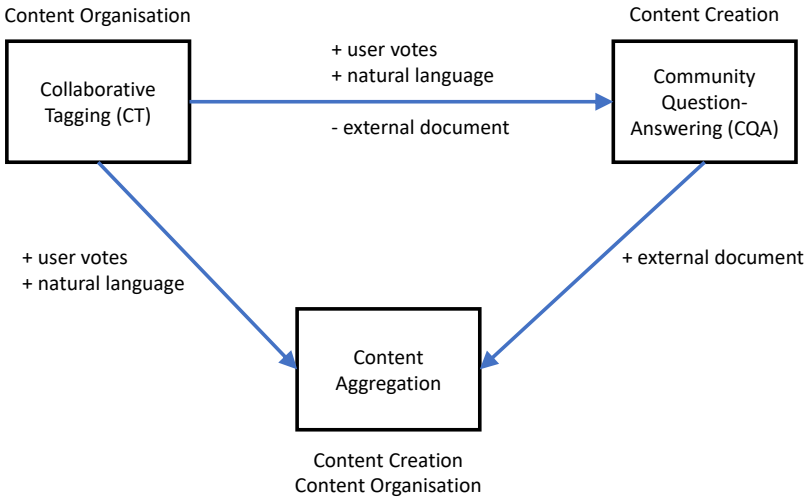


Figure 1.1: The selected UGC platform categories for the research and their distinct differences.

content information potential and the challenges faced in content management on those platforms.

For each part of the research for each UGC platform category, we propose and validate several possible enhancements for content management. We show that the improvements in content management on UGC platforms benefits information retrieval tasks within UGC platforms (and beyond) through improved:

- Content description
- Content categorisation and classification
- Content retrieval
- Content creation
- Expert search

Definition 2 (Information Retrieval, IR)**.** Finding material of an unstructured nature that satisfies an information need from within large collections [98].

1.2 Information Potential of User Generated Content

Users of UGC platforms are the core contributors of content [50] on the WWW, providing far more contents than expert authors and editors. The emergence of UGC platforms can be attributed to the willingness of users to share their knowledge and experience for the learning and/ or benefit of others [159]; especially with the ease of contributing contents on such platforms. This turns UGC platforms into rich repositories of knowledge [26].

The contents generated by these users have the information potential to enhance other information systems [105] despite the users themselves not being expert content creators. Studies have found that the contents generated by normal users have a high overlap with the contents of expert creators [17, 60]. Furthermore, users are able to describe [8], review and summarise large documents such as web pages [13] or organise these contents [151].

While it is true that the users of UGC platforms are not at the expertise level of expert authors or editors, the consensus opinion of the community on these platforms is

able to rival that of experts [132] or supplement them [111]. This collective effort is often known as the Wisdom of the Crowd (WotC) and is often employed in UGC platforms for the moderation of content. As the users of UGC platforms heavily outnumber the expert authors and editors, the WotC is able to achieve a good coverage of the WWW while keeping up with the latest relevant content [43].

Definition 3 (Wisdom of the Crowd, WotC)**.** The aggregated knowledge from the collective effort of a group of users in a community rather than that of a single expert [135].

In fact, users of UGC platforms have gained influence [12] and recognition for their contributions that rival those of experts. Today in the interconnected WWW, the users of UGC platforms and their content spreads beyond their home platforms [49, 145]

1.3 Challenges of User-Generated Content

UGC platforms allow users to easily create and share content. Studies have shown however that there is a large variance in content quality on UGC platforms [6, 68], with quality ranging from meaningful information that is valuable knowledge, to unwanted content such as spam. This variance can be attributed to the users having varying expertise and reliability [93].

Thus, there is a need for the moderation of content on UGC platform in order to ensure a healthy density of good content [124] in order to avoid noisy data [155] that could taint the integrity of UGC platforms [109]. The ease of accessibility however creates a large constant flow of diverse content, structure and quality onto the WWW [50]. The traditional approach to content curation by human moderators is no longer feasible and efficient to handle this high volume of content.

In this following subsections, we outline the main challenges to content curation on UGC platforms as our research problem before we delve deeper into the three selected platforms. These challenges motivate our research to improve content management on UGC platforms by overcoming these challenges.

1.3.1 Varying Information Quality of Content

The content generated by users of UGC platforms varies greatly in quality – with some contributions being better quality than the other [111] for the same topic or on the same issue. Content of high information quality is valuable for the consumption of other users; whereas content of low information quality can be distracting noise. Thus, there is a need to be able to estimate the information quality of user generated content in order to recognise good content and filter out bad content.

Definition 4 (Information Quality)**.** A measure of the significance, completeness, validity, accuracy, consistency, availability and timeliness of content [121] on information systems.

The services of expert editors and moderators are required for the curation of content. The number of users, however, far outnumbers the available experts especially on larger platforms. Thus, many UGC platforms are self-moderated (aka peer-moderated) by the community leveraging the WotC with tools such as votes, flags, reports and ratings. This research argues that this approach to moderation is still in its infancy with room for improvements due to the following:

- Not every member of the community is reliable and UGC platforms are unable profile the users [152]. Judgement from unreliable users or users of low expertise could negatively affect the collective judgement of the community especially on platforms which weights all users the same. Noise from the unreliable contributions would often muffle the good contributions from reliable users.

- The intentions of UGC platform users are unclear. There exist malicious users such as spammers who can deliberately disrupt the effort of the community in moderating the platform [109, 155].

This research tries to infer and possibly predict the information quality of user-generated content on UGC platforms. Good content is boosted or weighted up for retrieval and consumption by the users; whereas bad content is filtered out from the platform or identified to prevent the spread of false information.

This research aims to estimate user expertise based on the features of UGC platform. It is then possible to leverage the user expertise to weight user moderation differently according to users' reliability to improve the peer-moderation strategy on UGC platforms.

1.3.2 Unstructured Content

This research argues that there is a need to acknowledge the highly varied and unstructured nature of content on UGC platforms that complicates processing for information systems and applications. These variations include:

- Diversity in writing style with different vocabulary of terms. This could include memes, abbreviations and colloquialisms.
- Different formats of content on the WWW such as textual, hyperlinks (reference to other resources) and multimedia content.
- Presentation styles of content unlike conventional documents. For example, discussions between users on UGC platforms can be threaded.
- Dynamic content which changes over time with subsequent user contributions.

To account for this, we focus on content-agnostic approaches for the estimation of information quality for user-generated content. This research investigates readily available features (such as temporal information, user interaction history and user votes) found on UGC platforms to develop of the content-agnostic approaches. Our findings from Part I to III of the thesis found success in achieving good estimation performance while being efficient without the need for expensive content processing.

1.3.3 Lack of Content Metadata, Description and Classification

The large amount of content on UG platforms calls for an improved organisation for retrieval that is both effective and efficient. The diversity of user generated content (particularly the content format which ranges from text to hypertext and multimedia) challenges current automated approaches of content classification. Often, this content comes without accompanying metadata for organisation. On the other hand, traditional classification approaches using manual ontologies defined by experts are unable to keep up with the rapid growth of new data such as new interest domains (for example Reddit's AskMeAnything[12], FunnyGifs[13], AwwwPhotos[14]).

Fortunately, the hypertext nature of the WWW does bring together the content. User annotations such as tags on collaborative tagging platforms enable us to describe and classify other content. Such possibility motivates us to investigate the information potential and the reliability of annotations from CT platforms in Part I of the thesis.

[12]https://www.reddit.com/r/AMA/
[13]https://www.reddit.com/r/funnygifs
[14]https://www.reddit.com/r/aww/

1.4 Research Questions

This research is motivated by the information potential of user-generated content (Section 1.2) and acknowledges the challenges present to leverage and organise these content as a research problem (section 1.3). Thus, it is the objective of the research to bring order to user-generated content on the WWW. We identified several main research questions that helped guide our research towards the objective.

1.4.1 Can Information Quality be Estimated through Content-Agnostic Means?

There have been numerous attempts by researches in the past to estimate the information quality of user-generated content using content-based approaches [153, 157, 159] such as classification and sentiment analysis. Content features such as the length of the content, sentence structure (grammar) and typing mistakes provide indications of information quality as well. We argue that analysing and processing UGC content is expensive due to the amount of unstructured content on UGC platforms.

Therefore, the first task of the research is to explore the possibility of estimating information quality through content-agnostic approaches:- Do the features on UGC platforms contain sufficient signals for the estimation of information quality of content without looking at the content itself? Due to the diversity of UGC platforms, this would need to be a robust solution as well (discussed further in Section 1.4.6)

Upon answering this research question, the estimated information quality of content can be applied to enhance UGC platforms such as:

- Re-organising content where greater importance is given to content that is of higher information quality. This would improve retrieval performance by promoting high quality content that meets the information need of the users better.

- Filtering out noisy content that is of low information quality. This would allow the UGC platforms and their information systems to scale better with a high density of good content.

1.4.2 Is there a Relation between User Expertise and the Information Quality of their Content?

A challenge discussed earlier is the constantly high rate of content being created and shared on UGC platforms. This challenge contributes to the scalability concerns for information systems to manage, curate and process this content. For our research, we asked the following question – As users are the creators of content, is there a relation between the user expertise and the information quality of their generated content [18]? If so, are we able to infer the information quality of the content produced from the estimated expertise and reliability of the user?

Definition 5 (User Expertise)**.** A measure of a user's capabilities to make significant contribution towards content creation or curation. User expertise is domain or topic sensitive.

This research investigates the relation between user expertise and the information quality of their generated content on the selected three UGC platforms:

- Are expert users able to describe and annotate content on the WWW better? This is studied in Part I in the context of annotation-based retrieval.

- Are expert users able to produce answers that solve the questions of other users better? In Part II, we explore and evaluate the prediction of answer quality inferred from the estimated expertise of the answerers [18].

- Are expert users able to make significant contributions towards the discussion of a shared content or comments by other users? Many of the current UGC platforms facilitate discussions between users for shared contents such as content aggregation platforms discussed in Part III. This research evaluates how the curation of discussions could enhance shared contents.

1.4.3 Can User Expertise and Reliability be Estimated on UGC Platforms?

There is a diverse set of users on UGC platforms – ranging from reliable users (sometimes elected as moderators) and expert users who make good contributions, to active consumers of content and even malicious users who deliberately produce contents with low information quality [109, 155].

Users are valuable entities for UGC platforms, arguably more than the platforms' contents as both creators, consumers and curators of content. Thus, there is a need for UGC platforms to manage their users:

- Recognize and encourage expert users to contribute high quality content on the platform. Having high quality content on the platform attracts more consumers to the platform itself.

- Identify and respond to the threats from malicious users. UGC platforms need to avoid these malicious users from disrupting the other users' experience with the platform which would be detrimental [109]. For example, spammers could reduce the information density of a discussion board by introducing unrelated advertorial threads to the ire of other users or even harass the other users.

- Improve the self-moderation system for the community [135]. Currently, UGC platforms provide tools such as votes, ratings, reviews, reports, flags, dispute and many other tools for peer moderation. In addition, the community (or the platform itself) could elect key reliable users as moderators. The ability to estimate the reliability of each user would also enable the UGC platform community to better moderate itself by acknowledging the judgements of users with high reliability and reducing the impact of unreliable or malicious users (including bots).

1.4.4 Can User-Generated Content Describe and Classify Content Better?

Annotations are often used to describe content. On UGC platforms, users annotate WWW resources such as text, web pages and multimedia to better describe, organise and classify the content [132]; thus providing additional metadata for the content. These user generated annotations are however unstructured, unmoderated and are of varying quality with some annotations being better descriptors for the resources than others [111].

In Part I of this thesis, we attempt to estimate the expertise of users; then use the estimated expertise to infer the information quality of user annotations. The improved content descriptions can be applied to improve categorisation and classification of complex web resources [111, 132] or can be used for improved retrieval such as annotation-based retrieval [17, 28].

As highlighted in Section 1.3, the unstructured nature of user-generated content motivates the research to estimate user expertise through content-agnostic approaches by not

processing the annotations themselves. Instead, the user expertise would be estimated through signals from features of collaborative tagging platforms.

1.4.5 Can We Predict the Information Quality of User Generated Content?

To the best of our knowledge, many of the current UGC platforms still rely on peer moderation through WotC and moderators to curate content on their platforms as modern information systems still struggle with the unstructured nature of the content. This current approach however faces several challenges:

- A cold start issue: considerable time may be required before the peer moderation system is able to correctly judge and moderate content. This temporal gap is unwanted in UGC platforms – (1) in CQA platforms (Part II), the questioner would need to wait for a considerable time before knowing which is the responding answer (by user votes) that meets the information need; or (2) in A platforms (Part III), the visibility of unwanted content could spiral out of control.
- High amount of data that overwhelms the moderation team. Studies have found that users display positional bias [107, 118] and temporal bias [156] due to the large amount of content being generated on UGC platforms. Consequently, the community is not able to self-moderate the content on UGC platforms effectively.

This research proposed the prediction of information quality for content. The predicted weights can be used to then initially arrange and organise content on UGC platforms for user consumption. For the purpose of this, we evaluate the:

- Prediction for best answers to questions (Part II). Often, the questioner desires a single good answer that meets the information need. Unlike normal retrieval, the required content that meets the users' information need is yet to be created [95]. The received answers can be ranked according to the predicted answer quality for the consumption of the questioner or serve as a warning to the questioner to not trust a possibly harmful answer.
- Prediction for the number of votes a comment will receive in thread discussions (Part III). Unlike a question-answering environment, a discussion would require the aggregation of various user contributions together for a complete narrative.

A by-product from resolving this research question is the profiling of users with their expertise. Here, the proposed approaches could identify domain experts to encourage further contributions while limiting the influence of disruptive and malicious users.

1.4.6 Are the Approaches Robust Across Platforms?

In Section 1.1, we have seen a diverse categorisation of UGC platforms. The platforms differ from one another with their own distinct features, targeted audiences, content, structure and organization. Thus, the proposed approaches to achieving the research objective need to be robust and yet also be able to leverage on the unique key features of each platform to improve performance. This is true for the three selected UGC platforms visualised in Figure 1.1.

The common entity in all of the UGC platforms is the users as both the consumers and the contributors of content. Thus, the estimation of user expertise is the main focus of the research; which is then be used to infer information quality of content. For the selected UGC platforms explored in the research, we evaluate the performance of various estimation approaches proposed on applications such as:

- **CT platforms.**
  User generated annotations for the description and categorisation of available WWW resources. The inferred information quality of annotations are used for annotation-based retrieval.

- **CQA platforms.**
  User generated answers as to solve the problems raised by questioners. The inferred information quality of answers are used to predict the best answer to meet the information need expressed by each question.

- **CA platforms.**
  User generated comments in contributing towards the discussion and the critiquing of shared content. The vast quantity of comments can be organised to enrich the original content for the consumption of other users.

1.5 Thesis Organisation

The thesis is organised into the following three main parts followed by a conclusion in Chapter 9. Each part is dedicated to the investigation of a single UGC platform category to answer the research questions raised. Subsequent parts are often built upon the findings and discoveries from an earlier part. Each part contains between two and three chapters with discussions on:

- Introduction.
- An exploration on structure and representation.
- Information potential of content on the platform.
- Challenges faced on the platform.
- An investigative study on a UGC platform of the chosen category.
- Current approaches for the estimation of user expertise on the platform.
- Proposed approach for the estimation of user expertise.
- Inferring information quality from user expertise.
- Evaluation framework for the proposed approaches.
- Results, findings, discussion and analysis.
- Conclusion.

The main parts of the thesis and their chapters are as summarised in the following subsections.

1.5.1 Part I: Collaborative Tagging (CT)

Part I of the thesis deals with collaborative tagging platforms. On CT platforms, users are able to bookmark, organise and share various WWW resources through the use of annotations. User annotations are short terms, descriptors, keywords or tags attached to WWW resources on the platform [1].

In this part of the thesis, we attempt to improve the retrieval performance for WWW resources through the use of user annotations that they are associated with [60, 99]. Chapter 2 outlines the current state of classification and retrieval with CT platforms; all of which are based-on the content processing of annotations [2, 55, 104, 126]. We instead proposed an alternative – a content-agnostic approach which infers the information quality of user annotations according to the expertise of the annotators. Our approach is discussed and evaluated in Chapter 3. The findings from this part was published in:

Wern Han, Lim and Mark James Carman. Annotator Expertise and Information Quality in Annotation-based Retrieval. In *Proceedings of the 22nd Australasian Document Computing Symposium (ADCS '17)*. ACM, 2017.[15]

1.5.2 Part II: Community Question-Answering (CQA)

In Part II of the thesis, we evaluate the modern information retrieval potential from CQA platforms that allows users to perform IR with natural language questions instead of keywords; thereby allowing them to meet information needs [159] that could not otherwise be satisfied by traditional search-based IR [95].

First, we present an overview of the previous and current work done on CQA platforms in Chapter 4. Answer retrieval are often based on question similarity [26, 153] and expert search [93] to direct suitable answerer. Our study on the Yahoo! Chiebukuro [16] dataset enables us to better understand CQA platforms; enabling us to leverage a new feature in user votes for user expertise estimation.

The contributors of CQA platforms are diverse in expertise and reliability [93], resulting in a high amount of low quality answers [6, 159] (up to 30% [124]). Thus, in Chapter 5, we adapted the graph-based approaches from Part I to CQA platforms for the estimation of user expertise in order to infer answer quality. Our work would improve expert search and answer retrieval as a content agnostic approach that could further supplement the literature. The findings however motivated us to explore a new direction of user expertise estimation which we present in Chapter 6 – a novel pairwise comparison approach with competitive features. Findings from this part of the thesis was published:

Wern Han Lim, Mark James Carman, and Sze-Meng Jojo Wong. Estimating Domain-Specific User Expertise for Answer Retrieval in Community Question-Answering Platforms. In *Proceedings of the 21st Australasian Document Computing Symposium (ADCS '16)*, pages 33-40. ACM, 2016.[17]

1.5.3 Part III: Content Aggregation (CA) and Social Curation

The final part of the thesis investigates into the CA platforms where content is aggregated from the WWW with the additional functionality of discussion for its users. Modern CA platforms such as Reddit allow users to create and contribute their own content for sharing [92, 130]. Contents on such platforms are usually organised according to popularity [91] and peer recommendation [64, 74]. Contents of high popularity were however found to be of higher quality than less popular content [134]. The part begins with the study of current CA platforms and their evolution towards social curation in Chapter 7. This is aided by an investigative study of Reddit [18] to better understand the features and content management strategy of CA platforms.

The focus of our research is on user behaviours and content management by information quality for CA platforms, particularly the high amount of user comments within each discussion thread. Unlike the CQA platforms where we seek the single best answer (Part II), we attempt to identify and organise multiple user comments in shared content and discussion to enrich the platform for user consumption [134]. Furthermore, CA platforms allow users to contribute multiple content within the same discussion thread unlike the single answer on CQA platforms. In Chapter 8, we propose and evaluate an extension of the competitive pairwise comparison model to predict the information quality of user comments in Reddit threads. We published our findings from this part of the research at:

[15] DOI: https://doi.org/10.1145/3166072.3166075
[16] https://chiebukuro.yahoo.co.jp
[17] DOI: https://doi.org/10.1145/3015022.3015032
[18] https://www.reddit.com

Wern Han Lim, Mark James Carman, and Sze-Meng Jojo Wong. Estimating Relative User Expertise for Content Quality Prediction on Reddit. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17)*, pages 55-64. ACM, 2017.[19]

[19]DOI: https://doi.org/10.1145/3078714.3078720

Part I

Collaborative Tagging (CT)

Chapter 2

Classification and Retrieval with Collaborative Tagging (CT)

Collaborative Tagging (CT) platforms allow users to store and share WWW resources. These are organised by the users with the use of annotations – short terms, keywords or tags attached to WWW resources [1]. Examples of CT platforms on the WWW include:

- Bookmark platforms such as Delicious[1] and Pinterest[2].
- Image sharing/ hosting sites such as Flickr[3], Instagram[4] and Imgur[5].
- Video hosting and sharing sites such as Youtube[6] and Dailymotion[7].
- Audio streaming site such as Last.FM[8] and Soundcloud[9].

The user annotations on CT platforms have the information potential to enhance the performance of information systems [105]. The annotations can be used to describe the WWW well [13] or to classify the resources [151]. We look at the information potential of user annotation in Section 2.2 and how user annotations have been used to enhance information retrieval through annotation-based retrievals in Section 2.3.

In this chapter, we first outline the general structure and representation of a CT platform in Section 2.1. We introduce the three main entities that are of focus for our research in a CT platform – the users, the WWW resources and the annotations. This research proceeds to obtain and study data from the Delicious social bookmarking platform, a popular CT platform in Section 2.4. The highlights from the section are the studies on (1) how the WWW resources bookmarked in Delicious provide good representation of the WWW; and (2) how the user annotations in Delicious are suitable for use as query keywords for annotation-based retrieval. The chapter concludes with a summary, leading us to the estimation of user expertise and performance evaluation in annotation-based retrieval in Chapter 3.

[1] https://delicious.com/ or https://del.icio.us/
[2] https://www.pinterest.com/
[3] https://www.flickr.com/
[4] https://instagram.com/
[5] http://imgur.com/
[6] https://www.youtube.com/
[7] http://www.dailymotion.com/
[8] http://www.last.fm/
[9] https://soundcloud.com/

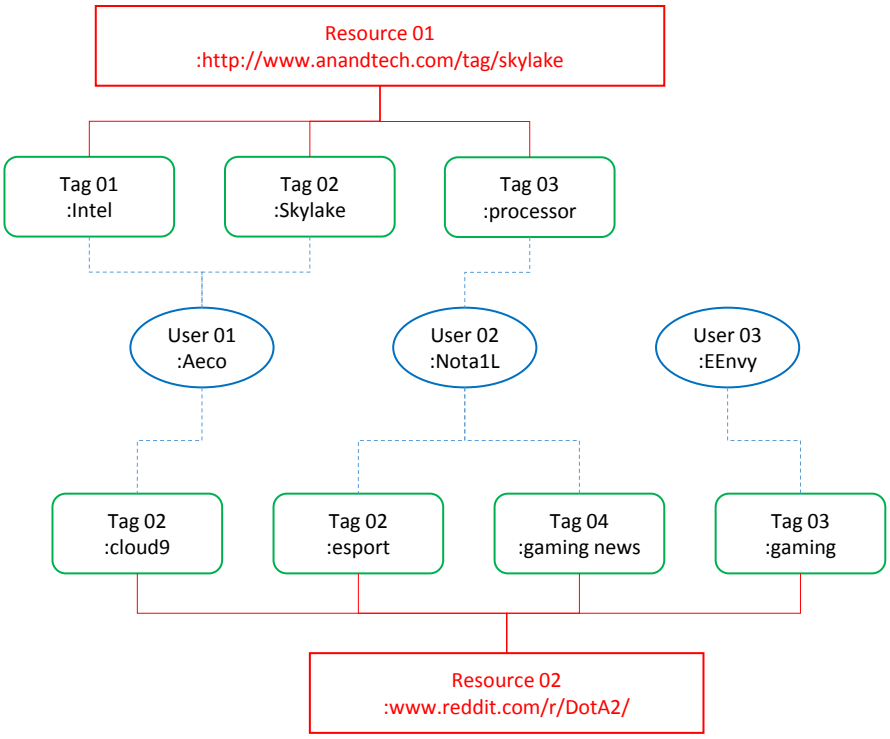


Figure 2.1: The structure of a general collaborative tagging (CT) platform.

2.1 Structure and Representation

CT platforms consist of three main entities – (1) the users of the platform, (2) the WWW resources that are stored, shared or organised; and (3) the user annotations. We visualise an example of a CT platform in Figure 2.1. In the figure, we observe that:

- A WWW resource can be annotated with one or more annotations where each annotation can be of a single keyword (tag) or a sequence of keywords (phrases). A single resource can be annotated with the same annotation multiple times, by different users.
- A WWW resource can be saved, shared or organised by one or more users. The same user can annotate a resource multiple times but only with a different annotation each time.
- A user is free to save, share or organise any WWW resource. Often, these actions are accompanied by user annotations of the user's choice.

A CT platform as illustrated earlier in Figure 2.1 can be modelled as a tripartite graph [86] with the assumption that each user and each WWW resource are unique entities on the platform. The tripartite graph (see Figure 2.2) consists of the three main entities for a CT platform that we shall discuss further in the following subsection.

2.1.1 User

Users play important roles on CT platform beyond being the consumers of content on CT platforms. Firstly, they are the discoverers of interesting WWW resources [86] which they save and share with the other users. Through user annotations that they attach to the WWW resources on the platform, they help to describe and organise the WWW resources.

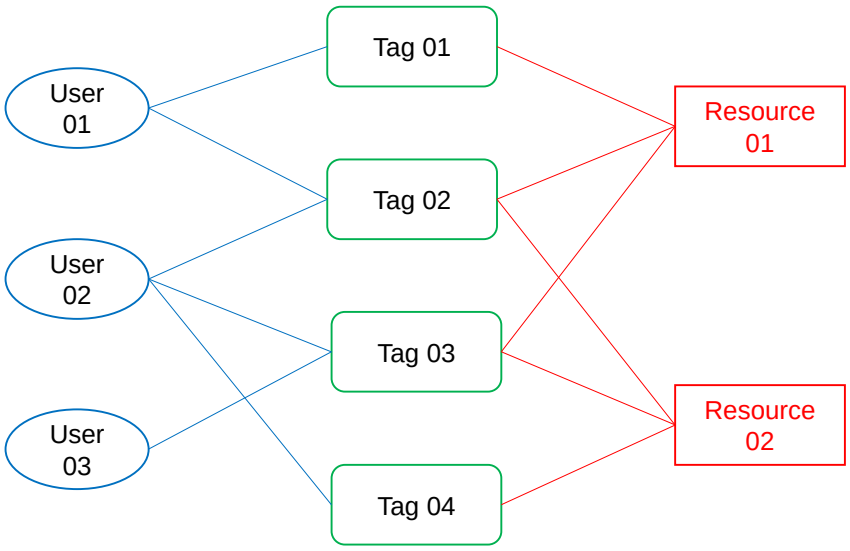


Figure 2.2: A tripartite representation of a collaborative tagging platform.

In our study of CT platforms, we assume each user to be unique within each platform, having a unique user ID. Each user has a different level of expertise (see Definition 5) which could affect their judgement of WWW resources and also their capabilities to attach good annotations to the resources.

2.1.2 Resource

The resources on CT platforms are from the WWW. Users would navigate through the WWW and then add the references (Uniform Resource Locators, URLs) to interesting resources via tools provided CT platforms. As they are WWW resources, they are diverse and can be of any format such as text documents, dynamic web pages or multimedia content.

Just like the users, each resource is unique within the tripartite graph representation of CT platforms; identified by their URL. Alternatively, CT platforms may assign each resource with a unique ID. Each resource can be of different information quality (see Definition 4) with one resource being better than the other on a particular topic [80, 109].

2.1.3 Annotation

Annotations are a core component for CT platforms for the users to share and organise content [139]. As discussed earlier, any user of any expertise is able to annotate any WWW resources. User annotations are as defined in Definition 6 and can be made up of a single keyword (a tag), W or a sequence of keywords (phrases).

Definition 6 (Annotation)**.** Annotations are short brief descriptions of the resources that they are associated with [60, 99]. On some platforms, annotations are interchangeable with tags.

The users select keywords from their own unrestricted vocabulary to what they think best describes or categorises the WWW resources. As such, annotations are not unique for the users or resources. A user's personal repository or vocabulary of annotations however can be personalised and is known as a personomy [65, 77].

Definition 7 (Personomy)**.** Personal vocabulary made out of terms used by the user for annotating resources [77, 147].

User annotations on CT platforms have a great information potential to enhance information systems within and beyond CT platforms. This thesis explores and discusses the information potential in Section 2.2.

2.2 Information Potential of User Annotations

User annotations are the user-generated content in focus for collaborative tagging platforms. Their primary use is for the users to describe, share or organise WWW resources. Being short and condensed for efficient processing, user annotations hold vast information potential [105].

2.2.1 Rivalling Experts

As discussed earlier, any user is able to annotate any resource on CT platforms; some or many of whom are non-experts. Research has has often shown that the user annotations of non-experts are very similar to the experts [17] on CT platforms such as Last.FM and Delicious. When aggregated together to reach consensus, the user annotations from non-experts are able to rival in quality with the experts' [132], with each annotation being able to increase the information quality [111] in describing WWW resources. On the other hand, the increase in redundant annotations does not reduce the description effectiveness [111].

With this in mind, there is no need for expert annotators to manually review or annotate large amounts of WWW resources on their own; a task seemingly impossible given the vast and ever increasing number of resources on the WWW. Instead, collection of user annotations from the large user community of CT platforms can be used to overcome this problem to attain a high and up-to-date coverage of the WWW [43].

2.2.2 Describing WWW Resources

User annotations are good descriptors of WWW resources – being able provide a reasonable summary of content [13]. In fact, user annotations were found to have a high overlap with both the title and content of WWW resources [60]. The short and condensed nature of annotations are more efficient to be processed than the actual large and varied content of the WWW resources themselves.

2.2.3 Organising WWW Resources

A common use case for user annotations is in the task of categorisation. As user annotations were found to be popular, frequent and stable [1], they can be used as features for classification tasks.

A notable information potential of user annotations is for the annotations to be used as categorisation classes directly without the need for content or feature processing. The Open Directory Project's (ODP[10]) [151] was able to successfully replace the use of traditional categorisation classes and ontologies.

[10]http://www.dmoz.org/ or updated to http://dmoztools.net/ as of March 2017.

2.3 Annotation-based Retrieval (AR)

Traditionally, users make use of web search engines to perform information retrieval on the web; retrieving WWW resources relevant to a given query. The retrieved resources are then ordered or ranked at query time based on their relevance to the query – often measured based on the similarity between query and content of resources [56].

Definition 8 (Query)**.** User queries are formulated by the users to describe their information need; based on the users' vocabulary and domain knowledge [144]. They are an approximation of users' information need during a specific information retrieval session [36].

The information potential of annotations provides us with an alternative to traditional IR. Annotations are short descriptors usable as features for similarity computation [60] or directly as classification classes [151]. In general, an annotation-based retrieval (AR) task would require:

- **User query.**
  User queries are as defined in Definition 8. User queries are unstructured and could contain some degree of ambiguity especially when the queries are short. Such characteristics are similar to that of user annotations.

- **User annotations for WWW resources.**
  User annotations are good descriptors of the WWW resources that they are attached to [60] on CT platforms. Such descriptors could provide an understanding of the WWW resources and their capabilities to meet the information need of the users. Alternatively, information seekers are able to explore and discover the WWW resources with a particular annotation of the seekers' choice without the need for a query formulation step.

An AR platform would measure the similarity between user queries and the annotations which describes the associated WWW resources. Just like in traditional IR, this similarity can be measured using word terms as there is a high overlap between the keywords or terms of user annotations and that of user queries [17, 28]; making them particularly useful for retrieval. In the following subsections, we discuss the current approaches to AR and also introduce a possible approach to AR from our research.

2.3.1 Current Approaches

Over the years, many annotation-based retrieval (AR) approaches were explored [8] to leverage the information potential of user annotations. The current approaches can be grouped into three distinct categories:

Simple Term Matching

The simplest approach would be the traditional method of matching terms between user queries and annotations attached to WWW resources. Any similarity measure can be used such as the highly successful Okapi BM-25 [120]. A key advantage of the simple approach is that the approach can be applied directly and adapted to any CT platform.

Annotation Context

Annotations are diverse especially when used on a wide variety of content; thus the contextual information of annotations can be used to improve annotation-based retrieval [2]. The annotations can be processed according to the:

- **Co-occurrence of annotations.**
  If two or more annotations are used in combination with each other (thus co-occur), then there exist some kind of semantic relationship between them [104]. Applying this same concept for user queries, then one could then better measure the similarity between user queries and the annotated WWW resources.

- **Order of annotations.**
  It is possible to estimate the context of annotations by considering the sequence of user annotations [126].

- **Clustering of annotations.**
  By grouping annotations together, it is possible to better measure the term similarity between annotations [119] or terms between user queries and annotations. In their work on the GroupMe! platform where users are able to group resources, Abel et al. [3] found that the grouping of annotations provided valuable semantic information.

- **Annotations as features for topic models.**
  Annotations could be processed to discover topics that categorise the user queries and the WWW resources [55] with algorithms such as the Latent Dirichlet Allocation (LDA).

Personalisation

There are also the personalised information retrieval (PIR) approaches to AR. In the PIR variant to AR, the user profile is taken into account for the relevance matching between WWW resources that the user queries [31, 78]. The user profiles are often built without asking the users for their personal information explicitly [108]; based on a user's personomy from their annotations [115] or the annotations of resources that they have spent time with [7].

Definition 9 (Personalised Information Retrieval, PIR)**.** An IR approach which takes into account a user's profile; for a tailored result that better understands and meets the information needs of the user [88].

2.3.2 Proposed Approaches with Information Quality

Our research proposes a novel approach to annotation-based retrieval. By estimating the user expertise (Definition 5), it is possible to infer the information quality (Definition 4) of their annotations or the WWW resources which they have interacted with (saving, sharing and organising). The inferred information quality can be directly incorporated into similarity measure calculation or resource scoring during retrieval as – (1) Score modifiers; or (2) Rank aggregation [5].

In Chapter 3 of the thesis, we explore the estimation of user expertise and how it can be used to infer the information quality of user annotations or WWW resources on CT platforms. By inferring the information quality of WWW resources, we are able to provide good WWW resources to the users to meet their information needs. Alternatively, we can leverage on the information quality of user annotations to better understand the content of WWW resources for an improved similarity measure between user queries and WWW resources.

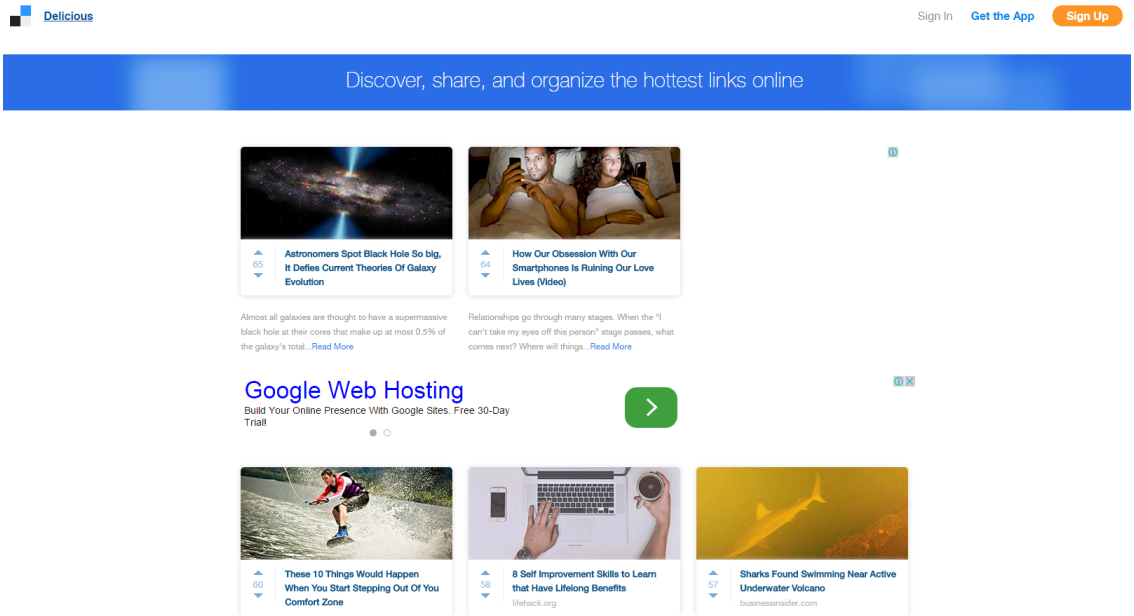


Figure 2.3: A screenshot of Delicious homepage. Web page visited on 19th July 2015.

2.4 Delicious Social Bookmarking Platform

In this section, we explore and study a collaborative tagging platform to better understand its structure, representation and features. Delicious[11] (or del.icio.us as it was formerly known) is a CT platform which provides social bookmarking services to its users. On Delicious, users are able to bookmark, organise or share WWW resources for the discovery of other users. A screenshot of the Delicious homepage is shown in Figure 2.3.

Delicious is chosen as the CT platform of choice for the research due to the following reasons:

- There are a wide variety of WWW resources on Delicious. On Delicious, the users are able to annotate any WWW resource with a URL to that resource regardless of the format or content. This differs from other CT platforms that are more specialised, such as Soundcloud[12] that only contain audio files or Flickr[13] for images.
- Delicious is an established CT platform that is widely used by its community; recording high quantity of annotations on its bookmarked WWW resources [103]. The annotations found in Delicious were also verified to be rich in contextual information that are suitable for information systems [55].
- At the time of the research in year 2012, it is possible to view and obtain public user annotations on the platform as illustrated in Figure 2.6. The service was however terminated on 1st June 2017 upon acquisition by Pinboard[14].

2.4.1 Design and Structure

Delicious is designed to help users explore interesting WWW resources as discovered by other users. This exploration can be done through their social feed on Delicious or through specific tags that are of interest to the user (see Figure 2.4).

[11] https://delicious.com/ or https://del.icio.us/
[12] https://soundcloud.com
[13] https://www.flickr.com
[14] https://pinboard.in/

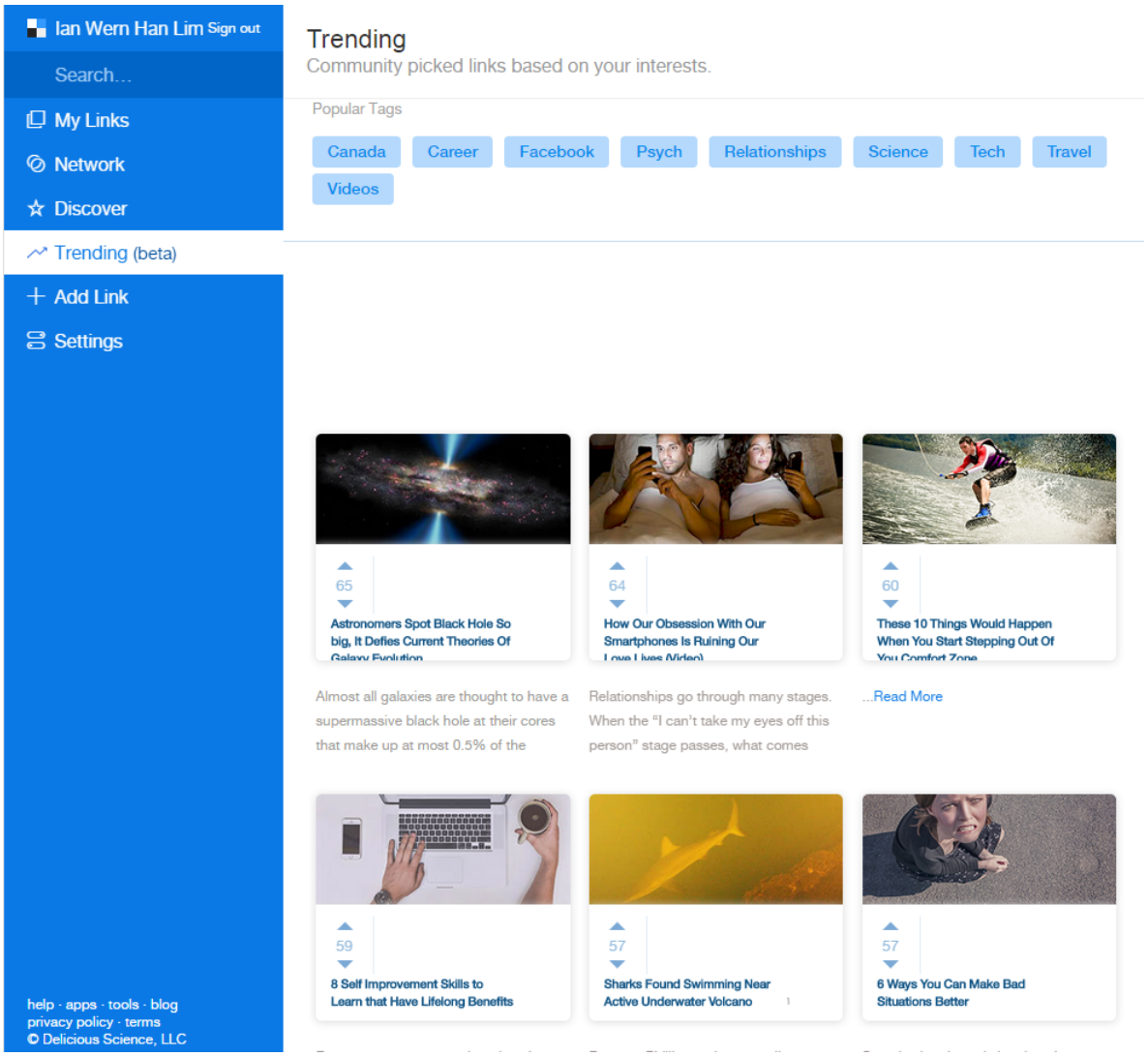


Figure 2.4: Screenshot of Delicious trend page. Web page visited on 19th July 2015.

On Delicious, the users are able to bookmark any of the WWW resources and share them on their social feed (see Figure 2.5). Users are able to annotate these bookmarks with user annotations such as tags or comments; from their unrestricted vocabulary which the users think best describe the WWW resources or for their own understanding and consumption.

2.4.2 A Delicious Dataset

A Delicious social bookmarking platform dataset was collected for the research – crawled for collection between 9th April 2012 to the 19th April 2012. While there exist public collaborative tagging datasets (such as those from Delicious) at the time of our study, these datasets were either outdated[15]; lacked the temporal information required[16] for the various explored approaches (to be discussed in Chapter 3); or were otherwise not suitable for our research[17]. The crawling was done through random sampling of the Delicious

[15]Delicious has undergone several changes over the years.
[16]Many of the datasets either contain only the user information or the tagging information [122].
[17]For example having a low number of users [25].

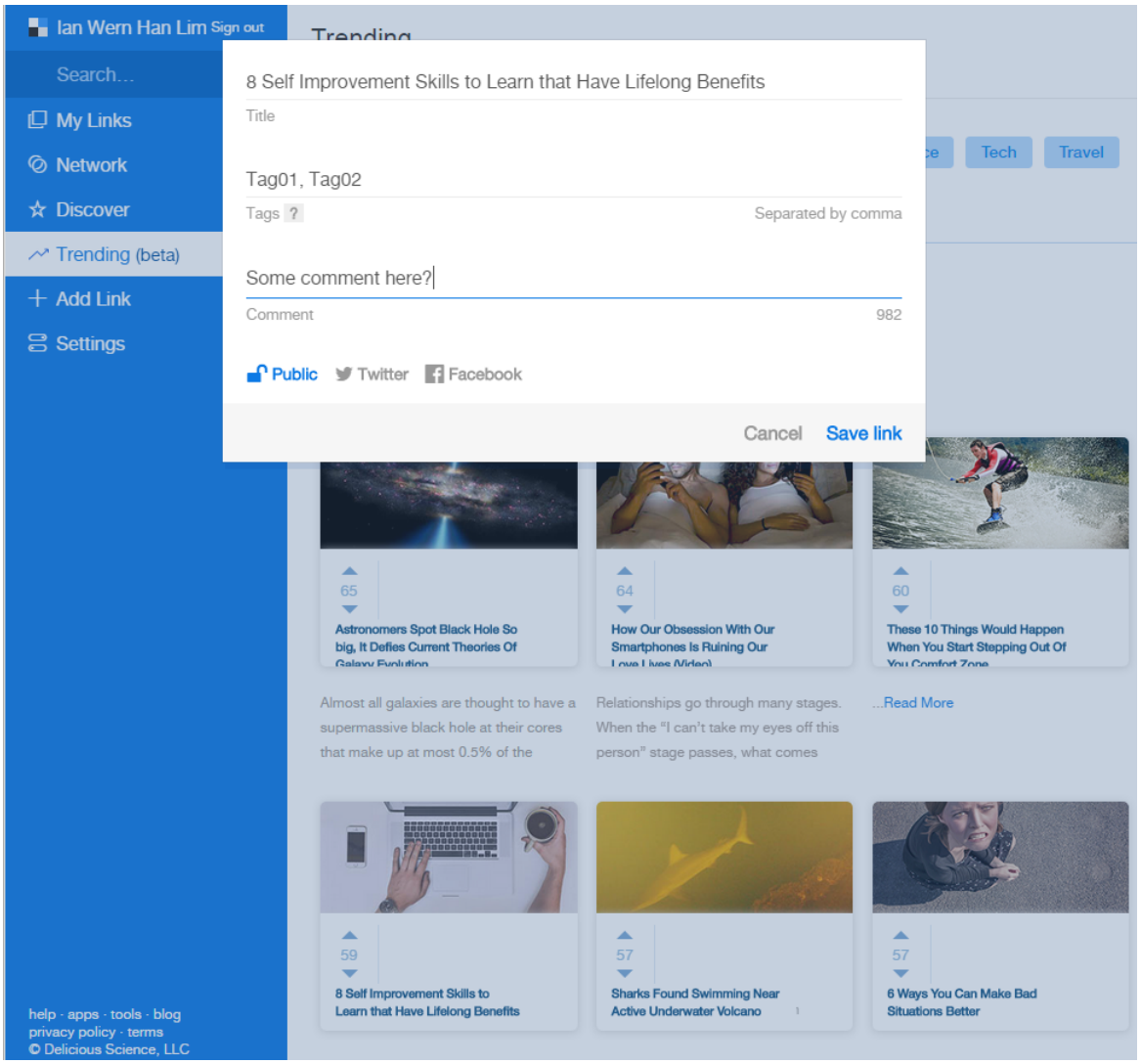


Figure 2.5: Bookmarking a World Wide Web (WWW) resource on Delicious. Web page visited on 19th July 2015.

Discover-100 pages which acted as the seed. The crawled public[18] bookmarks of the WWW resources were then downloaded and stored. The resulting dataset includes:

- 34,458 **unique WWW resources bookmarked.**
  Each of the WWW resources were identified according to their unique URL as stored in its Delicious page as shown in Figure 2.6. These resources were used for the retrieval tasks in Chapter 3.

- 191,624 **unique users profiled.**
  Unique users were identified according to their user ID such as user *@gerrypower* in Figure 2.6.

- 84,380 **unique annotation terms identified.**
  The unique annotation terms are annotated by the users to describe or organise WWW resources. In figure 2.6, these terms include '.en', 'open-data', 'tool' and 'refine'. The annotation terms are vital as features for annotation-based retrieval.

[18]No personal user information was collected.

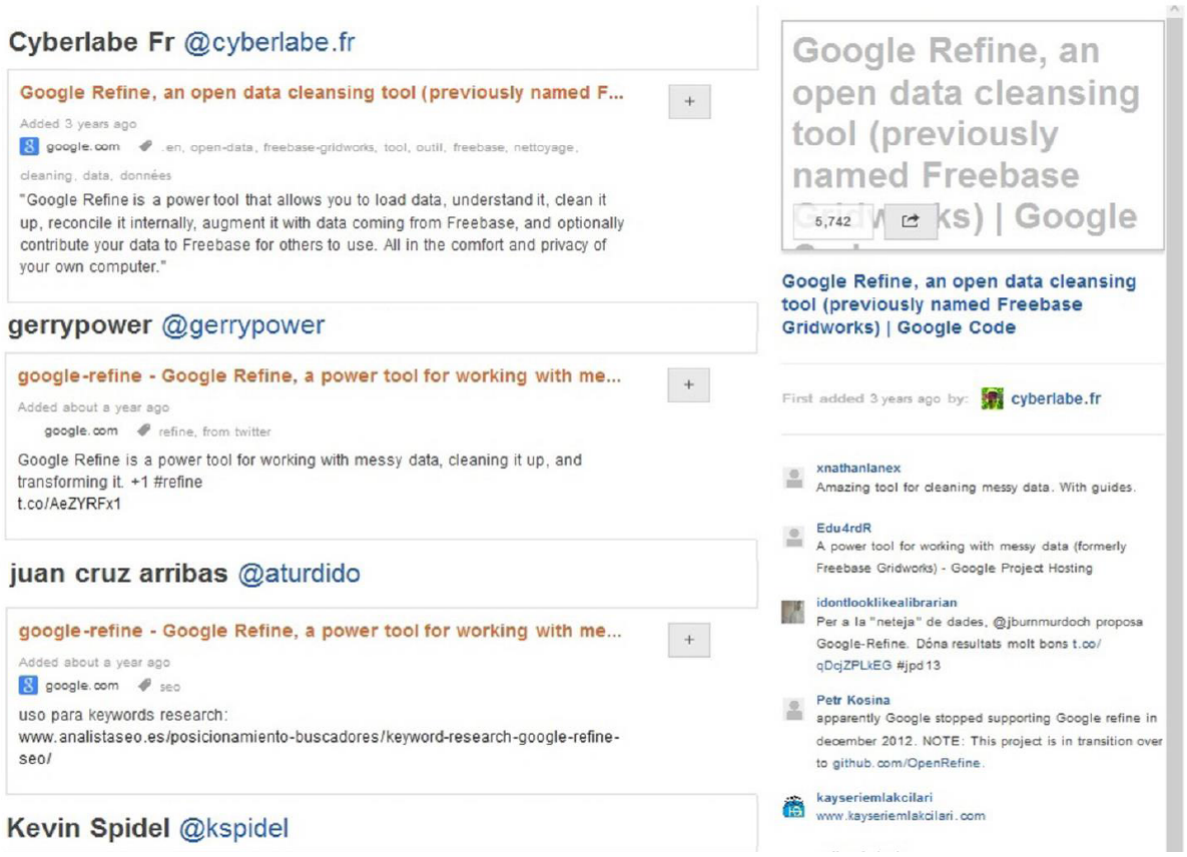


Figure 2.6: Metadata (user annotations and comments) on a World Wide Web (WWW) resource, bookmarked on Delicious. Web page visited on 20th June 2013.

- $256,262$ **unique canonical phrases discovered.**
  The phrases are attained from the sequence of annotation terms; where the terms in sequence have been sorted alphabetically to highlight the co-occurrence between terms. From the example in Figure 2.6 again, the phrases are 'refine, from Twitter', 'seo' and '.en, open-data, freebase-gridworks, tool, outfit, freebase, nettoyage, cleaning, data, donnees'.

This dataset is used for our study of CT platforms. Firstly, we explore how representative is the WWW resources of Delicious is for the WWW in Section 2.4.3; and also if user annotations are suitable to be used as queries for annotation-based retrieval in Section 2.4.4. Later in Chapter 3, the same dataset is used to evaluate the performance of AR with the explored algorithms.

2.4.3 Study: How Representative is Delicious?

To the best of our knowledge, there is no study on the representativeness of the WWW resources bookmarked on Delicious. Here, we ask the following question – Do the bookmarked WWW resources on Delicious provide a reasonable representation and coverage of the WWW as a whole; or are these resources just niche WWW resources?

From the mined dataset, we determine if the URL of bookmarked WWW resources overlap with those on the standard 2012 ClueWeb[19] via the 'url $->$ docno mapping'. ClueWeb has a large collection of web pages that is often used in information retrieval research. From $503,903,810$ WWW resources in the ClueWeb collection, we recorded

[19]http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=ClueWeb09%20Wiki

6,358 WWW resources from our dataset. This can be interpreted as 18.45% of WWW resources from our dataset that are found within the ClueWeb collection. Thus, the crawled dataset of Delicious can provide good coverage of the vast WWW since a large number of resources can be found in both datasets.

2.4.4 Study: Are Annotations Suitable for Queries?

In our discussion of the information potential for user annotations, we noted that others have observed that annotation terms have a high overlap with user query terms [17, 28]. This observation motivates our exploration into annotation-based retrieval later in Chapter 3 where the annotations from our crawled dataset of Section 2.4.2 would be used.

In this section, we compare the terms of the user annotations collected from the dataset with the terms of real world user queries provided in the TREC One Million Query Track (1MQ) of 2009[20]. This is the most recent 1MQ TREC available publicly at the time of the research with a total of 40,000 queries; and is commonly used in the exploration of ad-hoc retrieval on large collections.

The annotation and query terms are pre-processed with the removal of stop words and reducing terms to their root using word stemming [98]. The overlap between query terms and annotations are assessed based on the following criteria:

- **Exact annotation-query term match.**
  For a WWW resource r with annotations $a_1, a_2, ..., a_{|r|}$ and each annotation having terms $w_1, w_2, ..., w_{|a_x|}$; all terms must overlap with all terms of the query $w_1, w_2, ..., w_{|q|}$.

- **Single annotation-query term match.**
  For WWW resource r with annotations $a_1, a_2, ..., a_{|r|}$ and the resource can be described using k annotation terms $w_1, w_2, ..., w_k$; at least a single term must overlap with the query terms $w_1, w_2, ..., w_{|Q|}$.

We found that the annotation terms from the bookmarked WWW resources in the Delicious dataset is a valid representation of real world queries due to the following:

- There are 2,055 (5.14%) exact matches between WWW resources bookmarked in the dataset with the queries in the TREC OMQ 2009.
- There are 34,347 (74.74%) single matches between WWW resources bookmarked in the dataset with the queries in the TREC OMQ 2009.

2.5 Summary

In this chapter, we covered at the structure and representation of CT platforms; usually modelled as a tripartite graph (Figure 2.2). User annotations on CT platforms have the information potential to enhance information systems and are widely used for annotation-based retrieval – where the diverse WWW resources ranging from textual content to multimedia content that can challenge traditional content-based IR; can otherwise be described with annotations and retrieved effectively through annotation-based retrieval. We saw earlier that the bookmarked WWW resources can be a good representation of the WWW and there are sufficient user annotations to describe them.

In Chapter 3, we further explore the information potential of user annotations and their application for retrieval. This research aims to infer the information quality of annotations through the estimated user expertise of the annotators; as the inferred information quality can then be incorporated to improve AR.

[20] http://trec.nist.gov/data/million.query09.html

Chapter 3

Annotation-Based Retrieval with User Expertise and Information Quality

Motivated by the information potential of user annotations from collaborative tagging (CT) platforms in describing WWW resources; this research delves deeper into annotation-based retrieval (AR). In our exploration of AR, the user query terms are used for the similarity measure with the associated annotations of WWW resources [17, 28]. This is a content-agnostic approach as AR systems do not need to process the actual content of the WWW resource itself but could instead, rely on the annotations. Hence, making the technique suitable for the unstructured and unmoderated nature of the WWW.

First, we investigate some of the similarity measures that can be used for AR in Section 3.2 particularly the Okapi BM-25 similarity measure [120]. This research does not aim to improve on the current known similarity measure algorithms. Instead, the novel contribution of this research is the incorporation of estimated user expertise and inferred information quality to improve any similarity measure for retrieval. While there exist many different approaches to incorporate such measures (such as rank aggregations [5, 40]), the simple aggregated composite approaches explored here allow this research to best gauge the impact of such estimated additions.

To estimate the expertise of users, our research considered graph-based approaches. Section 3.3 outlines the graph model as a representation of CT platforms including the Spamming-resistant Expertise Analysis and Ranking (SPEAR) algorithm [109] in Section 3.4. This research expands on SPEAR with a Credit Graph model and several credit functions; discussed in Section 3.5.

An evaluation is then conducted on the proposed algorithms against the baselines – starting with a training phase for the algorithms in Section 3.7 and then followed by a testing phase discussed in Section 3.8. The observations and findings from the evaluation are then presented. Finally, we conclude this chapter and Part I of the thesis with a conclusion in Section 3.9.

3.1 Research Questions

In this chapter, we expand on the research questions from Section 1.4 for CT platforms. These research questions guide our research in this chapter for annotation-based retrieval.

3.1.1 Can the Information Quality of User Annotations be Estimated through Content-Agnostic Means?

On CT platforms, user annotations are used to describe the WWW resources that they are associated with. In Section 2.2, we saw the information potential of user annotations. The unstructured nature of UGC platforms however affects the reliability of user annotations as discussed in Section 1.3. Thus, this research attempts to infer the information quality of user annotations in describing WWW resources.

In this chapter, we explore the task of annotation-based retrieval (AR) using an evaluation framework outlined in Section 3.8. Given a query, AR algorithms attempt to retrieve relevant WWW resources according to their associated user annotations. If the user annotations are of higher quality in describing the WWW resources, then AR algorithms should be able to predict the resource relevance better. As user annotations are unstructured and vary according to the users' vocabulary [77], it can be a challenge to process these user annotations directly through content-based approaches based on natural language processing (NLP).

In our research, we attempt to estimate the expertise of the annotators themselves and then use their estimated expertise to infer the information quality of their annotations. Besides that, we attempt to infer the information quality of the WWW resources that they have interacted with.

3.1.2 Can User Expertise be Estimated on CT Platforms through Content-Agnostic Means?

The unstructured nature of UGC content challenges content-based approaches that have been traditionally used to profile users [31, 78] such as the processing the user content [24, 38]. This research instead attempts to leverage on the features of CT platforms to estimate the expertise of its users.

We have discussed in Section 2.1 the structure of a CT platform that could be modelled as a tripartite graph as illustrated in Figure 2.2. This graph representation of CT platforms can be used to calculate the authority of users on CT platforms for the estimation of user expertise, many of which we explore in Section 3.3. Inspired by the SPEAR algorithm [109], this research proposed the credit graph model in Section 3.5 model and with credit functions to weight user-resource links. The credit function variants introduced in Section 3.5.1 enable us to identify and evaluate possible CT platform features as signals for user expertise. We then study the user authority as a potential user expertise estimate.

3.1.3 Do Expert Users Better Annotate WWW Resources?

Following up on the estimated user expertise, this research attempts to infer the information quality of user annotations (in describing and categorising WWW resources) from the expertise of the annotators. We assume that the higher the user expertise, more likely the user would describe each WWW resource better with precise annotations. We study this impact by applying user expertise directly into the AR process – weighting the annotation terms according to the expertise of the annotators for similarity calculation during AR in Section 3.7 and Section 3.8.

3.2 Similarity Measure for Annotation-based Retrieval

Similarity measures provide a common strategy in IR – for attempting to predict the relevance of a document with respect to a user query [98]. The documents are ranked

according to their predicted relevance with the highest few presented to meet the user's information need. The same can be applied to WWW resources.

The earliest and simplest form of similarity measure is the Boolean model approach. In this model, a WWW resource is predicted as relevant to the user query if all terms from the query are found within the resource. On CT platforms, this is often used by default for the users to explore tagged WWW resources discovered by other users [13].

Often times, the annotation terms are cleaned through pre-processing for improved performance [98]. Terms which do not hold meaningful information are removed through stop word removal and terms are reduced to their root word through stemming. It should be noted, however, that the unstructured nature of annotations such as unwanted typos, colloquial terms and non-standard abbreviations could hamper the cleaning of annotation terms.

The bag of words model [160] can also be applied to the user annotations associated to WWW resources on CT platforms or the user query terms. Each WWW resource is described by the collection of words or terms in the bag; and then used to measure the similarity between resource and user query.

Often, the bag of words model is used with the Term Frequency Inverse Document Frequency (TF-iDF) approach to similarity matching [14]. The significance of a term in a collection is quantified as inversely proportional to the number of documents in which it occurs [71]. In this research, we incorporate two common [98] TF-iDF based similarity measures:

- Cosine Similarity [131] used to measure the similarity between user annotations on WWW resources for the Similarity Credit Function discussed further in Section 3.5.1.
- Okapi BM-25 [120] used to measure the similarity between user queries as annotation terms and user annotations on WWW resources for AR in Section 3.6.

Both approaches assume no interdependence between terms (or that the dependence between terms is not necessary) [98]. We acknowledge that modelling the dependency between annotations could be used to improve AR performance [104, 126]. But since this research aims to explore the impact of estimating of user expertise and inferred information quality, using a simple approach for retrieval enabled us to explore their impact better.

3.2.1 Cosine Similarity

Cosine similarity is commonly used to measure the similarity between documents, or the similarity between user queries and documents for retrieval [131]. It is a vector space model where each document is represented as a vector [125] and the cosine of the angle between the vectors indicates the similarity between documents.

In the context of our research, the cosine similarity can be used to measure the similarity between user-annotations by representing each user annotation action $a \in A$ as a vector over annotation terms W; according to Equation 3.1 for annotation a_x and annotation a_y.

$$\textbf{Similarity}(a_x, a_y) = \frac{\sum_{i=1}^{|W|} \textbf{Weight}(w_i, a_x)\textbf{Weight}(w_i, a_y)}{\sqrt{\sum_{i=1}^{|W|} \textbf{Weight}(w_i, a_x)^2}\sqrt{\sum_{i=1}^{|W|} \textbf{Weight}(w_i, a_y)^2}} \tag{3.1}$$

The weight for each annotation term $w \in W$ in an annotation action a is measured as the product of the term frequency (TF) and inverse document frequency (iDF) as shown in Function 3.2.

$$\textbf{Weight}(w, a) = \textbf{TF}(w, a) \cdot \textbf{iDF}(w) \tag{3.2}$$

Since user annotations are unstructured and vary according to a user's vocabulary [77], we perform log transformation on the term frequency values for the calculation of TF as seen in Function 3.3 with $\mathbf{Freq}(w, a)$ for the frequency of an annotation term w in a user annotation a.

$$\mathbf{TF}(w, a) = \log(\mathbf{Freq}(w, a) + 1) \tag{3.3}$$

The iDF as shown in Function 3.4 measures the significance of an annotation term – the more common the term w is across all user annotations in the collection $|A|$, the less significant the annotation is. This is relevant for the unstructured and unmoderated nature of UGC platforms which is vulnerable to spam.

$$\mathbf{iDF}(w) = 1 + \log \frac{|A|}{1 + |\{a \in A | w \in a\}|} \tag{3.4}$$

3.2.2 Okapi BM-25

The Okapi BM-25 [120] is a simple and effective similarity measure approach for probabilistic information retrieval [98]. In our study, the BM-25 is used to measure the similarity between user annotations attached to the bookmarked WWW resources on CT platforms and the annotation terms in user queries – higher similarity means higher relevance. This value is measured in a normalised range of $[0, 1]$; which can be used in conjunction with additional estimates such as resource popularity, information quality and user expertise (that is to be introduced in Section 3.6).

Function 3.5 details the function used for the Okapi BM-25 similarity measure for the relevance of resources in Collection R, given a user query with the following annotation terms $(w_1, ..., w_{|W|})$. The iDF function as shown in Function 3.6 is used to adjust the significance of each annotation term.

$$\mathbf{BM25}(r, W) = \sum_{w \in W} iDF(w) \cdot \frac{\text{Freq}(w, r) \cdot (k_1 + 1)}{\text{Freq}(w, r) + k_1 \cdot (1 - b + b \cdot \frac{|r|}{avgrl})} \tag{3.5}$$

where:

- W is the query consisting of a sequence of terms $(w_1, ..., w_{|W|})$.
- $\text{Freq}(w, r)$ is the frequency of the term w amongst the user annotations for resource r.
- $avgrl$ is the average resource length (in annotation words) for the collection.
- Free parameters values $k_1 = [1.2, 2.0]$ and $b = 0.75$ are generally adopted as reasonable parameter values in the absence of optimisation [98]. We however optimize these values in the training phase of the research in Section 3.7.

$$\mathbf{iDF}(w) = \log \frac{|R| - |R_w| + 0.5}{|R_w| + 0.5} \tag{3.6}$$

- $|R_w|$ is the number of resources in the collection R that annotated with the term w.
- $|r|$ is the total count of annotation words for resource r (i.e. the size of the document).

3.3 Graph Representation of Collaborative Tagging Platforms

As discussed in Section 2.1, CT platforms such as the Delicious social bookmarking can be modelled as a Tripartite graph (Figure 2.2) composing of the following entities – the bookmarked WWW resources, the users and their annotations. This graph representation enabled us to utilise various graph algorithms for key information [142].

Here in our research, we employ graph algorithms to estimate user expertise and then infer the information quality from their interactions. In the past, graph algorithms have been successfully used to measure the authority of graph entities particularly in IR [98] such as Google's PageRank [112] and the Hyperlink-Induced Topic Search (HITS) algorithm [80]. The approach consists of two main steps, which is to be discussed in the following subsections that can be adapted for CT platforms.

3.3.1 Entity Relationship

The tripartite model of CT platforms discussed in Figure 2.2 allows us to model the relationship between entities. These relationships between the entities can be weighted or unweighted; and directed or undirected. In general, the three main entity relationships are as followed:

User-User Link

User-user relations are typically used in graph algorithms for the estimation of authority-based user rankings. For example, the user-user relation can be used to measure the co-authorship consistency between document authors of similar experience [37]. Depending on the CT platform, it is possible for a user to directly interact with the other users in an explicit relation or indirectly in an implicit relation:

- **Direct or Explicit Links.**
  Many of the current CT platforms allow one user to follow another user directly, for example when a user is subscribed to another user on Youtube. Alternatively, the users could collaborate with one another for document co-authorship [37]. The first interaction can be modelled as a directed link, whereas the second be modelled as an undirected link.

- **Indirect or Implicit Links.**
  Indirect relations can be built through the identification of similar interactions between users. For example, if two users have recently interacted with the same WWW resources or added the same annotations [155], an implicit relation can be built between these users. Besides that, one could leverage on various signals to discover latent relations between users such as the topical information similarity between users or if two users are from the same organisation [37]. Based on the tripartite graph visualised in Figure 2.2, we could form implicit user-user links as shown in Figure 3.1 and Figure 3.2.

Resource-Resource Link

Like the user-user relations, the resource-resource relations could be built for the WWW resources on CT platforms. Traditionally, it is common for graph algorithms to be used to rank resources such as web pages [112] or research documents [37, 142]. Links can be built directly when a web page hyperlinks to another or when a research paper cites another.

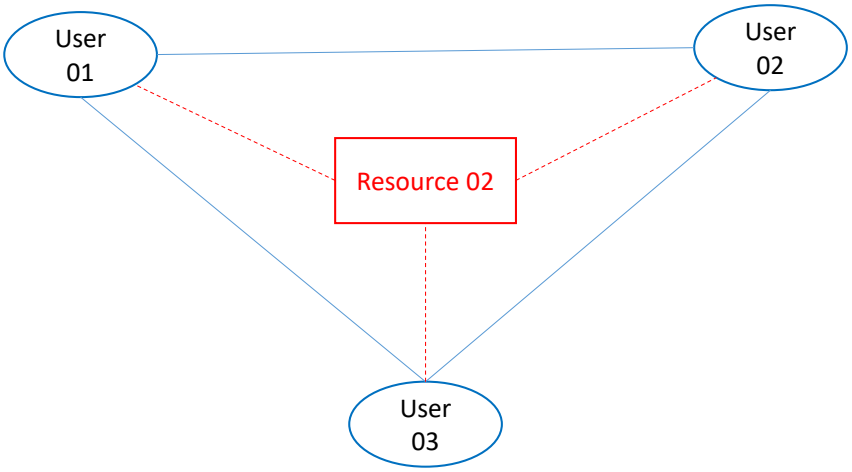


Figure 3.1: Example of indirect user-user link from similar annotated resources.

Links between WWW resources on CT platforms are often indirect – build through the similarity in the users who interact with the resources or described with the same annotations. Referring to the scenario from Figure 2.2, indirect relations can be built between resources if interacted by the similar users or annotated with the same annotation term giving us an undirected relation between *Resource 01* and *Resource 02*.

User-Resource Link

Relationships can be established based on user interactions with WWW resources. If a user discovers, shares, annotates, rates or comments on a WWW resource; there can be a relationship between the user and that resource. Such relationships hold valuable information such as earlier on the author-paper relation for research publications [37]. Here, the authors were able to show the inferred expertise of an author (a user in a bibliography system) is consistent in quality with that of the paper authored.

A graph-based approach which we explore deeper is the Spamming-resistant Expertise Analysis and Ranking (SPEAR) algorithm [109]. User-resource relation are built when a user bookmarks or annotates a WWW resource with an annotation. Under this model, links can be built between the user and the WWW resources as shown in Figure 3.3.

3.3.2 Link Analysis

The link analysis step leverages on the relations built between entities on the graph model such as the ones discussed in Section 3.3.1. It is a crucial step where the relation links, weighted or unweighted between entities can be used to determine various entity measures [80, 94] including:

- Impact of journals or research publications [37].
- Authority of a WWW resource [112, 162].
- Influence of users on social networks [53].

Link analysis is performed based on the random surfer model [72] which measures the probability of visiting an entity from another random entity [112]; otherwise known as a random walk [164]. This random walk is used to propagate values iteratively until convergence. If there are more than a single value being measured, then mutual recursion can be used [123] where the values mutually reinforce each other [16]. A common example

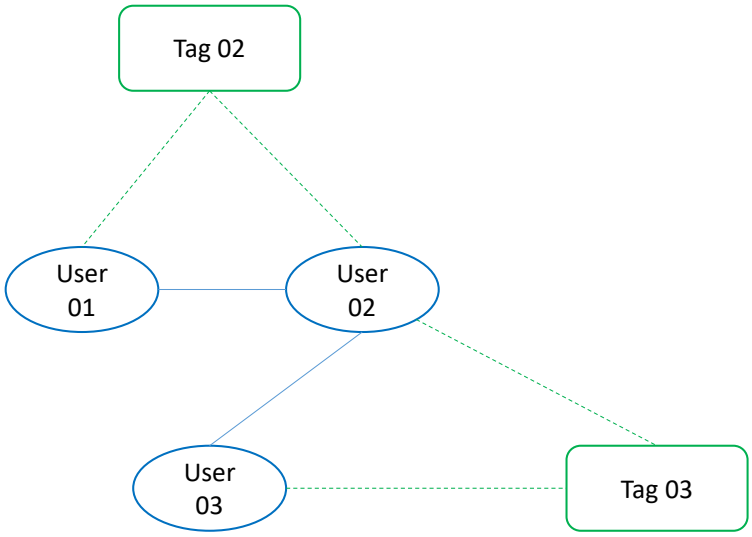


Figure 3.2: Example of indirect user-user link from similar tag used.

is the mutual recursion used to relate the hub and authority score of WWW resources in the HITS algorithm based on the assumption that a good hub would always lead to a WWW resource of high authority.

3.4 SPEAR Algorithm

The graph approaches explored in this research are inspired by the Spamming-resistant Expertise Analysis and Ranking (SPEAR) algorithm [109]. It is a link analysis algorithm based on the Hyperlink-Induced Topic Search (HITS) algorithm [80] for mutual reinforcement – each user acts as a hub to determine the authority (or quality) of resources. In this section, we describe the SPEAR algorithm for the estimation of user expertise and WWW resource quality.

3.4.1 User-Resource Entity Relationship

The SPEAR model relies on the user-resource entity relation where links are built between the users and the WWW resources that they have bookmarked [97] similar to that of Figure 3.3. This relation is based on the following assumptions:

- There is a direct mutual reinforcement relation [16] between user expertise and the WWW resource quality. Under this assumption, we can infer that the quality of a WWW resource should increase as the number of expert users who share or annotate it increases. Likewise, the expertise of a user should increase if that user share or annotate many high quality WWW resources.
- Users of high expertise do bookmark, share and annotate a large collection of WWW resources that are of high information quality. They tend to add more WWW resources that are of high quality than lower quality ones.

Relation Weights

The edges in the user-resource relations can be weighted. The weights are used for the propagation of values that determines the expertise of users and the quality of WWW

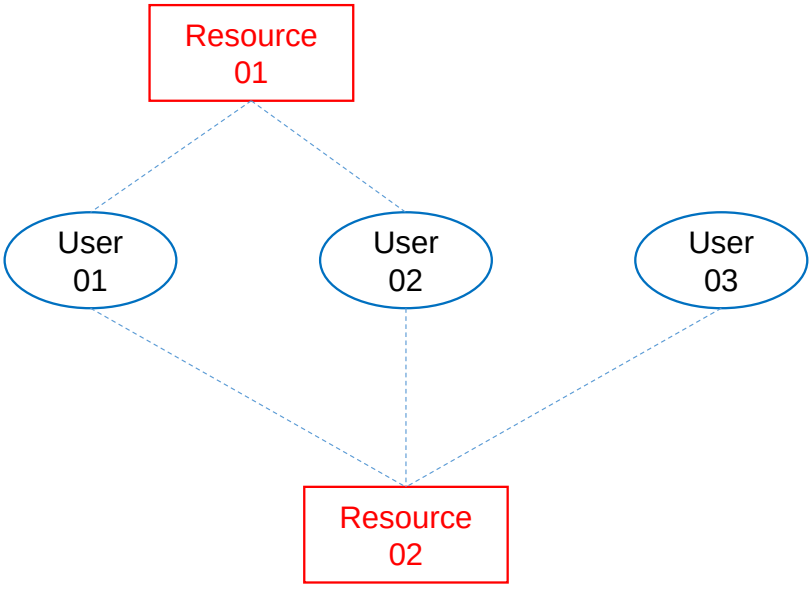


Figure 3.3: Example of user-resource links based-on user interactions.

resources. In SPEAR, the weights assigned for the relation are based on the Discover-Follower assumption that expert users are early discoverers of high quality or popular content, instead of just following the discovery of others [155]. Thus, user-resource relations for early users as discoverers are weighted higher than that of later users.

Based on the Discoverer-Follower concept, an interaction i_x that is used to build the user-resource relation between user u and resource r (with interactions I) can be weighted according to Function 3.7. The function assigns a value for each user interaction after interaction i_x (performed at time t) where these subsequent interactions $i_{t+1}, i_{t+2}, ..., i_{t+k}$ are performed by the followers. Thus, interactions by users as early discoverers would receive a higher weight than the later interactions[1].

$$\textbf{Weight}(i_x) = \sqrt{|\{i \in I | \textbf{Time}(i) > \textbf{Time}(i_x)\}|} \tag{3.7}$$

In Section 3.5.1, we explore several other ways to weight the user-resource relation – known as the Credit functions for how much credit should a user gain towards his or her expertise from that interaction. The implication from the Discoverer-Follower concept is then validated against the other credit functions in Section 3.7.

3.4.2 User-Resource Mutual Reinforcement with Links

The weighted user-resource links are then used for the mutual reinforcement of user expertise and WWW resource quality estimation. For SPEAR, the propagation of values is similar to the propagation between hubs-and-authority [80] of the HITS algorithm – expert users acting as hubs to locate WWW resources that are of high quality [109]. The user-resource relations in the graph model of CT platforms are stored in a user-resource adjacency matrix M; where the value for each cell in the matrix is 0 if the user has not interacted with that resource and is the weighted value from Function 3.7 if the user has.

[1]Note that user interactions performed at the same time t are not regarded as subsequent interactions.

The value propagation used here is that of a mutual recursion [123], a recursion for the mutual reinforcements of two attributes with each other [16] – the user expertise and the WWW resource quality. The relation for the propagation are as shown in Equation 3.8 for user expertise e and WWW resource quality q.

$$\vec{e} \leftarrow \vec{q} \cdot M^T \qquad \vec{q} \leftarrow \vec{e} \cdot M \tag{3.8}$$

The propagation process is then repeated until the user expertise and WWW resource quality values converge. At each iteration, normalisation is performed according to Equation 3.9 for the user expertise e_u and resource quality q_r. The square of the values would sum to 1 to ensure convergence [80].

$$\hat{e}_u = \frac{e_u}{\sqrt{\sum_{u'} e_{u'}^2}} \qquad \hat{q}_r = \frac{q_r}{\sqrt{\sum_{r'} q_{r'}^2}} \tag{3.9}$$

3.4.3 Algorithm

The entire process for the SPEAR algorithm can be described according to the Algorithm 1. The algorithm allows us to estimate user expertise and resource quality for collaborative tagging platforms.

Algorithm 1 Basic SPEAR Algorithm

Require: Users $U = \{u_1, u_2, ..., u_{|U|}\}$
Require: WWW Resource $R = \{r_1, r_2, ..., r_{|R|}\}$
Require: Set of interactions $I = \{i_1, i_2, ..., i_h, ..., i_{|I|}\} \in (\mathcal{U}, \mathcal{R}, \mathcal{A}, \mathcal{T})^{|I|}$

Set $\vec{e}$ to be the vector $(1, 1, ..., 1)$ for user expertise
Set $\vec{q}$ to be the vector $(1, 1, ..., 1)$ for WWW resource quality
$M \leftarrow GenerateAdjacencyMatrix(I, Credit(I))$
for $iterate = 1$ to $convergence$ **do**
  $\vec{e} \leftarrow \vec{q} \cdot M^T$
  $\vec{q} \leftarrow \vec{e} \cdot M$
  $\vec{e} \leftarrow \frac{\vec{e}}{||\vec{e}||}$
  $\vec{q} \leftarrow \frac{\vec{q}}{||\vec{q}||}$
end for
$L \leftarrow$ Sort users by their expertise in $\vec{e}$

return L

3.4.4 Analysis: WWW Resource Popularity vs Inferred Quality

The popularity of WWW resources is often used to measure the quality of WWW resources [140] – higher probability means greater presence, authority [143] and influence on the WWW. For CT platforms, the popularity of a WWW resource can be quantitatively measured by the number of user interactions – bookmarking, sharing and annotating.

Definition 10 (Popularity)**.** A state of content expressed by the amount of activity on or with the content itself [140].

From the SPEAR algorithm, it is possible to infer the information quality of WWW resources as a by-product of the user expertise estimation. However, are both measures the same? To understand the relation between the popularity of WWW resources and

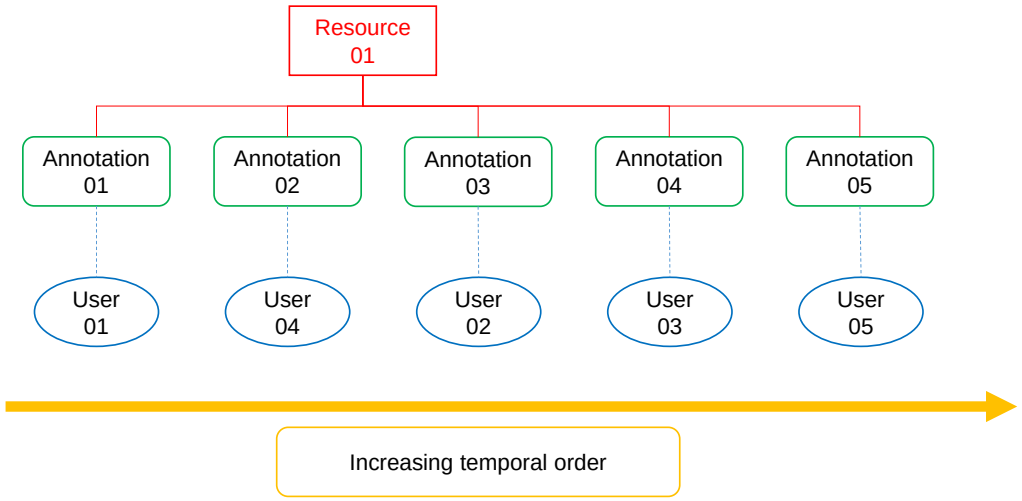


Figure 3.4: A bookmarked WWW resource example.

the inferred quality from the SPEAR algorithm, this research measures the correlation using the Delicious social bookmarking collected as detailed in Section 2.4.2. The findings include:

- Pearson's correlation coefficient of 0.19.
- Spearman's correlation coefficient of 0.78.

The findings from this quick study indicate that the popularity of WWW resources exhibits a strong but highly non-linear relationship with the inferred information quality of the resource. This relation can be generalised as an observation that a large number of popular WWW resources are of high quality. It should however be noted that there are high quality WWW resources that received average to high popularity. Besides that, we also find the several popular resources that are inferred to be of low quality.

One would of course expect that popular WWW resources would be of higher quality on average than less popular resource. Our findings however discover unpopular resources that are of good quality. Having established this relation, this research now explores the impact of these two measures on annotation-based retrieval in Section 3.8.

3.5 Credit Graph Model

Inspired by the SPEAR algorithm [109] discussed in Section 3.4, this research proposes the Credit Graph (CG) model for CT platforms. In general, the CG is a graph that connects all users in the platform to the WWW resources which they have interacted with. These interactions include bookmarking, sharing and annotating. The user-resource relation (see Section 3.3.1) are weighted according to the credit function of choice which measures the impact, significance or the contribution of that particular interaction.

Consider a CT platform scenario presented in Figure 3.4. The credit graph representation for this scenario is shown in Figure 3.5. The links are weighted based on the default Discoverer-Follower concept described in Section 3.4.1, given by the Function 3.7.

When all of the users and WWW resources on the CT platforms are accounted for, the resulting credit graph in Figure 3.6 is built based on the user interactions. In the following Section 3.5.1, we discuss several approaches to weight the links (or edges) of the model.

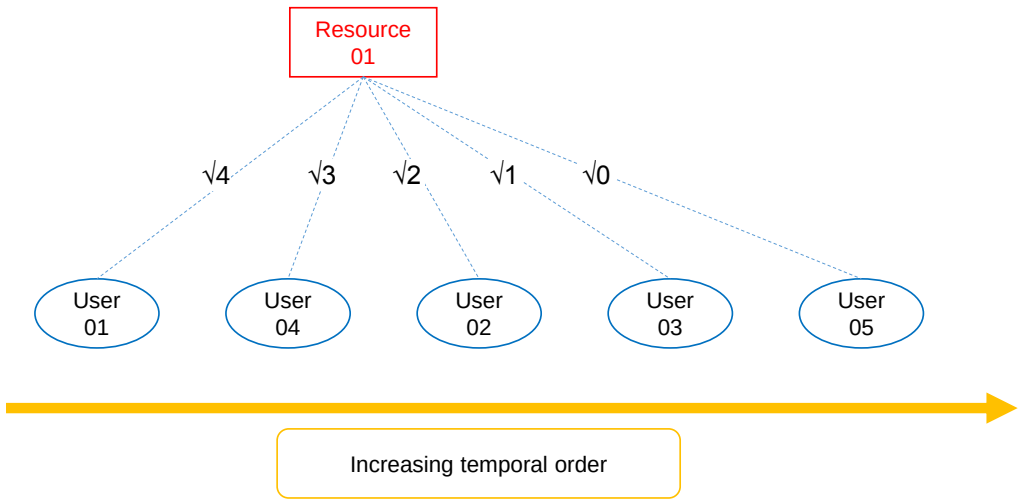


Figure 3.5: User-resource relation with SPEAR weights.

3.5.1 Credit Functions

The credit function **Credit**(I) measures the contribution or significance of each user interaction with WWW resources on CT platforms. The measure is used to weight the user-resource relation built based on the user interaction for the credit graph. This weight acts as the credit which a user would receive during the mutual reinforcement propagation (see Section 3.4.2) towards a higher expertise rating.

The credit function is inspired by the SPEAR algorithm [109] where user bookmark interactions are weighted based on the Discover-Follower concept (Section 3.4.1) as shown in Function 3.7. Here, credit is given based on the temporal ordering of the user interactions for a WWW resource with the earlier users receiving more credit as a discoverer (to be discussed further in Section 3.5.1. Following this, we explore several other possible variants for the credit function such as a popularity variant in Section 3.5.1 and a similarity-based variant in Section 3.5.1.

Optimisation of Credit

This research also explores the introduction of an optimisation parameter for the credit function in order to control the impact of the credit attributed to each user interaction. The credit exponent α is incorporated into Function 3.10 for the credit of an interaction i. For the original SPEAR algorithm, the authors chose a value of $\alpha = 0.5$ but did not disclose the reason for it or if any optimisation steps are taken.

$$\mathbf{Credit}'(i) = \mathbf{Credit}(i)^{\alpha} \tag{3.10}$$

Alternatively, we explore the logarithmic transformation of the credit function to obtain a less skewed distribution of credit for a WWW resource, as shown in Function 3.11.

$$\mathbf{Credit}'(i) = \log \mathbf{Credit}(i) \tag{3.11}$$

Popularity Credit

The popularity measure of WWW resources can be considered as an indicator of authority for retrieval [143] where resources of high quality are desired. Extending from this, we

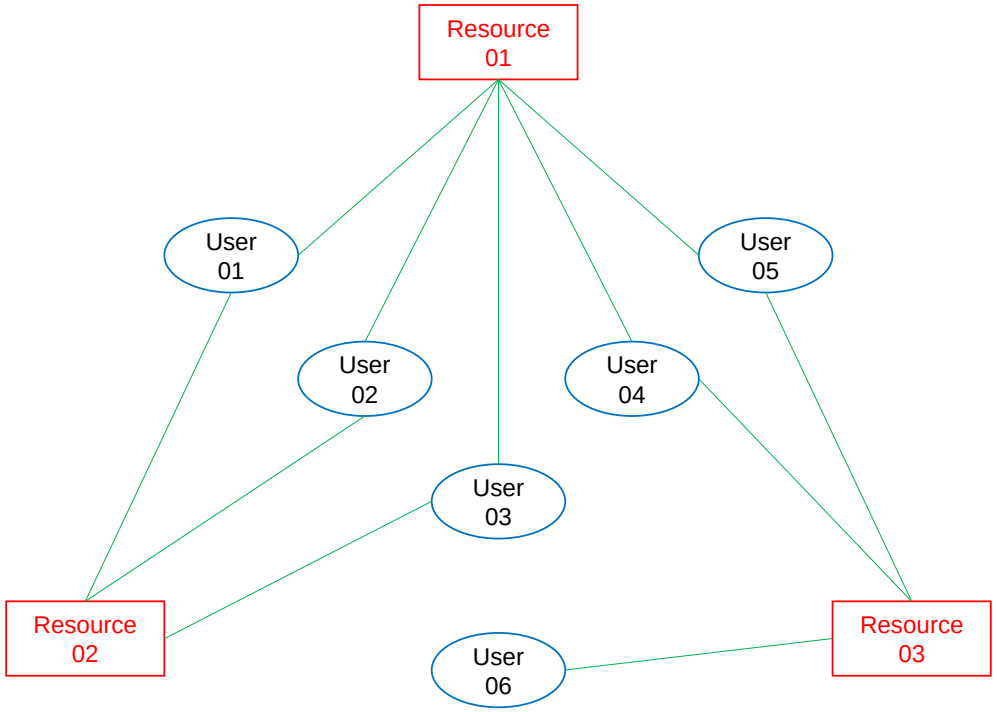


Figure 3.6: A Credit Graph model.

propose a credit function based on the popularity of WWW resources on CT platforms – users who have interacted with popular or well received WWW resources have a higher tendency to be experts themselves. Thus, these users should be given a higher credit for discovering and interacting with popular resources.

In the context of our research on CT platforms, the number of user annotations that a WWW resource received can be used as a signal for a popularity measure [140]. Alternatively, for social bookmarking CT platforms like Delicious, this can be also signalled via the number of unique users who have bookmarked a WWW resource. Thus, this research proposes a popularity variant of the credit function (Pop) that counts the number of interactions in a resource as the credit score. This is formulated as Function 3.12 for the user interaction i performed on WWW resource r with I user interactions.

$$\mathbf{Credit}_{\mathtt{Pop}}(i) = |I| \tag{3.12}$$

Unlike the other credit variants to be discussed later, the popularity variant would only require simple Boolean information – if users have interacted with a resource to assign credit without further processing. It is an approach which is content-agnostic of user interactions, thus suitable for platforms where user privacy is a concern or which are of limited access. Besides that, it is not challenged in scarceness of user interactions on some WWW resources as this scarceness denote lower popularity.

Temporal-Ordered Credit

Unlike the earlier popularity credit function, the temporal-ordered credit function requires user interactions to contain temporal information. If such information is available, it can be leveraged to improve the credit weights of user-resource links.

The credit function for the SPEAR algorithm [109] is temporal-based with the assumption that expert users are early discoverers of popular documents [155] – captured as the Discoverer-Follower concept discussed in Section 3.4.1. Thus, the function weights user-resource relations of early discovery higher than that of later ones. For a WWW resource r with user interactions I, the credit gained for interaction $i_x \in I$ can be determined according Function 3.13. It is possible to not directly process the temporal information of user interactions by just storing the order in which the user interactions are performed on the resources, maintaining a content-agnostic approach.

$$\mathbf{Credit}_{\texttt{Temp}}(i_x) = |\{i \in I | \mathbf{Time}(i) > \mathbf{Time}(i_x)\}| \tag{3.13}$$

Similarity Credit

While the earlier discussed two credit function variants are content agnostic for user interactions, the similarity credit approach utilises the additional information of annotation terms from the interactions to weight the user-resource relation. Here, this research measures the agreement between a user's annotation in the interaction with that of a general consensus in annotations of every other user interactions of the same WWW resource. The consensus is based on the 'Wisdom of the Crowd' [135] for user annotations as the ground truth.

A user contribution to the community through discovery and organisation with annotation needs to be in agreement with the community's judgement on the same WWW resource. Being just an early discoverer would not be sufficient to achieve high credit if the interaction does not agree with the community especially if the user annotations are spam. The proposed Function 3.14 measures the agreement of user interaction i_x with the user interactions I (with $i_x \in I$) by other users on WWW resource r.

$$\mathbf{Credit}_{\texttt{Sim}}(i_x) = \sum_{i \in I} \mathbf{Similarity}(i_x, i) \tag{3.14}$$

This agreement can be obtained through any content-based similarity measure. For our exploration, this research measures the agreement using the cosine similarity approach discussed in Section 3.2.1.

3.6 Annotation-Based Retrieval Algorithms

The information potential of user annotations discussed in Section 2.2 enables annotation-based retrieval. In the scope of our research, AR relies on the relevance measure between user query terms and the annotations associated to WWW resources [28] without the need to process the content itself. This content-agnostic retrieval approach can be performed using simple similarity measures like the Okapi BM-25 (see Section 3.2.2) used in this research for retrieval [98].

Using this simple but effective AR approach, this research evaluates the impact of estimated user expertise, inferred WWW resource quality and information quality of user annotations during retrieval. These measures are incorporated into the similarity measures as to be discussed in the later subsections. We acknowledge other available approaches to aggregate new measures to AR such as rank aggregation [5, 40] but wish to employ simpler approaches to better gauge the direct impact of the proposed measures.

3.6.1 Baseline: Okapi BM-25 Ranking

The simple but effective Okapi BM-25 [98, 120] ranking algorithm is used as the baseline for our evaluation. We detailed the algorithm in Section 3.2.2 and it will be used to predict

the relevance of WWW resources according to their attached user annotations given a user query. The BM-25 is chosen as our baseline for AR on CT platforms due to the following reasons:

- An effective approach for probabilistic IR [98].
- The inclusion of an inverse document frequency component helps to identify common or popular user annotations. This is relevant for the unstructured and unmoderated nature of UGC platforms that could be vulnerable to spam.
- It is expandable to include additional estimates directly such as WWW resource popularity, information quality of user annotations and user expertise.

3.6.2 Baseline: Popularity Ranking

Popular WWW resources are often desired during retrieval [143], possible due to their high authority [98, 140]. On CT platforms, the number of user annotations $|I|$ associated with a WWW resource r can be a signal of its popularity with Function 3.15. Within a given collection, the popularity of WWW resources are often normalised.

$$\textbf{Popularity}(r) = |I| \tag{3.15}$$

This research incorporates the popularity of WWW resources into the measure of relevance during AR as a composite score that extends from the BM-25 score. A WWW resource r would obtain a high rank in the results during retrieval if it is relevant to the user query A or if it is of high popularity. The linear composition approach in Function 3.16 helps us determine the impact of resource popularity during AR as we shift the popularity modifier β from one end to another. The resulting values are then normalised.

$$\textbf{PopularityBM25}(r, A) = (1 - \beta) \cdot \textbf{BM25}(r, A) + \beta \cdot \textbf{Popularity}(r) \tag{3.16}$$

We also explore a multiplicative approach for WWW resource popularity through Function 3.17. Under this approach, a WWW resource would need to be both relevant to the query and be of high popularity to be ranked higher during retrieval. The exponent β allows us to control the amount of influence the popularity measure for the resource has on the baseline BM-25 score.

$$\textbf{PopularityBM25}(r, A) = \textbf{BM25}(r, A) \cdot \textbf{Popularity}(r)^{\beta} \tag{3.17}$$

3.6.3 Proposed: Inferred Quality Ranking

The quality of WWW resource is a by-product from the link analysis of user-resource relations from SPEAR algorithm [97] and by extension, the proposed Credit Graph approach discussed in Section 3.5. To the best of our knowledge, this is not explored in any prior works due to the focus on user-user relations [155] with only the estimation of user expertise alone.

During retrieval, WWW resources of high quality are desired to better meet the information need of the user [98]. On the other hand, WWW resources of lower quality can contain misinformation due to the author's negligence or malicious intent [155] on the unmoderated environment of UGC platforms. Thus, relevant resources that are of high information quality should be ranked higher.

Similar to the earlier approach for the composite popularity discussed earlier (Section 3.6.2), this research introduces the composite scores with WWW resource quality – a linear

composition shown in Function 3.18 and a product composition in Function 3.19. These variables incorporate the inferred quality of WWW resource r from the credit graphs for the ranking of WWW resources during AR given the user query A.

$$\mathbf{QualityBM25}(r, A) = (1 - \beta) \cdot \mathbf{BM25}(r, A) + \beta \cdot \mathbf{Quality}(r) \tag{3.18}$$

$$\mathbf{QualityBM25}(r, A) = \mathbf{BM25}(r, A) \cdot \mathbf{Quality}(r)^{\beta} \tag{3.19}$$

3.6.4 Proposed: Expertise Weighed Annotation Ranking

For annotation-based retrieval, the relevance of WWW resources is measured according to the similarity between user annotations associated to the resource and that of user query terms [8] where the user annotations act as resource descriptors for the content-agnostic retrieval. Similarity measures with user annotations do however face the following challenges:

- User annotations from unreliable and non-expert users are noisy for information systems [109, 155].
- Annotations terms are varied from the users' unrestricted vocabulary.
- Term frequency could be unreliable due to – possible inflation of term frequency by spammers [59] or manipulation by malicious users such as Trojans to mislead users [155].

Given an estimation of user expertise, it is possible to weight the user annotations according to the expertise of the annotators. Besides that, the inclusion of user expertise would also carry forward an estimate of resource quality from mutual reinforcement relation [16] between user expertise and WWW resource quality.

For this research, the estimated user expertise from the credit graph of Section 3.5 is used in conjunction with the best credit function determined from Section 3.7 for AR. As the value of user expertise $\mathbf{Expertise}(u)$ lies in the range of $[0, 1]$, this research proposed the Function 3.20 to weight annotation terms a in resource r with user interactions I based on its annotator u. The user expertise $\mathbf{Expertise}(u)$ is scaled to a sum equivalent to the total term frequency and parameter γ is introduced to control the effect of user expertise.

$$\tilde{\mathbf{tf}}(a, r) = \sum_{i=(u,r,a,t)} \gamma \cdot \mathbf{Expertise}(u_i) \tag{3.20}$$

3.7 Annotation-Based Retrieval: Training Phase

In this section, we conducted training and evaluation of the various proposed AR approaches as seen in Section 3.6, including the possible credit functions introduced in Section 3.5.1 for the credit graph. The methodology for the training phase of the research is outlined in Section 3.7.1.

The training step enable this research to optimise, evaluate and discuss the values of several parameters which include:

- Okapi BM-25 free parameters of k and b.
- Credit modifier α.
- Popularity or quality modifier β.
- Expertise modifier γ.

3.7.1 Methodology

The training phase is modelled after an AR process – given a user query q, we need to rank and retrieve relevant WWW resources from the collection. The user annotations on WWW resources are used as user queries [8] which we found to be a good representation of user queries in Section 2.4.4. The annotation terms are pre-processed with stop word removals and also stemming to their root [98].

For each user annotation interaction $i = (u, r, a, t)$ by user u at time t, the algorithms try to retrieve the resource r using the user annotations a as the user query. The Mean Reciprocal Rank (MRR) performance of the AR approaches are then evaluated.

From the collection detailed in Section 2.4.2, the cut-off date of 1st January 2012 was selected – user annotation interactions before the date is used to estimate the user expertise and WWW resource quality through the credit graph variants. The remaining user interactions are used to generate user queries with a total of $81,441$. This temporal separation of the collection allows the research to model a real-world AR scenario.

Mean Reciprocal Rank (MRR)

The Mean Reciprocal Rank (MRR), a standard known-item-search evaluation measure is used to optimise each of the explored AR approach. Higher MRR values correspond to a better performance for retrieval. For each evaluated approach, the MMR performance is calculated according to Function 3.21 for a sample of Q queries.

$$\mathbf{MRR} = \frac{1}{Q} \sum_{n=1}^{|Q|} \frac{1}{\mathbf{Rank}(r)} \tag{3.21}$$

3.7.2 Optimisation of the Okapi BM-25 Parameters

The Okapi BM-25 [120] is used as the similarity measure for AR as discussed in Section 3.2.2. This research first determines the optimal values for the free parameters k and b [98] with a parameter sweep over the training data. These values are usually set to $k_1 = [1.2, 2.0]$ and $b = 0.75$ as reasonable values in absence of complex optimisation [98]. This research reports that the optimal values of the free parameters are $k = 0.1$ and $b = 0.1$ giving a MMR performance of 0.0739 for the BM-25 retrieval on the dataset over the MMR below 0.065 using the default values. This acts as the standard baseline for a simple generic AR.

3.7.3 Optimisation of the Credit Modifier

The credit modifier α controls the impact of the assigned credit for each user-resource link according to the users' annotation relation as shown in Function 3.10 and Function 3.11. The credit modifier is optimised individually for each of the credit function variants introduced in Section 3.5.1 in order to achieve the best performance.

The optimisation is based on the measured MRR performance for AR on the training dataset. The optimisation process for the credit modifiers also considers the other parameters such as the quality or popularity modifier β via a parameter sweep. For example, the optimisation for the temporal-ordered (Temp) credit function discussed in Section 3.5.1 is visualised in Figure 3.7.

To the best of our knowledge, the credit modifier was not discussed in the SPEAR algorithm [109] with the authors using a square root transformation (credit modifier value of $\alpha = 0.5$). From this finding, we saw that the best MRR performance of 0.0776 can be obtained with a log transformation for the credit modifier instead of explored value of

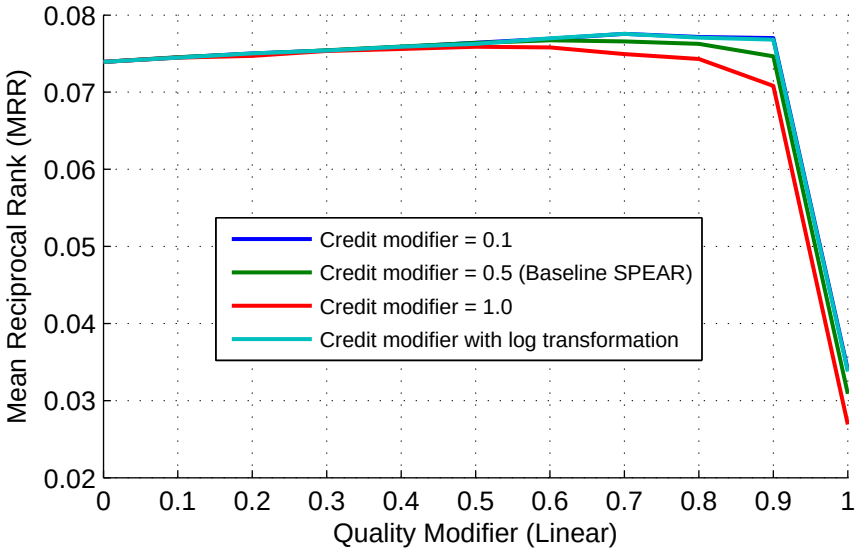


Figure 3.7: Mean Reciprocal Rank (MRR) against quality qodifiers (linear combination) for Temporal-Ordered (Temp) credit function.

$\alpha = 0.5$ by the authors [109] with several other credit modifier values outperforming the default values suggested by the authors.

To summarise, we found that the logarithmic transformation performs the best for the credit variants based-on the linear combination approach with varying quality modifiers for each of the credit function variants. Such observation can be attributed to the less skewed distribution of credit on WWW resources from the varied user interactions.

- $\beta = 0.2$ with the Popularity (Pop) credit function for a MRR performance of 0.0739.
- $\beta = 0.7$ with the Temporal-Ordered (Temp) credit function for a MRR performance of 0.0776.
- $\beta = 0.1$ with the Similarity (Sim) credit function for a MRR performance of 0.0739.

3.7.4 Optimisation of the QualityBM25 Parameters

The inferred WWW resource quality **Quality**(r) can be incorporated into AR as discussed in Section 3.6.3 – through a linear combination (Function 3.18) or a product combination (Function 3.19) with the Okapi BM-25 similarity measure. Following the optimisation of the free parameter in Section 3.7.2, we set the values of $k = 0.1$, and $b = 0.1$ in the optimisation of this QualityBM25 approach. The choice of credit modifier is set to the optimal logarithmic transformation from the findings in Section 3.7.3.

The baseline here is the approach with only the similarity measure from the Okapi BM-25 between the user query terms and the user annotations of WWW resources without any quality aggregation i.e. when the value of the credit modifier $\beta = 0$ in the linear combination. The results for the quality modifier optimisation with the linear combination aggregation are as plotted in Figure 3.8.

We also evaluate the performance of the product combination between the similarity measure and inferred WWW resource quality in Figure 3.9. As discussed earlier, the product combination requires both the similarity measure and the WWW resource quality to be accurate as a direct influence towards the ranking.

In both the linear and product combination for similarity measure with inferred WWW resource quality, we observe the temporal-ordered credit variant (Temp) to be the best

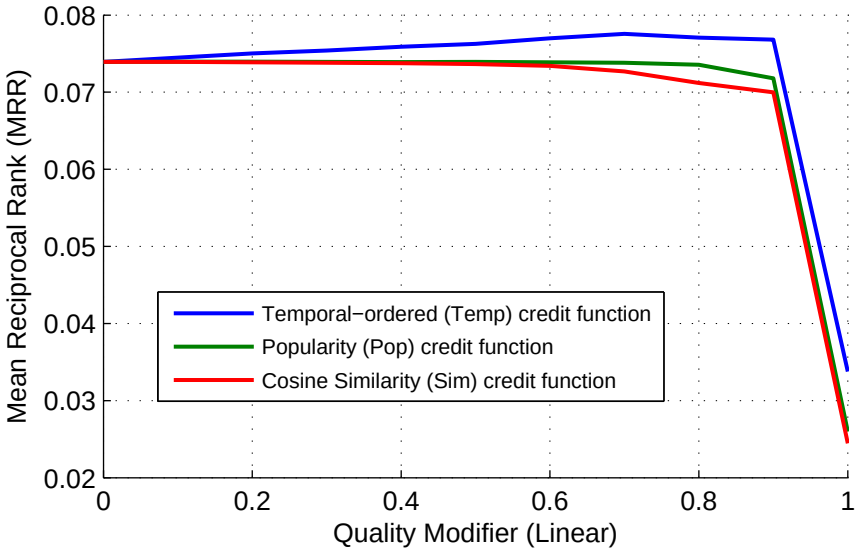


Figure 3.8: Mean Reciprocal Rank (MRR) against quality modifiers (linear combination) for credit functions.

performing variant at different values of quality modifiers. This observation is discussed further in the following Section 3.7.6 and 3.7.7.

3.7.5 Analysis: Baseline Okapi BM-25 vs QualityBM25 and Variants

Following the optimisation performed in the earlier sections, we analyse the impact in the addition of inferred WWW resource quality towards annotation-based retrieval. Our findings include:

- From the linear combination as seen in Figure 3.8, the MRR performance increases with the inclusion of the inferred WWW resource quality using the temporal-ordered credit variant – from 0.0739 (baseline without any quality) to 0.0776 when $\beta = 0.7$ (quality is dominant). The other credit function variants however recorded a performance drop when the quality modifier increases (resource quality being more dominant over the similarity measure). Overall, there is a positive impact for the estimated resource quality obtained through the credit graph approach with the Temp credit function variant on annotation-based retrieval performance over the baseline similarity measure of BM-25 only.

- For the product combination approach to QualityBM25, we again observe that the Temp credit function is the only credit function which performs above the BM-25 baseline. At the optimal quality modifier of $\beta = 0.1$, it performs at a MRR of 0.0773 above the baseline of 0.0739 by 4.6% indicating that Temp credit function is suitable in the estimation of resource document quality over the other credit functions.

3.7.6 Analysis: Temporal Information and Interaction Content

As seen from the findings in both Figure 3.8 and Figure 3.9, the Temporal-oredered (Temp) credit function outperforms the popularity-based (Pop) credit function and the Cosine-Similarity (Sim) credit function at their optimal point. This finding can be interpreted that:

- The count and/or the order of user interactions with resources are sufficient in estimating WWW resource quality. We saw the performance of the linear combination

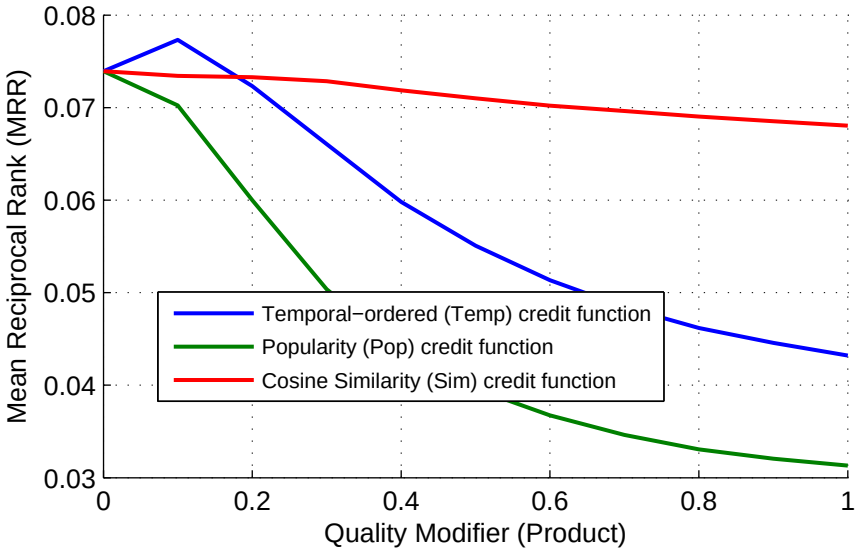


Figure 3.9: Mean Reciprocal Rank (MRR) against quality modifiers (product combination) for credit functions.

in Figure 3.8 for both the Temp and Pop credit functions when the influence of quality increases. This content-agnostic estimation of credit through the temporal feature is efficient without the need to process the annotation term similarity.

- The Sim credit function is however more stable when the impact of the similarity measure between user query and user annotations of WWW resources is reduced as seen in Figure 3.9. Besides that, it is able to outperform the Pop credit function in the product combination while remaining competitive in the linear combination.
- The Temp credit function outperforms the Pop credit function in both combinations by a notable margin. This suggests the need to differentiate user interactions despite them having interacted with the same WWW resource. Our findings here in fact support the proposed Discoverer-Follower concept of the SPEAR algorithm in Section 3.4.

3.7.7 Analysis: Temporal Ordering of User Interactions

Following the findings in the improved performance with the Temp credit function over the other credit function, this research asks the following question – Does the temporal ordering of user interactions based on the Discoverer-Follower concept matter?

To answer this question, we introduced a new temporal-ordered variant in Function 3.22. This is the reversed temporal-order (-Temp) variant where the late followers are rewarded with more credit over the early discoverers. Our observations found that this variant significantly reduce the AR performance with values below the baseline. This validates the significance of the discoverer-follower concept proposed by the SPEAR algorithm [109] and used in the Credit Graph.

$$\mathbf{Credit}_{\texttt{Temp}}(i_x) = |\{i \in I | \mathbf{Time}(i) < \mathbf{Time}(i_x)\}| \tag{3.22}$$

3.7.8 Optimisation of the ExpertiseBM25

The research proposed a novel new approach to AR with user expertise, the ExpertiseBM25 approach detailed in Section 3.6.4. Here, the estimated user expertise of the annotators is

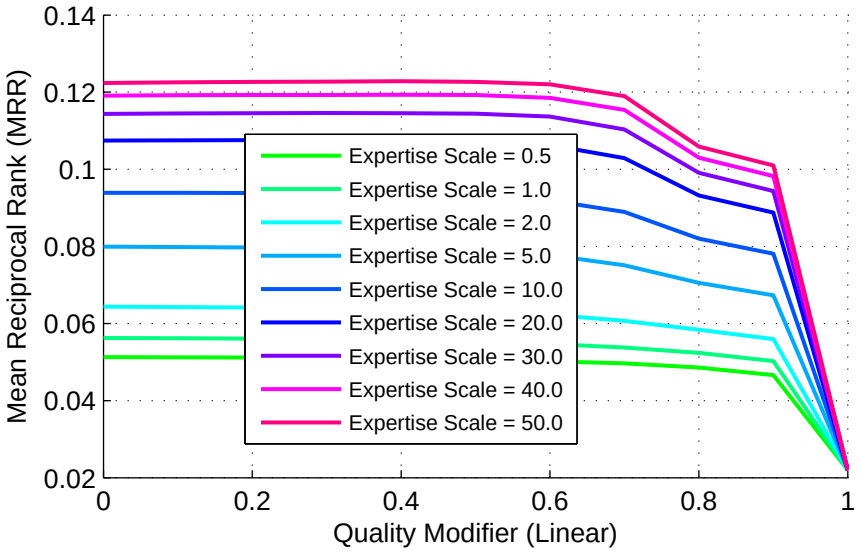


Figure 3.10: Mean Reciprocal Rank (MRR) against quality modifiers for Temporal-Ordered (Temp) credit function with ExpertiseBM25 for annotation weights and QualityBM25 for document scores.

used to score or weight annotation terms (see Equation 3.20) instead of the commonly used frequency [98]. We present our preliminary training evaluation and optimisation in Figure 3.10; using the best performing Temp credit function, linear combination and parameters from QualityBM25 discussed earlier.

We observe that the ExpertiseBM25 performs the best without the influence of inferred WWW resource quality from the credit graph when $\beta = 0$. On the other hand, scaling the impact of the estimated user expertise up by increasing the value of γ does improve AR performance. We explore and analyse this further in Section 3.7.9.

3.7.9 Analysis: Term Weights with User Expertise

From earlier findings, we saw an increase in MRR performance as the expertise modifier γ increases. In Figure 3.11 we looked to discover the optimal expertise modifier value and found this value to be when $\gamma = 191624$ before a dip in MRR performance.

This value is found to be the same as the number of users in the training collection for the credit graph, a total of $191,624$ users. Recall that the user expertise value $\textbf{Expertise}(u)$ is normalised at each iteration of the algorithm (see Function 3.9), resulting in the average expertise value of $\frac{1}{|U|} = \frac{1}{191624}$. Thus, the expertise modifier of $\gamma = 191624$ will result in an average user expertise of 1 which is the same as the term frequency of 1 during retrieval. This supports the optimal expertise modifier γ value observed in our optimisation.

3.7.10 Analysis: Baseline Okapi BM-25 vs Expertise

By using the user expertise of the annotations in the similarity measure instead of the term frequency, the score of annotation terms are redistributed – higher score for user annotations which better describe the WWW resources. This significantly reduces the noise of bad descriptors and also the unwanted personalised terms from users of lower expertise. Instead, the annotation terms from expert users are used to more reliably describe the WWW resources.

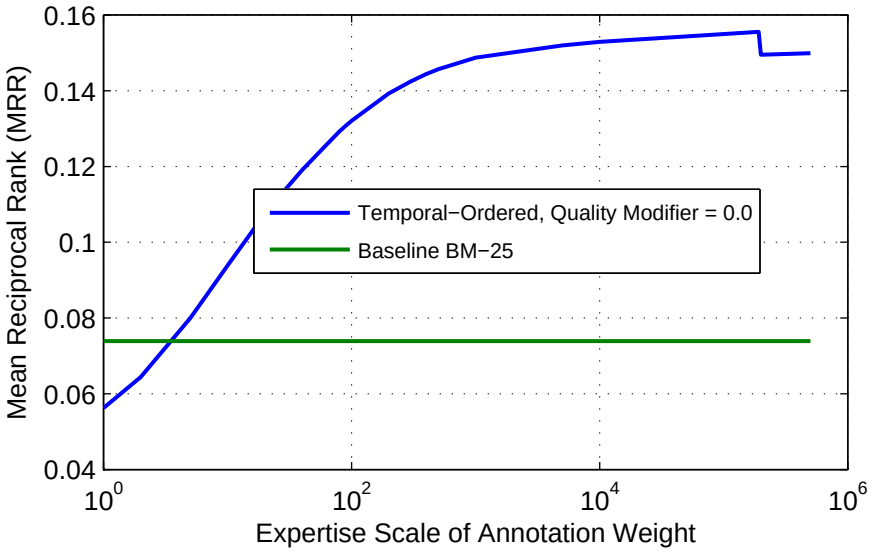


Figure 3.11: Mean Reciprocal Rank (MRR) against expertise modifiers for Temporal-Ordered (Temp) credit function.

Consider a possible scenario below: A WWW resource about basketball annotated by two users - an expert (**Expertise**$(u_x) = 0.00001$) with annotations 'Basketball, NBA, Sport' and a non-expert **Expertise**$(u_y) = 0.000001$ with annotations 'Game, Ball'. Using $\gamma = 191624$, the terms annotated by u_x will receive a weight of 1.916 (weighted according to the expertise of the annotator) which is close to 2 whereas the terms annotated by u_y will receive a weight of 0.191624. We can see that the annotations contributed by user u_x is more meaningful and thus will better describe the WWW resource when a basketball related query is given.

With that, we can deduce from our findings that the ExpertiseBM25 approach would outperform the baseline Okapi BM-25 measure drastically with a MRR performance of 0.1487 against 0.0739 of the baseline. This is an improvement of over 200%. With this finding, we conclude that the users of higher authority (expertise) are able to describe WWW resources better and directly contribute to improved annotation-based retrieval.

3.7.11 Overall Result: Popularity vs Quality or Expertise

Taking the best performance of every AR approach discussed in Section 3.6 and their variants, the AR performance can be summarised in table 3.1.

Approach	MRR	Statistical Significance over Baseline (P-value)
Baseline Okapi BM-25	0.074	Baseline
PopularityBM25, $\beta = 0.1$	0.076	0.028
QualityBM25 (Temp), $\beta = 0.7$	0.078	8.34e-06
ExpertiseBM25 (Temp), $\gamma = 191624$	0.156	2.2e-16

Table 3.1: Annotation-based retrieval (AR) performance: Mean Reciprocal Rank (MMR) with statistical significance.

We observe that while the popularity approach (PopularityBM25) outperforms the baseline where popular WWW resources are given higher ranks during retrieval, it still

performs below the proposed QualityBM25 and ExpertiseBM25. A possible reasoning for this is that popular WWW resources that are with high amount of user interactions are interacted by a wider range of users – possibly more non-expert users than expert users. These non-expert users may not describe the WWW resource well for AR. The QualityBM25 takes this into account where the inferred WWW resource quality takes into account the expertise of the users who have interacted with the WWW resource; and the measure do differ from that of popularity as seen in Section 3.4.4.

On the other hand, the ExpertiseBM25 approach is able to rank less popular WWW resources higher if the resources are of high authority; signalled by the experts who interact with the resources themselves. These resources of higher quality could be discovered and retrieved easier during retrieval. Thus, from the findings here, this research suggests for the proposed approaches to be used for AR to better meet the information need of the users better.

3.8 Annotation-Based Retrieval: Testing Phase

Following the optimisation performed in the training phase, this research evaluates the annotation retrieval performance of the proposed approaches introduced in Section 3.6. We outline the evaluation methodology and then discuss our findings.

3.8.1 Methodology

The Delicious dataset from Section 2.4.2 is used as the corpus collected for AR with the AR algorithms discussed earlier. Each approach is optimised according to the findings from the training phase in Section 3.7 for comparison.

100 user queries were randomly selected from the set of overlapping queries discussed in Section 2.4.4 without any edits in order to emulate real world user queries for AR. The queries were pre-processed for cleaning such as stop words removal and then stemmed to their root terms. Example of the queries explored include:

- 'health and fitness'
- 'job search'
- 'teaching resources'

For each query, WWW resources on the corpus are scored and ranked according to the optimal AR algorithms explored. These retrieved resources are then manually evaluated according to their relevance to the user queries as binary yes-no [98].

Precision Measure

The precision measure as shown in Function 3.23 is used as the evaluation measure for AR algorithms. Here, we looked at the precision up to 10 results **Precious@**10 as users are unlikely to look at search results beyond the first page [13]. The **Precious@**10 denotes the fraction of retrieved resources that are relevant in a ranked list (created with the explored AR algorithms) with the 10 best WWW resources for a given user query.

$$\textbf{Precision} = \frac{|\textbf{RelevantResource} \cap \textbf{RetrievedResource}|}{|\textbf{RetrievedResource}|} \tag{3.23}$$

The precision measure used for the evaluation here is different from the MRR measure used for the training phase in Section 3.7.1. Unlike the training phase where we attempt to retrieve the WWW resources that a user has annotated using that user annotation

as the query; we are using new real world queries[2] which could result in multiple relevant resources. Therefore, the precision measure is a more appropriate measure for our evaluation than MRR.

3.8.2 Results and Discussion

The AR performance for the explored algorithms are presented in Table 3.2 – using the **Precious@**10 measure, accompanied with a statistical significance of the results using the Paired-2-Sample T-Test with $\alpha = 0.05$, comparing the retrieval performance of the same query on the same collection.

Approach (Optimized)	Average Precision@10 (±Stderr)	Statistical Significance over Baseline (P-value)
Baseline Okapi BM-25	0.706 (±0.023)	Baseline
PopularityBM25	0.717 (±0.022)	0.092
QualityBM25	0.732 (±0.023)	0.010
ExpertiseBM25	0.754 (±0.021)	2.802E-06

Table 3.2: Annotation-based Retrieval: Precision@10 with Statistical Significance.

It can be observed that all the explored AR approaches are competitive amongst themselves with the ExpertiseBM25 approach as the best performer. All of the additional proposed enhancements such as the WWW resource popularity, WWW resource quality and estimated user expertise do improve the AR performance over the baseline Okapi BM-25 approach. This finding is consistent with the observation in the training phase (Section 3.7).

3.8.3 Analysis: Popularity Enhancements

By incorporating the WWW resource popularity into retrieval, it is possible to improve the AR performance as the popular WWW resources are often desired by the users [140]. We note that the PopularityBM25 only perform marginally better than the baseline BM25, unlike during the training. The performance increase is not significant. This could be due the wrongly inflated scores for popular WWW resources that are of lower relevance to the user queries.

3.8.4 Analysis: Quality Enhancements

Similarly, the QualityBM25 does outperform the baseline Okapi BM-25 approach and the PopularityBM25 approach. The inferred WWW resource quality from the credit graph algorithm can better identify WWW resources to meet the information need of the users better. We saw in Section 3.4.4 that the calculated popularity and the inferred quality measure of WWW resources are different, but the finding here supported the notion that the inferred quality measure could better meet the information need of the users. In fact, we saw that the optimal QualityBM25 measure is achieved when the quality inferred WWW resource quality is dominant with the quality modifier at $\beta = 0.7$.

[2]From TREC One Million Query 2009.

3.8.5 Analysis: Expertise Enhancements

From Table 3.2, we observe that ExpertiseBM25 is the best performing approach to AR. This performance is statistically significant over the baseline Okapi BM-25 algorithm while outperforming the other approach by a comfortable margin.

The ExpertiseBM25 approach incorporated the estimated user expertise directly into the similarity measure by having the user expertise weight the annotation terms instead of frequency-based weights. As user annotation terms are used as descriptors for WWW resources, weighing these annotations according to the expertise of annotators improves the similarity measure as it is now able to better judge the content of a WWW resource – expert users are better at describing content than non-expert users. Thus, WWW resources with higher relevance to the user queries can be ranked higher through the improved similarity measure. On the other hand, the other approaches such as PopularityBM25 and QualityBM25 are held back by the frequency-based similarity measure.

Besides, the ExpertiseBM25 indirectly incorporates the QualityBM25 strength as high quality WWW resources have a high number of expert user interactions. This is reflected during the similarity measure where the higher annotation term weights from expert users increases the relevance score of these WWW resources.

3.9 Conclusion

We investigated the annotation-based retrieval (AR) for WWW resources. In AR, the wisdom of the crowd is leveraged on through user annotations which has the information potential (Section 2.2) to describe and categorise WWW resources on collaborative tagging (CT) platforms. This enables the content-agnostic retrieval of diverse, complex and unstructured WWW resources as user annotations are representative of user queries (see Section 2.4.4). As user annotations are user-generated however they could be unreliable as discussed in 1.3. Instead of changing how AR works through complex algorithms, this research aims to improve the information quality of user annotations for AR and evaluate the possible improvements.

This research looks to estimate the expertise of users on CT platforms; and then use the estimated user expertise to better infer the information quality of their annotations. A Credit Graph model is proposed in Section 3.5, inspired by the Spamming-Resistant Expertise Analysis and Ranking (SPEAR) algorithm [109]. Our contribution here is the introduction and study of possible features for user expertise estimation through credit functions in Section 3.5.1. A by-product from our effort is the inferred quality of WWW resources as well. This research then incorporates all of these measures into AR – which to the best of our knowledge, there is no other works that explore such addition for AR.

The finding from the training phase in Section 3.7 provided our research with a few insights. Firstly, the addition of information quality measure and user expertise estimates into simple similarity measures like the Okapi BM-25 [98, 120] do improve AR performance. The inferred quality of WWW resources can help identify relevant WWW resources to meet the information need of the user better, even when compared to popular WWW resources as these measures do differ (see Section 3.4.4).

On the other hand, the user expertise is directly used to adjust the weights of annotation terms during retrieval instead of the commonly used term frequency as not all annotations are the same – where annotations of expert users are given a higher weight when compared to non-expert users. This answered our research questions detailed in Section 3.1 if expert users are able to describe or categorise WWW resources better.

The improved AR performance supports our explored approaches for the content-agnostic estimation of user expertise where CT features can be leveraged on instead of processing the user annotations themselves. We found that the temporal-ordering of user

interactions does provide a strong signal for user authority, over the popular of the WWW resource that they have interacted with or even the more complex processing for the agreement checking in user annotations with the community. Indirectly, our work here validates the Discoverer-Follow concept by the authors of the SPEAR algorithm discussed in Section 3.4.1. While this research found no improvement in the estimation of user expertise with the additional information from the content processing of user annotations, we believe that there is a potential in processing these information for the further enhancement such as topic models. Finally, we discover that the user expertise values should be normalised to a total average of 1 as we optimise the proposed ExpertiseBM25 approach.

The findings from the testing phase in Section 3.8 are consistent with the findings from the training phase. We saw the positive impact from the proposed WWW resource quality and user expertise measures towards AR over the baseline approaches.

It can be concluded that (1) it is possible to estimate the user expertise on CT platforms through content-agnostic means; (2) the information quality of user annotations on CT platforms can be inferred through the expertise of the users; and (3) expert users are able to better describe and categorise WWW resources. The results from our early research here motivates further work into the estimation of user expertise on other user-generated content (UGC) platforms such as community question-answering (CQA) platforms in Part II and content aggregation (CA) platforms in Part III of the thesis.

Part II

Community Question-Answering (CQA)

Chapter 4

Retrieval on Community Question-Answering (CQA)

Community Question-Answering (CQA) platforms enable the users to perform information retrieval (IR) with natural language questions instead of query keywords; thereby allowing them to meet information need that could not otherwise be satisfied by traditional search-based IR [95]. It is a user-generated content (UGC) platform where it is the users as answerers that create and contribute information as a response to the questioner's information need. The known CQA platforms on the World Wide Web (WWW) include:

- Quora[1]
- Stack Overflow[2] and the Stack Exchange[3] network
- Yahoo! Answers[4]

CQA leverages the inherent wisdom of the crowd (WoTC) to retrieve quality information from expert users amongst its community. Alternatively, since many users desire answers that are similar to that of previously answered questions, CQA platforms can also act as an excellent source of knowledge where traditional IR approaches are used to retrieve from the CQA archive [27, 68, 159]. These two distinct approaches are currently the commonly used approaches to IR on CQA platforms. The information potential of user questions and answers on CQA platforms are discussed further in Section 4.2.

The emergence of CQA platforms on the WWW can be attributed to the willingness of its users and the community to share their knowledge on the platform for the learning of others [159]. Despite the anonymity of the users, the contributors on CQA platforms [82, 84] include the expert users who are identified relative to the other users [113]. These expert contributors do differ from the average users [101] but it should be noted that active users regardless of their expertise are the main driving force of CQA platforms through their generated content [95].

In this Section 4.1, we first outline the general structure and representation of a CQA platform – the users with their generated questions or answers; and the user votes that moderate the content. Previously, there have been numerous attempts to infer the quality of user answers directly as discussed in Section 4.2.3. Alternatively, this research suggests that answer quality can be inferred from the expertise of the answerer. Thus, in Section 4.3, we discuss and explore the past literatures and attempts in the estimation of user expertise that have been applied for expert search [163].

[1] http://www.quora.com/
[2] http://stackoverflow.com/
[3] https://stackexchange.com/
[4] http://answers.yahoo.com/

We then proceed to obtain and study a previously used CQA dataset in Section 4.4.1 – the Chiebukuro Data (2nd edition)[5] that has been provided by the National Institute of Informatics by Yahoo Japan Corporation. The first edition of the dataset was used for the evaluation of Liu et al's research [95]. This dataset (the 2nd edition) is used later in Chapter 5 and Chapter 6 to evaluate the performance of our proposed approaches for the estimation of user expertise and answer quality.

4.1 Structure and Representation

Contents on CQA platforms are usually based on the questions asked by the users of the platform. Each question is accompanied by a list of answers (zero or more), contributed by other users in their attempt to meet the information need of the questioner. Some CQA platforms are more specialised such as StackOverflow for coding based question; whereas other CQA platforms are more general with questions organised according to their topics such as a question (stated with natural language) on Quora shown in Figure 4.1.

Thus, a CQA platform is generally consist of a community of users who are able to state their information need as natural language questions. Each question retrieves a set of answers produced by answerers in their attempt to solve the problem or meet the information need of the questioner. Many of the modern CQA platforms allow the both questions and answers on the platform to be voted – positive up-vote if the content is good and negative down-vote otherwise. The best answer in a question is inferred to meet the information need of the questioner [18] or as the gold standard [159], determined according to the highest vote or selected by the questioner himself/ herself [95].

The structure of a CQA platform with two questions can be visualised as shown in Figure 4.2. Relationships between the user, the question and the answer entity can be represented with a graph structure. In the following subsections, we discuss each CQA platform entity.

4.1.1 User

The users of CQA platforms U are important, playing the role of questioners, answerers or both. The relations between CQA users are based on their question-answering interactions on the platform. As questioners, users submit questions on the platform and other users respond to the question with answers. For each question, a user can only play a single role (questioner or answerer but not both).

The types of user roles on CQA platforms can be better understood in terms of the bow tie structure [21] of the users connection graph as illustrated in Figure 4.3 with the following components adapted for CQA platforms [158]:

- *Core* or the Strongly Connected Component (SCC). The core of CQA platforms are its users who play both roles as questioners and answerers. The users are strongly connected, often asking and answering questions amongst themselves. Studies on a JAVA forum found this component to be small with only 12.3% of the users [158]. We regard these users as active users who contribute large amounts of content on the platform and thus are our utmost priority for the estimation of user expertise.

- *In*. Users in this component usually play the questioner role. Their questions are answered by other users especially users who are placed in the *core*. Unlike the WWW, over half (54.9%) of the CQA platform users are here – telling us that there

[5] http://www.nii.ac.jp/cscenter/idr/en/yahoo/chiebkr2/Y_chiebukuro.html

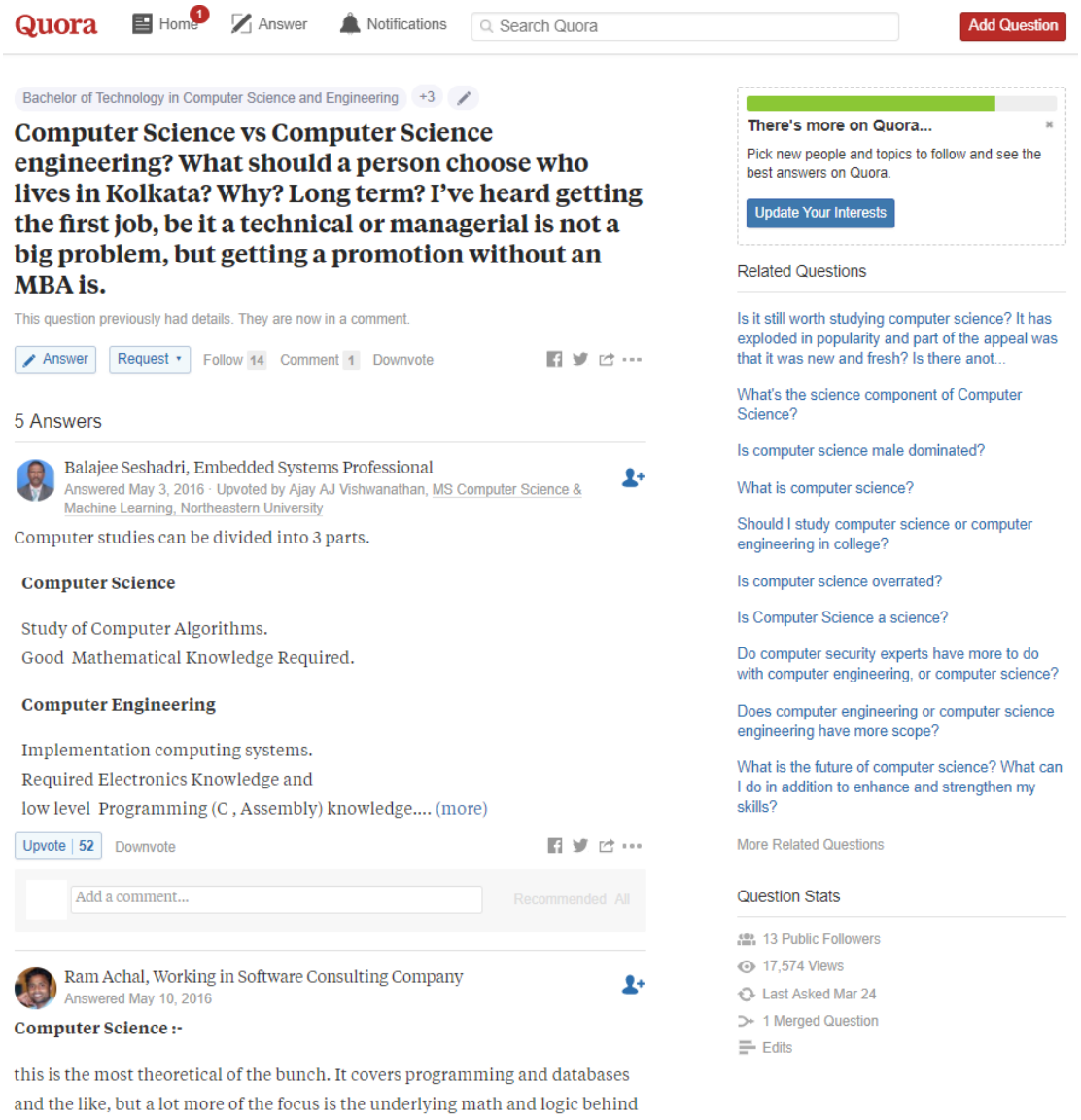


Figure 4.1: A screenshot of a natural language question asked on Quora. This is regarded as a question thread. Web page visited on 27th November 2017.

is a large need for answers and these answers are usually provided by the *core* users of the platform.

- *Out.* Users of the *out* component usually only the role of answerers. They would usually answer question by users in the *core* or the *tube*. We can assume these users to be expert contributors where they exist to meet the information need of other users without the need to gain knowledge from the other users via questions. These users are few by nature (only 13.0% of users) and further exploration found these users to be extremely active in answering questions.

- *Tendril.* Here, the users would answer the question by users from the *in* component only and thus is not connected to another user component. They made up around 17.5% of the community and can be regarded to be of low activity with little answers contributed.

- *Tube.* Users in the *tube* component are questioners who have their questions answered by the answerers of the *out* component and not by any other user. Only 0.4% of the

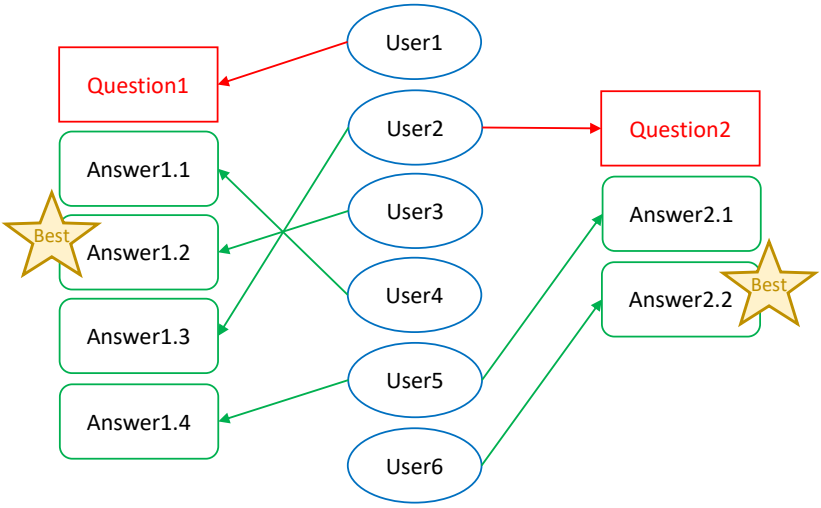


Figure 4.2: Example of a CQA structure with two question thread.

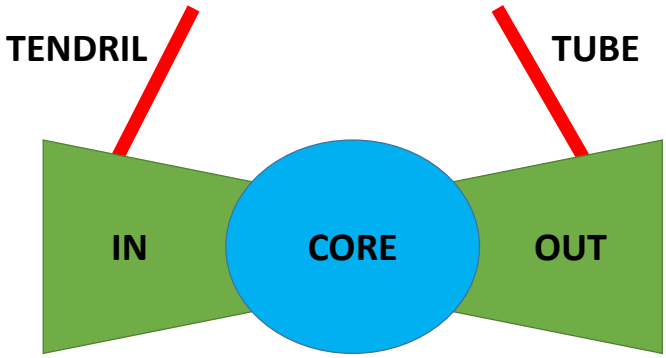


Figure 4.3: The Bow Tie structure.

users are found in this component [158] with their questions not answered by the core community. This can be due to the difficulty of their questions.

- *Disconnect.* The remaining 1.9% of CQA users are disconnected. These users are questioners where their questions are not answered by any other users of the platform, possibly due to – (1) badly asked questions; (2) difficult questions; (3) repeated questions which have been answered before; or (4) spam.

The CQA community is similar to that of other UGC platforms with a skewed distribution of user activity. There is a small amount of extreme active answerers with a high amount of answers whereas the other answerers only produce a few answers [158]. The same observation can be made for questioners.

User Expertise

The answerers on CQA platforms tend to answer questions that are asked by questioners of lower expertise with the top answerers having a adequate expertise to answer everyone's question [158]. This question-answering dynamic relies on the following assumptions about the users:

- A user answers the questions of other users according to his/ her own capabilities and expertise.
- A user knows the topic well if that user answers a lot of questions about the topic. We however note that there is an exception for trolls and spammers with their low community presence.
- A user knows the topic well if that user helps a large number of users for the topic.

This research aims to estimate the expertise of CQA users **Expertise**(U) as a quantitative measure for users as described in Definition 5. Users of higher expertise have the higher odds to produce content of information quality [93, 113]. As a relative measure, the user expertise enables the comparison between users for the following:

- Re-ranking user answers within a question to better meet the information need of the questioner. As a relative measure, a user of higher expertise would produce answers that are of higher information quality than that of another user of lower expertise [18]. Thus, we can suggest the best answer to the questioner with greater confidence.

- Identifying the expert users within the CQA community. Expert users could be encouraged to produce more high quality content on the platform [101]. Besides that, the identified experts could be directed to help questioners [93, 154] within their expertise level such as answering a difficult question that otherwise could not be answered or even discovered by other users.

4.1.2 Question

On CQA platforms, the questioners state their information needs with natural language as user questions [27, 95] instead of keyword-based queries. Upon submission, the question acts as a thread for a collection of answers by answerers in response to the question. Often on CQA platforms, the questions are organised according to the topic of the question which is selected the questioner himself/ herself [95] or determined automatically by the CQA platform using a topic model [159]. This collection of answers for each question can be beneficial to the other users who seek the same information as the questioners in future retrieval. The classification of questions for question similarity can benefit future retrieval tasks as discussed later in Section 4.2.2.

Definition 11 (Question)**.** User query stated with natural language [27, 95].

4.1.3 Answer

Answerers generate answers in their attempt to meet the information need of the questioners. Answers of high information quality are able to meet the information need of the questioners [159] as low quality answers are not able to do so [68]. We discuss the information quality of answers further in Section 4.2.3 and show how we can infer it according to the expertise of the answerers in Chapter 5 and Chapter 6.

As user-generated content, any user is able to play the answerer's role, and not all questions on CQA platforms are answered as discussed in Section 4.1.1. The information quality of user answers does vary [68, 124] for each question; with up to 30% of the answers in a question thread being unwanted [124].

4.1.4 Votes

A challenge for UGC platforms is the large amount of unstructured content and the unmoderated nature of UGC content as discussed in Section 1.3. Thus, UGC platforms

including CQA platforms rely on its community for content moderation [9]. Many of the CQA platforms of today allow its users to collectively place votes on user questions and answers as a measure of content quality [159]. These votes help the other users to identify which question to respond to or which answers to trust [68]. Later in Chapter 5 and Chapter 6, we shall look into how user votes can be used as signals for the estimation of user expertise.

Definition 12 (User Vote)**.** User's judgement on the quality or contribution of a particular content [9].

4.2 Information Potential of Questions and Answers

User questions and answers on CQA platforms contain vast amounts of content – generated by a large community of users. This UGC has the information potential to meet the information needs of its users as well as other non-participating users as discussed in the following sections.

4.2.1 Generation of New Information

Unlike traditional IR which retrieves content from a corpus such as a crawl of the WWW, users as questioners on CQA platforms play an important role in seeding the generation of new content. Questions on CQA platforms encourage new information to be generated by the answerers – which can then also be made available or retrievable through traditional keyword-based search [95]. For example, many developers turn to Stack Overflow [9, 81] for debugging new problems or bugs faced during development. Many of these new answers that help the questioners could then be retrieved in the future by other users who faced similar issues, leading us to the following discussion in Section 4.2.2.

4.2.2 Question Similarity

CQA platforms are rich repositories of information [159] suitable for retrieval especially when users seek information similar to a previously answered question. In CQA platforms, users state their queries as questions [95]; thus, meaningful information as valuable knowledge can be mined from the CQA platforms by seeking similar questions and then retrieving the answers of these questions for the users' consumption [26, 67, 153].

Often, the literature on CQA platforms attempts to measure the similarity of user questions – to organise and cluster similar questions together [27]; or to retrieve questions [159] for their answers. It is a known challenge that classic similarity measures such as the Okapi BM-25 [120] is not suitable to measure the similarity between questions due to the type of natural language used by the questioners [27]. Synonyms in natural language mean that there are few common terms between similar questions for language models [159].

Resulting from the lexical gap [67] that is problematic for similarity measures in traditional information retrieval models [159], much of the previous works focus on the similarity in the category of questions and answers [27, 79, 157]. The categorisation of questions and answers are usually based on topic models [157], often through supervised learning [159]. Alternatively, language translation models can be trained [153] and further enhanced with topic models [69] such as the latent topics from the Latent Dirichlet Allocation (LDA) [22]. Modern approaches to bridging the lexical gap also often leverage on the metadata of questions and answers [157, 159].

4.2.3 Answer Quality

Like many other UGC platforms, content on CQA platforms is often unmoderated by experts. This results in contents of varying quality on the platform which directly affects the performance of tasks such as information retrieval (through question similarity for example) that aims to leverage the knowledge on CQA platforms. The knowledge or meaningful information on CQA platforms is predominantly obtained from the answers that respond to questions; thus, answers with high information quality are desirable [159]. Similar question can be identified to leverage on the previously generated knowledge.

Studies have shown however that there is a large variance in answer quality on CQA platforms [6, 68, 124] ranging from meaningful information (knowledge) to fake answers, spam or noise. This can be attributed to the diverse user community as the main contributor of answers – with each user having different expertise and reliability [93]. As discussed in Section 4.1.1 however, there is a high number of users with low expertise [158] which results in a large portion of low quality answers [6] up to 30% [124] which further results in a reduction of knowledge density on CQA platforms.

Following such findings, work has been done to estimate the information quality of answers on CQA platforms [6, 32, 68, 96]. The authors aimed to identify low quality answers for filtering [68], while high quality answers are instead promoted during retrieval [136]. The improved estimate of answer quality can also be applied to improve information management and knowledge curation by forming valuable best question-answer pairs.

The approach of choice for the estimation of answer quality is through the metadata of the answers on CQA platforms as it is hard to process unstructured UGC content. For example, the best answer in a question can be known if CQA platforms allow the questioner to award the best answer [159] or by extracting the answer with the highest user votes upon stability being reached [68]. The extracted metadata features can be extracted to train answer quality models [128] or classifiers [6] to recognise good answers and predict the best answers.

These approaches do however suffer from the cold-start problem [127, 161] in the estimation of answer quality – for the user votes to stabilise as some answers are new (termed as cold) or for the questioner to attempt all of the answers in order to know what is the best answer. In our study of a CQA dataset in Section 4.4.7, we observe that it takes a several days for the questioners to identify the best answer. Thus later in Section 4.3, we explore how the estimation of user expertise can lead to good prediction of answer quality which overcomes the cold start problem.

Definition 13 (Cold Start Problem)**.** A situation where insufficient evidence has been obtained to ascertain the quality or topic of a new content or entity.

4.3 Estimation of User Expertise

The content on CQA platforms, both questions and answers are generated by the users of the platform playing the role of either the questioner or the answerer. As the expertise and reliability of each user differs [6, 68], the content generated by these users also differs in information quality especially with the uncertainty in user motivations [82, 84]. Studies have found a correlation between user expertise and the quality of their content, allowing for the calculation of the latter to be based on the former [18].

Expert users are more likely to contribute good content on CQA platforms, thus they are valuable in providing answers that meet the information needs of the questioners [113]. Besides that, the experts on CQA platforms are often the more active contributors [101]. While non-expert or users of lower expertise can be active as well, they tend to display a higher frequency in asking questions than answering them [158] – making them valuable as

a catalyst for new content to be generated. While we discussed the estimation of answer quality that has been written in Section 4.2.3; this section discusses the estimation of user expertise that would allow information systems to predict the information quality of answers which would in turn overcome the cold-start problem in user voting.

The expertise of CQA users can be estimated from their interactions with the platform [95, 154]. Simple approaches can be employed with basis statistics [18] such as:

- The count or ratio of questions asked.
- The number of answers contributed.
- The number of awards received such as the best answer award from the questioner.
- The number of votes received by the user from the community's long review.

In fact, such simple approaches to user expertise estimation can outperform more complicated approaches [18, 158] such as the graph-based approaches of the HITS algorithm [80] which we explore further in Chapter 5. Alternatively, the knowledge of the questioner-answerer relation enables direct comparison between user interactions such as pairwise comparison approaches [95] to build complex networks between users [10].

Much work has been done in the field of expert search and discovery – largely in the task of identifying and profiling [163] knowledgeable people for a given domain [34] within the CQA platform [93] or the user's social network [6]. Such approaches can be applied to overcome the poor expertise matching between the question and the answerer's domain or ability as it could route the questions to suitable experts [93, 154]. Expert search and discovery is a crucial component to improve knowledge creation by expert users on CQA platforms for future information retrieval tasks which we would also discuss in Chapter 5 and Chapter 6.

4.3.1 Simple Approaches

It is possible to estimate the expertise of CQA platform users using relatively simple measures. Despite the simplicity of the approaches, they are competitive in performance with other more complicated approaches such as graph-based approaches [158].

Traditional Information-Retrieval (IR)

The estimation of user expertise is often treated as an information retrieval task commonly known as expert search. User questioners are interpreted as user queries to search for the other users who best match the questions – according to their constructed user profiles [24, 38]. These user profiles are built by analysing mined user interactions on the platform particularly the content created by the users themselves. The users are then ranked based on the similarity between the questions and the user profiles [163]. Similarly, it is possible to retrieve similar answers by the similar users.

Z-Index

The Z-Index approach to scoring user expertise [158] is a simple approach that is able to outperform some complex graph-based algorithms and be competitive with the others [23]. The Z-Index considers how many times more a user answers questions than asking them. The Z-Index is based on the assumptions that:

- A user answers the questions of other users according to the user's ability. This ability corresponds to the user's expertise.
- A user knows the domain well if that user answers a lot of questions in the domain – with the exception of trolls and spammers. Thus, the Z-Index may perform poorly for platforms with high number of malicious users.

- Novice users tend to ask a lot of questions. When compared to expert users in the same domain, they have a higher probability to be asking questions. On the other hand, expert users within a domain have a high tendency to answer questions from the same domain.

- The expertise of an answerer is usually higher than the expertise of the questioner. The probability of this being the case increases as the answerer's answer gains more votes or if the answer is selected by the questioner to be the best answer for the question.

The Z-Index is suitable as a baseline approach for user expertise estimation due to the assumption that a random user is just as likely to ask a question as to answer one. It then calculates the number of standard deviations that the user's answer count lies above the expected value as described in Function 4.1.

$$\textbf{Z-Index}(u) = \frac{|\textbf{Answer}(u)| - |\textbf{Question}(u)|}{\sqrt{|\textbf{Answer}(u)| + |\textbf{Question}(u)|}} \tag{4.1}$$

where $|\textbf{Question}(u)|$ denotes the number of times the user u has asked a question and $\textbf{Answer}(u)$ denotes the number answers contributed by user u.

Vote Scores

Modern day CQA platforms allow users to place votes on questions or answers or both. These votes can be regarded as the community's long-term review of content [9] which can then be mined as a judgement of question and answer quality. Noting this characteristic, the user votes can then be used to as a signal regarding the user's contribution for estimating user expertise (e.g. by normalising or averaging user vote count for each user [95]).

Best Answer

A question thread can have one or more answers submitted by answerers. From these answers, many of the current CQA platforms are then able to determine the *best answer* after a period of time either from user votes or through selection by the questioners themselves (in the case of the Yahoo! Chiebukuro dataset) [95]. The best answers can then be used to estimate the user expertise score [18] by ranking users based on the number of best answers contributed. Alternatively, a best answer ratio for each user can be used effectively as a feature to predict the answer quality of answers contributed by that user [6, 68, 96]. The best answer approach is suitable for estimating the expertise of inactive users though outperformed by the state-of-the-art approaches in other cases [95].

4.3.2 Graph-based Approaches

Graph-based approaches have been successful in estimating the authority of entities such as the authority of web pages [80, 112]. The same concept can be applied for the authority of users as user expertise with the concept of peer assessments [158] through user interactions such as explicit question-answering relations. In the graph approach, the estimated user expertise of a user corresponds to the fraction of time a random walker would spend visiting that user as the walker iteratively follows the user interaction links from one user to another.

In the graph model, users are represented as nodes with directed edges from the questioner to the answerers in question threads [95]. The questioner acts as the hub to determine the answerers' authority [158] based on the mutual reinforcement between user

expertise and answer quality [16]. Many of the previous approaches however rely on the assumption that all answers are of equal quality. Once the user-user links are formed in the graph representation of CQA platforms, the users' expertise score are then propagated until convergence by successive iterations [76] when the score difference is below a set tolerance factor [112].

The earliest works on user expertise estimation with graph-based approaches focus on user emails through the discovery that the structure of links between users do contain more information than the content alone [24, 38]. This is a content-agnostic approach to user expertise estimation without the need to process the content of the user emails in order to generate the user profiles. Since then, graph-based approaches have been used for the estimation of user expertise such as the ExpertiseRank [158], slightly improved ExpertRank [70] from 2009 and the more recent 2013 version of ExpertRank [23]; each improving several disadvantages that are often associated with the graph-based approaches that will be discussed later in Section 4.3.2.

These newer graph-based approaches rebuke earlier findings of being outperformed by simpler approaches to user expertise estimation [158]; possibly due to the following reasons [10]:

- Different dataset behaviours and rule models.
- Improved evaluation methodology which we follow closely in this research.

Challenges for Graph-based Approaches

Graph-based algorithms faced several challenges when adapted for the estimation of user expertise on CQA platforms [23] which we attempt to address in the proposed Credit Graph approach to be discussed in Chapter 5. These challenges include:

- The assumption of transitivity between users (questioner-answerer) may not be valid due to the diversity in user domains, topics or categories. These domains could overlap for the answerers making the authority measure inaccurate.

- The quality of user content is not accounted for in the graph links of graph-based algorithms which would then directly affect the propagation of authority. For example, user-user graphs with links from the questioner to the answerer are built by treating all user answers to be the same [158]. We attempt to overcome this challenge by incorporating credit functions from Section 3.5.1, adapted for CQA platforms.

- The questioner-answerer (user-user) graph does not share the same intuition with citations or web page reference links used in successful traditional graph algorithms like PageRank [112]. Answerers can answer any questions regardless of their expertise and most systems do not allow the questioners to select the best answerers [158]. Thus, the propagation of authority is not precise.

- Unlike hyperlinks between web pages, there are often more than a single link between two users because there can be more than one questioner-answerer relationship between the two. These links can all be of different quality, weight or value.

Questioner-Answerer Graph

Responding to one of the challenges discussed earlier, the ExpertRank [23] approach models the multi-interaction between users and their questioner-answerer relations by aggregating the links with a concept called QFactor. The QFactor attempts to capture the familiarity between users in answering each other's questions using CQA features such

as answer votes, question-answer similarity and answer ranks (through learning to rank approaches). This factor is then used to weight or score user-user links.

Alternatively, the questioner-answerer links can be weighted according to the number of questions posted by the questioner that are answered by the answerers [154]. This approach assumes that the answerers to a questioner of high expertise tend to be users of high expertise as well. In the works of the authors for CQARank [154], the computed user authority is topical with the introduction of the Topic Expertise Model (TEM) where a random surfer would consider the topical similarity between users as well. Furthermore, the topics are also considered for the teleportation step similar to that of Topic-Sensitive PageRank [162] where the random surfer is more likely to teleport to other users of similar topical interest and expertise.

Answerer-Answerer Graph

Unlike the earlier graph approaches which make use of the question-answerer relation, there are works that look to extract answerer-answerer relations for the estimation of user expertise. If it is possible to identify the best answerer in a question thread through user votes or questioner's award [18, 95], then a directed link can be built from every answerer in that question to the best answerer of the question [10]. The limitation of this approach however is the exclusion of the questioners themselves from the constructed graph.

4.3.3 Pairwise Comparison Approaches

Pairwise comparison approaches model the question-answering interaction as a competition between the answerers of the question to best meet the information need of the questioner. This is possible through the implicit pairwise comparison between two answerers as players that is implied through their answering performance [95]. The resulting pairwise comparisons are then used in a competitive rating system or machine learning for the estimation of user expertise. A challenge for pairwise comparison approaches is the higher data requirement where user-user pairs are required to model the competition field in order to learn and estimate the user ratings as expertise scores. Thus, pairwise competition approaches are often not suitable for CQA platforms with high numbers of users but with low activity [95].

A landmark work for the estimation of user expertise through pairwise comparison is the competition-based approach of Liu et al. [95]. For each question thread, the user-user competition pairs are established between the best answerer (as the winner) against the other answerers and the questioner (as the losers) without any draws. The user ratings and their rating deviation are then updated for the winner and the losers after each question. The user rating is determined using the TrueSkill rating [57], a Bayesian skill rating system used for calculating the relative skill level of players in multi-player or team-based games. The rating system assumes that the performance of players follow a normal distribution with the mean μ_u as the average player skill and the deviation σ_u as the uncertainty in the player's skill. As more data are obtained for the user (from more game outcomes), the system is more certain of the player's skill and thus the deviation decreases.

4.4 Yahoo! Chiebukuro

The Yahoo! Chiebukuro[6] is a community-driven CQA platform. It is the Yahoo! Answers[7] question-answering service for Japan. Similar to many other CQA platforms, it allows

[6] https://chiebukuro.yahoo.co.jp/
[7] https://answers.yahoo.com/

Figure 4.4: A screenshot of the front page of the Yahoo! Chiebukuro platform. Web page visited on 20th September 2017.

users to create question threads as questioners to state their information need. The other users could then respond to the thread as answerers.

Unlike many specialised CQA platforms such as Stack Overflow, Yahoo! Chiebukuro is a general CQA platform covering a wide range of topic domains. These topic domains are organised as three major class levels – major, medium and minor determined by Yahoo!. Each question thread belongs to either a medium or a minor class domain; selected by the questioner on creation or automatically classified if not[8]. It should be noted that the categorisation schemes for the topic domains are periodically revised by Yahoo! with the current question threads moved accordingly.

Just like many of the other CQA platforms today, Yahoo! Chiebukuro is managed through peer-moderation. Users are able to vote on answers according to their perception of the answer quality – with the highest voted answer as the best answer for that question. A unique feature of the platform however is the prioritisation given to the questioner in determining the best answer for his or her question. The questioner has seven days to select the best answer before opening the best answer selection to the user votes. Noting this feature, we leverage the questioner-selected best answer as the gold standard for

[8]Automatic categorisation was introduced in 10 April 2017.

Figure 4.5: A screenshot for a Question thread with answers on the Yahoo! Chiebukuro platform. Web page visited on 20th September 2017.

our evaluation process later in Chapters 5 and 6. It should be noted that the question-best answer pair on Yahoo! Chiebukuro are highly regarded as useful *knowledge* on the platform.

4.4.1 A Yahoo! Chiebukuro Dataset

The Yahoo! Chiebukuro Data (2nd edition)[9] provided to National Institute of Informatics by Yahoo Japan Corporation is the CQA dataset for our research for the estimation of user expertise and answer quality; where the 1st edition of the dataset is used for the performance evaluation of the competition-based approach [95]. The dataset is in Japanese, making it suitable for the evaluation of the content-agnostic approaches of our research.

The dataset was collected between April 2004 and April 2009 over 411 topic domains. It contained a total of 16,257,422 solved questions, ignoring unsolved questions. These solved questions contained a total of 50,053,894 answers. From this dataset, we look to use the newest questions for the dataset (when automatic classification for questions into

[9] http://www.nii.ac.jp/cscenter/idr/en/yahoo/chiebkr2/Y_chiebukuro.html

topic domains were introduced). Thus, only using questions between April 2007 and April 2009, we have:

- 10,045,658 Questions
- 27,495,550 Answers
- 1,667,646 Users

This subset of the dataset is divided for training and testing in the evaluation of the proposed algorithms described in Chapters 5 and 6. These questions are also used for further analysis of the platform in the following studies.

Topic Domains

The questions in the dataset are organized into 411 separate categories (domain topics) with each question belonging to only a single category. Thus this dataset eliminates the need for topical component in the evaluated algorithms such as the Latent Dirichlet Allocation (LDA) to determine the topics, domains or categories of the questions. Besides that, this dataset allows the research to explore several proposed concepts and ideas consistently without being affected by topical components in the algorithm.

Training Data

Questions (and their answers) over a yearly period are extracted for the training of the proposed models that shall be discussed in Chapter 5 and Chapter 6. As we explore domain-specific user expertise, the models would also be trained on questions of specific domains – by randomly selecting 100 topic domains for the available 411. For our research, we would train the models using data from April 2007 up to April 2008.

Testing Data

The testing data used for evaluation are the questions from the subsequent year of the training data constrained to be in the same domain as the training data. For example, if the model is trained with questions from year 2007 to 2008 in the given domain, then the testing data built from the questions in year 2008 to 2009 of the same domain. For this research, we used the questions from April 2008 up to April 2009 for evaluation.

As it is possible for users in the evaluating year who were not active (without any interaction) in the training year, these users would be removed in the pre-processing stage so that such users do not affect the model evaluations as their expertise were not estimated.

4.4.2 Study: Consistent User Contributions

A general concern for any UGC platforms is the possible lack of user interactions on the platforms [15]. For CQA platforms, users are required to act as questioners in asking questions and as answerers to contribute information on the platform. We plot the moving average with a window of seven days for user contributions on the dataset in Figure 4.6.

From our dataset, each day has between 6,572 and 25,311 questions. There are a lot of questions contributed daily with an average of 13,742 questions and over half of the days having at least 14,292 questions. On the other hand, we found a greater variation in the amount of answers posted daily – in the range of 957 to 62,250. On average however, we saw the that the number of answers contributed by users are over double that of the questions contributed at 37,257 answers.

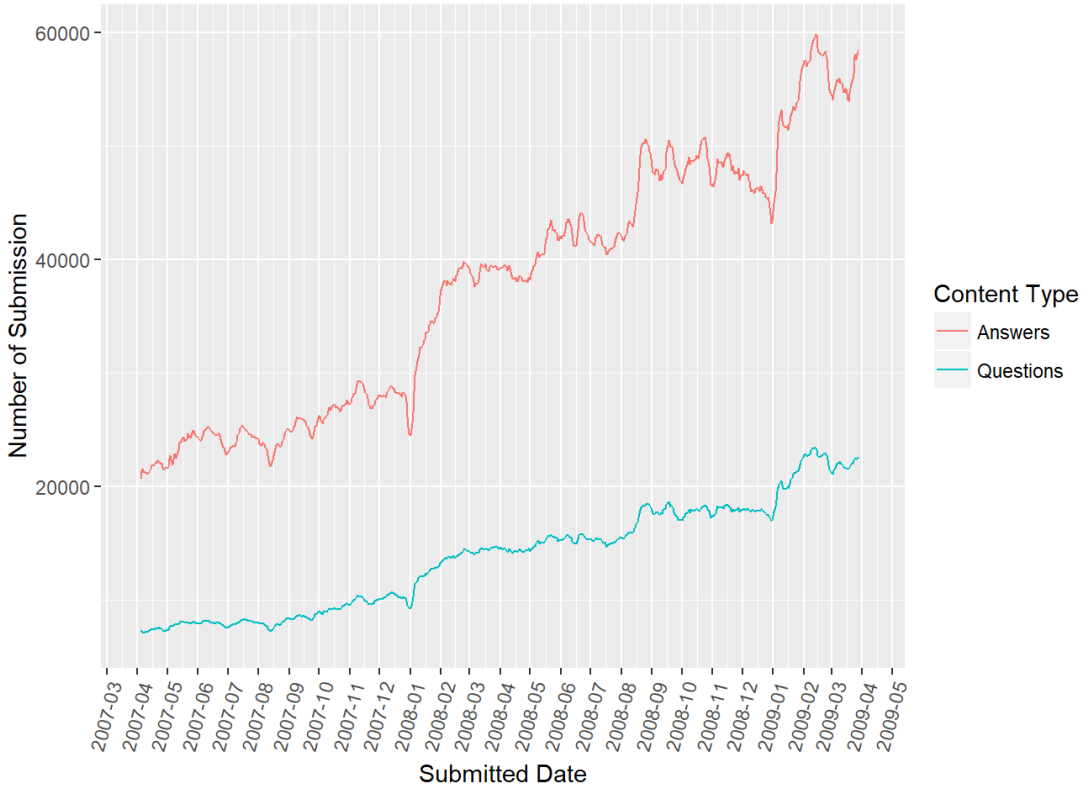


Figure 4.6: A moving average plot (7 days)) for the creation of questions and answers on the Yahoo! Chiebukuro CQA platform between April 2007 to April 2009.

4.4.3 Study: User Activity in Topic Domains

In general, we found a healthy amount of questions being posted for each of the topics – over three quarter of the Yahoo! Chiebukuro topics have at least $1,398$ questions and over half have at least $8,142$ questions over a period of 2 years in our dataset. The largest topic recorded a total of $518,067$ questions within the same period. Thus, there would be sufficient user questions within each topic for our study. The questions for each of the topics:

- Have between 1 to $3,387$ answers per question with an average of 2.44 answers for each question across all topics.
- Receive an average of 1.1269 votes for per answer on average across topic. The minimum vote for each answer in every topic is 0 votes. There is however a large variation in the maximum number of votes received by answers across topics ranging from 1 vote for a particular topic, up to $3,750$ votes for another (with a median of 34 votes).
- Require 10.21 days on average to solve for each question as an average across topic. The time to solve a question is similar across topic – many of the topics have their questions solved in 7 days as the fastest and 15 days as the slowest.

4.4.4 Study: User Votes

Yahoo! Chiebukuro allow its users to vote for answers under the condition that the questioners do not select the best answer within seven days. Thus, user answers are only open for voting by other users of the platform after the seventh day. In our analysis of the

dataset, we found that a majority of questions are selected by the best answerer – resulting in over half of the user questions having no votes with a median of 0 and mean of 1.189 answer votes for questions; and a average vote of 0.4118 for user answers. We however noted that Yahoo! Chiebukuro does not allow users to negatively vote the answers on the platform.

4.4.5 Study: Knowledge

As discussed earlier, the question-best answer pairs on Yahoo! Chiebukuro are highly regarded as *knowledge* on the platform. From the user questions, expert editors on the platform are given the authority to identify and award useful user questions for future retrieval. From the subset, we saw that $1,144$ questions are selected as useful question on the platform from a total of $10,045,658$ questions. Thus, the useful questions are in fact limited and of high value on the platform.

4.4.6 Study: Sufficient Answers

As the dataset contains only solved questions, we saw each question having at least one answer. We however identified that there are more than half of the user questions with at least two answers and a quarter of questions have at least three answers; with an average of 2.737 answers per question. Thus, there are sufficient answers for the consumption of the questioner. The high number of answers (maximum of $3,387$ in a question) would however challenge the questioners in identifying the answers that meet their information needs due to the varying answer quality as we shall discuss in Section 4.4.8.

4.4.7 Study: The Cold Start Problem

A motivation for our research is to overcome the cold-start problem in the recognition of the best answer to meet the information need of the questioner. From the dataset, we saw that the best answer for a question was only identified 10.13 days on average after the question was posted; with questions being solved on the 6th day the earliest and most questions solved from the 7th days onwards. In fact, there were more than a quarter of user questions that were only solved after 2 weeks.

Later in Chapters 5 and 6, this research attempts to predict the best answer for a question for the questioner's consumption. This is accomplished through the estimation of the answerers' expertise.

4.4.8 Study: Answer Quality

Answers on CQA platforms are of varying quality and the best answer is desired by the users of the platform [159]. From the questions collected, we saw the following types of answer:

- 19.50% of the answers are selected by the questioner as the best answer that meet their information needs. There is a remainder of 29.75% answers from the same questions.
- 17.04% of the answers are the best answers as voted by the community. There is a remainder of 30.95% answers from the same questions.
- 2.76% of the answers were revised by the answerer but these answers are still not the best answer for the question.

A total of 36.54% of the answers on CQA platforms were able to meet the information need of the users as answers of high information quality. Through the estimation of user expertise, this research attempts to identify the users who are able to contribute such

answers consistently and to do better than a guess for a best answer (well above the success rate above).

4.4.9 Study: Users as Questioners

Users of Yahoo! Chiebukuro are active questioners. From 1,667,646 users recorded in the dataset, we found a total of 1,390,636 (83.39%) of these users asked at least a single question. This observation strongly suggests that many of the users are information seekers with the questioners asking 7.224 questions on average – a large demand for information. Half of the questioners asked a total of 2 questions throughout the 2-year period, many over three thirds asking up to 4 questions. The remaining 25% of the questioners asked between 4 up to 10,159 questions, a heavily skewed behaviour.

We saw an increase in the number of questioners from April 2008 to April 2009 from 540,422 questioners to 1,066,490 with 850,214 of them being new questioners. This large growth suggests all of the following:

- There is an increase in information need for the users.
- There is an increase in information need that is not easily found or available on the WWW.
- There is an increase in preference for users to state their information needs with natural language instead of keyword-based queries.

4.4.10 Study: Users as Answerers

Unlike the questioners, we saw a smaller number of answerers – 775,773 (46.52%) of the users having answered at least one question. Compared to the questioners however, we also saw that answerers contributed more where they recorded a higher average answer of 35.44 answers with at least half of the answerers contributing at least three answers on the platform. The active answerers are once again skewed just like the questioners with between 12 to 34,835 answers contributed.

Similarly, we observed an increase in answerers as well from 284,609 to 600,139 with a total of 491,164 new answerers in the second period. We noted that the increase of answerers is not as significant as the questioners because of the users' reluctance to contribute answers rather than just consuming them – directly through question formulation or from the currently available answers.

4.4.11 Study: The Selfish Questioners

Analysing further, we looked at a particular group of questioners that are often known as *leechers* on many online communities. These questioners only asked questions without answering any and they are aplenty with a total of 891,873 users; making up 53.48% of the user community or 83.39% of the questioners on the platform.

Over the periods, we saw an increase in the number of such questioners. Here, we measured 599,420 (67.21%) who continued to ask more questions, with an average increase from an average of 0.8985 questions per questioner to 1.774 questioners. Only a small number of questioners (29.75%) recorded a drop in the number of questions asked, possibly due to them not returning to ask further questions once their information need has been met.

As for the questioners who are active in both of study period, they are few with only 85,600 (9.60%) questioners. An interesting observation can be made however, where there is an even split in questioners who posted more questions and less questions in the following period. Ignoring the extreme, many of such questioners posted either one additional or one less questions than the period before.

4.4.12 Study: The Helpful Answerers

On the other end of the spectrum, we have dedicated answerers – users who only answer questions on CQA platforms without asking any. Many online communities highly regard these users as the main contributors on their platform. From our dataset, we recorded 277,010 of such answerers which make up 16.61% of users on the platform or 35.71% of the answerers. This makes up a small subset of users unlike the questioners we saw in Section 4.4.12. Thus, there is a greater demand for information on CQA platforms than the supply of information from the community.

Similar to the questioners, we saw an increase in the activity of dedicated answerers over the period from 71,017 (9.15%) to 277,010 (26.31%). For dedicated answerers who are active during both periods, they are scarce with only 17,410 answerers with many of them answering more questions than the year before. Most of these users contributed up to 33 answers each which is valuable on the platform.

On the other end, we saw 2,923 answerers answer less questions than in the period before with them contributing an average of 10.15 answers each! This is a big loss in content creation for the Yahoo! Chiebukuro platform possibly due to – (1) Lack of motivation to contribute answers; (2) Having struggle against disruptive users; or (3) Lack of recognition for their contribution on the platform. Thus, this research attempts to identify and acknowledge the valuable contributors of quality content on the platform.

4.4.13 Study: The Active Samaritan

From the dataset, we also noted 498,763 (29.91%) users who contribute both questions and answers on the platform. These users ask more questions than answering – a majority asking up to 11 questions but only answering up to 3 questions. The activity of these users between periods can be summarised:

- 317,972 (63.75%) users have their activities increased between periods in asking more questions and also answering more; majority up to 10 questions and 12 answers each. Here, we note that these users tend to however ask between 1 to 8 more questions in the current period than the period earlier; but only answer 1 more.

- 101,389 (20.33%) users recorded reduced questioning and answering activity; with the majority answering up to 9 less questions and only asking up to 7 less questions each. These users do contribute a large amount of answers in the period before, up to 11 answers each.

- 14,176 (2.84%) users were found to be asking more questions but answering less questions between period. On average, they asked 4.346 more questions each but also answered 4.537 answers less. Many of these users (over 75%) did not answer any question in the latter period but instead asked up to 7 more questions each.

- 42,557 (8.53%) users did ask less (majority between 1 to 9 questions) but answer more questions in the current period than the period before. Over half of these users were questioners the period before, who then stopped asking questions in the current period but instead answered them.

4.5 Summary

In Section 4.1 of this chapter, we looked at the structure and representation of CQA platforms; with the addition of user votes as an indicator of content quality. As discussed in Section 4.2, the user questions and answers on CQA platform have the potential to

meet the information need of users that otherwise could not be met through traditional IR approaches [95]. Such potential does however raise the challenge in content processing of the user questions and answers. Thus, user expertise is often used to infer the information quality of content on CQA platforms. We discuss the current approaches to the estimation of user expertise in Section 4.3.

The Yahoo! Chiebukuro CQA platform is studied in Section 4.4 with a dataset provided to National Institute of Informatics by Yahoo Japan Corporation to better understand the content and user behaviours. In this study, we analyse the availability of user contributions such as questions, answers and user votes; together with their value as knowledge on the platform. Such content is of varying quality as discussed in Section 4.4.8. Besides that, we analyse the cold-start problem in Section 4.4.7 which motivates our research to quickly identify the best answer for user consumption. Finally, we study the user roles as questioners and answerers on the platform including their subgroups.

Later in Chapter 5, we first extend on the Credit Graph based approach proposed from Part I of our research after an overview of the various state-of-art graph-based approaches for user expertise estimation on CQA platforms. Their performances are evaluated according to their capabilities in predicting and retrieving the best answer for a question. Besides that, the approaches should be able to identify the experts in each topic domain to overcome the challenges identified through our user study. This is then repeated in Chapter 6 with a different direction – a competitive pairwise comparison approach.

Chapter 5

Credit Graphs for User Expertise Estimation

In Section 4.2, we discuss the information potential of CQA platforms in generating new information to meet the questioners' new information need [81], and also as a repository of knowledge for user consumption from earlier similar questions [26, 27]. As UGC however, the information quality of user answers does vary between users [68, 124, 159]. Thus, this research attempts to infer the information quality of user answers from the expertise of the answerers themselves; all without the need to process the textual content of the user answers. Instead, we explore at the content-agnostic approaches for the estimation of user expertise.

In this chapter, this research focuses on the graph-based approach for user expertise estimation. We discuss the current state-of-the-art graph-based approaches which are all based on user-user graph models [158] in Section 5.2. The approach relies on the concept of peer assessments between users [158], based on their similarity through user interactions, such as the explicit question-answering relationship.

Instead of extending on the user-user graph models, we extended the Credit-Graph model (user-resource relation) presented in Chapter 3 for adaptation to CQA platforms in Section 5.3. We argue that: (i) there is a need to account for the question difficulty relative to the question's domain, and (ii) the user's contribution to solving the questions matters where some answers are better than others.

Section 5.4 outlines our evaluation methodology as an answer retrieval task. Here, we attempt to identify and retrieve the best answer in a given question thread to best meet the information need of the questioner. The best answer can be predicted according to the inferred information quality from the estimated expertise of the answerer. The proposed Credit Graph approach is benchmarked against successful baselines – both simple [158] and complex [23]. We observe and discuss our findings.

The estimated user expertise is also applied for expertise search. In Section 5.5, we observe and analyse the capabilities of each user expertise algorithms in identifying expert users that are of great value on CQA platforms; many of which differ from the usual popular or active user groups. Finally, we conclude this chapter with a summary in Section 5.6.

5.1 Research Questions

Expanding on the research questions proposed in Section 1.4, we propose these research questions for our work into graph-based approaches for the estimation of user expertise, answer quality and for the prediction of the best answer on CQA platforms.

5.1.1 Can the Information Quality of User Answer be Estimated Through Content-Agnostic Means?

User answers on CQA platforms are unmoderated and are of varying information quality [6, 68, 124]. The users especially the questioners do however desire answers that are of high information quality [159] to meet their information needs. However, the amount of high quality answers on CQA platforms are limited as we illustrated earlier with our Yahoo! Chiebukuro study in Section 4.4.8. Thus, this research aims to estimate the information quality of user answers; ensuring that high quality user answers are consumed [136] while filtering out unwanted noise on the platform [6, 68].

Current literature on the information quality of user answers attempt to infer answer quality from the metadata of answers [68, 128] such as user votes. This is a content-agnostic approach without the need to process complex unstructured user answers. Such an approach does however suffer from the cold-start problem which we showed in Section 4.4.7 – with user questions taking 10.13 days on average to be solved. Thus, we explore an alternative to the estimation of user expertise.

5.1.2 How well can Graph-based Approaches Estimate User Expertise on CQA Platforms?

Users are of high importance on CQA platforms, playing the role of questioners, answerers or both. Through these roles, the users generate content within their capabilities [16, 18] – answerers who are of higher expertise answering the questions of questioners with lower expertise [154, 158]. We study the diverse users and their behaviour of the Yahoo! Chiebukuro CQA platforms in Section 4.4. Thus, much work has been done on the estimation of user expertise as discussed in Section 4.3.

One approach to the estimation of user expertise on CQA platform is the graph-based approach as discussed in Section 4.3.2. This is made possible by the representation of CQA platforms as a graph which we elaborate later; exploiting the user-user relations from their question-answering interactions [10, 23]. There has been disagreement however over the effectiveness of graph-based approaches particularly with findings that simple approaches are able to outperform or stay competitive with the more complex graph-based approaches [158].

In this chapter, we explore several graph-based approaches to user expertise estimation and their performance for the task of – (1) answer retrieval; and (2) expert search. Besides the current state-of-the-art graph approaches, this research adapts the previously proposed Credit Graph for CQA platforms as an alternative graph model with the inclusion of CQA features such as user votes as studied in Section 4.4.4

5.1.3 Do Expert Users Produce Better Answers?

Studies often mentioned the correlation between user expertise and the information quality of their generated content [18], particularly how users of low expertise are more likely to produce content that are of lower information quality [6, 158]. In our research, we define user expertise as a relative measure in Section 4.1.1; enabling us to compare the odds where a user would produce better content than another according to their user expertise. Content of higher information quality should be able to meet the information need of the content consumers better [113].

In this chapter, we attempt to estimate user expertise through graph-based approaches. The estimated user expertise is then used to infer the information quality of their generated content. In the context of CQA platforms, we attempt to compare the information quality between answers within a question thread in meeting the information need of the questioner; all according to the estimated user expertise of the answerers. Users of high

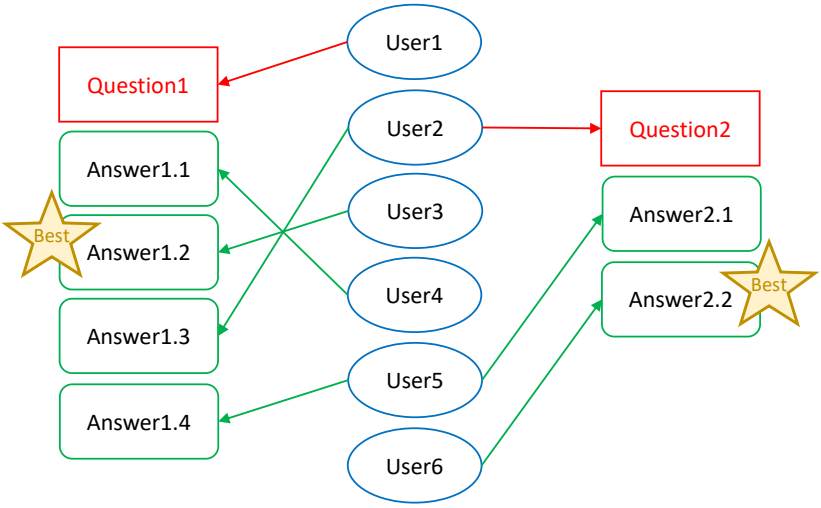


Figure 5.1: A CQA example scenario.

estimated expertise are then identified and encouraged to continue their contribution of high quality answers on the platform. This is vital as we saw an increase in the number of questioners as opposed to answerers in Section 4.4. Besides that, their contributed answers can be ranked higher for easy identification and consumption of the questioner.

5.1.4 Can We Predict the Best Answer in a Question Thread?

Users turn to CQA platforms to satisfy their information needs that could not be met through traditional IR, as such information were not available yet [9, 95]. Through the user questions, answerers are encouraged to generate new content in their attempt to meet this information need [159]. These new contents are however unknown and of varying quality (see Section 4.4.8) to the questioners, thus challenging them especially when facing a high number of answers (see Section 4.4.6).

Definition 14 (Best Answer)**.** The answer that best meets the information need of the questioner [18].

Modern CQA platforms attempt to guide the questioners through user votes in identifying the best answer for their desired consumption [68, 128]. Nonetheless, it takes time for the user votes to stabilise, as illustrated earlier in Section 4.4.7, creating a delay for the questioners to identify the best answer for their consumption. Thus, this research attempts to predict the best answer within a question thread by comparing the relative user expertise of its answerer as an answer retrieval tasks in Section 5.4.

5.2 User-User Graphs for Modelling CQA

The current state-of-the-art graph-based approaches to the estimation of user expertise on CQA platforms are modelled as user-user graphs. Such graphs focus on the links and relations between the users from their question-answering interactions. Consider the example scenario illustrated in Figure 5.1 – the question-answering interactions by six users on two questions. User-user graphs can be built according as the following – (1) Questioner-Answerer [23, 154]; or (2) Answerer-Answerer [10].

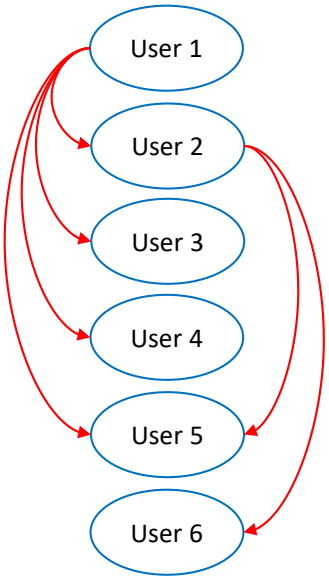


Figure 5.2: Questioner-Answerer graph model.

5.2.1 Questioner-Answerer Graph Model

The first graph model is the questioner-answerer model [23, 154] which models user roles as questioners and answerers. In this model, there is a directed edge from the questioner to the answerer based on the intuition that the answerers should be of higher expertise than the questioner to be able to contribute an answer [23]. Thus, there is a reinforcement in authority from the questioner to the answerer. Based on the earlier example in Figure 5.1, the following questioner-answerer graph can be built.

The directed edges can and should be weighted in a CQA scenario. As a user can answer one or more questions from another user, the edges can be weighted according to the number of answers that have been provided by the answerer to the questioner [154]. A variant of this is to just consider the questioner and the answerer with the best answer [18, 54] instead of all of the answerers. This approach however, would not provide us with the user expertise estimation for all of the users, because not all users would be represented within the graph.

5.2.2 Answerer-Answerer Graph Model

An alternative to the questioner-answerer graph model is the answerer-answerer graph model [10]. This graph model is used to represent the contribution of a user's answer in meeting the information need of the questioner in relation to the other answerers of the same question. In this model, there is a directed edge from all of the answerers of a question to the best answerer of the question. Thus, these answerers reinforce the authority of the best answerer who displayed a higher degree of expertise in order to best meet the information need of the questioner. The best answerer of a question can be easily identified in modern CQA platforms by extracting the user votes for the highest voted answer [159] or as selected by the questioner [95].

Unlike the earlier model however, the relation between the questioner and the answerer is not accounted for. Instead, this model can be regarded as a competition-based model between the answerers to produce the best answer within a question.

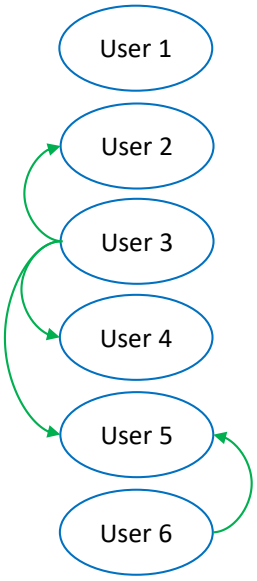


Figure 5.3: Answerer-Answerer graph model.

5.2.3 Link Weight Variant

User-user relations can be weighted – where some relationships are stronger than others. As discussed earlier, these relations are directed. In the user-user graph model, these weights would determine the propagation of authority from one user to another. From the literature, there are multiple ways to weight the annotations which we would discuss further and would act as our baseline for the evaluation later in Section 5.4:

- Count. The links between users can be weighted according to the number of interactions between them such as the interaction of – (1) one user answering the question from another [154]; or (2) both answering questions within the same thread [10].

- Vote. As CQA platforms enable users to vote answers as an indicator of answer quality, the user votes can be used to weight the user-user relations. If two users provided answers in the same thread, then the links between them can be weighted according to the difference in vote gained for the answers. This value would be positive for a directed edge from an answerer to the best answerer of the question [159]. The questioner-answerer graph model, the user questions themselves are often unweighted. Thus, the weight from the questioner to the answerer can just the vote of the answers themselves.

5.2.4 Transfer of Probability

Once the user-user graphs are built, the authority would be transferred between the users. Traditionally, graph models would propagate authority on a global scale such as in the PageRank algorithm [112] with a transfer probability of 0.5. Due to the sheer adoption of CQA platforms amongst the users, these user-user graphs can be really large regardless of the graph models. Thus, a smaller transfer probability can be used 0.05 for a more local propagation or scope of the graph [23]. As user-user graphs are used as the baseline for the evaluation of our research in Section 5.4, we study the performance of both the global and local propagation.

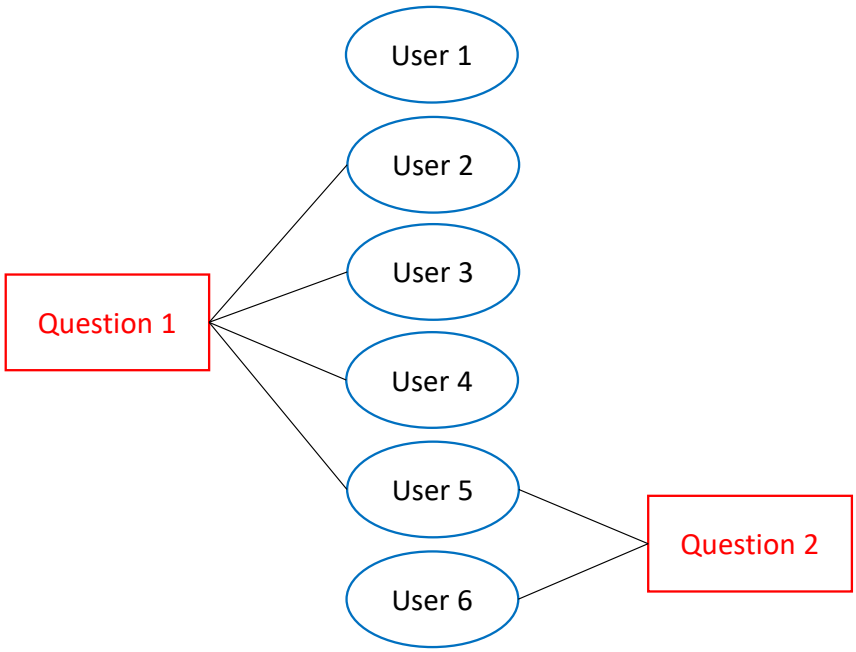


Figure 5.4: Proposed Question-Answerer credit graph model.

5.3 Credit Graphs for Modelling CQA

The current state-of-art graph approaches to the estimation user expertise as discussed earlier are modelled according to user-user relations as discussed in Section 5.2. Our research however proposes that the graph model should not only include the user-user relations; but instead to also include user questions as a question-answerer graph model. This proposed model is similar to the Credit Graph model introduced in Part I of the research, inspired by the Spamming-resistant Expertise Analysis and Ranking (SPEAR) algorithm [109] used to model user interactions (annotation) of bookmark-user relations in collaborative tagging (CT) platforms. The proposed credit graph model is as illustrated in Figure 5.4, for the same scenario in Figure 5.1.

The credit graph model's inclusion of questions is motivated by the following assumptions which were not accounted for in the user-user graph models:

- There should be a direct mutual reinforcement between the expertise of the users and the difficulty of the questions which they have answered. The answerer of a question would require an expertise higher than that of the question difficulty in order to solve it; with the best answerer having the highest expertise among the other answerers and the questioner [18]. The current state-of-the-art graph approaches only focus on the user-user relation without accounting for the difference between question difficulty asked by users of varying expertise on CQA platforms.

- There is a correlation between the number of user answers and that of question difficulty or importance. A question is deemed to be of higher difficulty or importance where a large number of users are trying to work together in order to solve it, building upon the earlier answers. As discussed in earlier sections, the aggregation of user-generated content can rival that of expert content themselves [132]. Besides that, expert users tend to contribute improved answers towards questions if the previous answers for that question are poor [113].

5.3.1 Credit Functions for Answer Contribution

Following the Credit Graph model introduced in Figure 5.4, each user's contribution in a question can be measured and weighted separately. This is unlike the earlier user-user graphs where the edges are weighted according to the aggregated interaction scores between users [23, 154]. To the best of our knowledge, the current state-of-the-art user-user graph approaches regard each answer to be of the same contribution [23, 154] or only the best answer to be different [10].

Thus, the proposed Credit Graph model treats each user interaction to be different, and to also differ between questions. Each user interaction is modelled as a user-question link, weighted according to the contribution of that interaction towards meeting the information need of the questioner. It should be noted that each user is only allowed to interact with each question once, and not contribute multiple answers for a single question. This removes the need to aggregate the user-question relations in the graph. For the Credit Graph model, we make the following assumptions:

- Each user answer is different in content, resulting in a difference in the level of contribution towards solving the question and meeting the information need of the questioner. This contribution is indicated by the number of votes received from the other users [9, 95].

- The best answerer has the highest expertise amongst the other answerers of the question, being able to produce the highest contribution towards solving the question. The best answerer can be identified through the votes received or as selected by the questioner. The questioner would meanwhile have the lowest expertise [18].

- Expert users are users who solve questions within the topic of their interest or expertise. Their answers are often of high quality with very high user votes or the best answer for the question.

As a result, this research proposed and evaluated various measures of user contributions within questions; borrowing the concept of credits from the proposed Credit Graph model in part I and adapting it according to the features on CQA platforms. We discuss the credit functions in the following sections.

Voting-based Credit

Modern CQA platforms allow the users to vote for content on the platform as a measure of content quality [159]. The user votes can be extracted as a signal of user contribution for the credit functions in the Credit graph model. Thus, this research proposes a simple vote-based credit function as an indicator of user contribution. For answer a in question q, the vote gain can be represented as $V(a)$ where the number of votes ranges from 0 to infinity. If there are negative votes, the votes can be adjusted to be positive values by shifting the scale so that 0 is the lowest value.

As the total number of user votes for user answers differs from question to question $V(q)$, we may need to normalise the votes as shown in Function 5.1. An optimisation parameter, the credit exponent α is introduced to control the impact of the credit gained by answers.

$$\mathbf{Credit_{vote}}(a) = \left(\frac{V(a)}{V(q)}\right)^{\alpha} \tag{5.1}$$

Temporal-based Credit

The other credit function variant is the temporal-based credit function which accounts for the temporal ordering of answers as credit. This variant is based on the assumption that newer answers are of higher quality, thus deserving more credit because expert users tend to contribute improved answers if the prior answers to the question question were not able to meet solve the question [113].

The credit for answer a_x in question q can be estimated by counting the number of earlier answers in question q, $A_{a:T(a)<T(a_x),Q(a)=Q(a_x)}$. Just like the earlier variant, an optimisation parameter α is introduced to control the impact of the credit function.

$$\mathbf{Credit_{temp}}(a_x) = |A_{a:T(a)<T(aa_x),Q(a)=Q(aa_x)}|^{\alpha} \tag{5.2}$$

5.3.2 Mutual Reinforcement Propagation

The user-question relations in the Credit Graph model are weighted according to the selected credit variants discussed earlier. These calculated credits are stored in the user-question adjacency matrix M, where the value of each cell is determined by the chosen credit function if the user has answered the question or 0 otherwise.

The values in the matrix are then used for the propagation of authority with expert users acting a hubs to locate question of higher importance or difficulty. Questions of higher difficulty would require more expert users to answer. The propagation is similar to that of the HITS algorithm [80] as shown in Equations 5.3 for user expertise e and question difficulty d.

$$\vec{e} \leftarrow \vec{d} \cdot M^T \qquad \vec{d} \leftarrow \vec{e} \cdot M \tag{5.3}$$

Normalisation is then performed at each iteration to ensure convergence as shown in Equation 5.4 with e_u as the expertise of user u and d_q as the difficulty of question q. The algorithm for the proposed propagation is as described in Algorithm 2. It should be noted that this algorithm is similar to Algorithm 1 from Chapter 3 but adapted for CQA platforms.

$$\hat{e}_u = \frac{e_u}{\sqrt{\sum_{u'} e_{u'}^2}} \qquad \hat{d}_q = \frac{d_q}{\sqrt{\sum_{q'} d_{q'}^2}} \tag{5.4}$$

Algorithm 2 Basic Credit-Graph Algorithm

Require: Users $U = \{u_1, u_2, ..., u_{|U|}\}$
Require: Questions $Q = \{q_1, q_2, ..., q_{|Q|}\}$
Require: Answers $A = \{a_1, a_2, ..., a_{|A|}\}, a_i \in (U, Q, V, T)$
  Set $\vec{e}$ to be the vector $(1, 1, ..., 1)$ for user expertise
  Set $\vec{d}$ to be the vector $(1, 1, ..., 1)$ for question difficulty
  $M \leftarrow GenerateAdjacencyMatrix(A, Credit(A))$
  for *iterate* $= 1$ to *convergence* **do**
    $\vec{e} \leftarrow \vec{d} \cdot M^T$
    $\vec{d} \leftarrow \vec{e} \cdot M$
    $\vec{e} \leftarrow \frac{\vec{e}}{||\vec{e}||}$
    $\vec{d} \leftarrow \frac{\vec{d}}{||\vec{d}||}$
  end for

5.4 Answer Retrieval Evaluation

CQA platforms are valuable as a rich source of information which leverage on the knowledge of its community. Users of CQA platforms are able to formulate queries using natural language in the form of questions [95], to retrieve future information from experts when they answer the questions. Alternative, the users can seek information from currently available answers on the platform by identifying the similar answered questions from the past [26, 67, 153, 159].

Both approaches discussed earlier rely on the user generated answers as the target for retrieval. Thus, the evaluation for this part of the research is set-up as an answer retrieval tasks to – (1) infer the information quality of answers; and (2) predict the best answer amongst the answer without a cold-start. This can be achieved by estimating the expertise of the answerers which is then used to infer the information quality of their answers. The estimated expertise can be both general and domain-sensitive (to a topic); both of which are explored further in this evaluation.

5.4.1 Dataset

The CQA dataset collection (see Section 4.4.1) is separated into two sets from 100 randomly selected topic domains for evaluation. These two sets are used for training where the user expertise is estimated, and testing where the user expertise are used to infer the information quality of their answers to then predict the best answer of each question.

Training

The first set includes questions from April 2007 to April 2008 as the training set which estimated the expertise of answerers in the period. In total, there are $3,027,450$ questions which are used to estimate the expertise of $492,888$ active users using the various approaches to be introduced later.

Testing

The following set of questions from April 2008 to April 2009 is then used as the testing set for evaluation. This set includes questions from the same domain for a total of $1,230,461$ questions. Pre-processing is performed on this set to remove inactive users without their expertise estimated from the training set. Besides that, questions without any ground truth for the best answer (see Section 5.4.2) are removed as they would not be used for evaluation. This leaves us with the following remaining questions with meets the ground truth requirements:

- All questions: $53,871$ questions.
- Questions with at least 3 answers: $17,609$ questions.
- Questions with at least 5 answers: $5,057$ questions.

As the application of user expertise for inferring answer quality is of greater importance for questions with large amount of answers, we prioritize the performance of models on questions with at least five answers.

5.4.2 Evaluation Criteria and the Ground Truth

The Yahoo! Chiebukuro CQA platform enables questioners to select the best answer. The questioner's selected best answers are used as the ground truth for evaluation because these answers are manually selected by the questioner to be the best answers which meets the questioner's information need best without room for misinterpretation (unlike user votes

which relies on the judgement of other users). Thus, it is often used as the gold standard [159] for evaluation. Besides that, many of these user expertise estimation approaches are built upon user votes. A different measure as the ground truth would further challenge these models during evaluation.

5.4.3 Evaluation Measure

The Mean Reciprocal Rank (MRR) is used for evaluation where we model the question-answering process of CQA platforms as a information retrieval (IR) problem. Users submit queries in the form of questions Q, true to the nature of CQA platforms. The questions would retrieve answers posted by answerers. Thus, we rank the answers according to the user expertise for our evaluation to predict and retrieve the best answers for each question (from the ground truth as a_{best}) for evaluation. The MRR measure for evaluation is as shown in Equation 5.5.

$$\mathbf{MRR} = \frac{1}{Q}\sum_{n=1}^{|Q|}\frac{1}{\mathbf{Rank}(a_{best,q_n})} \tag{5.5}$$

The evaluation considers the pre-processing performed earlier for the testing set of questions; and considers the varying number of answers for the question. This research notes that the difference in the measure MRR performance is significant even when the value is little as it is harder to get significant improvement if evaluation is done on a large dataset [101] like ours where all questions are considered before being filtered towards questions with more answers.

5.4.4 Evaluated Approaches

This research evaluates several approaches towards the estimation of user expertise. These approaches have been discussed earlier – (1) simple approaches introduced in Section 4.3.1; and (2) graph-based approaches in this chapter.

Baseline: Z-Index

The main baseline of choice is the Z-Score discussed in Section 4.3.1 which is regarded to be an algorithm that is competitive to complex algorithms [158]. Thus, we employ the Z-Index for users according to the number of questions which the users have both asked and answered.

Baseline: Vote Scores

Many CQA platforms support user voting. Thus, a simple baseline algorithm which utilizes such votes to estimate user expertise would be used. We apply the simple baseline in Equation 5.6 to obtain the expertise of user u at time t with β as the optimisation parameter to control the impact of votes.

$$e_u^t = \sum_{a \in A} V(A_{a_i:(u_{a_i}=u)\wedge(t_{a_i}<t)})^{\beta} \tag{5.6}$$

State-of-the-Art: User-User Graphs

Graph algorithms are widely used as the state-of-the-art approaches for the estimation of authority such as user expertise. The current state-of-the-art graph models for CQA platforms are user-user graphs as described earlier such as the questioner-answerer graph (Figure 5.4) and the answerer-answerer graph (Figure 5.3).

For our evaluation, the answerer-answerer graph [10] model is used. The answerer-answerer graph model can estimate the expertise of most if not all of the users on CQA platforms; which is not possible with the questioner-answerer model due to the lack of in-degrees for the answerers. This selected model considers the credit function variants with both global transfer probability (= 0.5) and local transfer probability (0.05):

- Count where the weight of the edge scales with the number of interactions between the two users.
- Votes where the weight of the edge scales with the difference in normalised votes between the users for the same question they have answered.

Proposed: Credit Graphs

The proposed credit graph model is discussed in Section 5.3. For our evaluation, we consider the credit function variants with the credit modifier of $\alpha = 0.5$.

- Temporal-based. Newer answers to receive higher credit due to assumptions that newer answers are only posted if the question do not already have a good answer [113].
- Vote-based. Answers with higher votes to receive credit proportional to the votes that the answers receive.

5.4.5 Results: Answer Retrieval

The answer retrieval performance for the evaluated approaches is shown in Table 5.1. We drilled down the retrieval performance down to questions with at least three answers and at least five answers. The performance was compared against the baseline Z-Index approach. The results showed that it is possible to predict the best answer in a question by inferring the information quality of answers according to the estimated user expertise of the answerer; performing well above a random guess of answers (1/ number of answers). It should also be noted that these approaches are all content-agnostic.

If the estimated user expertise is not domain-sensitive, the simple Votes Scores approach as the estimated user expertise performed the best for all question. When drilled down to questions with at least five answers, it recorded a MRR performance of 0.4885 which is a 4.81% improvement over the baseline. This performance was followed by graph-based approaches with the proposed Credit Graph approach performing better than the state-of-the-art user-user graphs.

The findings showed that the estimated user expertise should be domain-sensitive with performance improvements over each of the general domain-less variants. The best performing approach is the proposed Credit Graph model with a MMR performance of 0.4950 and an improvement of 6.19% over the Z-Index baseline. It is the best performing approach in the prediction of the best answer through the estimated user expertise. Such improvement was however found to not be statistically significantly better than the baseline than the baseline ($t = 2.750$, $p = 0.00615$, $df = 602$)[1].

In the following parts of this sections, we would analyse the performance of each of the explored approaches with respect to the research questions raised in Section 5.1.

5.4.6 Analysis: Inferring Answer Quality with User Expertise

The findings from this evaluation step answered the research questions proposed in Section 5.1 – that it is possible to infer the information quality without processing the answers

[1]Using a Paired-2-Sample T-Test with Bonferroni [39] correction for multiple comparisons (α = 0.05/#comparisons = 0.00217)

themselves. We achieved this by inferring the information quality of answers from the estimated expertise of the answerer themselves, with the expert users producing better answers. Thus, this enabled us to then predict the best answerer in a question by just identifying the answerer with the best expertise.

Table 5.1: Mean Reciprocal Rank (MRR) performance for the evaluated approaches with questions from April 2008 to April 2009. Best performer bolded for each category.

Approach / Model	Topic	Mean Reciprocal Rank (MRR)					
		All Questions		At least 3 Answers		At least 5 Answers	
		Selected	vs Baseline	Selected	vs Baseline	Selected	vs Baseline
Z-Index (Baseline)	General	0.9439	0%	0.7008	0%	0.4661	0%
Votes Score	General	**0.9452**	**+0.14%**	**0.7077**	**+0.98%**	**0.4885**	**+4.81%**
User-User (Count, Global)	General	0.9438	-0.02%	0.7000	-0.12%	0.4691	+0.64%
User-User (Count, Local)	General	0.9438	-0.02%	0.6999	-0.13%	0.4677	+0.34%
User-User (Votes, Global)	General	0.9438	-0.02%	0.7000	-0.12%	0.4700	+0.84%
User-User (Votes, Local)	General	0.9441	+0.02%	0.7017	+0.12%	0.4736	+1.60%
Credit Graph (Temporal Credit)	General	0.9439	-0.01%	0.7003	-0.07%	0.4674	+0.28%
Credit Graph (Votes Credit)	General	0.9447	+0.08%	0.7046	+0.54%	0.4829	+3.60%
Z-Index (Baseline)	Topical	0.9442	+0.02%	0.7020	+0.16%	0.4798	+2.93%
Votes Score	Topical	0.9439	0.0%	0.7008	0.0%	0.4889	+4.90%
User-User (Count, Global)	Topical	0.9453	+0.14%	0.7079	+1.01%	0.4851	+4.07%
User-User (Count, Local)	Topical	0.9455	+0.16%	0.7082	+1.05%	0.4803	+3.83%
User-User (Votes, Global)	Topical	0.9452	+0.14%	0.7077	+0.97%	0.4884	+4.78%
User-User (Votes, Local)	Topical	0.9455	+0.17%	0.7084	+1.08%	0.4866	+4.38%
Credit Graph (Temporal Credit)	Topical	0.9458	+0.20%	**0.7109**	**+1.43%**	0.4779	+2.53%
Credit Graph (Votes Credit)	Topical	**0.9420**	**+0.21%**	0.6904	-1.49%	**0.4950**	**+6.19%**

5.4.7 Analysis: User Expertise Should be Domain-Sensitive

Our findings here suggested that the estimated user expertise should be domain-sensitive with each of the approaches performing better than their general non-topical counterparts though some improvements were minor. The improvement in performance was most notable for questions with at least five answers. This finding is consistent with the notion that users are not experts in every topic or domain [69]; but instead specialises in certain domains and topics which they are interested in. Users are then able to make significant contribution (content of higher quality) in these domains. Thus, there is a need to identify and differentiate the domains which the users are specialised in instead of a single expertise measure.

The challenge however lies in identifying the domains or topics for the users' question-answering interactions. This would often require some form of content-analysis and models if the questioners themselves do not select the topic, domain or category of the question. For such occasion, CQA platforms such as the Yahoo! Chiebukuro platform would resort to automatic classification according to the content of the questions. While the goal of this research is towards the content-agnostic approaches; the algorithms explored in this research can be incorporated, complimented or enhanced with topic models for an improved domain-sensitive user expertise estimation.

5.4.8 Analysis: Information Requirements

All of the approaches towards the estimation of user expertise would require prior user interaction information. Often, there is a correlation between the amount of information available for the approaches and the performance of the approaches; where additional information would help to train better models for the estimation of user expertise. There are more question-answering interactions available for the estimation of user expertise if the expertise is non-topical.

We however observe that despite having access to fewer user interactions for the estimation of user expertise, the evaluated approaches do estimate user expertise better when it is domain-sensitive. It is likely that user interactions which are of different domain or topics turned into unwanted noises for the user expertise estimation process. Still, all of the models performed consistently well for all of the 100 domains evaluated.

5.4.9 Analysis: Are Graph Approaches Needed?

Table 5.1 shows the performance evaluation of simple approaches against the more complex graph-based approaches. Our findings here are consistent with earlier findings that simple approaches are competitive in performance to that of more complicated graph approaches [158]. This observation can be made for both the estimation of general and topical expertise of the users:

- General user expertise. The simple Vote scores approach is the best approach, outperforming the current state-of-the-art user-user graph approaches. It can also be observed that there are occasions where the graph algorithms performed below the baseline Z-Index although by a marginal amount. The proposed Credit Graph approach does however perform better than the state-of-the-art graph algorithms and is thus more competitive against the best performing voting-based approach.

- Topical user expertise. The simple algorithms remained competitive to the state-of-the-art algorithms for the topical models as well, especially with the voting-based approach outperforming all of the User-User graph models for questions with at least five answers. We note however that the proposed Credit Graph approach

does produce the best performance over both the simple and state-of-the-art graph approaches for questions with at least five answers, a situation where answer ranking is needed. Still the performance improvement is still minor and insignificant when compared to the simple approaches.

5.4.10 Analysis: User-User vs Credit Graph

The findings suggest that the proposed Credit Graph model for CQA platforms outperform the commonly used state-of-the-art User-User graph model baseline and also the simple approaches [158]. Each user interaction (asking a question or answering a question) provides a different in contribution; and moreover, the questions themselves are of varying difficulty, importance and domain. Thus, the generalized graph introduced with the Credit Graph appear to be more suitable for modelling CQA platforms.

5.4.11 Analysis: Significance of User Interactions

User interactions on CQA platforms can be good indicators for the estimation of user expertise in the prediction of answer quality. Introduced with the Credit Graph model, the credit functions weights each user interaction according to their contribution in answering the question. A contribution of this research is identifying user interaction features that indicates contribution – (1) user votes; and (2) temporal order of user answers.

Findings from Table 5.1 showed that the user votes gained by the user answers can be indicators of contribution; as the community review of the content [159]. The user votes gained are used directly as a measure of user expertise[2] or as a measure of contribution significance[3]. We do however note that the user votes may not reflect the correct assessment of content due to the possible temporal bias during the voting process [156] where newer answers are only submitted after votes have been assigned to earlier answers. This possibly explains the inconsistency or drop in some of the explored user votes variants.

Thus, we also introduced the use of an answer's temporal ordering in a question to infer the significance of the answer. The findings suggest that the temporal ordering of answers as a structural information of CQA platforms is useful to estimate both the general and domain-sensitive relative user expertise where we observed the temporal-ordered credit function performed well. This validates the earlier assumptions introduced in Section 5.3 that newer answers would be of higher quality as users especially experts exhibit the behaviour of answering questions that do not already provide a good answer [113].

5.4.12 Conclusion

The findings from this section suggests that it is possible to infer the information quality of user answers according to the answerers' expertise for answer retrieval. The best answer retrieval performance measured with the mean reciprocal rank (MRR) solely according to the estimated expertise of the answerer is well above a random guess for the best answer, even for questions with high number of answers[4].

This research explored several state-of-the-art approaches in the estimation of user expertise through content-agnostic means; while also proposing new additions and adaptation for CQA platforms such as the credit graph. User answering interactions have the information potential to reflect their expertise by measuring the contribution of their answers as indicated by features such as user votes and temporal order.

[2]Vote score approach, user-user graphs with vote scores and the vote-based credit graphs.
[3]Z-Index approach.
[4]Questions with at least five answers.

5.5 Expert Search Study

Expert search is the task of identifying and profiling [163] knowledgeable users on a given domain [34]. By identifying reliable experts on CQA platforms, questions can be routed to suitable experts within the questions' domain [93] for improved matching in getting answers that meets the information need of the questioners. This approach is unlike the common practice of retrieving information from previously answered questions based on question similarity. The benefit of such practice includes:

- Increase in question response rate where questions can be directed quickly to reliable experts instead of waiting for the question to be discovered by the suitable experts themselves; especially when there is a large amount of user activity on the platform. Besides that, this can be expanded for the questioners themselves to be directed straight to the suitable experts or their past content which would be of interest.

- Better effectiveness in meeting the information need of the questioners, where the questions can be answered directly by reliable experts in the domain of the problem. Furthermore, the matching can be adjusted to identify suitable experts within a difficulty range which is suitable for the questioner's consumption as the research's proposed user expertise is a relative measure between users. For example, questions from a novice user should be answered by reliable users who are of close to the expertise level of the questioner instead of a high level expert user whose answers may be too convoluted for the questioner.

- To the best of our knowledge, current answer retrieval approaches on CQA platform only measure the similarity between a new question and previously answered questions on the platform as discussed in Section 4.2.2 without considering the difficulty of the question. Plus, the expertise of the questioner is not considered in identifying the similar questions. With the estimation of user expertise, it is possible to better measure the similarity between questions by considering the expertise of the questioners and the answerers who respond to the questions.

The challenge lies in identifying the topical experts on CQA platforms. In Section 4.3, we discuss the many attempts in user profiling and estimation of user expertise. Earlier in this chapter however, we showed that user expertise can be estimated through content agnostic means and then be successfully utilised for answer retrieval – where users with the higher expertise providing a higher probability in contributing good answers that meet the information need of the questioner. Thus, we further explore the estimated user expertise from earlier sections for the task of expert search.

5.5.1 Methodology

This research explores the user expertise estimation for expert search as a follow-up to the answer retrieval evaluation discussed in Section 5.4. The topical models are trained based on the data between year 2007 and year 2008 of the Yahoo! Chiebukuro corpus (Section 4.4.1 for a total of 100 randomly selected domains. The evaluation and analysis are then performed on the data between year 2008 to year 2009 for the same domains used during training. In the topical expert search study, the models explored here are the same as the ones explored for the answer retrieval evaluation earlier.

Study Measures

The main measure of this study is the Pearson Correlation [114] which measures how well two sets of data are related. This research fixed the first set of data to be the estimated

relative user expertise (from the model in the previous year) and the other to be the measures obtained from the dataset (in the following year):

- **Total answer posted.**
  This measures the activeness of a user in answering questions on the CQA platform. The correlation between the estimated user expertise and the activeness of the user would provide us with a better insight on the behaviour of identified experts through each algorithm.

- **Selected best answer.**
  The selected best answer is the answer credited by the questioner himself/ herself that best meet the questioner's information need. The correlation here measures the relationship between the estimated user expertise and the ability of the user to meet the information need of the questioner.

- **Voted best answer.**
  The voted best answer measure is the answer with the highest votes for the question. The correlation here measures the relationship between the estimated user expertise and the contribution of the user as judged by the community.

- **Effectiveness of selected best answer.**
  Measuring the correlation between the estimated user expertise and the count of the best answer contributed by that user is not sufficient not every user is equally active. Thus, this correlation measures the effectiveness of the users in provided the best answers as selected by the questioner by accounting for the total number of answers provided by the users.

- **Effectiveness of voted best answer.**
  Like the above, for the best answers voted by the community instead of questioner selected best answers.

- **Effectiveness of best answer.** This measures the relation between the estimated user expertise from each model with the ability of the users to contribute answers that best meet the information need of the questioner or deemed to contain the highest contribution.

5.5.2 Findings

The findings of the measures from the study detailed earlier are reported in Table 5.2. We then proceed to analyse the findings from the study for each of the user expertise estimation approaches.

5.5.3 Observation and Analysis: Z-Index

We observe that the user expertise estimated from the Z-Index model have the highest correlation with the number of answers posted. These experts, identified through the Z-Index model are active contributors of answers towards questions. This finding is consistent with the Z-Index approach which calculates the ratio of answers contributed against the questions posted [158]. Thus, these estimated experts were active in the year prior (during the training year) to deserve high expertise estimate from the model and continue to be active in the following year (evaluation year).

Despite the large number of answers contributed by the model estimated experts, we observe a healthy amount of these answers can meet the information need of the other users especially the questioners themselves with the best recorded correlation for the effective

selected best answer. A high number of these best answers were however from questions with low number of answers (situations which answer retrieval is less important), consistent with the earlier findings from Section 5.4.

We however do note that these estimated experts are already active users and thus would be contributing towards the platforms and their community without the need to identify, search and encourage these users to further participate on the platform.

5.5.4 Observation and Analysis: Votes

It can be observed that unlike the Z-Index model, the inferred experts from the Vote Score model do not have a high correlation with their activeness in the evaluation year. This is possibly due to the model relying on the votes which the users garnered instead of only the number of answers provided; an approach that do not penalized inactive users unlike the Z-Index model.

Despite recording the lower correlation with the number of best answers selected and voted for their contributions, the experts inferred were slightly more effective in meeting the information need of the users especially the community (high effective best answer votes). This is further supported from the earlier answer retrieval findings that the Votes Score approach is competitive for questions with high amount of answers explored in Section 5.4. With the same efficiency but lower number of answers posted; the identified experts from the model could be encouraged to participate actively and contribute more towards the improvement of the platform.

5.5.5 Observation and Analysis: User-User Graphs

The performance of the User-User graph models varies according to the variants. Generally, the inferred experts from the model perform similar to the experts inferred from the Votes Score approach where the identified experts are not active in the evaluation year but are as effective with their answers. The finding here is also consistent with the answer retrieval performance for questions with at least five answers of these state-of-the-art User-User graphs being competitive with the Votes approach for user expertise estimation.

5.5.6 Observation and Analysis: Credit Graphs

Despite performing the best during the answer retrieval evaluation (Section 5.4), the experts estimated from Credit Graph models recorded low correlation with the user activity of answer posted in the evaluation year. Thus, it is natural to encourage the identified experts to contribute more towards the platform and its community.

This study however shows low effectiveness in identifying the users that produce the best answers. Such behaviour is possible as the Credit Graph approach only performs well above the competition for questions with at least 5 answers – a scenario where answer search is important (possibly difficult questions) which are not dominant on CQA platforms. Thus, we can utilise this approach to locate inactive users who are yet able to provide good answers in a competitive environment.

Table 5.2: Study of Pearson Correlation between the Estimated User Expertise from the Approaches/ Models and User Contributions on Yahoo! Chiebukuro.

Approach / Model	Pearsons Correlation between User Expertise and:					
	Total Answer Posted	Selected Best Answer	Voted Best Answer	Effectiveness of Selected Best Answer	Effectiveness of Voted Best Answer	Effectiveness of Best Answer
Z-Index (Baseline)	0.6530	0.6254	0.5845	0.0462	0.0292	0.0569
Votes Score	0.2680	0.2756	0.2672	0.0421	0.0349	0.0582
User-User (Count, Global)	0.2299	0.2220	0.2099	0.0353	0.0263	0.0466
User-User (Count, Local)	0.2536	0.2447	0.2299	0.0354	0.0273	0.0474
User-User (Votes, Global)	0.2452	0.2118	0.2042	0.0385	0.0291	0.0511
User-User (Votes, Local)	0.2343	0.2445	0.2394	0.0402	0.0326	0.0551
Credit Graph (Temporal Credit)	0.2083	0.2136	0.2091	0.0144	0.0111	0.0193
Credit Graph (Votes Credit)	0.1856	0.2179	0.2271	0.0204	0.0171	0.0283

5.5.7 Conclusion

The explored user expertise estimation approaches can identify expert users who are able to contribute good answers in the future. The findings suggest that the Z-Index baseline approach is able to identify expert users who are active, whereas the other approaches are able to better identify inactive expert users which is a greater feat as there are less information about these users. In fact, these experts are highly effective in providing the best answers in meeting the information need of the questioner.

Despite the credit graph approach achieving the best answer retrieval performance (in Section 5.4) and it being able to identify inactive experts; the identified experts were found to not be effective enough in providing the best answer when compared to the baseline. Once again, we saw that the simple approaches are competitive in terms of performance with that of the state-of-the-art graph-based approaches for the estimation of user expertise. Thus, this drives our research to explore a new direction – the estimation of user expertise later in Chapter 6.

5.6 Summary

In this chapter, we explore several approaches to estimate user expertise on CQA platforms with a focus on graph-based approaches. As discussed in Section 5.2, the current state-of-the-art approaches to user expertise estimation models CQA platforms as user-user graphs and propagating authority as user expertise; according to the questioner-answerer or answerer-answerer relations.

Our research proposed an alternative graph model in Section 5.3 – the credit graph model which accounts for the difference in user answering relations; which is adapted from our earlier success in Part I. This is overlooked in the user-user graph models which assumes each user answering interaction to be the same when aggregated as the user relation weights. We argued that user questions are of varying difficulty and thus require a need for a mutual reinforcement between user expertise and question difficulty; instead of considering each question and answering action to be the same. Thus, providing the better answer in a difficult question should reward the answerer with a greater expertise measure. In order to do so, the credit graph model each user answering interaction as a user-question relation. These relations would however need to be weighted. Our research explored how answerers' contribution can be measured by extracting content-agnostic features on CQA platforms; presented as credit functions in Section 5.3.1.

The estimated user expertise of answerers is then used to infer the information quality of their answers; relying on the assumption that expert users have a higher probability in contributing better answers that meets the information need of the question [18]. In Section 5.4, we showed how the information quality of answers can be inferred directly from the expertise of the answerers without processing the answer content directly as a content-agnostic approach to the problem. All of the explored approaches were competitive in the prediction of answer though we note that simple approaches such as the Z-Index and Credit Scores are in fact competitive to the more sophisticated graph-based approaches (both User-User graphs and Credit graph); consistent with earlier findings from the literature [158]. Besides that, we identified features such as user votes and the temporal ordering of answers as signals for the estimation of user expertise.

This chapter concludes with a study on expert search – how experts can be identified and retrieved [163] according to their estimated expertise from the discussed approaches. Our study showed that the simple approaches to the estimation of user expertise such as the Z-Index approach is suitable to estimate the expertise of active users; with the more complex approaches otherwise. Despite the good answer retrieval performance, the credit graph approach were not performing better than the single approaches for expert search.

This leads us to explore new approaches to the estimation of user expertise in the following Chapter 6.

Chapter 6

Competitive Pairwise Comparison for User Expertise Estimation

Graph-based approaches to user expertise estimation on CQA platforms propagate authority according to users' answering interaction; with higher propagation if the answer is of greater contribution according to the mutual reinforcement between user expertise and answer quality [16]. In the earlier Chapter 5, these approaches are successful in estimating user expertise for answer retrieval (Section 5.4 for best answer). In performing expert search (Section 5.5), we saw these approaches being able to identify experts of lower activity; though the identified experts from the Credit Graph approach were not as competitive as the User-User graphs in identifying effective experts.

In this chapter, we explore the alternative to graph-based approaches in utilising the CQA features for the content-agnostic estimation of user expertise with the goal of addressing the research questions raised in Section 6.1. If the answerers' contribution can be measured, then it can be compared between answerers – with the answerer of higher expertise being able to contribute better answers. In hindsight, this is the motivation for the user-user graphs where authority is propagated from an answerer to be better answerer in the answerer-best answerer graph and from the questioner to the answerer in the questioner-answerer graph.

Thus, it is possible to form user-user pairs for performance comparisons; often known as pairwise comparison approaches used in many domains such as sporting competition [129]. Here, the implicit pairwise comparison between two answerers is implied through their answering performance[1] [95] and these users are rated within a rating system as a measure of expertise which can be easily updated unlike the earlier graph-based approaches. We discuss this approach further in Section 6.2 of this chapter. Many of the current state-of-the-art pairwise comparisons are only concerned in the outcome and not the actual performance difference; similar to that of real-world sports like tennis. This research argues otherwise in Section 6.3 through a proposed competitive pairwise comparison model with the addition of win-margins and temporal rating period.

To evaluate the performance of expertise estimation approaches, this research once again sets up the same answer retrieval task used in Section 5.4 as the evaluation methodology in Section 6.4. Similarly, we perform the same study for expert search in Section 6.5. Finally, we conclude this chapter and our study on CQA platforms with a summary in Section 6.6.

[1] How well their answer meet the information need of the questioner.

6.1 Research Questions

Further expanding on the research questions introduced in Section 1.4 and accounting for the answered research questions from Section 5.1, the following research questions guide our research into the pairwise comparison approach to the estimation of user expertise on CQA platforms.

6.1.1 Can User Expertise be Estimated on CQA Platforms Through Pairwise Comparison Approaches?

In Section 4.3.3, we saw that user expertise can be estimated on CQA platforms with graph-based approaches. In this chapter, this research attempts to perform the same with pairwise comparison approaches; exploring how to best do so for CQA platforms – how to best build the comparison pairs from user answering interactions on the platform and which features are suitable to be used as the performance measure. We present our proposed pairwise comparison approach in Section 6.3 with several key enhancements to the current state-of-art approaches.

Similar to player rating in sports, user ratings can be used to indicate the expertise of users; where users of similar expertise perform the same [101]. Unlike the expertise measure from graph-based approaches, user ratings are able to account for the change in user expertise over a temporal period [150]. Besides that, the expertise rating can be updated directly between the users in the comparison pairs going forward.

6.1.2 Do Expert Users Produce Better Answers?

The estimated user expertise is once again used to infer the information quality of their answers as a content-agnostic approach. This is motivated by the observation that users of higher expertise tend to produce content which meets the information needs of consumers [113, 158]; enabling the questioners to quickly[2] identify the good answers amongst a large amount of sub-par answers[3].

Unlike Chapter 3, we now estimate user expertise through several pairwise comparison variants. Our goal here is to better estimate user expertise for an improved answer retrieval performance later in Section 6.4 when compared against all of the earlier explored approaches. Besides that, we also study the effectiveness of identified experts for answer contribution as an expert search task in section 6.5.

6.1.3 Can We Predict the Best Answer in a Question Thread?

Just like Chapter 3, this research attempts to predict the best answer that could meet the information need of the questioner even when there is large amount of answers of varying quality in the question thread itself. As the estimated user expertise are still relative measures between the users, we can infer the information quality of their answers as a relative measure in order to suggest the best answer during answer retrieval in Section 6.4.

6.2 Pairwise Comparison Approaches

Pairwise comparison approaches on CQA platforms do model the question-answering interaction of its users as a competition between the answerers. Each answerer would attempt to perform best to meet the information need of the answerer. This enables pairs to

[2]See Section 4.4.7 on answer cold-start.
[3]See Section 4.4.8 on varying answer quality.

be formed between answerers in each question thread and these answerers are compared according to their answering performance [95]. Depending on their performance, the expertise of these answerers would be estimated – with the better answerer having a higher expertise analogously how we evaluate a tennis player's performance during a match.

6.2.1 Rating as Expertise

Rating systems are commonly used for competitive sports and games to rank their players. One of the earliest notable rating systems is the Elo rating system which builds a simple statistical model which estimates the ability levels (or commonly known as the skill) of its players [41]. Elo is initially used for chess [47] and then expanded to many other sports[4] and games[5]. Elo has then evolved over the years with the introduction of newer rating systems such as Glicko [48] and TrueSkill [57]. Many of these ratings are relative measures which enable the comparison between users.

With user ratings, it is possible to predict the outcome of a match by considering the difference in ratings between two competitors. A higher rated player is expected to win when matched against another player of lower rating. Thus, modern rating systems such as Glicko and TrueSkill are used to improve matchmaking that puts players or teams of similar skill levels together for a satisfying competitive experience [51][6].

Definition 15 (User Rating)**.** A measure of strength to reflect a user's true capabilities [46].

The user ratings are updated according to the outcome of matches such as the work of Liu et al. [95] with the Trueskill rating system for CQA platform. These matches are the pairs formed between the answerers as a competition within each question thread; with the outcome determined according to their answering performance. For both Glicko and TrueSkill, a player's performance follows an approximately normal distribution according to the player's average skill (mean μ_u) with an uncertainty factor (deviation σ_u). These measures are then updated for the winner and the losers after each question. A user's expertise can be inferred as Function 6.1 as a conservative estimate of user skill [95].

$$\textbf{Expertise}(u) = \mu_u - 3\sigma_u \tag{6.1}$$

6.2.2 User Pairings

For pairwise comparison approaches, user-user pairs are needed to compare the users according to their performance. This is akin to a match between players and the outcome of the match is then used to update the user ratings which we consider to be user expertise. In the literature, user-user pairs are formed according to the user interactions within a single thread, usually between the answerers and the best answerer of the thread (for rating with TrueSkill) [95] similar to the edges on answerer-best answerer graph [10]. Recall the scenario earlier from Figure 5.1 with two questions involving six users. User-user pairs can be formed within each question, with the best answerer as the winner in a pair:

- **Question**1**.**
  3 user pairs between User3 as the best answerer against User2, User4 and User5.
- **Question**2**.**
  A single user pair between User6 as the best answerer against User5.

[4]Notably the National Football League (NFL).
[5]Notable Magic the Gathering, World of Warcraft and League of Legends.
[6]Notably Glicko in Guild Wars 2 and TrueSkill in the Halo franchise.

The motivation for such pairing is due to the ease in identifying the best answerer on CQA platforms through the highest voted answer [159] or questioner's best answer selection (as seen on Yahoo! Chiebukuro). Besides that, such pairing is favourable for the result oriented approach to pairwise comparisons with the best answerer as the ultimate winner.

6.2.3 Result or Outcome Oriented

Many of the current user rating systems are only concerned with the outcome of a match as an indicator of user performance. The higher rated user should perform better than the lower rated user, coming out on top in the match. If the outcome is not as expected, then it is an upset. The actual performance or win-margin which resulted in the outcome is not considered when the user rating is updated. For example, in Glicko, the winner will receive an outcome score of 1 for winning and the loser a score of 0; both receiving a score of 0.5 if a draw occur. As we discussed earlier that the user-user pairs are built to always include the best answerer, there is a winner and a loser with no possibility of a draw occurring.

6.2.4 Ratings Update

User ratings are updated according the outcome of a match [41] or a pairwise comparison in our context. Often, the user ratings are updated directly after a match [95] which is also known as a live rating. This is often the preferred method as it enables the ratings to self-correct as soon as possible for the next match, especially when the user ratings are used for matchmaking [57]. Alternatively, the user ratings can be updated at the end of a competition – after a series of matches especially when there is a temporal gap between the competition to account for the potential user growth [101].

6.3 Competitive Rating Approach

The competing rating is this research's adaptation of the pairwise comparison approach for the estimation of user expertise on CQA platform. It is based on the competition approach discussed in Section 6.2 with several changes which we shall discuss in the later sections. The proposed competitive rating approach is modelled according to the following:

- Each question is a competitive tournament organised by the questioner as the user who open the question thread for answerers to participate. As the organiser, the questioner will not be participating in the competition, thus not having their expertise estimated or updated. Our reasoning for this model is due to the difference in intention between the questioner and the answerers of the question which prevents reasonable comparison[7].

- The participants of the competitive tournament are the users who chose to answer or respond to the question. These users as answerers would be judged and have their expertise estimated or updated.

- The performance of each participant is correlated with the number of votes received by their answer. This would be collected after a certain period of time as user votes can be regarded as the community's long-term judgement of content [9], as a measure of the user's contribution towards meeting the information need of the questioner. User performances are comparable within the same competition and the best performing user is the best answerer within the question thread.

[7]Unless we were to model the questioner as the ultimate loser.

- The user expertise level does not change much within the short period of time [95, 101]. Thus, their answering performance is consistent according to their capabilities for questions within the same period.

6.3.1 Glicko-2 Rating

Pairwise comparison approaches have used TrueSkill as a measure of user expertise on CQA platforms [95]. It is a Bayesian skill rating system used for calculating the relative skill level of players in multi-player or team games such as the Halo franchise. It is patented [58] and limited to use by Microsoft.

To incorporate the competitive scores and the rating period into our competitive rating model for user expertise estimation, we explore the Glicko-2 [46] rating model, an extension on a previous Glicko [48] model. Glicko-2 uses a logistic distribution unlike the Gaussian distribution of TrueSkill which is similar in application and is computationally efficient for pairwise comparison between two users. Besides that, Glicko-2 is able to handle the sudden change in merits for the user rating and variance stochastically [46]. In the following sections, we shall discuss Glicko-2 in greater detail.

User Rating

The user rating μ_u is a measurement of strength. In our application, the user's strength is regarded as the user's expertise. The user rating is updated after each rating period as shown in Equation 6.2 according to the outcome of matches between user u_x and his/her opponents during the same rating period t.

$$\mu_{u_x}^{t+1} = \mu_{u_x}^{t} + (\sigma_{u_x}^{t+1})^2 \cdot \sum_{j=1}^{max} g(\sigma_{u_y}^{t})(w_{u_y} - P(\mu_{u_x}^{t}, \mu_{u_y}^{t}, \sigma_{u_y}^{t})) \tag{6.2}$$

In the traditional Glicko rating, w_{u_y} represents the total outcome (or scores) of the pairwise comparison of user u_x against user u_y with each game producing a value of 1 if u_x won, 0.5 if it is a tie and 0 if u_y won. As we introduced the competitive scores (Section 6.3.3), w_{u_y} represents the total aggregated win-margin scores of u_x against u_y in the rating period t.

The strength uncertainty confidence, $g(\sigma)$ distinguishes the uncertainty in a user's strength within the Glicko rating system. It is mainly used for measuring the opponent's strength (Equation 6.3). From the equation, we can see that the strength confidence increases when the rating deviation σ is low. When applied into Equation 6.2, opponents with lower rating deviation σ_{u_y} amplify the outcome of the match to update the user's rating μ_{u_x}.

$$g(\sigma) = \frac{1}{\sqrt{1 + 3\sigma^2/\pi^2}} \tag{6.3}$$

The approximate probability of user u_x defeating opponent u_y as a function of estimated strength is described in Equation 6.4. From the equation, we can see that the probability of defeating an opponent is dependent on the difference in user rating $\mu_{u_x}^{t} - \mu_{u_y}^{t}$ where it would be a large probability if the difference is positive and a small probability if the difference is negative. The uncertainty in opponent's strength $g(\sigma_{u_y}^{t})$ is accounted for as well.

$$P(\mu_{u_x}^t, \mu_{u_y}^t, \sigma_{u_y}^t)) = \frac{1}{1 + exp(-g(\sigma_{u_y}^t)(\mu_{u_x}^t - \mu_{u_y}^t))} \tag{6.4}$$

User Rating Deviation

The rating deviation σ_u is a feature of the Glicko algorithm as an extension of Elo. The deviation can be used as a standard deviation which measures the uncertainty in the user rating where a higher value signifies a higher uncertainty. The user rating and the user rating deviation can be used together to estimate the strength (expertise in our application) of the user better by reporting the 95% interval in the range of $[\mu - \sigma, \mu + \sigma]$. We can take a conservative approach (similar to that of TrueSkill [95]) to take the lower limit of the strength estimate.

Just like user rating, user rating deviation is updated only after each rating period with the equation shown in Equation 6.5. If a user did not participate in the rating period, the user's rating deviation is updated as Equation 6.6.

$$\sigma^{t+1} = \sqrt{\left(\frac{1}{(\sigma^t)^2 + (\phi^{t+1})^2} + \frac{1}{\delta^2}\right)^{-1}} \tag{6.5}$$

$$\sigma^{t+1} = \sqrt{(\sigma^t)^2 + (\phi^t)^2} \tag{6.6}$$

with the δ^2 as the game outcome variance as shown below:

$$\delta^2 = \left(\sum_{j=1}^{m} (g(\sigma_{u_y}^t))^2 P(\mu_{u_x}^t, \mu_{u_y}^t, \sigma_{u_y}^t)(1 - P(\mu_{u_x}^t, \mu_{u_y}^t, \sigma_{u_y}^t))\right)^{-1} \tag{6.7}$$

The volatility component, ϕ is the main improvement of Glicko-2 over the original Glicko rating system. It is this volatility which describes the change in merits where high volatility ϕ denote sudden shifts. A higher volatility will result in a higher user rating deviation which suits the goal of the Glicko-2 rating system. It usually starts at a default value of $\phi^0 = 0.06$ and ϕ is updated by an iterative procedure [46].

6.3.2 Proposed: Formatted Pairings

Instead of only looking the answerer-best answerer pairs as discussed in Section 6.2.2, this research proposes alternative pairs for comparison. We model each question thread as a Round-Robin league, resulting in pairs between every answerer in the thread. Thus, from the example scenario in Figure 5.1, we would form the following competitive pairs:

- **Question1.**
  6 user pairs: User2-User3, User2-User4, User2-User5, User3-User4, User3-User5 and User4-User5.
- **Question2.**
  1 user pair: User5-User6.

The questioner is excluded from the competitive pairs, the User1 in Question1 and User 2 in Question2 because they are the organisers as introduced earlier for the competitive model. This differs from the landmark approach [95], which models the questioner as the ultimate loser in the question.

It is possible to build the competitive pairs in many ways using various competition model for a question such as Brackets, Double Elimination and Bubble Brackets. As of now, our current research only focuses on the Round-Robin format to get every user involved in order to secure enough user interactions for the estimation process. Without sufficient user information, pairwise comparison approaches often struggle [95].

6.3.3 Proposed: Win-Margins

We discussed in Section 6.2.3 that many of the traditional approaches only consider the outcome or the results of a match. These approaches [10, 95] only treat the best answerers[8] as the winners and every other answerer as the losers without considering the actual performance or match score which led to the outcome. This could result in two different answerers with varying expertise being updated the same if they suffer a loss in the same question thread. Furthermore, draws are often neglected as there are always winners as long as the best answerer can be identified.

Many of the modern CQA platforms do allow users to vote. Thus, we propose the use of votes as contribution scores for measuring the users' performance in the competitive pairs built according to Section 6.3.2. This enables us to compare the performance between each answerer competitively through the votes gained by their answers in the question – with the answerer who gained more votes as the winner of the match. But instead of only focusing on the outcome of the match, we incorporate the user votes to determine the win-margin as the performance difference between the users.

Definition 16 (Win-Margin)**.** The performance difference that resulted in the observed outcome.

Due to the varying number of user votes within a question thread, this research propose a win-margin function to score the performance of the users in match as shown in Figure 6.8. This is inspired by the Bradley-Terry model which correlates the win probability of a player in a game with the player's real rating μ_x against the opponent's rating μ_y to be $\frac{\mu_x}{\mu_x+\mu_y}$ (for winning) and $\frac{\mu_y}{\mu_x+\mu_y}$ (for losing) [20]. Function 6.8 can be adjusted to account for the negative votes on some CQA platforms by rescaling the votes during pre-processing.

$$\mathbf{WinMargin}_{q_k}(u_x, u_y) = \frac{V_{q_k}(a_{u_x})}{V_{q_k}(a_{u_x}) + V_{q_k}(a_{u_y})} \tag{6.8}$$

6.3.4 Proposed: Rating Period

User ratings are often live – where the user ratings are directly updated at the end of every match [95] or at the end of every competition (question in our CQA context). The user expertise is updated live to immediate use particularly for matchmaking in competitive games [57]. The challenge in estimating user expertise on CQA platform is that users do enter CQA platform at different level of expertise. By observing user interactions over a period, it is possible to approach even closer towards the true expertise of the users.

Users can also gain expertise over time with each user going through a different growth rate [150]. It should however be noted that user expertise and behaviour do not change much within a short period [83, 95]; with the user performance being relatively consistent within the same period [101].

Thus, for efficiency purposes, this research proposes using a rating period t_k where all outcome within the rating period is collected before being used to update the user rating at the end of the rating period. If users have multiple confrontations within the same rating period, then their scores are aggregated; enabled by Glicko-2 treating matches within the

[8]Based on user votes or questioner selection.

same period as matches occurring simultaneousness [46]. For our research, we explore a number of rating periods for updating – live rating, one day, one week (seven days) and one month (thirty days).

Definition 17 (Rating Period)**.** The amount of time between competitions.

6.4 Answer Retrieval Evaluation

Having estimated user expertise using various pairwise comparison approaches, we put the expertise value to the test for answer retrieval using similar evaluation methodology to that detailed in Section 5.4. The estimated user expertise is directly applied to infer the information quality of user answers in a question thread in order the predict the best answer that would meet the information need of the questioner.

6.4.1 Evaluated Approaches

As we already observed that topical domain expertise is able to improve answer retrieval, our evaluation process now focuses on the domain-specific expertise estimated with pairwise comparison approaches. All the evaluated approaches from Section 5.4.4 make their return for evaluation here, using the Z-Index approach as the baseline again.

Proposed: Competitive Rating with Glicko

This research proposed some changes to the pairwise comparison approach for the estimation of user expertise as the competitive rating approach, discussed in Section 6.3. For our evaluation, we set-up the question threads from the dataset as a Round-Robin league where each user will face each other once in their same question; and their performance in the match is indicated by the number of votes gained by their answer.

The estimated user expertise are domain-specific; thus, the question threads are separated according to their assigned topic[9]. The default volatility of $\phi^0 = 0.06$ is used for the competitive rating approach. The explored variants are designed to compare the state-of-the-art competition inspired pairwise comparison approach [95] using the Glicko-2 rating system against our proposed competitive rating approach:

- **Scoreless (competition) vs win-margin (competitive).**
  To the best of our knowledge, the current pairwise comparison approaches as discussed in Section 6.2.4 would update the user ratings according to the outcome of the comparison without regard to the actual user performance which led to the outcome – the winner getting a score of 1 and the loser a score of 0 or both getting a score of 0.5 when it is a draw. We introduced the win-margin outcome measure in Section 6.3.3 as an alternative where the actual user performance does matter in updating the expertise rating of the users.

- **Live update (competition) vs rating period (competitive).**
  User ratings of answerers are directly updated following the outcome of a match or a competition (see Section 6.2.4), known as a live update. The proposed competitive approach discussed in Section 6.3.4 would only update user ratings between rating periods such as one day, one week and one month.

[9] See Section 4.4.1 for the topics in the dataset.

6.4.2 Results

The answer retrieval performance for the best answer prediction of all evaluated approaches and their variants are presented in Table 6.1. In this chapter, we focused on how pairwise comparison approaches perform against the earlier explored approaches from Section 5.4 of Chapter 3. When compared against the baseline Z-Index simple approach for topic sensitive user expertise, we saw all of the pairwise comparison approaches (both the state-of-art and our proposed variants) performing very well – up to an improvement of 21.14% with a MRR of 0.5646 for questions with at least 5 answers. This performance improvement difference is statistically significant ($t = 5.818$, $p = 9.65779E - 09$, $df = 602$).

The various pairwise comparison approaches do also outperform all of the graph-based approaches; including the best performing credit graph approach which was proposed in Section 5.3. The credit graph approach obtained a MRR performance of 0.4950 with 6.19% improvement over Z-Index was not statistically significantly better than the baseline ($t = 2.750$, $p = 0.00615$, $df = 602$) and is 14.06 worse than the best performing pairwise comparison approach with the proposed competitive rating. In the follow sections, we analyse the performance of pairwise comparison approaches further.

6.4.3 Analysis: Win-Margin Matters

The best performing pairwise comparison variants for all questions are the proposed competitive rating variants with the win-margins instead of the ones without win-margin or scoreless outcome-focused variant [95]. The win-margin variant detailed in Section 6.3.3 accounts for the actual user performance which led to the outcome of the pairwise comparison; providing additional essential information to the relative nature of user expertise – how much better is a user compared to another. Such additional information to better reach convergence in estimating the actual expertise of users; provided sufficient user interactions are available within rating periods (one day or one week) and not overfit over a long rating period.

6.4.4 Analysis: Periodical Rating Updates

The current state-of-the-art pairwise comparison competition approach with TrueSkill [95] updates the user expertise as a live update. As discussed in Section 6.3.4, user expertise does not change much within a short period of time [95] and produce content of similar information quality within the same temporal frame [101]; motivating the introduction of temporal period[10] for the competitive rating. User ratings are only updated between rating periods, aggregating the outcome from the same rating period. This provides the pairwise comparison approaches with more user information for accurate user rating updates.

As presented in Table 6.1, the pairwise comparison variants with rating periods outperform the variants with live updates of user expertise. The performance difference is around 5% over the baseline in the MRR measure for best answer prediction. From the explored values for rating periods, we observe that the best performing rating periods are one day and one week. The live update (zero days) would update the user expertise too quickly, resulting in inaccurate expertise estimation from incomplete information. A long rating period of one month is not suitable either, as it is not able to capture the change in user expertise as the users grow over time [101, 150], since they consume more knowledge.

6.4.5 Conclusion

In Section 6.1, we ask if user expertise can be estimated on CQA platforms through pairwise comparison approaches. Findings from the answer retrieval evaluation showed

[10]Which is already easily adjustable in Glicko-2 [46].

that pairwise comparison approaches are in fact able to do so, even outperforming both the simple approaches and also the more complex graph-based approaches. As a relative measure, the estimated user expertise enabled us to compare the inferred information quality of contributed answers to better predict the best answer for a question; based on the assumption expert users do produce better answers.

This research suggests modelling question threads as a league, pitting answerers in the question thread against each other in a Round-Robin league format. The outcome from their matches are used to update their expertise rating. Instead of only looking at the outcome of the match, we now compare the performance of these users which led to the outcome as the user-user pairs are no longer built just against the best answerer as the ultimate winner of the league; thus, enabling us to reach accurate user expertise estimation quicker. The performance of each user against their opponent is measured according to the win-margin introduced in Section 6.3.3 to estimate the competitive rating of the users; improving performance over the competition outcome-based scoreless variants. This rating is updated between rating periods, providing sufficient user information for accurate updates while accounting for the potential user expertise growth of the user. The combination of these suggestions does improve the user expertise estimation capabilities of pairwise comparison approaches over the other explored approaches up to 21.14% over the baseline Z-Index approach.

Table 6.1: Mean Reciprocal Rank (MRR) performance for the evaluated algorithms for 53,871 questions in year 2008-2009. Best performance in bold.

Approach	Mean Reciprocal Rank (MRR)					
	All Questions		At least 3 Answers		At least 5 Answers	
	Selected	vs Baseline	Selected	vs Baseline	Selected	vs Baseline
Z-Index (Baseline)	0.9442	+0.02%	0.7020	+0.16%	0.4798	+2.93%
Votes	0.9439	0.0%	0.7008	0.0%	0.4889	+4.90%
User-User (Count, Global)	0.9453	+0.14%	0.7079	+1.01%	0.4851	+4.07%
User-User (Count, Local)	0.9455	+0.16%	0.7082	+1.05%	0.4803	+3.83%
User-User (Votes, Global)	0.9452	+0.14%	0.7077	+0.97%	0.4884	+4.78%
User-User (Votes, Local)	0.9455	+0.17%	0.7084	+1.08%	0.4866	+4.38%
Credit Graph (Temporal Credit)	0.9458	+0.20%	0.7109	+1.43%	0.4779	+2.53%
Credit Graph (Votes Credit)	0.9420	+0.21%	0.6904	-1.49%	0.4950	+6.19%
Competitive (Scoreless, RatingPeriod 0)	0.9507	+0.71%	0.7366	+5.11%	0.5344	+14.65%
Competitive (Scoreless, RatingPeriod 1)	0.9514	+0.79%	0.7408	+5.71%	0.5597	+20.07%
Competitive (Scoreless, RatingPeriod 7)	0.9512	+0.77%	0.7397	+5.55%	0.5539	+18.82%
Competitive (Scoreless, RatingPeriod 30)	0.9512	+0.77%	0.7396	+5.53%	0.5626	+20.70%
Competitive (WinMargin, RatingPeriod 0)	0.9506	+0.70%	0.7362	+5.05%	0.5321	+14.16%
Competitive (WinMargin, RatingPeriod 1)	0.9516	+0.81%	**0.7419**	**+5.86%**	**0.5646**	**+21.14%**
Competitive (WinMargin, RatingPeriod 7)	**0.9516**	**+0.82%**	**0.7419**	**+5.86%**	0.5564	+19.39%
Competitive (WinMargin, RatingPeriod 30)	0.9513	+0.78%	0.7403	+5.62%	0.5590	+19.92%

6.5 Expert Search Study

Similar to the user study in Section 5.5 of Chapter 5 for expert search, we study the identified experts from pairwise comparison approaches and their variants. We follow the same methodology and study measures. The goal from this study is to identify expert users who would be effective in contributing high quality answers in the future; preferably identifying expert users who are inactive for further encouragements to contribute on the platform.

6.5.1 Findings

The user study findings are presented in Table 6.2 for the pairwise comparison variants, aggregated with the other approaches from the earlier study of Table 5.2.

6.5.2 Observation and Analysis: Pairwise Comparison

Compared to the other explored approaches, the experts identified from the pairwise comparison approaches were found to have low correlation with the number of answers posted. These users are hard to spot due to their limited user activity. But despite the lower number of answers contributed, these answerers were able to produce the best answer in a question thread for answer retrieval as we saw in Section 6.4 especially when the questions do have a higher number of answers. Thus, these expert users are wanted and should be encouraged to contribute more on CQA platforms.

6.5.3 Observation and Analysis: Impact of Rating Period

The rating period for the proposed competitive pairwise comparison variant has a significant impact on our expert search study. We saw that as the rating period increase, the identified experts are more active; but still much less active when compared to the experts identified through the simple and graph-based approaches. Besides that, identified experts from a long rating period of thirty days still perform really well for the best answer prediction in answer retrieval – better than the other approaches while remaining competitive with the other rating period.

Interesting observations can made for the effectiveness of the identified experts in contributing the best answers. Despite the fact that identified experts are relatively inactive, these experts are really effective when the rating period is set to thirty days. The effectiveness of these experts decreases as the rating period decreases. The explanation for such an observation is that the Glicko-2 rating decays the user ratings if the users were not active within the rating period. This supports the earlier finding that pairwise comparison approaches are not suitable for CQA platforms with high number of users but with low activity [95]. A possible workaround for this in the future is to adjust the decay portion of the Glicko-2 update functions in order to not penalise inactive users or use a longer rating period to supplement a shorter one.

Table 6.2: Study of Pearson Correlation between the Estimated User Expertise from the Approaches/ Models and User Contributions on Yahoo! Chiebukuro.

Approach / Model	Pearsons Correlation between User Expertise and: Total Answer Posted	Selected Best Answer	Voted Best Answer	Effectiveness of Selected Best Answer	Effectiveness of Voted Best Answer	Effectiveness of Best Answer
Z-Index (Baseline)	0.6530	0.6254	0.5845	0.0462	0.0292	0.0569
Votes Score	0.2680	0.2756	0.2672	0.0421	0.0349	0.0582
User-User (Count, Global)	0.2299	0.2220	0.2099	0.0353	0.0263	0.0466
User-User (Count, Local)	0.2536	0.2447	0.2299	0.0354	0.0273	0.0474
User-User (Votes, Global)	0.2452	0.2118	0.2042	0.0385	0.0291	0.0511
User-User (Votes, Local)	0.2343	0.2445	0.2394	0.0402	0.0326	0.0551
Credit Graph (Temporal Credit)	0.2083	0.2136	0.2091	0.0144	0.0111	0.0193
Credit Graph (Votes Credit)	0.1856	0.2179	0.2271	0.0204	0.0171	0.0283
Competitive (Scoreless, RatingPeriod 0)	-0.0005	0.0001	0.0003	0.0051	-0.0024	0.0021
Competitive (Scoreless, RatingPeriod 1)	0.0004	0.0005	0.0005	0.0020	0.0019	0.0029
Competitive (Scoreless, RatingPeriod 7)	0.0061	0.0014	0.0078	-0.0011	0.0016	0.0004
Competitive (Scoreless, RatingPeriod 30)	0.0069	0.0172	0.0173	0.0417	0.0327	0.0562
Competitive (WinMargin, RatingPeriod 0)	0.0004	0.0011	0.0014	0.0031	0.0010	0.0031
Competitive (WinMargin, RatingPeriod 1)	-0.0005	-0.0005	-0.0005	-0.0020	-0.0019	-0.0029
Competitive (WinMargin, RatingPeriod 7)	0.0026	0.0057	0.0056	0.0144	0.0098	0.0183
Competitive (WinMargin, RatingPeriod 30)	0.0065	0.0164	0.0162	0.0405	0.0316	0.0545

6.6 Summary

In this chapter, we explored an alternative method to estimate the expertise of users on CQA platforms following our exploration on simple and graph-based approaches. This alternative approach is the pairwise comparison approach which forms comparison pairs between users in order to estimate the relative expertise between them. The goal in this chapter is to estimate the expertise of the users in order for us to infer the information quality of their answer contribution – with the user having the higher expertise producing the better answers so that the best answer can be predicted for expert search. Our goals are stated in the research questions of the chapter in Section 6.1 which we attempt to answer.

First, we looked at the current state-of-the-art approaches to user expertise estimation with pairwise comparison [95] in Section 6.2. User-user pairs are formed for comparison if both users are answerer within the same question with one of the answerer being the best answerer, similar to the answerer-best answerer graph model [10]. Each question is regarded as a competition and each pair as a match; thus user ratings from statistical rating systems such as Elo [41] and TrueSkill [57] can be regarded as the estimated user expertise.

Our research proposed the competitive rating approach in Section 6.3, in our attempt to improve the user expertise estimation process. The competitive rating approach includes several differences from the state-of-art competition approach such as – (1) a Round-Robin styled league for user pairs; (2) incorporating the performance of users in each match measured from the win-margin of each match to update user ratings; and (3) only updating user ratings between rating period. We justified the motivation between these variants in order to improve the estimation process. The rating system of choice for the proposed competitive rating approach is the Glicko-2 [46] rating system due to the additional features suitable for our suggested variants and TrueSkill being patented [58].

In our answer retrieval evaluation with the task of best answer retrieval (Section 6.4), we saw the pairwise comparison of all variants to perform really well. All the variants outperformed the earlier explored simple and graph-based approaches with the our competitive pairwise comparison approach with win-margins and rating period significantly ($p = 9.65779E-09$) outperforming the Z-Index baseline by 21.14% for answer retrieval in questions with at least five answers. Besides that, the proposed win-margin extension does comfortably outperform the outcome-focused scoreless state-of-art competition approach; suggesting that the user performance which led to the pairwise match outcome do matter to better estimate the true expertise of the users. The user ratings should not be updated live as done in the competition approach but should instead be updated between rating period; requiring a suitable temporal time frame where there is sufficient user information for accurate rating updates (long enough period) without neglecting the potential user expertise growth (but not too long).

For our expert search study in Section 6.5, we made several interesting observations. Firstly, we found the identified experts from the pairwise comparison approaches are all of low activity despite them being able to provide the best answer for our expert search. As we increase the rating period for our competitive rating approach, the identified experts were found to be more active. These experts are however much less active than the ones identified through simple or graph-based approaches; suggesting the need to find and motivate these users more. The second key observation made is that the effectiveness of the identified experts increases as the rating period increases. User expertise are updated between rating period with a decay for users who did not participate the period itself. As the identified experts are often inactive, their expertise decay could result in a lower effectiveness correlation and prevent us from identifying these experts. Thus, a potential

future work is to explore the impact of decays towards the estimation of user expertise especially for inactive users[11].

[11] We incorporate this study later in Part III of the research.

Part III

Content Aggregation (CA) and Social Curation

Chapter 7

Content Aggregation on the World Wide Web

Content aggregation (CA) platforms are the one-stop-fits-all destination for information – where quality content from the World Wide Web (WWW) is identified and curated for the consumption of users. As such, many of the content aggregation platforms have prospered since the beginning of the WWW, commanding a large amount of internet traffic [134]. Notable content aggregation platforms include:

- Slashdot[1]
- Digg[2]
- Reddit[3]

Over the years, these platforms have evolved with the loosening of responsibility from hired editors to the user community of the platforms. Many CA platforms of today now rely on users to identify and share interesting content from the WWW on their platform for the consumption of other users [137]. This shift towards the social curation (SC) model is discussed further in Section 7.1 through an overview of notable CA platforms.

Today, CA platforms have an undeniable impact on the WWW as presented in Section 7.2. Instead of just content curation, users are encouraged to generate and contribute new content for the consumption of other users. The SC shift poses new challenges in the organisation of large and diverse content on CA platforms; especially with user-generated content (UGC) that are of varying quality. In Section 7.3, we look at content management on the CA platforms of today and discuss the challenges faced as CA platforms are socially curated by their users in Section 7.4.

One of such platform is Reddit, which has gained immense popularity over the years [130] to the point where it is sometimes referred to as the "Front Page of the Internet". Today, Reddit has built its reputation as one of the most popular sites on the WWW with an Alexa[4] rank of 27 as of June 2016. We conducted a study of Reddit in Section 7.5 to better understand its structure (in Section 7.5.1) and content management on the platform.

Our study in this chapter motivates our work in Chapter 8 where we attempt to predict user contributions on Reddit in order to improve content management on the platform. As with Part I and Part II of the research, we once again look to estimate the expertise of CA users and then proceed to infer the information quality of their content.

[1] https://slashdot.org/
[2] http://digg.com/
[3] https://www.reddit.com/
[4] http://www.alexa.com/

7.1 Shift Towards Social Curation

Just as the Web 2.0 of user-generated content evolved from the author-generated content era; CA platforms evolved into their current state of collaborative-based social curation from editor-based curation. This evolution moved from editors having full control over the content of CA platforms (SlashDot) to editors determining the front-page (Digg) and finally to its current form with the user community having full autonomy of content on the platform (Reddit).

7.1.1 SlashDot

One of the earliest CA platform is Slashdot which began in the year 1997. On Slashdot, editors would select, edit and aggregate content such as stories or news for the consumptions of the users on the platform [87]. The users could just consume or comment on the threads created by the moderators, moderated through peer-assessments [117].

7.1.2 Digg

Following the success of Slashdot, Digg rise in popularity in year 2004. Unlike Slashdot which relies on editors to aggregate content from the WWW to the platform, Digg allows users to submit content to the platform [63]. The submitted content can then be voted on the platform – actions coined as 'Digging' with a up-vote known as a 'dig' and a down-vote known as a 'bury'. These votes describe how interesting is a piece of content [89] and when a user votes for a piece of content, the content is shared to the voter's followers as recommendations [61] sorted by in reverse chronological order.

Content on the highly visible front-page of Digg was still to a certain extend selected and managed by the editors of Digg, with users competing for attention towards their contributed content as voted by other users. Here, it can be observed the loosening of responsibility in content curation from editors to the users.

7.1.3 Reddit

Recently, Reddit which started in year 2005 has gained popularity among users of the web with approximately 450 million page-views in December 2014 alone [134]. The growth in Reddit's adoptions through a large migration of users from Digg (in year 2010 following Digg v4) can be credited to the empowerment of its users on its platform with the capability to create community (subreddits), aggregate and moderate content.

Users on Reddit are equal, unlike the segregated roles on Digg such as editors, moderators, power users[5], anonymous users and many more [91]. Each of these user roles comes with their own privileges including number of votes possible and the magnitude of the votes. Without such roles, it is up to the user community of Reddit to aggregate, curate and moderate the content including determining which content are pushed to the front-page of Reddit without the need for editors.

7.2 Information Potential and Impact on the WWW

As mentioned earlier, CA platforms of today have a sizeable impact on the WWW. They are important to the Web 2.0 due to their large user base which commands a significant amount of traffic to both CA platforms [134] and other platforms [29, 42]. CA platforms affect the WWW by:

[5] who are viewed by users to be having too much power in Digg v4.

- Identifying interesting and popular content on the web [134] that could meet the information need to the users on the platforms. The content are curated with increased visibility [62] for popular and novel items [92, 149].

- Directing a large volume of user traffic to content on another platforms or webpages. One of the terms coined for this is the Slashdot Effect – a Flash Crowd occurrence [29] where websites linked from Slashdot's stories receive sudden swell in traffic. The servers of these websites could collapse under the heavy traffic if not prepared to handle the surge in traffic [42].

- Playing a large role in the dissemination of information [44, 62], especially through promotion by influential users [12] to great effect [49]. Their influence spreads beyond their own platforms and linked sites; shaping how stories can be viewed, enhancing content and creating new viral content such as memes [145]. Often, discussions on Reddit are picked up by mainstream media and reported to the masses.

- From a sociological perspective, CA platforms give us a great deal of understanding with regard to the interest and importance that the users place on different topics – this characterises the internet zeitgeist.

Due to the large influence which CA platforms have on the WWW and its society today; it is vital for their contents to be managed properly so that harmful false information does not spread.

7.3 Content Management on Content Aggregation Platform

A very large amount of content exists on CA platform. For example, over 60 million threads were submitted to Reddit between 2008 and 2012 [130] and we found that popular subreddits such as the *r/gaming* have an average of $1,257$ threads submitted daily[6]. The volume of data creates a need for the management and ordering of content on these platforms for user consumption and peer-moderation.

The ordering of content matters. Users were found to have a positional bias for links [35], where users were observed to display consumption patterns from the top of a web page [33] or a list [73] before proceeding down to the bottom [92] – top items having greater visibility for user consumption. Thus, the content on the users' stream should be sorted to better direct the users' attention towards content of good quality [92].

Thus, it is preferred to have these contents organised according to their quality to better direct the users' attention towards content of good quality [92]. We do note that a thorough study of how best to rank content on a user's stream is currently lacking for Reddit [134].

7.3.1 Peer Recommendation

Prior works on CA platforms mainly looked into the recommendation of content (particularly on Digg) and focuses on analysing the visibility and attraction of shared content. All contents competes for the general users' attention to be on the front-page of the platform.

In Digg, recommendations are fed directly to a user's followers' content stream whenever the user votes for review of a piece of content [61]. The items are sorted according to temporal order. The motivation for such a feed is from the discovery that users do follow other users for content [90]; similar to many other social media platforms such as Twitter [11, 66] which make use of follower-based recommendations [74]. Thus, researchers have

[6]Figure from our dataset in Section 7.5.3

looked to improve on the recommendation performance of Digg with user models [64] for identifying similar users. Users would then be pushed content that is highly rated by similar users as personalised recommendation [91], via a peer recommendation process [92] to better meet these information needs.

7.3.2 Content Popularity

A common approach in the ordering of content is to sort the content according to popularity. Studies have found that the popularity of items on Digg is largely influenced by quality [91], with content of high popularity found to be of higher quality than less popular content [134]. It should however be noted that the novelty of content does degrade over time [149] and that the popularity of Reddit content does not equate to quality. For example, popular image threads are often reposting images from earlier threads with identical link that had previously been ignored by the community before achieving popularity [45].

Content popularity can be determined by looking at the number of votes garnered by the content on the CA platform [89]. It is however difficult to predict the future popularity of an item [138], particularly new content due to the cold-start problem, even by studying early user reception [91, 116].

The only available data initially comes from the content itself such as the content title with certain title leading to greater exposure with a positive effect on popularity [85]. These features are however weak features for prediction [30] and can be computationally expensive.

7.4 Challenges for CA Platform with Social Curation

As CA platforms moved towards the direction of social curation without expert editors as discussed in Section 7.1, content management on CA platform proved to be challenging with many problems which we attempt to address later in Chapter 8. These challenges include:

7.4.1 Managing High Amounts of Unstructured Content with Varying Quality

Users are of different expertise levels and produce or submit content based on their capabilities. This creates a diverse spread of content – some of which may not be suitable for consumption. If not properly managed, such misinformation could spiral out of control [12, 44, 145]. Thus, there is a need to infer the information quality of content for improved content management.

7.4.2 Vulnerability to Malicious Users

As an open platform, new users are welcome to become part of the communities – to contribute and help moderate the platform through peer assessment [92, 117]. The existence of malicious users however can undo such efforts. For example, these users can game the system by creating proxy accounts for self-promotion or coordinate attacks for propaganda purposes [106]. Through early identification of such users, social curation platforms can defend against such attacks.

7.4.3 Obtaining Sufficient User Interactions

As discussed earlier, a challenge faced by UGC platforms is in maintaining a healthy amount of user interactions on the platform [15]. Often, traditional social platforms rely

on the user's personal motivation for contributing content [45] with the majority of users only consuming content without submitting any content on Reddit [110].

The lack of user activity is always a concern for social platforms [15] due to the decrease in data and information to work with [87]. For example, a CA platform would not be able to obtain a good coverage of the web if there are insufficient users submitting content; or if there is a lack of user votes making it difficult to judge a content's popularity and quality.

In our study of Reddit in Section 7.5, we see a high amount of user activity in creating threads and commenting on them. While large subreddits are immune to this, smaller subreddits and the community do suffer from the lack of user activity as the user interactions are centred around the larger subreddits [130]. We also note from the same study however that there are limited user votes for the moderation of content. This deficiency in user review greatly affects the content moderation itself [89] on Reddit. Through the estimation of user expertise of content authors, it is possible to predict the information quality of the content despite the lack of user votes.

7.4.4 Vote Bias

User bias does exist on social platforms, particularly in the judging of content through their votes. Users tend to vote for the content which they have viewed early [146] and might not continue to browse the remaining content; resulting in the content not being judged well when it is sorted by temporal order. Besides that, users can be biased by the "herding" effect – where the community's previous view on the content [118] (such as the current votes gained by the content or an early swell in user votes) affects the behaviour of other users in future voting [116].

7.5 Reddit

Reddit is a hybrid content-aggregation and discussion board platform – aggregating content for the users' consumption while providing a platform for user discussion. Over the years, it has gained popularity with an exponential growth to the point that it is sometimes referred to as the Front Page of the Internet. As of 2nd February 2014, it was ranked as the 69th most popular website in the world according to Alexa with 112 million unique visits per month and over 60 million threads submitted between the year 2008 and the year 2012 [130]. In December 2014, Reddit receives approximately 450 million page views [134]. Figure 7.1 showed the front page of Reddit.

As a CA platform, Reddit is a suitable platform for our research into the estimation of user expertise and information quality of UGC through the use of content-agnostic approaches. Unlike our earlier work in Part I and Part II, the complex structure of Reddit will further challenge the robustness of our proposed approaches. To the best of our knowledge, there are not many studies conducted on Reddit [134] with earlier works focused on earlier pure CA platforms without user discussions such as Slashdot and Digg.

7.5.1 Structure and Representation

Users are free to create threads on any user-defined topics or community, known as subreddits. Threads are created with the submission of an external link or user written text [130]. Once the thread is created, users of Reddit including the thread starter can then at any time, participate in the thread discussions through comments, critics, questions and many more.

As CA platforms evolve over time, the structure of each platform does differ. For the discussion of this chapter and the following Chapter 8, Reddit will be used as the example

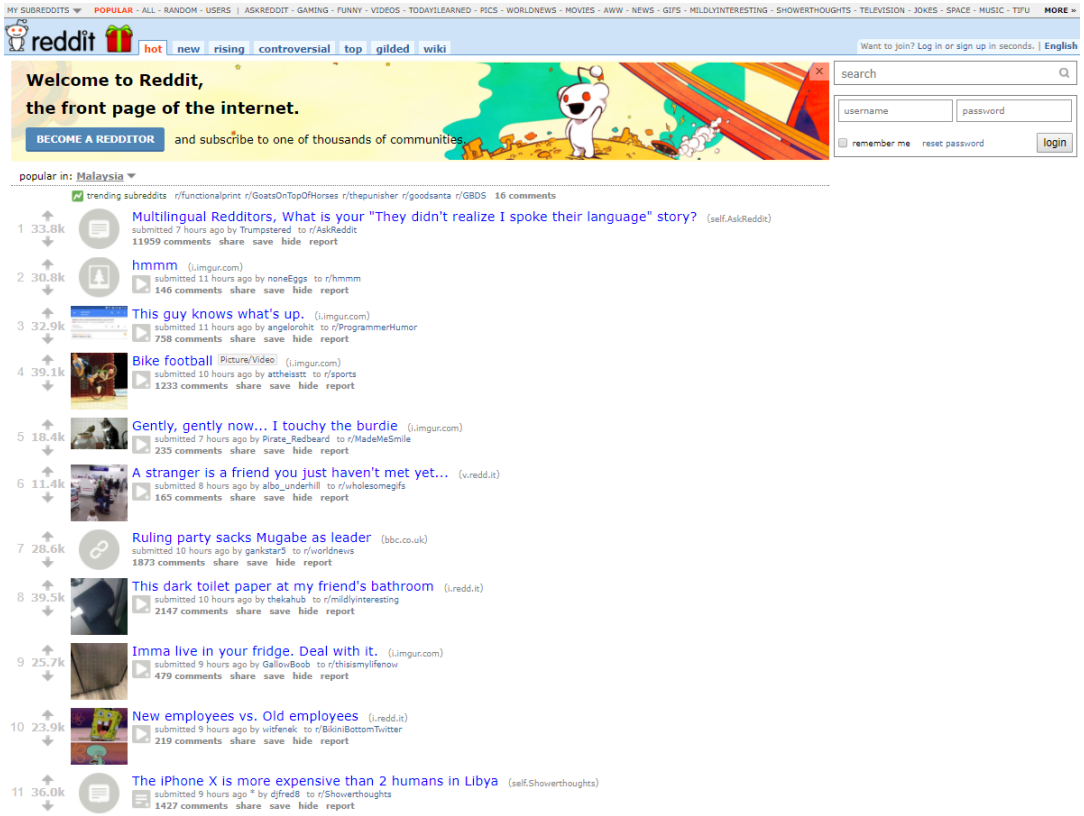


Figure 7.1: A screenshot of the front page of Reddit. Web page visited on 20th November 2017.

structure of a CA platform. In the following sections, we discuss the key components of Reddit.

User Community (Subreddit)

As discussed in Section 7.1, Reddit provides autonomy to the users. This enables the natural building of communities on Reddit known as subreddits – each being independent, dedicated to a specific topic and moderated by volunteers [130] such as the *r/politic* subreddit in Figure 7.2. It is within these subreddits that users would create threads and submit comments to.

The organisation of Reddit is based on the subreddits as shown below in Figure 7.3. Users navigate through the subreddits via pages. Pages list threads sorted according to the aspect chosen by the user such as the temporal ordering, number of votes, highest activity and so on. One important note on the organisation of Reddit is that each Reddit thread can only be in one subreddit but could appear in one or more pages of that subreddit as the pages merely sort the subreddit threads according to the users' requirements.

The distinction between subreddits is not clear however, especially with the ever-increasing number of subreddits. Still, studies have shown that these communities are diverse enough with submissions that aggregate content from all over the web. Besides that, it should also be noted that some subreddits thrive on certain submission types; for example, *r/pics* encourages link submission of images whereas *r/ama*[7] encourages text-based self-submissions.

Nonetheless, there is a clear distinction in user activity between the top subreddits and the other subreddits [130]. For example, the top 20 subreddits make up more than 70%

[7] ask me anything

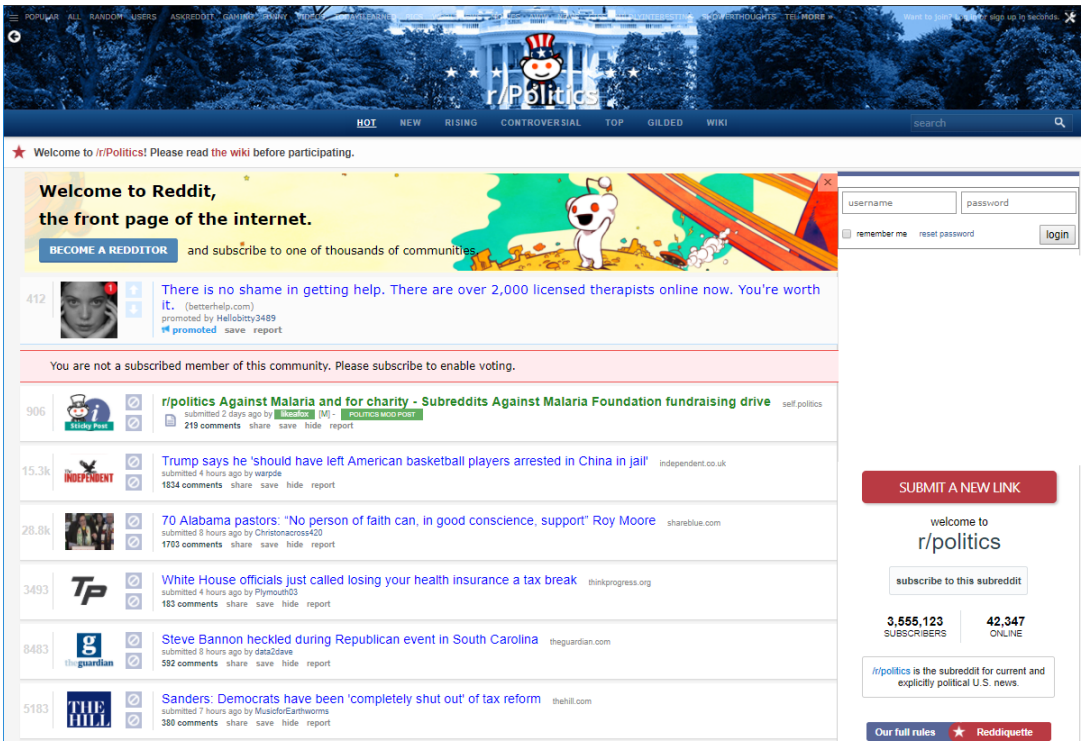


Figure 7.2: A screenshot of the Politic subreddit (*r/politic*) of Reddit. Web page visited on 20th November 2017.

of the user votes on Reddit. A similar observation can be made for the thread comments. It should be noted that active users are in fact active on several subreddits and not only on the top subreddits; benefiting more on social platforms.

Thread

User-created threads are the main source of content on Reddit and are organised according to the related subreddits. A successful thread submission on Reddit must be of high quality, with a good title which gives exposure and has a positive effect on popularity [85].

Unlike many other UGC platforms, Reddit allows users to directly reply to other user comments in threads, creating a tree-like structure of various levels, depths and heights. The general structure of a Reddit thread with multiple user interactions is as illustrated in Figure 7.4.

These threads are created and submitted by users through two main processes:

- **Link submission.**
  Thread starters can create threads by submitting a link to external sources such as news articles or image hosting services. Such submissions make up the main population of threads on Reddit; contributing towards Reddit's role as an aggregator of news, stories and multimedia. The link submissions are diverse enough to aggregate content from all over thee WWW [130]; notably to Imgur[8], an online image repository that complements the text-based nature of Reddit as an image extension. An example of a thread submission is shown in Figure 7.5

- **Self-submission.**
  Reddit users can generate textual content for sharing or discussion with other Reddit users as displayed in Figure 7.6. Such content is often known as self-submissions

[8]https://imgur.com/

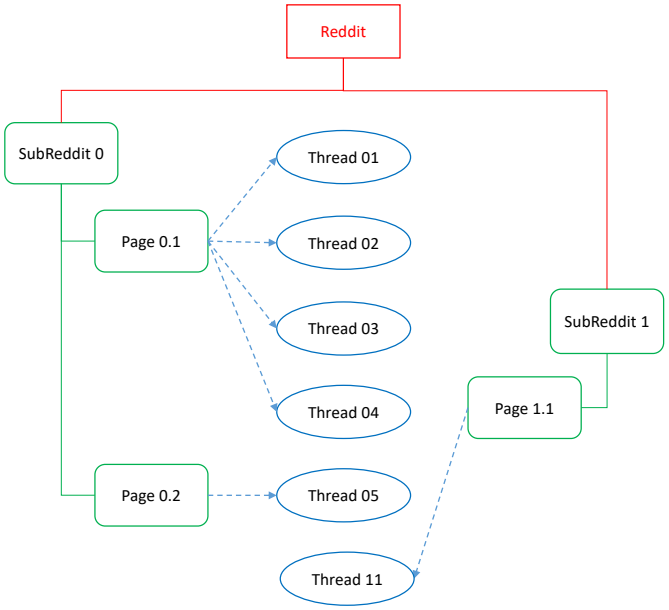


Figure 7.3: Organisation of Reddit – subReddit, pages and threads.

which is then stored directly on Reddit itself. Over the years, Reddit has undergone a transition from a heavy aggregation platform with link submissions to a hybrid discussion board with the rise of a self-referential community with self-submissions [130]. Self-submission content includes personal stories from users, user questions, user opinions, user-written articles and even open-ended write-ups to inspire discussions amongst users of the subreddits. Such content often results in a higher comment count than link submissions.

Comments

As illustrated in Figure 7.4, users can interact within a Reddit thread through comments. Such interactions enrich the thread by providing additional meaningful information [134] such as clarifications, further explanations, questions or even critic to the thread content itself (see Figure 7.7). Any user including the thread starter himself/herself can contribute comments within the thread. There are however a high number user threads on Reddit with a low number of comments [45].

User comments for Reddit threads can be categorised into two distinct types as illustrated in Figure 7.8:

- **Direct Comments.**
  Direct comments are the main drivers of discussion in a Reddit thread. These comments directly respond to the thread content itself regardless of whether it is a link submission or self-submission. Other users could also directly address the thread starter through these comments.

- **Indirect Comments.**
  The indirect comments are comments which respond to other comments, contributing towards the discussions driven by the direct comments. These comments can either be of agreement with the direct comments, supplementing it with additional information; or argue against those comments with valid contradictions.

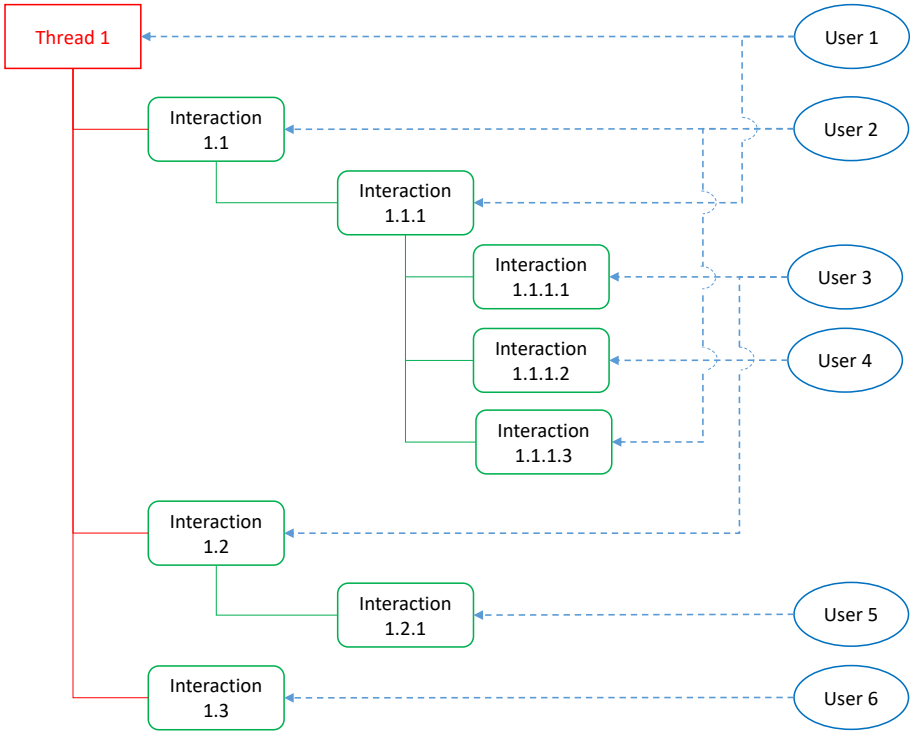


Figure 7.4: Structure of a Reddit thread.

User

Reddit users are integral to Reddit as a social curation platform – users share interesting or popular external content on the WWW to Reddit as link submissions. The sheer number of users which Reddit enjoys now ensure a good coverage of the WWW [45]. Alternatively, users can generate new original content as text submissions. As a discussion board as well, users are provided with the tools to partake in discussions on Reddit by commenting in the threads themselves. Such comments add value to the shared content and we can view these commenters as collaborators on Reddit. In fact, there are significantly more users who comment within threads than users who create threads [45]. All of these contents are consumed by the wide audience of Reddit [110], otherwise known as *lurkers*. Figure 7.9 showed the profile of a Reddit user.

Activity on social platforms can be further promoted by motivating the users for various reasons such as reputation gain [45]. In Reddit, users are encouraged to scavenge the WWW for interesting resources to be submitted as link submissions instead of producing their own submission by only rewarding *karma* points for the former.

Modern CA platforms like Reddit empower users for content curation – where users as a collective[9] moderate content on the platform. The users can vote for content on the platform as an indication of quality [89]. For example, studies have shown that users do vote for shared content where over half of Reddit threads are voted mode than once [45].

Votes

Reddit facilitates the peer moderation of content through user votes which are often regarded as the community's long-term judgement on the quality of content on many UGC platforms [9, 89]. These votes are given by Reddit users as their judgement of the content

[9]See Wisdom of the Crowd in Definition 3.

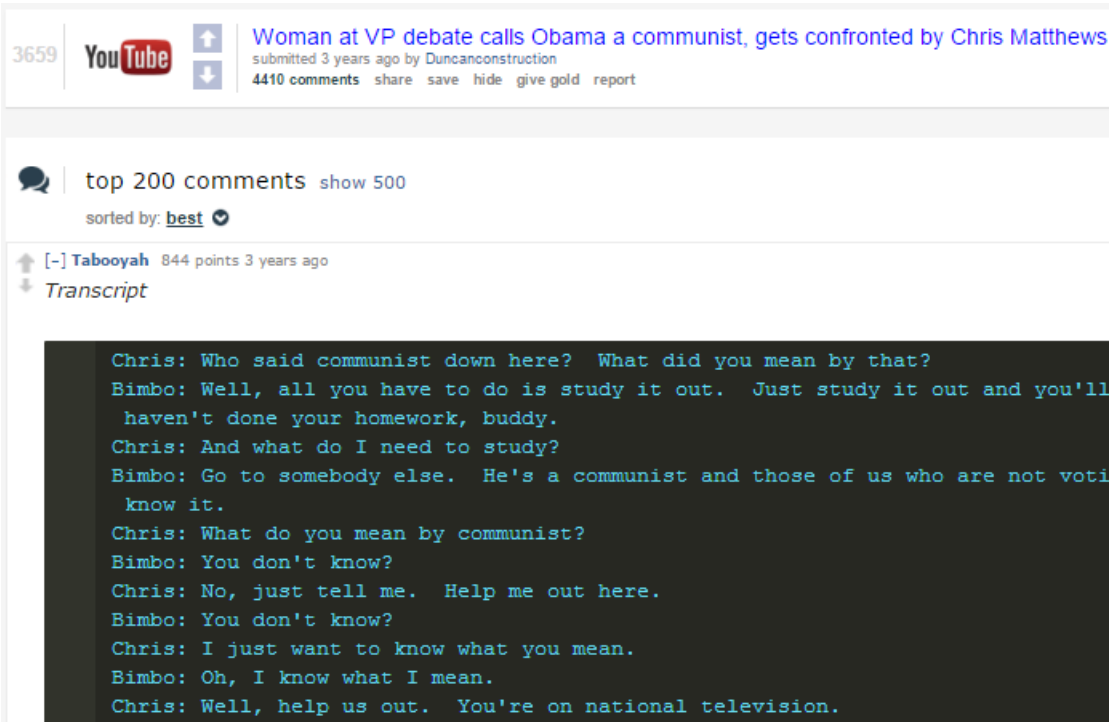


Figure 7.5: A screenshot of a link submission thread with an external link from Youtube. Web page visited on 1st December 2015.

[130] – Up-votes (positive votes) to indicate content which they deem to be good, and down-votes (negative votes) to be poor.

Thus, user votes can be used to indicate the quality of a content as perceived by the community (illustrated in Figure 7.10). Just like Digg, the user votes directly affect the ranking of 'hot' content and how the content is presented to the consumer. Unlike Digg however, there is no distinction between users, thus, all users have the freedom to vote with the same voting mechanism without any limitations.

As a mean of preventing vote abuse by malicious users, Reddit is no longer displays the number of up-votes and down-votes gained by a piece of content (since June 2014) [75]. Instead, content votes are calculated internally in Reddit according to the number of up-votes and down-votes received through a function they called Vote Fuzzing (VF). The VF mechanism artificially inflates the number of up and down-votes for a resource, while maintaining the true difference between them which is then attached to the content itself as *points* (i.e. a random value is added to the up-vote and down-vote counts). This can be viewed by all users as shown in Figure 7.11 and Figure 7.12. In this paper we extract user vote difference, Vote(I), as a content-agnostic indicator to estimate user expertise.

It should however be noted that there exists bias for user votes. The works of Weninger et al. [146] discover that the highest voted comments on Reddit are comments submitted close to the creation of the thread as the users mainly only partake in early discussions.

Negative Vote Difference

In general, content a with negative vote difference indicates that the content does not contribute towards the discussion of the thread or is disagreed by the community as a whole (in the thread). Often, negative votes are abused particularly for cyberbullying and thus, are removed from many platforms [52]. Previous studies have concluded however that the existence of down-votes does not negatively affect Reddit [106].

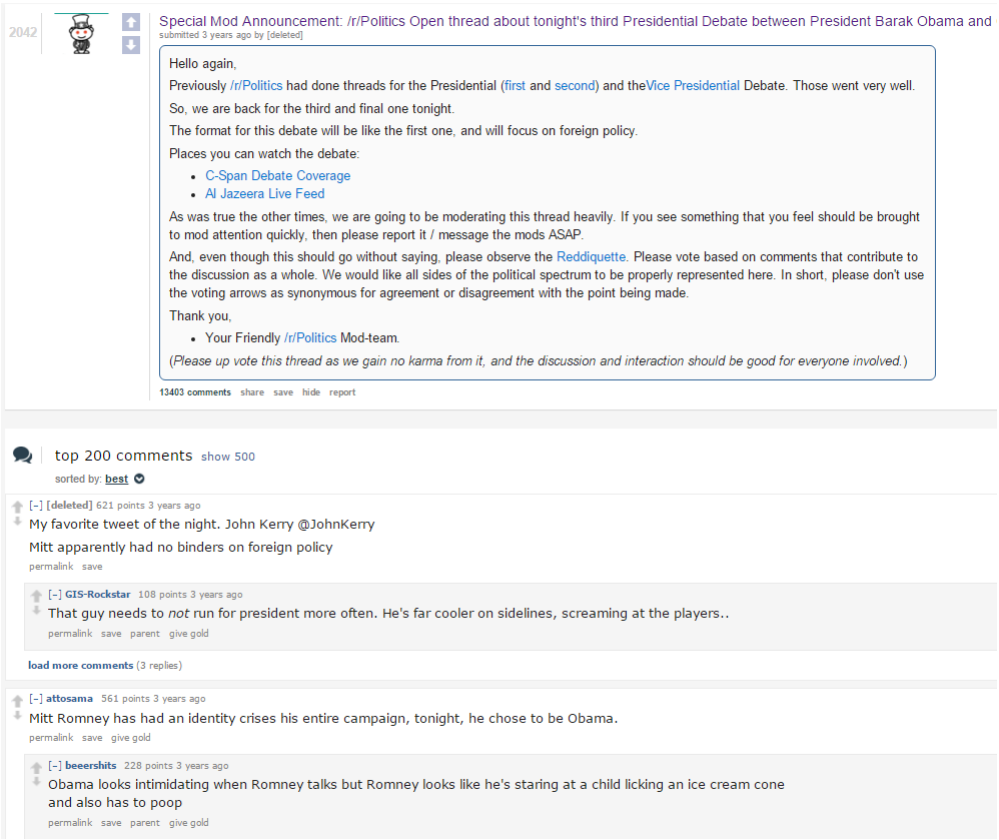


Figure 7.6: A screenshot of a self-submission thread with textual content written by the thread starter. Web page visited on 1st December 2015.

Many of the current traditional and state-of-the-art approaches to user expertise estimation (including Part II of this thesis) were not designed to handle negative votes directly; requiring some form of modification such as:

- **Ignoring or filtering out negative interactions.**
  Content with negative vote differences can be inferred to be of no contribution to the discussion thread and thus ignored by removing them or setting the vote difference to a value of zero. Ignoring these interactions improves the scalability of any user expertise estimation with less data for processing though it is possible to suffer from the lack of information for the estimation of user expertise.

- **Transforming or normalising the negative vote difference into positive votes.**
  The interactions with negative vote difference in a thread can be transform into a suitable positive value through commonly used data handling techniques. Alternatively, they can be normalised relative to the other interactions in the thread.

It should be noted however that the negative vote difference on Reddit is important. Thus, they should be acknowledged and handled differently from the commonly used approaches discussed above. Negatively voted interactions can imply disagreement with the community of that subreddit which do contribute towards the discussion of the thread. For example, the disagreement between left-wing and right-wing ideology or between liberal

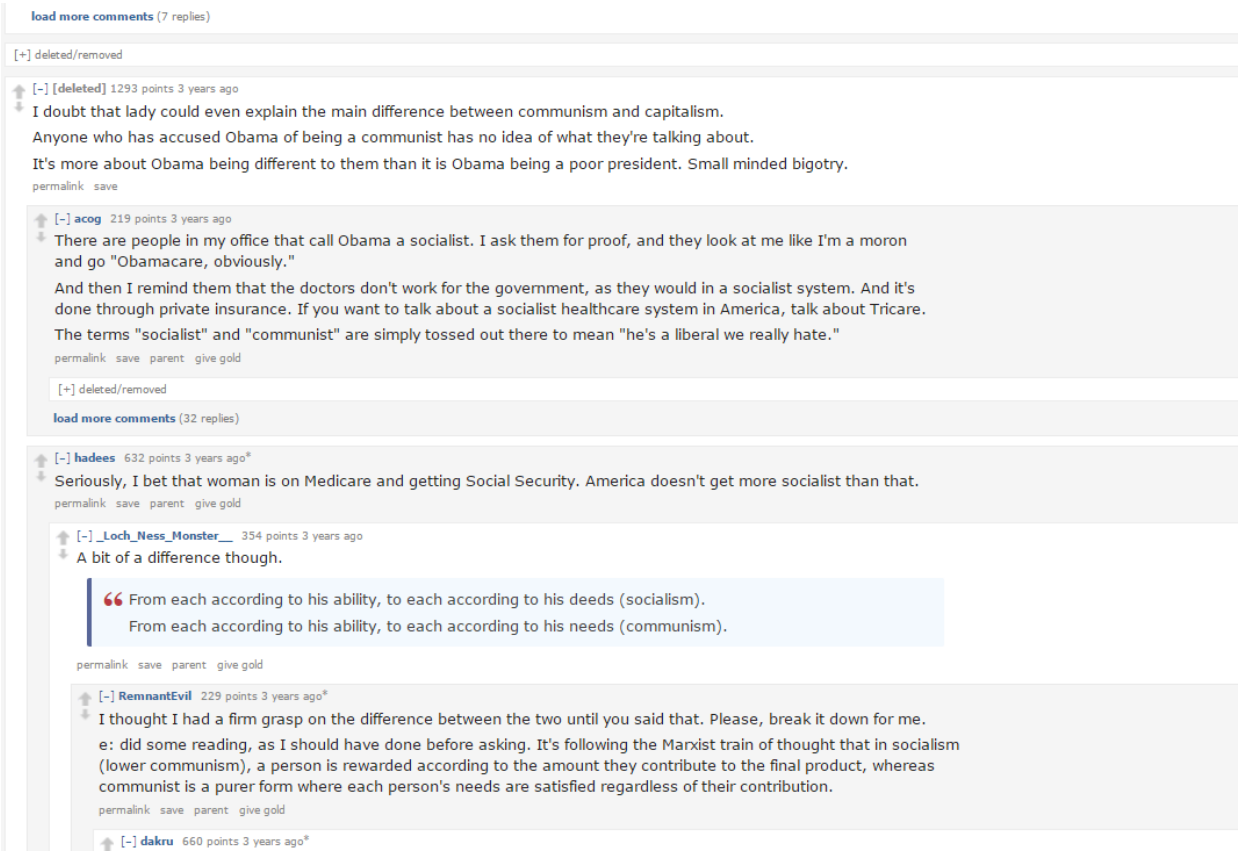


Figure 7.7: A screenshot of comments in a Reddit thread. Web page visited on 1st December 2015.

and conservative believers may result in one of the sides being severely down-voted. Similarly, these interactions provide the alternate voice of reasoning against the *hive-mind*[10] or *circle-jerk*[11] communities on Reddit.

One of the indicator for such occurrence is when an interaction and the interaction which it responds to are voted on the opposite polarity (negative vote difference in one and positive vote difference in another). Once resolved, such disagreement could reinforce the contribution of one or both interactions. The capability to be able to identify and account for such scenarios might be beneficial for some of the explored user expertise estimation algorithms for Reddit users. The challenge here lies in how the negative interactions can be modelled with the tree-like structure of Reddit thread as visualised in Figure 7.4.

This research acknowledges the opportunity in this area to improve the process of user expertise estimation. Thus, in the following chapter, we attempt to explore, account for and manage interactions with negative vote difference.

7.5.2 Reception, Popularity and Quality

The popularity of content can be implied through user actions such as user votes or comments that describe the users perception of the content itself. The implied popularity then can be used to estimate the quality of content as there exist strong correlation between the popularity and quality of content – it is found that the most popular content on Reddit are of higher quality than the less popular ones [134].

[10] Accepted mainstream opinion accepted and shared amongst users.

[11] Members of a community who share the same opinion and reinforces each other's opinions while denouncing others.

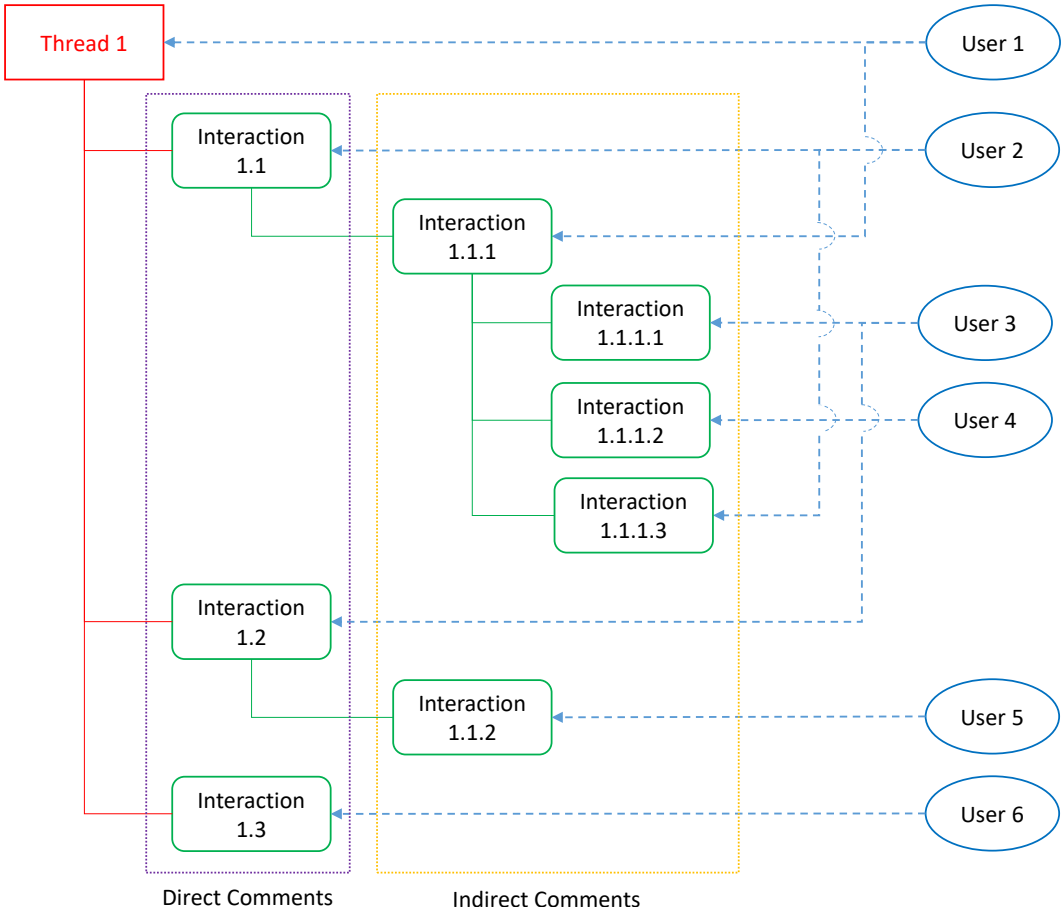


Figure 7.8: Direct and indirect interactions of Reddit.

We observed however that this is a directional relationship since the popularity of Reddit content does not always equate to quality where studies have discovered that popular image threads are actually repost of earlier threads [45]. In fact, identical links could be ignored by the community for some time before later achieving popularity [45].

Comment Ranking

The high popularity of Reddit today results in large number of threads for each subreddit with a large amount of user interactions in comments. Thus, there is a need to curate this user-generated content on Reddit. The sorting of user comments is performed mainly on the direct comments with indirect comments attached in nested form. The users can choose how comments on the Reddit threads are sorted as seen in Figure 7.13.

At the time of writing, the interactions in a thread can be sorted according to the following:

- **Best.**
  The *best* sorting is based on the statistical sampling of user votes (95% confidence) for the interactions. In this sorting approach, the ratio of up-votes to down-votes matters more than the actual vote difference if there are sufficient votes to increase the confidence level. The current implementation here is based on the lower bound of Wilson score confidence interval [148] for a Bernoulli parameter.

- **Top.**
  The *top* sorting is a simple sorting approach which ranks the highest vote difference higher than the lower vote difference.

Figure 7.9: A screenshot of a user profile on Reddit, American actor Wil Wheaton. Web page visited on 20th November 2017.

- **New.**
  A temporal sorting approach ranks newer content higher than older content. This ranking approach is useful for the users to easily identify, keep-up with and consume new content in a thread.

- **Old.**
  This is the opposite of the new sorting above where older content is ranked higher than newer content for users to better follow the development of a thread discussion.

- **Controversial.**
  The controversial sorting ranks content according to values of both up-votes and down-votes; implemented according to the following equation:

$$\textbf{Score}(i) = \frac{\textbf{upvotes}(i) - \textbf{downvotes}(i)}{\textbf{max}(\textbf{abs}(\textbf{upvotes}(I) - \textbf{downvotes}(I)), 1)} \tag{7.1}$$

- **Question and Answer (Q&A).**
  This is a new sorting approach introduced recently. This sorting is based on the best sorting discussed earlier but adjusted to suit a question-answering environment,

Figure 7.10: A screenshot of the front page of Reddit, with us voting a thread up and another thread down. Web page visited on 20th November 2017.

focusing on the thread starter – the choice of thread format for the popular r/ama subreddit.

- **Hot.**
  This approach sorts content according to how recent an interaction was made and discussed. This sorting approach has however been removed from the current live version of Reddit.

To the best of our knowledge, the sorting approaches discussed in Section 7.5.2 – that are currently or have been used on Reddit do not consider the users who generate the content. Many of these approaches in fact just make use of the current voting data for interactions which can be abused by malicious users especially since the algorithm is available for view as Reddit is open-source with only the Vote Fuzzing mechanism hidden from the public. With the additional information from the estimated user expertise and reliability explored in Chapter 8, the impact from these malicious users can be reduced or eliminated.

Furthermore, it can be noted that these sorting algorithms only look at each comment in isolation without considering relations with the other comments or the type of comment. We look to infer the information quality of user comments, relative to each other together within the same thread.

7.5.3 A Reddit Dataset

To better understand the nature of content and users on Reddit, this research performed a study on Reddit based on the publicly available dataset hosted on a Google BigData repository[12] from the *r/datasets* subreddit[13]. From the collection, we made use of threads and comments from various subreddits dated between 2015 and July 2016. The dataset was downloaded on the 27th September 2016 which left a temporal gap for the threads and comments to stabilise before conducting our study.

From the repository, this research investigated threads and comments from the highly subscribed subreddits[14] with healthy number of threads. As we shall discuss later in this

[12] Compiled by Reddit user *u/Stuck_In_the_Matrix*.
[13] https://www.reddit.com/r/datasets/comments/3bxlg7.
[14] As of 2nd September 2016 according to http://redditmetrics.com/top.

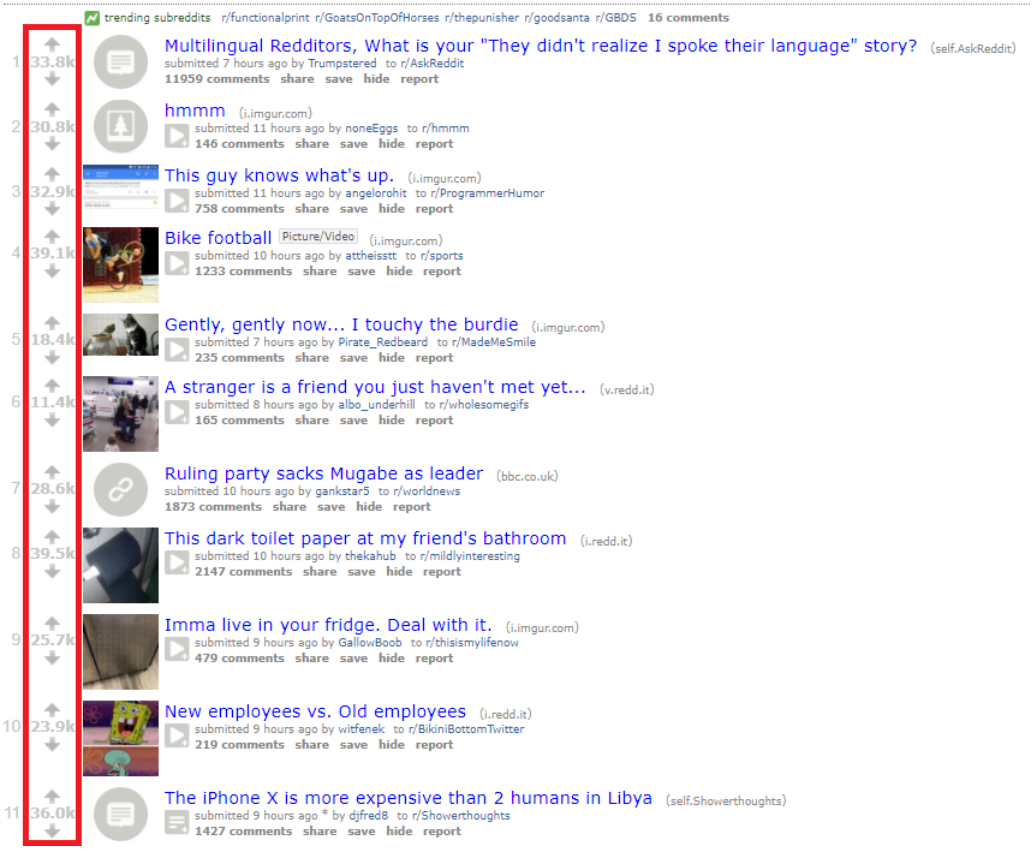


Figure 7.11: A screenshot of the front page of Reddit, highlighting the vote difference for each thread with a red box. Web page visited on 20th November 2017.

section, user behaviour on these subreddits differs from one to another. Thus, the study of this dataset would provide us with valuable insight on the varied users and content:

- r/science[15] ranked 5th.
- r/worldnews[16] ranked 6th.
- r/gaming[17] ranked 10th.
- r/elifive[18] (explain like I'm five) ranked 17th.
- r/politics[19] ranked 55th.

7.5.4 Study: Sufficient Threads

The quantity of threads created daily differs between subreddits as shown in Table 7.1 and Figure 7.14. We note that the *r/science* subreddit recorded the lowest number of threads created daily and *r/gaming* the highest. Such difference can be credited to the type of content and the moderating community of the subreddit. For example, users are able to share their gaming experience on *r/gaming* whereas only interesting articles or findings are posted to *r/science*.

The number of threads created on these subreddits are relatively stable except for *r/politics* which fluctuates with regards to the political scene. Upon further inspection, we observed that this subreddit, despite its name, is a subreddit only for the United States

[15]https://www.reddit.com/
[16]https://www.reddit.com/r/worldnews
[17]https://www.reddit.com/r/gaming
[18]https://www.reddit.com/r/elifive
[19]https://www.reddit.com/r/politics

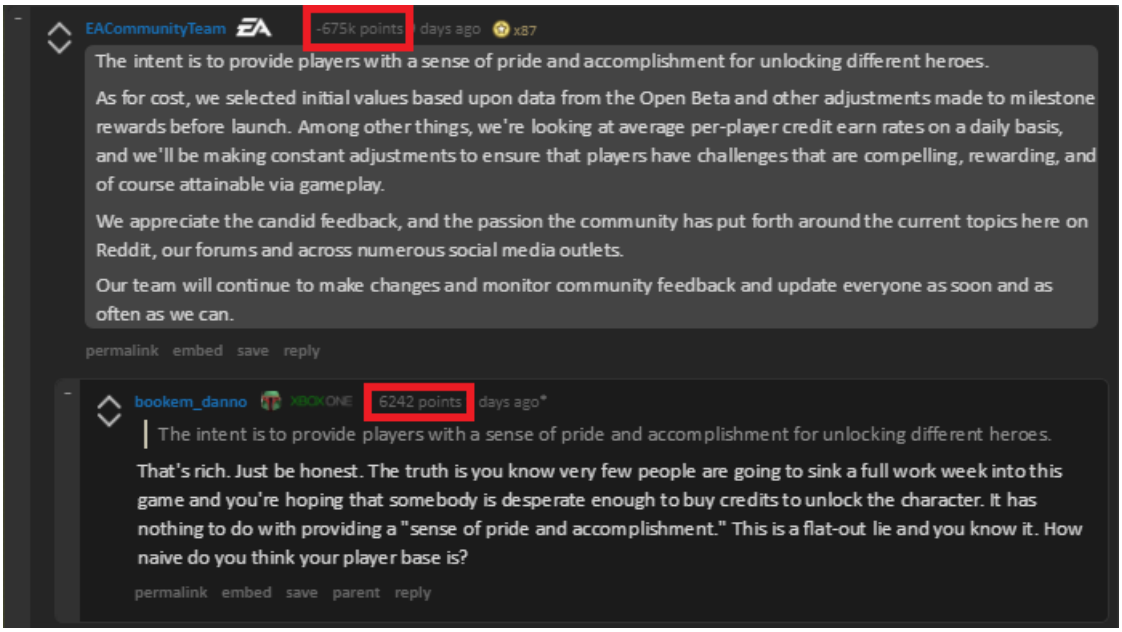


Figure 7.12: A screenshot of comments in a Reddit thread, highlighting the positive and negative vote difference. Web page visited on 22nd November 2017.

Figure 7.13: Sorting options for comments on Reddit threads.

(US) political news. Thus, this contrasting behaviour would be an interesting exploration point against that of the global *r/worldnews* subreddit in future studies. The number of threads created is very stable for the *r/science* subreddit with minimal fluctuation when compared to the other subreddits.

From the plot, a consistent dip in the number of threads created towards the end of December 2015 and earlier January 2016 can be observed. Presumably, the holiday season has an effect on user activity – users are less active during such period with the exception of the *r/gaming* subreddit with holiday releases.

Such observations encourage the research to explore the possible performance difference these subreddits; each with varying amount of user activity and available information.

7.5.5 Study: Sufficient Comments

We present the distribution of user comments for Reddit threads in Table 7.2. This user comment distribution is extremely skewed to the right. The number of comments for each thread differs between subreddits and the difference can be attributed to the subreddit topic itself – *r/politics* with a high amount of argument between users.

An interesting observation to be made here is that the threads in *r/explainlikeimfive* and *r/politics* would often receive at least one user comment; and on the other hand,

Table 7.1: Thread Density of Subreddits by Day.

	Number of Threads by Day						
Subreddit	Minimum	1st Q	Median	Mean	3rd Q	Maximum	Total
elifive	446.0	593.0	648.0	663.2	717.8	1484.0	161827
gaming	973.0	1161.0	1227.0	1257.0	1328.0	2355.0	306735
politics	190.0	673.8	904.0	894.3	1061.8	1728.0	218210
science	65.0	118.8	174.0	164.8	199.0	263.0	40215
worldnews	503.0	774.8	922.0	892.8	1006.5	1567.0	217834

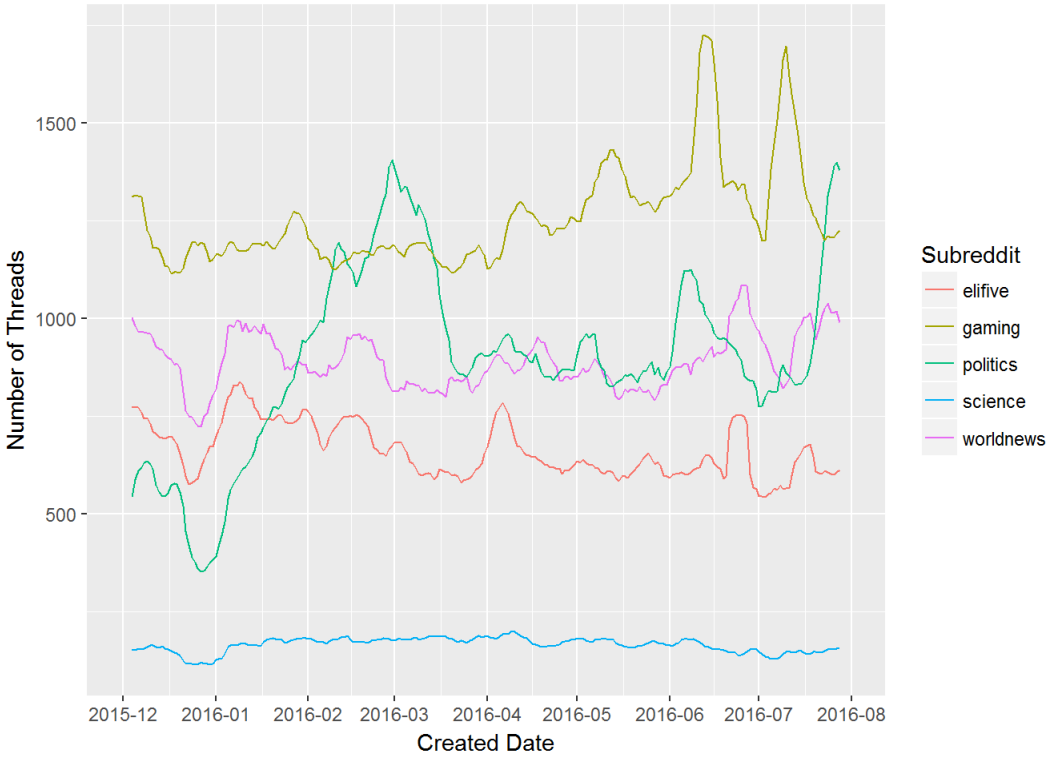


Figure 7.14: Moving average (7 Days) of Reddit threads created.

at least half of the *r/gaming* threads have no user response despite having the highest number of threads submitted daily.

Study: Comment Length

Exploring further into the comment word count as presented in Table 7.3, it can be noted that the comments for *r/gaming* threads are usually short. On the other hand, the question-answering nature of threads in *r/explainlikeimfive* encourages more complete and elaborate responses which resulted in a higher word count for user comments.

Study: Comment Type

Direct comments respond directly to the submitted content or the thread starter; commenting on, discussing or arguing with the content shared by the thread starter. If the thread itself is a question, the comments may attempt to answer it. In Table 7.4, we observe that the comments in *r/politics* and *r/worldnews* are often responses to other comments in a discussion thread unlike the other subreddits here.

Table 7.2: Comment Distribution of Reddit Threads.

	Number of Comments on Reddit threads					
Subreddit	Minimum	1st Q	Median	Mean	3rd Q	Maximum
elifive	0.0	1.0	2.0	8.5	5.0	6105.0
gaming	0.0	0.0	0.0	9.9	2.0	9502.0
politics	0.0	1.8	3.0	43.8	15.0	47831.0
science	0.0	0.0	1.0	15.8	2.0	6370.0
worldnews	0.0	0.0	1.0	22.5	2.0	32901.0

Table 7.3: Word Count Distribution of Comments.

	Number of Words for Thread Comments					
Subreddit	Minimum	1st Q	Median	Mean	3rd Q	Maximum
elifive	1.0	9.0	29.0	51.7	64.0	1837.0
gaming	1.0	5.0	12.0	22.8	27.0	2831.0
politics	1.0	9.0	20.0	36.8	43.0	1933.0
science	1.0	10.0	10.0	29.5	38.0	1609.0
worldnews	1.0	7.0	18.0	36.1	41.0	1980.0

There is a strong correlation in the number of direct comments with the number of comments in every Reddit thread. An interesting observation here is that despite the threads in *r/worldnews* having a much lower number of direct comments when compared to the other subreddits, the subreddit does record the highest correlation here telling us that it is very discussion-focused with users replying to each other a lot more.

7.5.6 Study: Vote Distribution

We discussed the importance of user votes for the moderation of Reddit content in Section 7.5.1. On Reddit, users contribute towards the peer-moderation of content within each subreddit by voting on threads and comments.

Study: Thread Votes

Starting with the threads themselves in Table 7.5, it can be observed that there are no Reddit threads with negative vote differences. A possible explanation for this phenomenon is that users tend to just ignore threads that are of lower quality instead of judging the submitted content themselves. Thus, the removal of negative votes [52] would not affect the organisation of content.

The vote difference distribution for Reddit threads are heavily skewed to the right particularly for the *r/explainlikeimfive* and *r/gaming*. Contrarily, the *r/politics* has almost half of its threads without any vote difference (value of 0) but the voted threads have a higher average vote difference when compared to the other subreddits.

Study: Comment Votes

Unlike the threads, we observed Reddit comments with negative vote differences as shown in Table 7.6. These negative votes are however not dominant as they account for less than a quarter of the comments. As users do respond to each other through comments, the users are more involved, and it is possible the discussions between the users could lead to disagreement. Thus, the users would then assign negative votes to comments which they deemed to be incorrect or to be of low quality.

Table 7.4: Comment Types of Reddit Threads.

	Number of Comments		Correlation between Direct and All Comments in Threads	
Subreddit	Direct	All	Pearson's	P-value
elifive	502010 (34.29%)	1463800	0.8957	$< 2.2 \times 10^{-16}$
gaming	524804 (36.23%)	1448582	0.9178	$< 2.2 \times 10^{-16}$
politics	1842373 (18.17%)	10137258	0.8371	$< 2.2 \times 10^{-16}$
science	210202 (30.78%)	682810	0.9267	$< 2.2 \times 10^{-16}$
worldnews	969053 (18.41%)	5263228	0.9413	$< 2.2 \times 10^{-16}$

Table 7.5: Vote Difference Distribution of Reddit threads.

	Scores of Reddit threads					
Subreddit	Minimum	1st Q	Median	Mean	3rd Q	Maximum
elifive	0.0	1.0	1.0	17.5	1.0	7074.0
gaming	0.0	1.0	1.0	57.7	1.0	9942.0
politics	0.0	0.0	1.0	111.2	12.0	9605.0
science	0.0	1.0	1.0	108.6	5.0	10924.0
worldnews	0.0	1.0	1.0	62.5	4.0	11850.0

In general, across all subreddits, most of the comments tend to have a relatively low vote difference value within the range of one to three.

Table 7.6: Vote Difference Distribution of Reddit comments.

	Scores of Reddit comments					
Subreddit	Minimum	1st Q	Median	Mean	3rd Q	Maximum
elifive	-3276.0	1.0	1.0	6.8	2.0	8537.0
gaming	-865.0	1.0	1.0	9.5	3.0	6243.0
politics	-1763.0	1.0	1.0	5.2	3.0	7098.0
science	-206.0	1.0	1.0	6.9	3.0	5313.0
worldnews	-3032.0	1.0	1.0	8.0	3.0	6917.0

7.5.7 Study: User Activity

Users are an integral component for Reddit as a social curation platform – users sharing external content as link submissions or producing their own textual content as text submissions and comments. In table 7.7, we observe a healthy number of users for each subreddit.

We however observed that user behaviour differs between subreddits. Unlike the other subreddits, the users of *r/explainlikeimfive* and *r/gaming* are active thread contributors, with almost 20% of the users of *r/gaming* only contribute to threads without any comments. Most users are active commenters on these subreddits, particularly the users of *r/politics*, *r/science* and *r/worldnews*.

Only a small portion of users in *r/science* and *r/worldnews* are active contributors of threads and comments. It is interesting that despite the question-answering nature of *r/explainlikeimfive*, 12.72% of its users are both thread starters and commenters; a phenomenon explainable through the assumption that the thread starter themselves would post comments to further seek information, such as clarification from the other comments.

Table 7.7: Activities of Reddit Users.

	Number of Users				
SubReddit	Threads	Comments	Threads only	Comments only	Total
elifive	81965 (27.55%)	253324 (85.16%)	44139 (14.84%)	215499 (72.45%)	297464
gaming	124630 (29.84%)	335491 (80.32%)	82221 (19.68%)	293083 (70.16%)	417713
politics	37914 (11.82%)	304619 (94.99%)	16067 (5.01%)	282773 (88.18%)	320687
science	13387 (9.47%)	131269 (92.84%)	10126 (7.16%)	128009 (90.53%)	141396
worldnews	48566 (9.56%)	476480 (93.81%)	31453 (6.19%)	459368 (90.44%)	507934

Study: User Activity Type

The distribution in the amount of content contributed by the users is heavily skewed to the right. Here, as we observe at least 60% of the users of the explored subreddits created only one thread or made only one comment. At the 90th user percentile, the users only created between $2-6$ threads and $3-4$ comments.

7.6 Summary

Content aggregation (CA) platforms have a large impact on the World Wide Web (WWW) today as a content-rich platform. Users can rely on CA platforms to aggregate interesting or popular WWW resources for consumption; directing a large volume of traffic to those resources. Much work has been done to manage content on CA platforms to better meet the information need of the users such as peer-moderation, personalised recommendations and ranking of content.

Over the years, CA platforms have moved from an editor centric curation model to a social curation model with the loosening of responsibility from selected experts to the user community of the platform. This is motivated by the preference of users towards CA platforms with greater user autonomy which we explored in Section 7.1.

This change creates new challenges for CA platforms from the unstructured and unmoderated nature of user-generated content (UGC) that are of varying information quality. Socially curated CA platforms can be vulnerable to malicious users for self-promotion and vote manipulation. Besides that, user vote bias can hamper peer-moderation efforts. In Section 7.4, we discussed these challenges in greater detail.

Responding to the challenges above, this research aims to improve the content management of socially curated CA platforms. Reddit is one of such platform which we explore in Section 7.5. Unlike other CA platform, Reddit encourages its users to generate new content on the platform – resulting in additional UGC on the platform according to the user-defined communities known as subreddits. Users can curate resources from the WWW with a link submission or contribute new textual content as a self-submissions. Besides that, users can also interact within Reddit threads by commenting with additional information or state their arguments; creating a complex tree-like structure illustrated in Figure 7.4 and organised according to the approaches detailed in Section 7.5.2.

This research performed a study on a Reddit dataset detailed in Section 7.5.3 in order to better understand the complex structure, content and users of Reddit. Our study explored five large subreddits namely *r/science*, *r/worldnews*, *r/gaming*, *r/elifive* and *r/politics*. We observed diversity in content generated and user behaviour between subreddits:

- Different number of threads submitted daily and the distribution of comments in those threads.

- Type of comments[20] do differ; but we note that the length of user comments is fairly consistent.
- The distribution of user votes for peer-moderation do also vary between subreddits.
- Users are mainly commenters on Reddit but they tendency to submit threads do differ between subreddits.

In Chapter 8, this research attempts to improve content curation on Reddit. Similar to our efforts in the earlier Part II of the research, we propose that the information quality of content to be inferred according to the expertise of the content contributor. Thus, this enable us to predict the future user votes for the peer-moderation of content contribution which would overcome the cold-start problem and provide an alternative organisation for user comments in Reddit threads.

[20]Direct and indirect comments.

Chapter 8

Predicting User Contributions on Reddit

We saw the revolution for Content Aggregation (CA) platforms towards a social curation model in Section 7.1, creating new challenges for content management as discussed in Section 7.4. It is however crucial for us to overcome these challenges due to the large impact of CA platforms on the World Wide Web (WWW).

In this chapter, we attempt to improve content management on Reddit. On Reddit, users are encouraged to not only curate content from the WWW for sharing but to also generate new content for the users – either as (1) self-submission threads; or (2) comments within threads. Similar to our earlier efforts, we attempt to estimate the expertise of Reddit users in order to infer the information quality of their generated content. To the best of our knowledge, there exists no research that aims to predict the contribution quality of Reddit users through their estimated user expertise; and instead all prior work focuses on the processing of the content itself [85, 134]. At the current stage of our research, we infer the information quality of user comments to help facilitate discussions within threads.

The thread-based user interactions illustrated in Figure 7.4 create a new challenge for us in the estimation of user expertise which is unlike our earlier research on Community Question-Answering (CQA) platforms in Part II. Users could have multiple interactions within a single thread as they contribute multiple comments when participating in a thread discussion – replying directly to the thread or other comments. Besides that, the existence of down-votes on Reddit would require our earlier proposed approach to be modified for this change in peer-moderation later in Section 7.5.1.

First, we detail the research questions which guide our exploration in Section 8.1. Next in Section 8.2, we detail our motivation for the estimation of user expertise for applications in user modelling and content organisation. To do so, we look at several user estimation approaches inspired from our earlier works – Contribution Count in Section 8.3.1, Z-Index in Section 8.3.3, Contribution Scores in Section 8.4, and Competitive Rating approach in Section 8.5.

We evaluate the performance of the explored approaches as a contribution prediction tasks for user comments in Section 8.6, followed by an analysis of the performance. This is performed on the Reddit dataset from Section 7.5.3 which is studied to better understand user behaviours and their contributed content over several subreddits. Finally, we conclude this chapter with a summary in Chapter 8.7.

8.1 Research Questions

We adapt the research questions introduced in Section 1.4 for our exploration into the estimation of user expertise on Reddit.

8.1.1 Can User Expertise be Estimated on Reddit Through Content-Agnostic Means?

As a socially curated CA platform, Reddit encourages its users to contribute – either by sharing content from the WWW, creating content or participating in discussions. Besides that, the users are free to form communities (subreddits) with the responsibility to moderate content through user votes [130]. But it should be noted that any user is able to be a part of the Reddit and the subreddits of their choice. The users are diverse as we saw from subreddits to subreddits as we studied in Section 7.5.7 with varying degree of expertise [75].

A draw towards Reddit from other CA platforms such as Digg is shared autonomy between users [91]. In this chapter, we attempt to estimate the expertise of Reddit users with the goal of identifying reliable expert users to enhance content curation and also the dissemination of interesting information [44]. Similarly, users of low expertise could be malicious users who aim to disrupt the platform. Thus, their influence should be reduced. As of the time of writing, all Reddit users are empowered the same, with the same capabilities to impact the platform.

Similar to our earlier approach in Part I and Part II, the user expertise would be estimated through content-agnostic means in order to circumvent the unstructured nature of UGC. To accomplish this, we performed a study of Reddit in Section 7.5 and identified the suitable features for user expertise estimation such as vote difference.

8.1.2 Do Expert Users Produce Better Comments?

In this stage of our research, we look to explore the correlation between user expertise and the information quality of their comments. Our goal here is to identify user comments that are of significant contribution to the thread discussion; to see whether users of higher expertise produce better comments [75].

In this chapter, we explore the estimation of user expertise which is then used to infer the information quality of their comments. In the context of Reddit threads, we attempt to compare the information quality between user comments within a given thread by comparing the relative user expertise between the commenters. Our exploration takes into account the two types of user comments noted in Section 7.5.1 – direct and indirect comments. This exploration could ease user consumption of content especially the noticeable large discussions in threads throughout subreddits as seen in Table 7.2 by identifying high quality comments by reliable users.

8.1.3 Can We Predict the Information Quality of User Comments?

As a discussion board, users of Reddit comment within threads. These comments discuss the content shared in the submitted thread – either agreeing with the thread content and providing additional information; or disagreeing with the thread content with counter arguments. Such discussion adds value to Reddit and is popular amongst the users of Reddit where over 80% of Reddit users are commenters (refer to Table 7.7).

Any user can participate in thread discussions and this number increases as the popularity of the threads increases, causing an information overload for other users. Furthermore, these users can vary in capabilities and intention which further complicates the organisation of comments. Reddit already provides several organisations for thread comments as discussed in Section 7.5.2, but this research argues the following:

- The ranking approaches are vulnerable to disruption and manipulation by malicious users such as trolls on the platform [102].

- Many of the comment rankings rely on user votes. There can be a cold-start problem in obtaining a sufficient number of user votes [89] and these votes are susceptible to user bias [92, 107, 118, 146]. Peer-moderation would require sufficient community judgement in order to function well [135].
- Many of the current literature only consider each user comment in isolation without considering the other comments in the thread.

Thus, this research aims to predict the information quality of user comments by inferring it from the commenters itself. Comments of higher information quality contribute more towards the thread discussion. Thus, the prediction is then used to sort user comments for improved user consumption by ranking the better comments higher [73, 92]. This prediction task is used for the evaluation of various user expertise estimation approaches in Section 8.6.

8.2 Application of User Contribution Prediction

The challenges discussed earlier can be tackled with the prediction of Reddit users' contribution to the platform. To the best of our current knowledge, there are no research that aims to predict the contribution of Reddit users through their estimated user expertise and instead focuses on the processing of the content itself [85, 134]. The following sections detailed how the estimation of user expertise and the resulting user contribution prediction can improve Reddit as a whole.

8.2.1 User Modelling

Our research explores the content-agnostic approaches to improve information retrieval and management on UGC platforms such as collaboration tagging (Part I) and community question-answering (Part II). Due to the vast amount and unstructured nature of UGC, this research took a user-centric approach by understanding the users who produce the content. By profiling the users according to their expertise, we are able to deduce the information quality of their contribution in meeting the information needs of other users. Thus, in this chapter, the workings and findings from earlier parts are used to model users of Reddit as our selected CA platform in order to overcome the challenges raised in Section 7.4. The accuracy of the explored approaches to the estimation of user expertise is validated in Section 8.6.

Identifying Domain Experts and Influencers

Users of higher expertise would produce content that meet the information need of other users better [75]. Researches have looked to identify users of high expertise – a process known as expert search [163] to improve the quality of the content on their platform such as the re-routing of questions to identified expert [93]. On Reddit, users of high expertise are able to aggregate and submit interesting links from the web for their area of expertise; or contribute to a discussion in submitted threads with additional information.

Similarly, influential users [12] in their own areas of expertise can be identified to enhance the dissemination of interesting information [44]. For example, the content interacted by these celebrated users on Reddit could be emphasized on the platform as there is a higher tendency of these users to produce content that is of the interest to the community and their active subreddit.

With the capability to identify these users, Reddit can focus on motivating these users to contribute [84] towards the platform by either (1) the production of high quality content (thread submissions or comments); or the moderation of content on Reddit. It is through

these users that platforms are able to maintain the density of high quality content on the platform for the consumptions of other users. Do note that instead of computationally identifying and empowering these users, it can be reasonable to instead bring these users to the attention of appointed moderators of each community.

A by-product in the identification of reliable experts is the early detection of malicious or unreliable users. User reliability and expertise are relative measures; thus, users that are rated much lower than the other users (on the opposite end of the scale) as measured through the expertise estimation algorithms should be cautioned for as having a high tendency to be malicious users.

User-Reliability Aware Voting

As a social curation platform, Reddit depends on its users to moderate the content on its platform. This is a challenge where the vulnerabilities in the moderation process can be exploited to disrupt the integrity and accuracy of the communities' moderation. For example, malicious users [102] can boost and promote their content by having proxy users up-voting their content.

To the best of our knowledge, there is no prior research that investigated the varying reliability and expertise of users on these platforms. The studies into the prediction of content popularity and quality through user vote analysis [89, 91, 116] all weight every vote the same regardless of the users who provide them. This results in the vulnerability discussed above. Besides that, the voice of highly reliable experts could be drowned out by the average users within the community; resulting the hive-mind or circle-jerk phenomena [107] where content of lower quality such as misinformation could spread and spiral out of control [44].

On Reddit, the developers rolled out the vote-fuzzing mechanism to try and eliminate possible content bias [107, 118] by hiding the actual votes gained by the content itself; instead only display the vote difference or hiding the vote counts for the first few moments. Self-appointed moderators could shadow-ban[1] identified malicious users by stripping away their voting weights. This research however argues that it is possible to still beat the mechanism by:

- Coordinated attacks to vote for a particular content as there is no need to know its current vote count that is hidden via the vote-fuzzing mechanism to boost up a piece of content. In fact, the early swell in user votes could cause a chain reaction towards popularity of the content [107, 116].

- Having multiple proxy accounts or bots to manipulate the votes for content – accounts that lie low away from the human moderators to avoid being shadow-banned.

As user votes do play a big role on the peer-moderation on social platforms such as Reddit, an improved voting mechanism could improve the content management. Digg weights user votes according to their user groups such as power users [91] but these are not done computationally.

This research proposes the possibility of weighting user votes according to their estimated expertise – higher weight for identified experts, average weight for the average users and low to little weight for malicious users or users with low expertise estimates. These weights should be adjusted dynamically as the users' behaviour changes and also varies according to the topic or community in which the voted content is found. As a result, this increase the impact of reliable users while reducing the impact that malicious users have on the platform. This process is similar to Part I where we weighted annotation terms by the expertise of their annotator to improve the resource description.

[1]The banned individual is not aware of the ban.

8.2.2 Content Organisation

From the estimated user expertise, the contribution of the users towards the platform can be predicted. This is based on the hypothesis that users of greater expertise and reliability would produce content of higher information quality [75] for both thread submissions and comments in threads. With this additional information, contents such as submission threads and user comments can be sorted to better meet the information needs of the users and also improve the peer moderation process of Reddit.

Overcoming the Cold Start

Previous studies into the estimation of content quality have relied on the votes and voting patterns associated with the content itself [89] – an approach vulnerable to the cold-start problem (see Definition 13) when there is no or few votes available. As new content are constantly being produced on the platform, how should the content be organized in order to improve content consumption and peer moderation even if there is little to no peer-moderation until more users view and judge the content?

This research proposes the organisation of new content dynamically according to the estimated expertise of the users who contribute the content. It is more likely for reliable expert users to produce the content of high quality. Thus, this content can be promoted in the list for the consumption of users as users do consume content from the highly visibly top of a list to the bottom [33, 73, 92]. This is the alternative to the temporal sorting of content that does not overcome the cold-start problem and is instead suspicable to temporal bias [45]. As more user votes are available, the sorting of content can then be adjusted to the optimal order.

Overcoming the User Bias

A challenge in the peer moderation of content on social platforms such as Reddit is the prevalence of user bias in the judgement of content through their votes. Users tend to vote for the content which they have viewed early [146] and might not continue to view and judge the remaining content. This is a challenge faced in peer moderation of content when the content is ordered according to their temporal submission order.

Furthermore, users can be biased from the influence of herding effect – where the community's views on the content [107, 118] such as current votes gained by the content affects the opinions of other users. For example, users are more likely agree with the content that is accepted by the other users and up-vote these contents with already highly voted by the community. The vote fuzzing mechanism implemented in Reddit cannot overcome this bias as the vote difference displays the community's agreement of the content.

This research proposes a change in the ordering of content to account for the user bias. Content can be ordered according to the estimated user expertise of its contributors as an alternative to the temporal ordering. The votes garnered are to be hidden early on as the ordering changes dynamically through a combination of predicted quality, temporal creation and votes gained to better collect user judgements without any bias.

8.3 Simple Counting Approaches

Our earlier work in Part II for the estimation of user expertise saw simple that approaches were competitive in performance when compared to state-of-the-art estimation approaches [158]. Thus, we begin the estimation of user expertise with simple counting-based approaches – (1) the C-Count; and (2) the Z-Index.

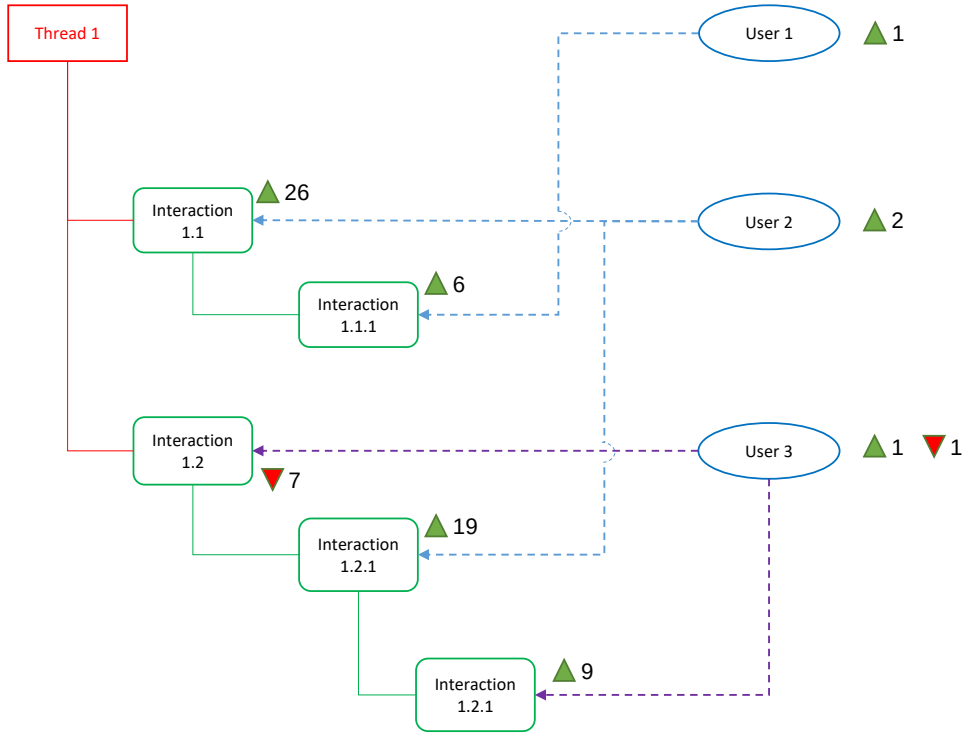


Figure 8.1: The Contribution Count (C-Count) approach for user expertise. Interaction vote difference is listed beside each interaction.

8.3.1 Baseline: Contribution Count (C-Count)

This is the simplest approach. It makes the assumption that expert users are those users who make a large number of significant contributions as judged by the vote difference. Thus, the approach counts the number of good contributions made by the users as their estimated user expertise as shown in Figure 8.1.

Identifying Contributions Significance

Reddit is moderated through peer assessments that empowers the users with the ability to vote [130, 134] as judgements [89] for content. The votes can be used to infer the contribution of the user interactions towards the thread discussion. Thus, the vote difference gained is used to measure the contribution of each comment. The contribution significance of each comment, Sig(I) can be judged within each thread according to:

- **Polarity.**
  A user comment is a "good contribution" if the vote difference gained by that comment is positive; and negative otherwise. This measure of polarity would however suffer from user voting bias where users are inclined to vote positively for good content but not negatively for poor content [118] – motivating us to explore the other variants.

- **Median.**
  The median vote difference over all comments in a thread can act as threshold to judge if a comment's contribution is significant. A contribution is good if it is "more significant" than half of the other contributions of the thread; consistent with the relative nature of our estimated user expertise. Moreover, this variant overcomes

the imbalanced ratio of positive vs. negative vote difference in separating good and bad contributions.

- **Median Direct.**
  This variant is an extension of the median approach which uses the median vote difference of all direct comments (refer to Section 7.5.1 for comment types) instead.

- **Mean.**
  Similar to median variant, we adjust the threshold to consider the significance of the user comment. This approach could however punish the users when there is a non-uniform distribution over votes.

- **Mean Direct.**
  The threshold is adjusted according to the vote difference mean of direct user comments.

Definition 18 (Contribution Significance)**.** The information quality of a user's contribution towards the thread discussion.

Punishing Bad Contributions?

If a user produced a bad contribution $\text{Sig}(I) < 0$, the user could be punished. This research determines the effect from punishing users for their bad contributions by counting the following as the user's estimated expertise – (1) Only the number of good user contributions as in Function 8.1; or (2) The difference between the number of good vs. bad contributions as in Function 8.2 with our findings discussed in Section 8.6.6.

$$\text{C-Count}(u) = |\text{Sig}(I_u) \geq 0| \tag{8.1}$$

$$\text{C-Count}(u) = |\text{Sig}(I_u) \geq 0| - |\text{Sig}(I_u) < 0| \tag{8.2}$$

Decay

User expertise changes overtime as users grow to become better experts in their area. Thus, the user's latest interactions would provide a better judgement for expertise estimation. We update a user's estimated expertise at temporal time t, $\text{C-Count}_t(u)$ by a power of the decay factor $\lambda = 1.0, 0.9, 0.5, 0.1$ on the earlier measure as shown in Function 8.3.

$$\text{C-Count}_{t+1}(u) = \text{C-Count}_t(u)^{\lambda} + |\text{Sig}_t(I_u) \geq 0| \tag{8.3}$$

8.3.2 Multiple Interactions

A Reddit user can contribute one or multiple comments in a thread. Here, this research takes into account all of these interactions. If a user is active, then it is up to the user to ensure that the comments made are 'good contributions', or suffer the consequences otherwise. This research takes this stance as to not discourage user activity which is vital towards to generation of content on the platform. Besides that, this enabled the research to study the effect of user activity on user expertise estimation.

8.3.3 Baseline: Z-Index

The Contribution Z-Index is another baseline approach for this research as a possible improvement over the earlier Contribution Count approach. It is based on the Z-Index used in Community Question-Answering (CQA) platform [158] from Part II adapted for

Reddit – it considers how many more times a user has made a good contribution rather than a bad one; based on the assumption that a random user is just as likely to make a good contribution as they are to make a negative one. Thus, the count of positive vs negative contribution should follow a Binomial distribution. The Z-Index measure is used as the estimated user expertise. This approach also includes all the variants discussed earlier.

$$\text{Z-Index}(u) = \frac{|\text{Sig}(I_u) \geq 0| - |\text{Sig}(I_u) < 0|}{\sqrt{|\text{Sig}(I_u) \geq 0| + |\text{Sig}(I_u) < 0|}} \tag{8.4}$$

8.4 Contribution Scores (C-Scores)

In this baseline, this research measures the significance of user contributions as a score. The vote difference of user contributions $\text{Vote}(I)$ is in the range of $(-\inf, \inf)$ which is then used to infer contribution [95] according to:

- Is the content interesting [89]?
- How popular is the content [91, 116]?
- Does the content contribute towards the discussion?

Experts are users with high collected scores from their comments. We utilize this additional information as a measure of how good or bad user interactions are in contributing towards the estimation of user expertise.

8.4.1 Scores as Expertise

From the collected contribution scores, the user expertise can be estimated. Here, we explore two variants:

- Sum of contribution scores as shown in Function 8.5.

$$\text{C-Score}(u) = \sum_{k=0}^{|I_u|} \text{Vote}(i_k) \tag{8.5}$$

- Average of contribution scores shown in Function 8.6.

$$\text{C-Score}(u) = \frac{\sum_{k=0}^{|I_u|} \text{Vote}(i_k)}{|I_u|} \tag{8.6}$$

In the sum variant, users are rewarded for being active as displayed in Figure 8.2. Similar to the Contribution Count and Contribution Z-Index; we decay the contribution scores of the earlier user comments before including the newest user comments for the estimation of user expertise.

8.4.2 Contribution Score Adjustments

The implementation of contribution scores can be tricky due to the varying number of user comments and the number of votes within a Reddit thread. Thus, there would be a need to adjust the contribution scores according to the thread where the interaction is. The variants explored here are:

- **Raw.**
  The vote difference is taken as the user's contribution score without any adjustments,

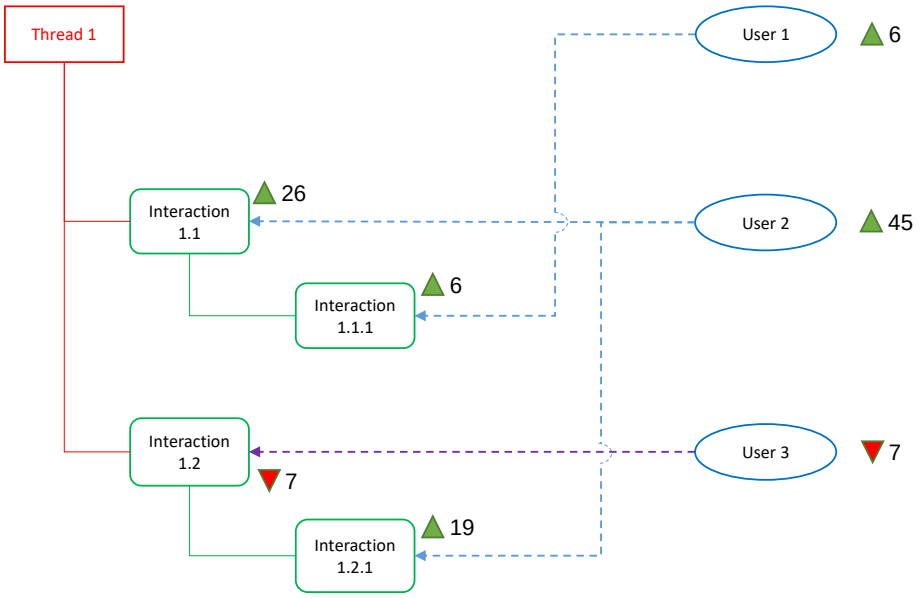


Figure 8.2: The Contribution Score (C-Scores) approach with the sum-variant for user expertise.

enabling us to explore the effect of popular threads[2] on the contribution scores for user expertise estimation.

- **Mean and Mean Direct.**
  The mean variant attempts to normalise the contribution scores gained for each user interaction according to the mean vote difference of all comments or direct comments only.

- **Median and Median Direct.**
  Similar to the mean adjustment, here we normalise the contribution scores by the median instead.

8.5 Proposed: Competitive Rating (C-Rating)

The C-rating approach is a pairwise comparison approach for user comments that was inspired from its success in answer quality prediction on CQA platforms in Chapter 6 of Part II. Several adjustments have been made to adapt the pairwise comparison approach for Reddit comments.

8.5.1 A Competitive Model for Reddit

First, we would need to model each Reddit thread as a competition between the users who comment on it. The performance of users is measured according to the significance of their contribution. The estimated expertise are the user ratings updated according to the users' performance in each thread, compared to that of other users in the thread. A user of higher rating is expected to be more likely to contribute significant contributions than users of lower rating.

[2]High number of comments and user votes

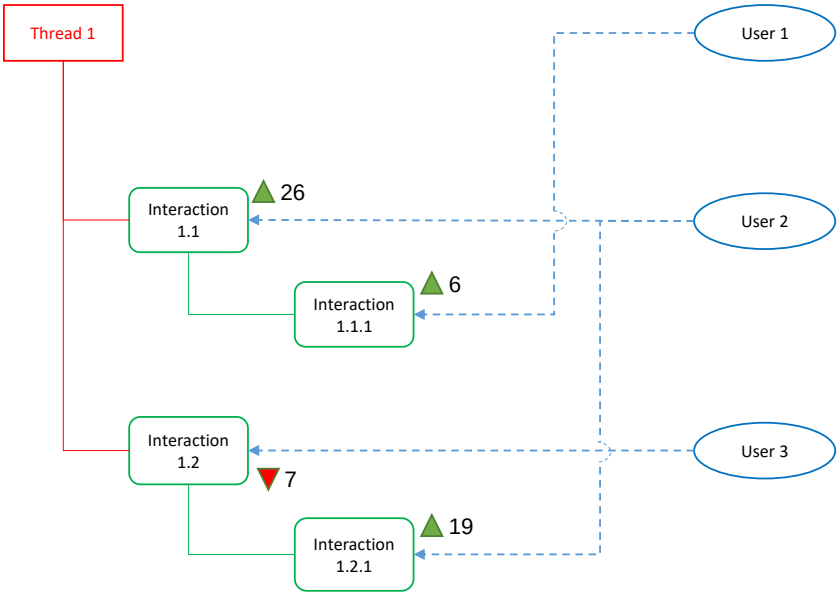


Figure 8.3: An example scenario for Reddit thread. The vote difference of user interaction is listed beside each interaction.

Comparison Pairs

Comparison pairs are built between user comments within each Reddit thread such as the Round-Robin format used in Chapter 6 between every user comment. Unlike our work on CQA platform, users can interact multiple times with a single thread, contributing multiple comments as they partake in the thread discussion. This complicates the building of comparison pairs, akin to a competition scenario where a player faces the same opponent multiple times, each time with a new result. Thus, we propose the possible variants to be made:

- **Multiple interaction-based competition.**
  This variant form competitive pairs between users according to their interactions. If there are multiple interactions or content produced by the users within a single thread, then multiple pairs are formed here.

- **Best interaction-based competition.**
  If a user has multiple interactions within the same thread, then the best interaction (according to the credit that interaction gained) will be used as that user's performance measure in that thread in the pairwise comparisons with the other users.

- **Average interaction-based competition.**
  The credit received by all of the user's multiple interactions in the thread itself is averaged as the user's performance measure in the thread for the competition between the user and other users.

- **Summed interaction-based competition.**
  All of the user's interactions in the thread are summed together to be a single contribution to the thread. While the approach is simple, it has the downside of benefiting and inflating the contribution of active users in the thread.

For our evaluation later in Section 8.6, we model each Reddit thread according to the multiple interaction-based competition variant in order to keep it similar with our work

in Chapter 6 though we looked to explore the other variants in the future. Comparison pairs are built between each user comment; given that the authors of the comments are not the same. Consider a Reddit thread illustrated in Figure 8.3, the following possible pairs include:

- Interaction 1.1 and Interaction 1.1.1
- Interaction 1.1 and Interaction 1.2
- Interaction 1.1 and Interaction 1.2.1 cannot be built as they are both contributed by the same User 2.
- Interaction 1.2 and Interaction 1.2.1
- Interaction 1.1.1 and Interaction 1.2.1

We however note that such an approach may not be efficient as popular threads do contain a large number of user comments as we saw in Table 7.2. Thus, we made adjustments to the Round-Robin format to only form comparison pairs between a comment and all of the other direct comments in the thread in order to reduce the runtime complexity. Therefore, under this adjustment, the following pairs are formed:

- Interaction 1.1 and Interaction 1.1.1
- Interaction 1.1 and Interaction 1.2
- Interaction 1.2 and Interaction 1.2.1

Independent Subreddits

The subreddits are independent of each other under the competitive approach just like in sports – how well you do in one sport is considered independent of another (think of basketball vs tennis). While a user can be active in multiple subreddits [130], that user would have different expertise score for each of the subreddits that he or she has participated in as an expertise vector. A general rating can be used to measure a user's reliability across all threads however and this would be studied in our future work.

8.5.2 User Contribution Performance

The performance of users in each pairwise comparison can be measured according to the vote difference gained by their comments similar to how the contributions are measured in Section 8.4 as contribution scores. This is the inferred performance of the users in contributing towards the thread discussion which would then be used to update the user ratings.

Glicko-2 Rating

This research once again chose the Glicko-2 rating system [46] for the estimation of user expertise. We detailed the Glicko-2 rating system in Section 6.3.1 and the same model is adapted for Reddit.

Rating Period

The resulting outcome from the comparison pairs generated according to Section 8.5.1 are collected if they from threads that are submitted within the same rating period. As we evaluated in Section 6.4.4, the addition of rating periods does improve the performance of competitive pairwise comparison as opposed to the commonly used live-update. For our evaluation in this chapter, we set the rating period to a period one day (daily) due to the high amount of threads submitted daily (Table 7.1) and comments contributed (Table 7.2) on those threads.

Win-Margin

Traditionally, competitions including real world sports are only concerned with the outcome of a match-up and the winner takes all. Thus, pairwise comparison approaches are often scoreless without looking at the margin of victory and often disregard draws [10, 95]. The win-margin variant is inspired by the Bradley-Terry model which correlates the win probability of a player in a game with the player's real rating μ_1. This win probability is $\frac{\mu_1}{\mu_1+\mu_2}$ when the player is against an opponent with a rating of μ_2 [20].

Our earlier work on CQA platform in Chapter 6 found improvements towards the estimation of user expertise if we were to consider the win-margin between the winner and the loser of a pairwise comparison. Since Reddit is a different platform (with negative vote difference) we once again explore the scoreless and the win-margin variants. To account for the negative vote difference in comments, modifications to the win-margin are required. We propose the following adjustment:

- If the vote difference for both comments are positive, there is no change to the win-margin calculation.
- If there is a negative vote difference, then it is a total victory (win-margin of one) for the commenter with the higher vote difference. The justification for this approach is that comment with negative vote difference can be perceived to be of no contribution to the thread discussion.
- If the vote difference for both comments are negative, the values are swapped and then converted to positive before applying the same win-margin calculations.

We leave to future work as to how the negative vote difference can be incorporated better into our proposed competitive rating approach especially for the difference-based competitive models of Reddit.

8.6 Contribution Prediction Evaluation

The performance of the approaches discussed earlier are evaluated for the estimation of user expertise on Reddit. The user expertise estimated from the approaches are used to predict the information quality of comments (future vote difference gained) contributed by the users. The evaluated approaches include:

- C-Count detailed in Section 8.3.1
- Z-Index detailed in Section 8.3.3
- C-Score detailed in Section 8.4
- C-Rating detailed in Section 8.5

8.6.1 Methodology

The evaluation process is conducted on the collected Reddit dataset detailed in Section 7.5.3. User expertise estimation performance is investigated on four subreddits – *r/explainlikeimfive*, *r/gaming*, *r/science* and *r/worldnews*. The fifth *r/politics* is not evaluated pending more data collection for future work due to the volatility surrounding the US 2016 Presidential Elections; unlike the global *r/worldnews*.

Training and Testing Cycles

Unlike traditional Web platforms with static content, a social curation platform like Reddit expands rapidly with the constant flow of user content. Besides building high accuracy models for the estimation of user expertise, the approaches explored are fast and efficient in taking in new content for the estimation process. Thus, the performance evaluation is

modelled as a continuous cycle, beginning with a training cycle according to the timestamp of Reddit threads.

- **Training Cycle.**
  Once the threads and the comments have been used for evaluation, they are added to the training data. New users will have their expertise estimated for the first time and existing users have their expertise updated with the additional new data. The updated user expertise is then used for prediction in the following testing cycle.

- **Testing Cycle.**
  In the testing cycle, the threads and comments of the threads are used for evaluation. Here, the estimated user expertise from the explored approaches are used to predict the information quality of comments according to the truth and evaluation measures. Only comments of known users with the expertise estimated from their prior interactions are evaluated.

Ground Truth

The actual vote difference gained is used as the ground truth judgement of information quality for thread comments. User comments within each Reddit thread are ranked according to this judgement. The explored approaches attempt to rank the same comments according to the estimated expertise of comments' author; to match the ground truth rank order to predict the future community assessment of content quality. Our experiment looked at predicting the ordering of – (1) direct comments; and (2) all comments in a thread. We then measure the Kendall's Tau correlation between the ground truth and the predicted ordering.

Kendall's Tau Rank Coefficient, Tau-B Measure

This research selects (up to) the 10 best comments for each Reddit thread[3] according to the ground truth. Joint observation pairs are then formulated for each of these comments in a Reddit thread. We then measure the number of concordant and discordant pairs for the observation pairs between the ground truth ranking and the ranking for each of the explored approaches. As it is possible for two comments to have the same vote difference or two users to have the same estimated expertise, we handle these ties with the Kendall's Tau-B, τ_B statistic.

8.6.2 Results

As detailed in Section 8.6.1, the evaluation process is based on an online training-test cycle. From a total of $724,505$ threads, $333,175$ threads have at least 1 direct comment created by a user with prior contribution (for the expertise to be estimated) and $344,375$ threads with at least 1 comment. Hence, 45.99% and 47.53% of the threads are suitable for direct comments and all comments evaluation respectively.

We note that this value is much smaller for the *r/gaming* subreddit with only 23.03% and 25.70% respectively; possibly due to the low number of comments per thread (Table 7.2) and almost 20% of its users are thread contributors only (Table 7.7). On the other hand, the *r/explainlikeimfive* subreddit recorded a higher number of suitable threads for evaluation with 78.78% and 79.51% of threads for direct comments and all comments respectively; possibly due to a higher density of threads in the subreddit with at least one interaction. The high number of threads and the comments in each thread ensures the significance of the evaluation measure.

[3]This is to otherwise avoid expensive computations for threads with a very high number of comments

Prediction Performance

Table 8.1 summarises the comment quality prediction based on the estimated user expertise of the explored approaches at their best variant; measured with the Kendall's Tau-B Rank Coefficient against the ground truth. It can be observed that the best performing approach is the C-Rating approach followed by the C-Count, C-Score and finally the Z-Index.

Table 8.1: Kendall's Tau-B Rank Coefficient for Direct Comments Ordering based on Quality Prediction with User Expertise Estimation with the Evaluated Approaches (Best Variant). Best performance in bold.

	Direct Comments for Subreddit				
Approach	elifive	gaming	science	worldnews	All
C-Count (mean threshold, no decay)	0.9164	0.7923	0.9148	0.8890	0.8797
C-Score (average score, raw value, no decay)	0.7816	0.7014	0.8873	0.7810	0.7720
Z-Index (mean threshold)	0.7400	0.6928	0.8710	0.7088	0.7280
C-Rating (scoreless)	**0.9796**	**0.9298**	**0.9760**	**0.9867**	**0.9716**

Table 8.2: Kendall's Tau-B Rank Coefficient for All Comments Ordering based on Quality Prediction with User Expertise Estimation with the Evaluated Approaches (Best Variant). Best performance in bold.

	All Comments for Subreddit				
Approach	elifive	gaming	science	worldnews	All
C-Count (mean threshold, no decay)	0.8284	0.6923	0.8610	0.8512	0.8084
C-Score (average score, raw value, no decay)	0.6423	0.5702	0.8098	0.6941	0.6566
Z-Index (mean threshold)	0.6258	0.5709	0.7975	0.6265	0.6256
C-Rating (scoreless)	**0.9472**	**0.8786**	**0.9469**	**0.9777**	**0.9434**

Generally, the explored approaches perform well on all subreddits except for the *r/gaming* subreddit where the threads lack user comments as discussed in Section 7.5.5. The C-Rating approach however performed very well in this subreddit as a robust approach.

8.6.3 Analysis: Differentiating Users

The estimated user expertise is a relative measure which allows us to compare the frequency of quality contributions between two users. For example, in Reddit threads, it would be ideal for the expertise of the authors of the comments to not be the same unless they are all contributing comments of the same quality. This is to allow us to rank their comments according to the difference of their estimated expertise.

We present in Table 8.3, the evaluated approaches at their best variant by not having an order according to the estimated user expertise when there is an order for the ground truth. Here, we observe that the C-Count approach is unable to differentiate user expertise well within an evaluated thread when compared to the other approaches. Upon further inspection, we note that the C-Rating (scored variant) has the highest number of ties in order when the ground truth is a tie as well with only 0.24% of the ties going undetected.

8.6.4 Analysis: Vote Difference as a Contribution Measure

Similar to the earlier work on CQA platforms in Part II, the user vote information should not be directly used as an estimate of user expertise. The C-Score approach recorded a lower performance measure in comparison with the other approaches; with the average value variant outperforming the sum variant. Instead this information should be used as

Table 8.3: Evaluated Approaches (Best Variant) with the Least Number of Orderless User Expertise in Reddit Thread Comments.

Approach	Percentage of Thread	
	Direct Comments	All Comments
C-Count (mean threshold, no decay)	5.88%	5.12%
C-Score (averaged score, raw value, no decay)	5.31%	4.70%
Z-Index (mean threshold)	5.58%	4.88%
C-Rating (scoreless)	5.46%	4.77%

an indication of contribution significance – by counting the significance of a contribution using the C-Count or as a performance indicator for pairwise comparison in the C-Rating. This is crucial for an environment with sparse user activity (a lower comment count in thread, a lower word count per comment) such as the *r/gaming* subreddit where we observe the largest performance gain.

8.6.5 Analysis: Vote Difference as a Relative Performance Measure

The high performance of the C-Rating approach is consistent with our earlier findings on CQA platforms in Chapter 6 that it is possible to apply rating systems in user expertise estimation through pairwise comparisons. For Reddit, we showed that this is possible even when comparing against direct comments of threads only; a decision we made to improve the efficiency of the expertise estimation process.

User votes on content can be used as a judgement of user performance including the negative vote difference. The additional information from the existence of negative vote difference however does not benefit the successful win-margin (scored) variant (unlike in Chapter 6) as the scoreless variant recorded a slightly better performance with the exception on for *r/science* and *r/worldnews* subreddit.

Raw Values as Measures

Another interesting observation here is that the best performing variant for the C-Score uses the vote difference as a raw value; a value which we had feared would be inflated and thus, should be normalised within the context of the thread itself. For example, popular threads tend to have higher number of user votes that inflate the vote difference of comments.

What this finding suggests is that users who commented on such threads and recorded a high number of vote difference should be rewarded for being able to make an impact on such a challenging thread in the first place. This is further supported by the C-Score having the average score value as its best variant, implying users should be providing consistently high-quality content instead of more content of a lower quality.

Significance Threshold

The vote difference gained by a content is useful as an indicator of content contribution for a thread discussion. Often, the polarity of vote difference is used to differentiate contribution significance especially from the users' perspective – positive for being significant or of high information quality; and negative for being insignificant or to be of low information quality.

Our findings here showed that the best performing variant for both the C-Count and the Z-Index have the mean of vote difference for all comments in a thread as a threshold for content significance – a comment is considered significant for a thread if the vote

difference of that comment is above the mean. This finding is in line with earlier studies on user voting bias where users more are inclined to vote positively for good content but not negatively for poor content [118].

8.6.6 Analysis: Penalising Bad Content

This research explored the possibility to ignore bad contributions by users which may be caused by user bias or personal opinion; and instead only reward good contributions. The findings however suggest otherwise where all such variants in the C-Count and C-Score were outperformed by variants which take the bad contribution into account. Besides that, the C-Rating which naturally account for bad contributions as a losing performance measure recorded the highest evaluation measure. Thus, the users need to be penalised for making bad contributions that are disruptive in Reddit thread discussion.

8.6.7 Analysis: Decaying User Expertise

Literature suggests that the user expertise improves overtime [101, 150] as they consume more knowledge and earlier contributions should be discounted as it does not reflect the users' current expertise. A decay factor is introduced to reduce the impact of earlier contributions towards the estimation of user expertise. Our findings however do not agree with this notion where the variants with decay for the C-Count and C-Score were outperformed by the non-decay variants. This can possibly be due to:

- There is a lack of consistent user contribution or activity on Reddit, as discussed in Section 7.5.7. Decaying these user interactions would reduce the amount of information available for the estimation of user expertise.
- There should be an impact from the earlier user interaction which rewards good contributions and penalises the disruptive discussions. In this fashion, the users would need to be making significant contributions in the present and future to atone for the earlier bad interactions.

8.6.8 Conclusion

The findings suggest that it is possible to estimate user expertise for the prediction of content quality, with an average Kendall's Tau-B rank correlation of 0.9434 using the proposed C-Rating approach to rank thread comments in comparison against the ground truth of unseen actual user moderation. This performance is consistently better than the other explored baseline approaches even in subreddits with sparse amount of user comment interactions such as the *r/gaming* subreddit; showcasing the robustness of the approach for the unpredictable nature of UGC platforms.

The vote difference is a good feature to indicate prior content quality for the explored approaches. This research discovers however that the vote difference should not be used directly as an indication of user expertise as the C-Score approach performs poorly. Instead, the vote difference can be used to identify the content of significance contribution in a thread discussion using the mean vote difference of comments in a thread in the C-Count approach; or as a relative measure for the pairwise comparison of user performance in the C-Rating approach. Both approaches were able to estimate user expertise for the prediction of content quality.

Findings from the research also suggest that the despite the fact that raw vote difference values are inflated within the popular threads, the values should not be normalised. Users should be rewarded for interacting and making an impact on such competitive threads. Besides that, this research discovers that the polarity of vote difference gained by content

is not a suitable threshold in judging contribution significance of content due to user vote bias. Instead, a piece of content is deemed to be significant if it registered a vote difference above the average of vote difference for other content within the same scope.

8.7 Summary

In this chapter of the research, we estimate the expertise of Reddit users in order to predict their future contribution. The goal as discussed in Section 8.2 is to achieve an improve content curation mechanism on Reddit to improve the organisation of user comments in Reddit threads – users of greater expertise would often generate content that contributes better to the thread discussion. Our proposed direction overcomes the challenge of peer-moderation such as the cold-start problem in garnering user votes and also the potential vote bias. As a by-product, the influence of users of lower expertise and malicious users can be reduced so as to not disrupt the discussion on Reddit.

Similar to our earlier efforts in Part I and Part II, our approach to the estimation of user expertise is content-agnostic as to circumvent the challenge in processing unstructured user-generated content. The vote difference feature of Reddit can be extracted to indicate user contribution significance as we introduced in Section 8.3.1. By identifying significant contributions, user expertise can be estimated through the simple count of net-positive voted contributions (C-Count) and also the Z-Index approach. The measure of significance can also be incorporated as an estimate of user expertise with the C-Score approach in Section 8.4. Alternatively, Reddit can be modelled as competitive environment between users. In Section 8.5, we adapt the successful competitive pairwise approach from Chapter 6 to Reddit – accounting for the multiple interactions by a single user and the existence of negative votes.

With the estimated user expertise from all of the explored approaches, we attempt to predict the contribution significance of future user comments in Reddit threads in Section 8.6. Our findings suggest that it is possible to predict the information quality of user comments from the estimated user expertise – where users of higher expertise tend to contribute better towards thread discussions.

The proposed C-Rating approach is robust with a good Kendall's Tau-B rank correlation performance of 0.9434 across the explored large subreddits in predicting user contribution. This prediction is made solely using the estimated user expertise without processing the comment content itself. Also, it is able to differentiate users and their performance well despite the high number of Reddit users.

The vote difference for user comments can be a good indicator of contribution significance, providing the mean vote difference in the Reddit thread is used as the significance threshold. Bad contribution by users should be penalised when used in the expertise estimation process and not to be decayed due to the sparseness in user interactions as we observed in Section 7.5.7 earlier.

Chapter 9

Conclusion

Today, users on the World Wide Web (WWW) are no longer just consumers of content. Instead, they are valuable contributors of user-generated content (UGC) [50]. UGC has the information potential to rival that of experts [132] or supplement them [111]. This capacity has turned UGC platforms into rich repositories of knowledge [26] with the Wisdom of the Crowd (WotC) [135] including the three UGC platforms which we explored in parts of this thesis:

- Delicious social bookmarking website, a Collaborative Tagging (CT) platform.
- Yahoo! Chiebukuro site, a Community Question-Answering (CQA) platform.
- Reddit, a Content Aggregation (CA) platform with social curation and discussion board.

The advent of UGC platforms however creates new challenges for the WWW. Any user on UGC platforms can generate, contribute and share content. These users vary in reliability and expertise [93]; resulting in some user contributions being better than others [111]. Moreover, the intentions of these users [109] are not known and malicious users can severely disrupt the experience of other benevolent users leading in some cases spiralling out of control [44].

The unstructured nature of UGC complicates computational techniques for content-based judgement of content quality and also the use of user profiling approaches [152] especially considering the large constant flow of UGC [50]. Thus, many UGC platforms leverage on the WotC by having their user communities peer-moderate content on the platform [135]. User votes as the communitys long-term review of content quality [9] can be used to identify content of low information quality. This research however argues that:

- The community is not made from reliable users with many affected by user biases [107, 118].
- There is a cold-start problem in peer-moderation, which can be further challenged by temporal bias during judgement [45, 146].
- Peer-moderation systems are vulnerable to manipulation by malicious users [102].

Hence, this thesis detailed the research undertaken to address these challenges. We attempt to infer the information quality of UGC to improve content management on various platforms. Instead of processing the diverse, unstructured and large volume of UGC; we infer the information quality of UGC from the contributing users themselves. By estimating the expertise of users, it is possible to infer the information quality of their generated content [18] – expert users producing content that are of higher quality than non-experts [155].

To accomplish such a feat, we explore the estimation of user expertise. This research studies distinct UGC platform features as signals to indicate user expertise; as a content-agnostic approach to circumvent the unstructured noisy nature of UGC. The explored

estimation approaches need to be robust in adapting to the structure and representation used in UGC platforms. In return, the estimated user expertise is applied to improve information systems with UGC platforms by:

- Identifying popular and high-quality WWW resources during information retrieval (IR).
- Categorisation of WWW resources through user annotations.
- Improving similarity measure calculations for annotation-based retrieval (AR) through better WWW resource descriptors.
- Predicting the best answer for a user question, therefore overcoming the cold-start problem.
- Identifying domain-sensitive experts, an expert search task to direct and motivate further high-quality contributions.
- Introducing measures for user contribution significance on Reddit.
- Predicting future user comment votes to facilitate discussions in Reddit threads.
- Identifying influential users on social communities such as subreddits.

Our work is documented through the three parts of the thesis, addressing the research questions proposed earlier in Section 1.4.

9.1 Addressing Research Questions

We summarise responses to the research questions introduced in Section 1.4, which we answered in each part of the thesis.

9.1.1 Information Quality of UGC can be Estimated through Content-Agnostic Means

The information quality of content can be estimated through content-based analysis that takes into account the sentiment, sentence structure, grammar and terms present in the content [153, 157, 159]. This research argues however that these approaches are not suitable for the unstructured nature of UGC nor is it computationally efficient with the constantly high number of UGC. In our study of Yahoo! Chiebukuro from Chapter 4, we saw questions with large number of answers. Similarly, in the case study of Reddit (a CA platform with social curation) from Chapter 7, we observe large numbers of thread submissions daily that is accompanied by a considerable number of user comments.

In this research, we explored a diverse range of user expertise estimation approaches; all adapted to the different structure and organisation of each UGC platform. The user expertise is then used to infer the information quality on these platforms including the:

- Accuracy of user annotations of CT platforms in describing WWW resources.
- Capability of user answers in CQA platforms for meeting the information needs of the questioners.
- Significance of user comments in contributing towards the thread discussion on Reddit.

There is a Relation between User Expertise and the Information Quality of the Content

The information quality of UGCs can be inferred from the expertise of users who generate them without the need to process the content themselves. Expert users usually generate better content than other users of low expertise. On the other hand, malicious users generate unwanted noisy content which disrupts the experience of other users. By identifying the experts, it is possible to distinguish:

- High quality user annotations in CT platforms to improve description of WWW resources.
- The best answers in CQA platforms that best meet the information need of the questioner.
- Significant user comments that contribute towards thread discussions on Reddit.

Such an approach addresses the cold-start problem often associated with peer-moderation approaches; and the resulting user votes can then be used to further improve the user expertise estimation. Besides that, the content can be pre-emptively organised to counteract both the user bias and the temporal bias during peer-moderation.

On UGC platforms, there are still more content consumers than content contributors; as we discover from our Yahoo! Chiebukuro case study in Section 4.4 where we observe a much higher number of questioners than answerers. Through expert search [34, 163], it is possible to identify knowledgeable expert users who should be motivated by UGC platforms for additional content contribution.

9.1.2 User Expertise can be Estimated on UGC Platforms

It is possible to estimate the user expertise on UGC platforms through content-agnostic means. In this thesis, we discussed various simple and state-of-the-art approaches before proposing several approaches on our own. These approaches are explored and evaluated on several UGC platforms. The explored approaches are summarised in the following subsections.

It should be noted that all of the explored approaches are content-agnostic approaches to user expertise estimation. Here, we studied the features, structure and representation of UGC platforms in order to identify suitable signal of user expertise such as the temporal ordering of user interactions in Section 3.4.1 and user votes.

Simple Approach

It is possible for simple approaches to user expertise estimation to be competitive against more complex approaches [158]. Thus, we explored several simple approaches:

- **Simple Count.**
  As the simplest approach, the count of good or significant user contributions can be used as a measure of user expertise. For example, we can count the number of best answers contributed by answerers or the number of significant comments by Reddit users in a thread discussion, such as the C-Count approach from Section 8.3.1.

- **Z-Index.**
  The Z-Index approach is based on the assumption that a random user is as likely to produce a good contribution than that of a bad contribution. This approach is applied to estimate user expertise on both CQA platforms in Section 4.3.1 and CT platforms in Section 8.3.3.

- **Vote-based Scores.**
  User votes are found in peer-moderated UGC platforms [9]. UGCs can be voted by the community with the resulting votes as the community's long-term judgement of quality [159]. Thus, the votes gained by the users through their generated content can be used to indicate user expertise such as the Vote Scores approach in Section 5.4.4 and the C-Score approach detailed in Section 8.4.

Graph-based Approach

By studying the structure of UGC platforms, it is possible to represent each platform with a graph model. This is explored throughout Parts I and II of the thesis, such as the tripartite graph model visualised in Figure 2.2 for the CT platforms and the questioning-answering relations visualised in Figure 5.1 for the CQA platform.

With a graph model, authority can be propagated between the entities, similar to the PageRank algorithm [112] and Hyperlink-Induced Topic Search (HITS) algorithm [80]. Such authority can be regarded as a representation of user expertise.

We argue that the current state-of-the-art user-user graph approaches do not differentiate user interactions. For example, two users might answer different questions with one being more difficult than the other. Thus, we proposed the Credit Graph model for CT platforms in Section 3.5 and CQA platforms in Section 5.3. The Credit Graph model edges are weighted according to the significance of the user interaction which resulted in the edge creation. A research contribution is the exploration of various credit functions in Section 3.5.1 and Section 5.3.1 to further improve the estimation performance. The expert users identified using the Credit Graph model are able to outperform those selected by the user-user graph in producing the best answer as evaluated in Section 5.4.

Pairwise Comparison Approach

User expertise is a relative measure [93, 113] which can be used to compare the different level of contribution between different user interactions. Thus, comparison pairs can be built between user interactions which are then used for user expertise estimation. In a competition model, users compete to make the best contribution such as the best answer in a question thread discussed earlier in Section 6.2. The outcome from the pairwise comparison is then used to update the expertise of the users. For pairwise comparison approaches, the user expertise is indicated according to the user rating from rating systems such as ELO [41], TrueSkill [95] and Glicko [46].

This research improved on the pairwise comparison model by introducing the win-margin performance indicator in Section 6.3.3 as a replacement for the scoreless outcome-oriented approach. This technique is also adopted for Reddit as detailed in Section 8.5.2. Besides that, we incorporate rating periods to further improve the performance and increase the efficiency of estimation process. Our evaluation in Chapter 6 and Chapter 8 found the proposed competitive pairwise comparison approach to be the best expertise estimation approach. Moreover, this approach is efficient as an online approach without the need to retrain models; where new information can be added directly to further improve performance.

The Estimation Approaches are Robust

All of the user expertise estimation approaches were explored and adapted for a diverse set of UGC platforms. These approaches had no trouble in estimating user expertise though some approaches do significantly outperform the others especially the proposed Credit Graph model and Competitive Pairwise Comparison.

This is accomplished by studying and identifying suitable features in each UGC platform which can be a signal of user expertise while ensuring that these features can be easily adapted to the proposed approaches. The features include:

- **The temporal-ordering of user interactions.**
  The proposed Credit Graph model incorporated the temporal-ordering of user interactions for its credit function. In Section 3.7.7, we validate the Discoverer-Follower

concept and it lead to an improved performance on CT platforms. We attempted the same for CQA platforms in Section 5.3.1 which performed considerably well.

- **User votes.**
  User votes exhibit strong signals for the estimation of user expertise. These votes can be used directly as scores to indicate user expertise. Alternatively, they can be used as an indicator of performance – signalling whether the user contribution is significant or not. For example, in Section 8.6.5 the significance of a user comment in a Reddit thread was determined by comparing the number of votes received with the average number of votes for items in the thread (instead of the commonly used polarity measure). This indicator is then applied to many user expertise estimation approaches including the graph-based and pairwise comparison approaches.

Despite the diversity of content and user behaviours between subreddits that we studied in Section 7.5, our findings in Section 8.6.2 highlighted the robustness of the proposed Competitive Rating approach.

9.1.3 User Annotations Can Describe and Classify Content Better

By weighting user annotations according to the expertise of the annotators in addition to the commonly used term-frequency based description, WWW resources can be described in a more useful manner as we proposed in Section 3.6.4. The improved descriptions are then used to improve AR as we evaluated in Section 3.8.

9.1.4 It is Possible to Predict the Information Quality of User-Generated Content

The information quality of UGC can be inferred from the expertise of the users who contribute them. Users with higher expertise will contribute content of greater quality. In Chapter 5 and Chapter 6, this research used this relation to predict the best answer in question threads. Then in Chapter 8 instead, we were able to predict which user comment would gain more vote by comparing the relative user expertise of the commenters.

9.2 Future Work

Our research for each part in this thesis can be expanded further. Some of the extensions include:

9.2.1 User Annotations for Training Machine Learning

In Part I of the thesis, this research incorporate user expertise to improve annotation-based retrieval (AR) where the user annotations act as the descriptor for WWW resources. A possible extension from here on is to explore how this change in weight for user annotation could affect the training for machine learning algorithms such as the learning the content of an image from the annotations associated to it. With the popularity of image-based social platforms such as Instagram[1], there exists a large repository of valuable data for the machine learning.

9.2.2 Expertise-based Peer Moderation

Currently, peer-moderation on social platforms regard each user to be the same; and thus, weighting their votes equally. This approach to peer-moderation can be exploited and

[1] https://www.instagram.com/

gamed by malicious users which motivates changes to UGC platforms such as Reddit's vote fuzzing mechanism. Such change does not however completely overcome this challenge as Reddit is still plagued by down-vote bots.

Since we are able to estimate the expertise of users on UGC platforms, we are able to improve on the peer-moderation process by weighting user votes according to the expertise and reliability of the voters. Thus, reliable users can have more impact on the moderation of content while reducing the unwanted disruptions from malicious users. As a by-product, users may then be motivated to contribute good content in order to maintain their autonomy and influence on the platform.

9.2.3 Interactions between Threaded Comments

Direct comments on Reddit directly address the content or the thread starter; whereas indirect comments respond to the other comments in the same thread. Currently as done in Section 8.5.1, we do not differentiate the relation between different users if they are responding to each other in a threaded discussion. Likewise, it is the same for the other users who reply to the same comments. Hence, an extension to our proposed C-Rating approach is to study, explore and incorporate such relations. At the time of writing, work has been started to research this further.

9.2.4 Improved Expert Search

Currently, many expert search solutions retrieve the experts within each field on query. We propose that with the estimated expertise of users, it is possible to better match users during expert search. For example, if there is a new programmer seeking help, he/she should be assisted by someone slightly above his/her capabilities rather than a programming guru with highly advanced convoluted solutions beyond his/her understanding or capabilities.

9.2.5 Expertise of Crowdsource Workers

There are plans to expand our research to other Web 2.0 platforms such as to crowdsourcing services. Given the capabilities to estimate a user's domain sensitive expertise, are we able to identify suitable experts for a given task? If so, how can we reward such users? Besides, a model can be proposed to aggregate the jobs done by different crowdsource workers together for improved results.

9.3 Conclusion

This research has achieved its goal of developing content-agnostic mechanism for bringing order to three distinct user-generated content (UGC) platforms in this thesis. The UGCs have the information potential to improve information systems – (1) user annotations as WWW resource descriptors; (2) user answers for questions as natural language queries; and (3) enhancing available content through comments and discussions.

Despite the challenges in the unstructured nature and high volume of UGC, we were able to infer the information quality of these contents by inferring this from the estimated expertise of the content contributors. The user expertise is estimated through content-agnostic approaches which is efficient and shown to be effective in various tasks. Expert users can be identified and encourage to contribute more meaningful content; whereas unwanted malicious users can be weeded out and have their influence reduced. This leads to improved content management on UGC platforms that can better meet the information needs of the users.

References

[1] Scott A. Golder and Bernardo Huberman. The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32, 2005.

[2] Fabian Abel, Matteo Baldoni, Cristina Baroglio, Nicola Henze, Daniel Krause, and Viviana Patti. Context-based Ranking in Folksonomies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 209–218. ACM, 2009.

[3] Fabian Abel, Nicola Henze, and Daniel Krause. Ranking in Folksonomy Systems: Can Context Help? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1429–1430. ACM, 2008.

[4] Mark S. Ackerman, Volker Wulf, and Volkmar Pipek. *Sharing Expertise: Beyond Knowledge Management*. MIT Press, 2002.

[5] Sibel Adali, Eis Hill, and Malik Magdon-ismail. Information vs. Robustness in Rank Aggregation: Models, Algorithms and a Statistical Framework for Evaluation. *Journal of Digital Information Management (JDIM), special issue on Web information retrieval*, 5, 2007.

[6] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding High-quality Content in Social Media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194. ACM, 2008.

[7] Abdul Manan Ahmad and Mohd. Hanafi Ahmad Hijazi. *Web Page Recommendation Model for Web Personalization*, pages 587–593. Springer Berlin Heidelberg, 2004.

[8] Einat Amitay, David Carmel, Nadav Har'El, Shila Ofek-Koifman, Aya Soffer, Sivan Yogev, and Nadav Golbandi. Social Search and Discovery Using a Unified Approach. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 199–208. ACM, 2009.

[9] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '12, pages 850–858. ACM, 2012.

[10] Çiğdem Aslay, Neil O'Hare, Luca Maria Aiello, and Alejandro Jaimes. Competition-based Networks for Expert Finding. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1033–1036. ACM, 2013.

[11] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth ACM*

International Conference on Web Search and Data Mining, WSDM '11, pages 65–74. ACM, 2011.

[12] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The Role of Social Networks in Information Diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 519–528. ACM, 2012.

[13] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing Web Search Using Social Annotations. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 501–510. ACM, 2007.

[14] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper Recommender Systems: A Literature Survey. *Int. J. Digit. Libr.*, 17(4):305–338, 2016.

[15] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.

[16] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 51–60. ACM, 2009.

[17] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can All Tags Be Used for Search? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 193–202. ACM, 2008.

[18] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. Identifying Authoritative Actors in Question-answering Forums: The Case of Yahoo! Answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '08, pages 866–874. ACM, 2008.

[19] Danah m. Boyd and Nicole B. Ellison. Social Network Sites: Definition, History, and Scholarship. *J. Comp.-Med. Commun.*, 13(1):210–230, 2007.

[20] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[21] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph Structure in the Web. In *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, pages 309–320. North-Holland Publishing Co., 2000.

[22] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Learning the Latent Topics for Question Retrieval in Community QA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011.

[23] Yuanzhe Cai and Sharma Chakravarthy. Expertise Ranking of Users in QA Community. In *Database Systems for Advanced Applications: 18th International Conference*, pages 25–40. Springer Berlin Heidelberg, 2013.

[24] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. Expertise Identification Using Email Communications. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 528–531. ACM, 2003.

[25] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011. ACM, 2011.

[26] Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 201–210. ACM, 2010.

[27] Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Quan Yuan. Approaches to Exploring Category Information for Question Retrieval in Community Question-Answer Archives. *ACM Trans. Inf. Syst.*, 30(2):7:1–7:38, 2012.

[28] Mark J. Carman, Mark Baillie, Robert Gwadera, and Fabio Crestani. A Statistical Comparison of Tag and Query Logs. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 123–130. ACM, 2009.

[29] Xuan Chen and John Heidemann. Flash Crowd Mitigation via Adaptive Admission Control Based on Application-level Observations. *ACM Trans. Internet Technol.*, 5(3):532–569, 2005.

[30] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can Cascades Be Predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936. ACM, 2014.

[31] Paul-Alexandru Chirita, Daniel Olmedilla, and Wolfgang Nejdl. PROS: A Personalized Ranking Platform for Web Search. In *Adaptive Hypermedia and Adaptive Web-Based Systems: Third International Conference, AH*, pages 34–43. Springer Berlin Heidelberg, 2004.

[32] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding Question-answer Pairs from Online Forums. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '08, pages 467–474. ACM, 2008.

[33] Scott Counts and Kristie Fisher. Taking It All In? Visual Attention in Microblog Consumption. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2011.

[34] Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC 2005 Enterprise Track. In *The Fourteenth Text REtrieval Conf. Proc. (TREC)*, volume Special Publication 500-266. National Institute of Standards and Technology (NIST), 2005.

[35] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An Experimental Comparison of Click Position-bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94. ACM, 2008.

[36] Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem. Learning User Interests for a Session-based Personalized Search. In *Proceedings of the Second International Symposium on Information Interaction in Context*, IIiX '08, pages 57–64. ACM, 2008.

[37] Hongbo Deng, Jiawei Han, Michael R. Lyu, and Irwin King. Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In *Proceedings of the*

12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, pages 71–80. ACM, 2012.

[38] Byron Dom, Iris Eiron, Alex Cozzi, and Yi Zhang. Graph-based Ranking Algorithms for e-Mail Expertise Analysis. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, pages 42–48. ACM, 2003.

[39] Jean Dunn and Olive Jean Dunn. Multiple Comparisons Among Means. *American Statistical Association*, pages 52–64, 1961.

[40] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622. ACM, 2001.

[41] Arpad E. Elo. *The Rating of Chessplayers, Past and Present.* Arco Pub., 1978.

[42] Jeremy Elson and Jon Howell. Handling Flash Crowds from Your Garage. In *USENIX 2008 Annual Technical Conference*, ATC'08, pages 171–184. USENIX Association, 2008.

[43] Flavio Figueiredo, Fabiano Belém, Henrique Pinto, Jussara Almeida, Marcos Gonçalves, David Fernandes, Edleno Moura, and Marco Cristo. Evidence of Quality of Textual Features on the Web 2.0. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 909–918. ACM, 2009.

[44] Adrien Friggeri, Lada A. Adamic, Dean Eckles, and Justin Cheng. Rumor Cascades. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*, pages 101–110. Association for the Advancement of Artificial Intelligence, 2014.

[45] Eric Gilbert. Widespread Underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 803–808. ACM, 2013.

[46] Mark Glickman. Dynamic Paired Comparison Models with Stochastic Variances. *Journal of Applied Statistics*, 28(6):673–689, 2001.

[47] Mark E. Glickman. A Comprehensive Guide to Chess Ratings. *American Chess Journal*, 3:59–102, 1995.

[48] Mark E. Glickman. Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Applied Statistics*, 48:377–394, 1999.

[49] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The Structure of Online Diffusion Networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 623–638. ACM, 2012.

[50] Scott A. Golder and Bernardo A. Huberman. Usage Patterns of Collaborative Tagging Systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[51] Thore Graepel and Ralf Herbrich. Ranking and Matchmaking. *Game Developer Magazine*, 2006.

[52] Joshua Guberman, Carol Schmitz, and Libby Hemphill. Quantifying Toxicity and Verbal Violence on Twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW '16 Companion, pages 277–280. ACM, 2016.

[53] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. WTF: The Who to Follow Service at Twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 505–514. International World Wide Web Conferences Steering Committee, 2013.

[54] Zoltan Gyongyi, Georgia Koutrika, Jan Pedersen, and Hector Garcia-Molina. Questioning Yahoo! Answers. Technical Report 2007-35, Stanford InfoLab, 2007.

[55] Morgan Harvey, Ian Ruthven, and Mark J. Carman. Improving Social Bookmark Search Using Personalised Latent Variable Language Models. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 485–494. ACM, 2011.

[56] Taher H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 517–526. ACM, 2002.

[57] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, 2007.

[58] Ralf Herbrich, Tom Minka, and Thore Graepel. Determining Relative Skills of Players, 2009.

[59] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

[60] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can Social Bookmarking Improve Web Search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 195–206. ACM, 2008.

[61] Nathan O. Hodas and Kristina Lerman. The Simple Rules of Social Contagion. *Scientific Reports*, 4:4343, 2014.

[62] Nathan Oken Hodas and Kristina Lerman. How Visibility and Divided Attention Constrain Social Contagion. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 249–257. IEEE Computer Society, 2012.

[63] Tad Hogg and Kristina Lerman. Social Dynamics of Digg. *EPJ Data Science*, 1(1):5, 2012.

[64] Tad Hogg, Kristina Lerman, and Laura M. Smith. Stochastic Models Predict User Behavior in Social Media. *CoRR*, abs/1308.2705, 2013.

[65] Andreas Hotho, Robert Jschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A Social Bookmark and Publication Sharing System. In *Proceedings of the First Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102. Aalborg Universitetsforlag, 2006.

[66] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised Sentiment Analysis with Emotional Signals. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

[67] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 84–90. ACM, 2005.

[68] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. A Framework to Predict the Quality of Answers with Non-textual Features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 228–235. ACM, 2006.

[69] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer Topic Model for Question Retrieval in Community Question Answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2471–2474. ACM, 2012.

[70] J. Jiao, J. Yan, H. Zhao, and W. Fan. ExpertRank: An Expert User Ranking Algorithm in Online Communities. In *2009 International Conference on New Trends in Information and Service Science*, pages 674–679, 2009.

[71] Karen Sparck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[72] Audun Jøsang, Roslan Ismail, and Colin Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decis. Support Syst.*, 43(2):618–644, 2007.

[73] Jeon-Hyung Kang and Kristina Lerman. VIP: Incorporating Human Cognitive Biases in a Probabilistic Model of Retweeting. In *Social Computing, Behavioral-Cultural Modeling, and Prediction: 8th International Conference*, pages 101–110. Springer International Publishing, 2015.

[74] Jeon-Hyung Kang, Kristina Lerman, and Lise Getoor. LA-LDA: A Limited Attention Topic Model for Social Recommendation. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 211–220. Springer-Verlag, 2013.

[75] Simon Kassing, Jasper Oosterman, Alessandro Bozzon, and Geert-Jan Houben. Locating Domain-specific Contents and Experts on Social Bookmarking Communities. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 747–752. ACM, 2015.

[76] Leo Katz. A New Status Index Derived from Sociometric Analysis. *Psychometrika*, 18(1):39–43, 1953.

[77] Heung-Nam Kim, Inay Ha, Jin-Guk Jung, and Geun-Sik Jo. User Preference Modeling from Positive Contents for Personalized Recommendation. In *Proceedings of the 10th International Conference on Discovery Science*, DS'07, pages 116–126. Springer-Verlag, 2007.

[78] Hyoung R. Kim and Philip K. Chan. Learning Implicit User Interest Hierarchy for Context in Personalization. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 101–108. ACM, 2003.

[79] Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 192–202, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[80] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, 46(5):604–632, 1999.

[81] Pavneet Singh Kochhar. Mining Testing Questions on Stack Overflow. In *Proceedings of the 5th International Workshop on Software Mining*, SoftwareMining 2016, pages 32–38. ACM, 2016.

[82] Peter Kollock. *The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace*, chapter 9, pages 220–239. Routledge, 1999.

[83] Yehuda Koren. Collaborative Filtering with Temporal Dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '09, pages 447–456. ACM, 2009.

[84] Karim R Lakhani and Eric von Hippel. How Open Source Software Works: Free User-to-User Assistance. *Research Policy*, 32(6):923 – 943, 2003.

[85] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 311–320. The AAAI Press, 2013.

[86] Renaud Lambiotte and Marcel Ausloos. Collaborative Tagging As a Tripartite Network. In *Proceedings of the 6th International Conference on Computational Science - Volume Part III*, ICCS'06, pages 1114–1117. Springer-Verlag, 2006.

[87] Cliff Lampe and Paul Resnick. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 543–550. ACM, 2004.

[88] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic Identification of User Goals in Web Search. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 391–400. ACM, 2005.

[89] Kristina Lerman and Aram Galstyan. Analysis of Social Voting Patterns on Digg. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 7–12. ACM, 2008.

[90] Kristina Lerman and Tad Hogg. Using a Model of Social Dynamics to Predict Popularity of News. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 621–630. ACM, 2010.

[91] Kristina Lerman and Tad Hogg. Using Stochastic Models to Describe and Predict Social Dynamics of Web Users. *ACM Trans. Intell. Syst. Technol.*, 3(4):62:1–62:33, 2012.

[92] Kristina Lerman and Tad Hogg. Leveraging Position Bias to Improve Peer Recommendation. *PLoS ONE*, 9(6):e98914+, 2014.

[93] Baichuan Li and Irwin King. Routing Questions to Appropriate Answerers in Community Question Answering Services. In *Proceedings of the 19th ACM International Conference on Information & Knowledge Management*, CIKM '10, pages 1585–1588. ACM, 2010.

[94] Longzhuang Li, Yi Shang, and Wei Zhang. Improvement of HITS-based Algorithms on Web Documents. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 527–535. ACM, 2002.

[95] Jing Liu, Young-In Song, and Chin-Yew Lin. Competition-based User Expertise Score Estimation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 425–434. ACM, 2011.

[96] Yandong Liu, Jiang Bian, and Eugene Agichtein. Predicting Information Seeker Satisfaction in Community Question Answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 483–490. ACM, 2008.

[97] C. m. A. Yeung, M. G. Noll, C. Meinel, N. Gibbins, and N. Shadbolt. Measuring Expertise in Online Communities. *IEEE Intelligent Systems*, 26(1):26–32, 2011.

[98] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[99] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, HYPERTEXT '06, pages 31–40. ACM, 2006.

[100] Cameron A Marlow. Linking without Thinking: Weblogs, Readership and Online Social Capital Formation. In *Proceedings of the 56th Annual Conference of the International Communication Association*, pages 6–7. International Communication Association, 2006.

[101] Julian John McAuley and Jure Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908. ACM, 2013.

[102] Emily Merritt. *An Analysis of the Discourse of Internet Trolling: A Case Study of Reddit.com*. PhD thesis, Mount Holyoke College, 2012.

[103] Elke Michlmayr and Steve Cayzer. Learning User Profiles from Tagging Data and Leveraging Them for Personal(ized) Information Access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)*, pages 1–7, 2007.

[104] Elke Michlmayr, Steve Cayzer, and Paul Shabajee. Add-A-Tag: Learning Adaptive User Profiles from Bookmark Collections. In *International Conference on weblogs and social media*, 2007.

[105] Aleksandra Klasnja Milicevic, Alexandros Nanopoulos, and Mirjana Ivanovic. Social Tagging in Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *Artif. Intell. Rev.*, 33(3):187–209, 2010.

[106] Richard Mills. Researching Social News - Is reddit.com a mouthpiece for the 'Hive Mind', or a Collective Intelligence approach to Information Overload? In *Proceedings of the ETHICOMP 2011*. Sheffield Hallam University, 2011.

[107] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[108] Maurice D. Mulvenna, Sarabjot S. Anand, and Alex G. Büchner. Personalization on the Net Using Web Mining: Introduction. *Commun. ACM*, 43(8):122–125, 2000.

[109] Michael G. Noll, Ching-man Au Yeung, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. Telling Experts from Spammers: Expertise Ranking in Folksonomies. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 612–619. ACM, 2009.

[110] Blair Nonnecke and Jenny Preece. Lurker Demographics: Counting the Silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 73–80. ACM, 2000.

[111] Scott Novotney and Chris Callison-Burch. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-expert Transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 207–215. Association for Computational Linguistics, 2010.

[112] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.

[113] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Trans. Inf. Syst.*, 30(2):10:1–10:28, 2012.

[114] K. Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242, 1895.

[115] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou, and Constantine D. Spyropoulos. Web Usage Mining As a Tool for Personalization: A Survey. *User Modeling and User-Adapted Interaction*, 13(4):311–372, 2003.

[116] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. Using Early View Patterns to Predict the Popularity of Youtube Videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 365–374. ACM, 2013.

[117] Nathaniel Poor. Mechanisms of an Online Public Sphere: The Website Slashdot. *Journal of Computer-Mediated Communication*, 10(2):00–00, 2005.

[118] Maria Priestley and Alex Mesoudi. Do Online Voting Patterns Reflect Evolved Features of Human Cognition? An Exploratory Empirical Investigation. *PLoS ONE*, 10(6):e0129703, 2015.

[119] Joni Radelaar, Aart-Jan Boor, Damir Vandic, Jan-Willem Van Dam, and Flavius Fasincar. Improving search and exploration in tag spaces using automated tag clustering. *J. Web Eng.*, 13(3-4):277–301, July 2014.

[120] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC3. In *Overview of the Third Text REtrieval Conference (TREC3)*, page 109126. Gaithersburg, MD: NIST, 1995.

[121] K. Roebuck. *Data Quality: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors.* Lightning Source, 2011.

[122] Filipe Roseiro Cgo, Srgio Roberto Pereira da Silva, and Roberto Pereira. AuthorityRank: Cognitive Authority and Information Retrieval in the Web. In *Proceedings of the IADIS International Conference on WWW/Internet*, 2012.

[123] Manuel Rubio-Sánchez, Jaime Urquiza-Fuentes, and Cristóbal Pareja-Flores. A Gentle Introduction to Mutual Recursion. *SIGCSE Bull.*, 40(3):235–239, 2008.

[124] Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin. Using Graded-relevance Metrics for Evaluating Community QA Answer Selection. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 187–196. ACM, 2011.

[125] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, 1975.

[126] Nikos Sarkas, Gautam Das, and Nick Koudas. Improved Search for Socially Annotated Data. *Proc. VLDB Endow.*, 2(1):778–789, 2009.

[127] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and Metrics for Cold-start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260. ACM, 2002.

[128] Chirag Shah and Jefferey Pomerantz. Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 411–418. ACM, 2010.

[129] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. *J. Mach. Learn. Res.*, 17(1):2049–2095, 2016.

[130] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 517–522. International World Wide Web Conferences Steering Committee, 2014.

[131] Amit Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43, 2001.

[132] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263. Association for Computational Linguistics, 2008.

[133] Wolfgang G. Stock. Folksonomies and Science Communication: A Mash-up of Professional Science Databases and Web 2.0 Services. *Inf. Serv. Use*, 27(3):97–103, 2007.

[134] Greg Stoddard. Popularity and Quality in Social News Aggregators: A Study of Reddit and Hacker News. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 815–818. ACM, 2015.

[135] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[136] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. Quality-aware Collaborative Question Answering: Methods and Evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 142–151. ACM, 2009.

[137] Olivia L. Sylvester. Read Less, Know More?: The Effect of News Aggregators on Quality Journalism. *CMC Senior Theses*, 1(604), 2013.

[138] Gabor Szabo and Bernardo A. Huberman. Predicting the Popularity of Online Content. *Commun. ACM*, 53(8):80–88, 2010.

[139] Martin N. Szomszor, Iván Cantador, and Harith Alani. Correlating User Profiles from Multiple Folksonomies. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 33–42. ACM, 2008.

[140] Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. A Survey on Predicting the Popularity of Web Content. *Journal of Internet Services and Applications*, 5(1):8, 2014.

[141] Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, and Subbaraj Shakthikumar. Sentiment Analysis of Movie Reviews on Discussion Boards Using a Linguistic Approach. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 81–84. ACM, 2009.

[142] Chi-Yao Tseng and Ming-Syan Chen. Significant Node Identification in Social Networks. In *Proceedings of the 15th International Conference on New Frontiers in Applied Data Mining*, PAKDD'11, pages 459–470. Springer-Verlag, 2012.

[143] Liwen Vaughan and Mike Thelwall. Search Engine Coverage Bias: Evidence and Possible Causes. *Inf. Process. Manage.*, 40(4):693–707, 2004.

[144] Jian Wang and Brian D. Davison. Explorations in Tag Suggestion and Query Expansion. In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, SSM '08, pages 43–50. ACM, 2008.

[145] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality Prediction and Community Structure in Social Networks. *Scientific reports*, 3(2522), 2013.

[146] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 579–583. ACM, 2013.

[147] Robert Wetzker, Carsten Zimmermann, Christian Bauckhage, and Sahin Albayrak. I Tag, You Tag: Translating Tags for Advanced User Models. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 71–80. ACM, 2010.

[148] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[149] Fang Wu and Bernardo A. Huberman. Novelty and Collective Attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[150] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. Temporal Recommendation on Graphs via Long- and Short-term Preference Fusion. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '10, pages 723–732. ACM, 2010.

[151] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring Folksonomy for Personalized Search. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 155–162. ACM, 2008.

[152] Yabo Xu, Ke Wang, Benyu Zhang, and Zheng Chen. Privacy-enhancing Personalized Web Search. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 591–600. ACM, 2007.

[153] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval Models for Question and Answer Archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 475–482. ACM, 2008.

[154] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. CQArank: Jointly Model Topics and Expertise in Community Question Answering. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 99–108. ACM, 2013.

[155] Junjie Yao, Bin Cui, Qiaosha Han, Ce Zhang, and Yanhong Zhou. Modeling User Expertise in Folksonomies by Fusing Multi-type Features. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications - Volume Part I*, DASFAA'11, pages 53–67. Springer-Verlag, 2011.

[156] Reyyan Yeniterzi and Jamie Callan. Analyzing Bias in CQA-based Expert Finding Test Sets. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 967–970. ACM, 2014.

[157] Chengxiang Zhai and John Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[158] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 221–230. ACM, 2007.

[159] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. Question Retrieval with High Quality Answers in Community Question Answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 371–380. ACM, 2014.

[160] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding Bag-of-words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[161] Zhou Zhao, James Cheng, Furu Wei, Ming Zhou, Wilfred Ng, and Yingjun Wu. SocialTransfer: Transferring Social Knowledge for Cold-Start Cowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 779–788. ACM, 2014.

[162] Guangyou Zhou, Siwei Lai, Kang Liu, and Jun Zhao. Topic-sensitive Probabilistic Model for Expert Finding in Question Answer Communities. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1662–1666. ACM, 2012.

[163] Zhi-Min Zhou, Man Lan, Zheng-Yu Niu, and Yue Lu. Exploiting User Profile Information for Answer Ranking in cQA. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 767–774. ACM, 2012.

[164] Yangbo Zhu, Shaozhi Ye, and Xing Li. Distributed PageRank Computation Based on Iterative Aggregation-disaggregation Methods. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 578–585. ACM, 2005.