Graph based techniques for community expert ranking in online discussion forums

PHD Thesis

By
Ch. Muhammad Shahzad Faisal
67-FBAS/PHDCS/F11

Supervised By
Asst. Prof. Dr. Ali Daud
Dept. of CS&SE, IIU, Islamabad

Department of Computer Science & Software Engineering
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad, Pakistan
2016

This dissertation is submitted to
International Islamic University Islamabad, Pakistan
In partial fulfillment of the requirement of the degree of
Doctor of Philosophy (Computer Science)

Acknowledgement

I would like to acknowledge many people who have supported, guided and advised me during my PhD research work. First, I would like to express my sincere gratitude to my supervisor Dr. Ali Daud, Assistant Professor, Department of Computer Science & Software Engineering, International Islamic University, Islamabad. Through the course of my studies, I have had the great fortune to get to know and interact with him. Their comments during writing PhD thesis are valuable to me. I am also thankful to Prof. Dr. Muhammad Sher, Dean, Faculty of Basic and Applied Sciences, for his consistent guidance. I would like to extend thanks to my friends Dr. Khalid Iqbal, Dr. Irfan, Dr. Saqib Iqbal, Dr. Hikmatullah khan, Mehmood-ul-Hassan, Shahid Iqbal and Abu Bakr Usman Akram for their valuable comments and support. Last but not the least; I would like to thank all my beloved family members for their support and sparing me for the studies.

Declaration

I hereby declare and affirm that this thesis neither as a whole nor as part thereof has been copied out from any source. It is further declared that I have completed this thesis entirely on the basis of my personal effort, made under the sincere guidance of my supervisor. If any part of this report is proven to be copied or found to be a reproduction of some other, I shall stand by the consequences. No portion of the work presented in this report has been submitted in support of an application for other degree or qualification of this or any other university or institute of learning.

Ch. Muhammad Shahzad Faisal
67-FBAS/PHDCS/F11

Dedication

This work is dedicated to my beloved Sheikh Hazrat-e-Allama Pir Alaudin Siddiqi sahib, my parents and family who always prayed, motivated, encouraged and guided me to achieve this milestone.

Ch. Muhammad Shahzad Faisal

INTERNATIONAL ISLAMIC UNIVERSITY ISLAMABAD
FACULTY OF BASIC & APPLIED SCIENCES
DEPARTMENT OF COMPUTER SCIENCE & SOFTWARE ENGINEERING

Date: 18-1-2017

Final Approval

It is certified that we have read this thesis, entitled "**Graph Based Techniques for Community Expert Ranking in Online Discussion Forums**" submitted by **Ch. Muhammad Shahzad Faisal,** Registration No.**67-FBAS/PHDCS/F11**. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University Islamabad for the award of the degree of Doctor of Philosophy in Computer Science.

Committee

External Examiner:
Prof. Dr. Muhammad Afzal,
Director,
Dr. A. Q. Khan Institute of Computer Science and Information Technology,
Kahuta, Distt. Rawalpindi

18/1-17

External Examiner:
Dr. Waseem Shahzad,
Associate Professor,
Department of Computer Science
FAST, H-11, Islamabad.

18/01/2017

Internal Examiner
Dr. Syed Husnain Abbas Naqvi,
Assistant Professor /Chairman,
Department of Computer Science & Software Engineering
FBAS, IIUI

Supervisor (on Leave)
Dr. Ali Daud
Assistant Professor,
Department of Computer Science & Software Engineering
FBAS, IIUI

Chairman:
Dr. Syed Husnain Abbas Naqvi,
Chairman,
Department of Computer Science & Software Engineering
FBAS, IIUI

Dean:
Prof. Dr. Muhammad Sher,
Dean,
Faculty of Basic & Applied Sciences,
International Islamic University Islamabad

Abstract

Social web or Web 2.0 has gain popularity since last decade due to its valuable services such as social networking, blogs, online forums, through which users can easily produce and consume information. Online discussion forums are an emerging service of social web, provide an excellent opportunity for knowledge exchange and sharing of ideas. In online forums, collaborations occur when questioning-answering take place among online forum members. Expert finding in online discussion forums, such as BBC, StackOverflow, is a specialized problem of information retrieval. Previously, the expert finding approaches in online forums were based on content and link based features. The link based expert ranking techniques are based on users' social network authority and can be measured through link analysis techniques such as PageRank and HITS. Content based techniques utilize the answers content to measure user's reputation or expertise. Posts contents quality can be measured through textual and non-textual features. Textual similarity is measured through standard similarity techniques such as cosine and semantic similarity. Non-textual features include post length, position, references and sentiments etc. Users expertise are measured through their self-reputation scores, however, users performance is not evaluated on the basis of their neighbors' or co-existing participants' reputation scores. Moreover, important features such as user's activity, participation strength, discussion quality and consistent performance have not been utilized for expert finding problem. Thread ranking is another specialized problem of information retrieval in online discussion forums with the aim of finding relevant and quality threads for a given query. Thread ranking problem is addressed through structure and content-similarity features, however features such as semantic similarity, participants' reputation and thread structure have not been utilized.

In this research work, we propose improved expert finding techniques for both rated and non-rated discussion forums such as BBC and StackOverflow. In case of non-rated forums like BBC, we measure the users' expertise through their co-existing users' reputation. Users who answer together in multiple threads are termed as Co-existing users. For expert finding in rated-forums like StackOverflow, our techniques consider the element of consistent performance of a user. Reputation features are derived from StackOverflow dataset which are based on voter reputation, vote ratio and tags popularity. We have validated our both expert ranking techniques (for rated and non-rated forums) against a link based expert finding technique and achieved quality results. Lastly, we have addressed the thread ranking problem in BBC forums. Threads quality has been measured through structure, content quality and participant reputation. Experiments on BBC forum dataset show that our thread ranking technique outperforms the baseline technique.

Achievements and Contribution

Journal Publications

1. Ch. Muhammad Shahzad Faisal, Ali Daud, Abubakr Usman, "Expert ranking using reputation and answer quality of co-existing users", *International Arab Journal of Information Technology,* vol.14 no.2 (Impact factor & ISI Thomson Reuters Indexed), 2017.

2. Ch. Muhammad Shahzad Faisal, Ali Daud, Faisal Imran, "A novel framework for social web forums' thread ranking based on semantics and post quality features", *The Journal of Supercomputing* (Impact factor & ISI Thomson Reuters Indexed), 2016

3. A.Irshad, M Sher, MS Faisal, A Ghani, M Ul Hassan, S Ashraf Ch, "A secure authentication scheme for session initiation protocol by using ECC on the basis of the Tang and Liu scheme", *Security and Communication Networks* (Impact factor & ISI Thomson Reuters Indexed) vol. 7, no. 8, pp:1210-1218, 2016.

4. Ch. Muhammad Shahzad Faisal, Ali Daud, Khalid Iqbal, "Impact of User reputation, social quality and temporal features on Movie Recommendation", Multimedia Tools and applications, (Under review).

5. Ch. Muhammad Shahzad Faisal, Ali Daud, Abubakr Usman akram, Khalid Iqbal, Teh Ying Wah, "Expert finding techniques for rated forums", WWW (Under review).

Conference Publications

6. Abubakr Usman Akram, Khalid Iqbal, Ch. Muhammad Shahzad Faisal, Umer Ishfaq, "An effective experts mining technique in online discussion forums", in proceedings of IEEE, ICE CUBE conference, Quetta, Pakistan, 2016.

7. Shumaila Mushtaq, Ch. Muhammad Shahzad Faisal and Khalid Iqbal, "Review Spam Detection using Sentiments and Novel Features", in proceedings of International Conference on Advanced Computer Theory and Engineering (ICACTE 2016), Hong Kong, 2016.

8. W Ahmad, CMS Faisal, "Context based image search", in proceedings of 14th International IEEE Multitopic Conference, pp.67-70, 2011.

Table of Contents

Table of Figures

Table of Tables

Table of Abbreviations

Abbreviations	Descriptions
5W1H	(who, what, where, when, why, how), a method of asking questions
Ans-Post	Answer post
ATT	Answer-type taxonomy
BOP	Bag of posts
BOW	Bag of words
CEF	Context based expert finding
Central	Central post of a thread
Cosine-sim	Cosine similarity
CQA	Community question answering sites
ER	ExpertiseRank
Exp-PC	Expert ranking based on performance consistency (g-index)
ExpRank-AQCS	An extension of ExpertiseRank with our proposed ExpRank-FB technique
ExpRank-COM	An extension of the ExpertiseRank with our proposed ExpRank-CRF technique
ExpRank-CRF	Expert ranking based on co-existing user reputation
ExpRank-FB	Expert ranking based on features like answer count, links, quotes and answer length etc.
Head-Post	Thread's title and its first Ans-Post in combination
HITS	Hypertext Induced topic search
IMDb	Internet Movie Database
LDA	Latent Dirichlet allocation
MAP	Mean average precision
MAX	Maximum semantic similarity score
MPAA	Motion Picture Association of America film rating system
MQP	Movie Quality prediction
O_{sim}	Overlap similarity
PR	PageRank
Rep-FS	Expert ranking based on user reputation (votes ratio, voter reputation, tag quality and participant's reputation)
SC	Sentiment classification
Sim	Content similarity between posts
SO	StackOverflow programming forum
SO-forum	StackOverflow programing forum

SVM	Support vector machine
T-CRRank	Post quality and participants' reputation based thread ranking
Thread-Rank$_{cq}$	Thread's content quality rank
Thread-Rank$_{pr}$	Thread's participant reputation rank
Top-k	Top-k similar posts
T-SimRank	Thread to post semantic similarity (Thread ranking)
VSM	Vector space model
W-Net	WordNet
WWW	World Wide Web

Table of Notations

Symbols	Sets	Description
U	$U= \{u_1, u_2, u_3.....u_n\}$	Set of users u ∈ U
T	$T= \{t_1, t_2, t_3......t_m\}$	Set of threads t ∈ T
P	$P=\{p_1,p_2,p_3........p_n\}$	Set of posts for a given thread p ∈ P
E	$E= \{e_1, e_2, e_3.....e_n\}$	Set of experts e ∈ E
R	$R= \{r_1, r_2, r_3.......r_n\}$	Set of repliers r ∈ R
Q	$Q= \{q_1, q_2, q_3 ....q_n\}$	Set of questions q ∈ Q
A	$A= \{a_1, a_2, a_3 ......a_n\}$	Set of answers a ∈ A
CE	$CE=\{ce_1.ce_2.t_1, ce_1.ce_2.t_2, ce_1.ce_2.t_3.....ce_n.ce_k.t_m\}$	Set of users who co-exist in different threads ce ∈ U
U_{CTi}		Thread support count for each user
Sim		Post content similarity
Sim_{cosine}		Cosine similarity between posts
Sim_{sem}		Semantic similarity between posts
UCT_k		thread count of user i
$Support$		The proportion of transactions in the database which contains the item-set X
$Confidence$		The proportion the transactions that contains X which also contains Y.
Sup		Support count
UR_i		Self-reputation of a user
CE_{rep}		Reputation of a co-existing user
PR		PageRank of a user
I		Incoming link
d		Damping factor
C		Number of outgoing links

$ER(X)$		ExpertiseRank of a user X
L		the total number of users who helped U
β		A threshold to estimate their answer relevance of a user by computing $Sim_{semantic}$ score of co-existing users C_e's post content
ψ		Support threshold
p		Spearman's rho correlation coefficient
tf-idf		Term frequency and Inverse document frequency
W		Weight of a particular term within a thread
Tf		Term frequency

Chapter-1
Introduction

Preliminary

The increasing popularity of social networking sites is witnessed by large number of users for short time such as Yahoo Answers! , Quora, Ask MetaFilter, Baidu Knows, StackOverflow, BBC discussion forums, IMDB and MovieLens. The growing accessibility of social web allows users to collaborate with each other in a flexible way. Due to the massive usage of social web applications by the public, the volume of accumulated content is increased to gigantic levels by raising questions on user authenticity and content credibility. The development of new social networking platforms increased with the induction of new features in order to attract and facilitate public. Several research efforts have been made to extract useful information from social web content such as mining crowd wisdom, knowledge extraction, mining user expertise and finding useful information. However, extracted knowledge is still insufficient from large volume of online information. Therefore, importance of research in social web domain is increased and demanding because the most of the organizations and public are now considering the usage of social web applications as an important activity for improving their lives and businesses.

Online discussion forums such as BBC and StackOverflow, becomes the tool of choice for finding useful information in several domains such as education, politics, religion, science, entertainment, sports etc. The essential feature of online forums is the asker-replier relationship between participants where users post their questions and domain experts answer them. Finding quality answers and domain experts in the forum is the primary objective of the user. Therefore information retrieval and expert discovery are the potential research problems in online forums. In this thesis, using BBC and StackOverflow forum dataset, thread's content quality, structure and users reputation using co-occurrence features, performance consistency and semantic similarity techniques have been analyzed.

More generally, this thesis, addresses three research problems namely, (1) In chapter 3, Expert ranking techniques for online non-rated forums such as BBC are described. (2) In chapter 4, Expert ranking techniques for online rated-forums such as StackOverflow are described, (3) In chapter 5, Thread ranking techniques in online forums are described. Introduction and significance of social web applications and particularly, conversational media are discussed below.

1.1. Social web applications

Conventional World Wide Web (WWW) is based on static pages, created by individuals and organizations, facilitates in publishing the information. User interactions hardly take place in such environment due to the lack of rich interaction services. In the beginning, scripting languages like html, java script and css did not support content sharing with little interaction support for users. Users' feedback options such as likes, votes and ratings features were also difficult to implement. Differently from conventional web, social web introduced sophisticated technologies to allow better user interaction. Social web considered social relations to link people through World Wide Web. Web 2.0 applications popularity has been increased due to their ease of use and user friendly features. Social web usage trends[1] can be seen due to its significance and popularity for the last decades. Facebook[2], Twitter[3] , Wikipedia, blogs and online discussion forums are prominent examples of Web 2.0 services for sharing views and exchange of ideas among friends by the users. Web 2.0 have improved the way of searching and getting information from internet where user can give feedback in form of sharing, liking, rating and reviewing items. In addition, opportunities for mobile devices to access user-generated data are opened for text, audio and videos. Several popular social web applications are shown in figure 1.1. Social Web applications can be classified into two categories and discussed in the subsequent sub-sections.

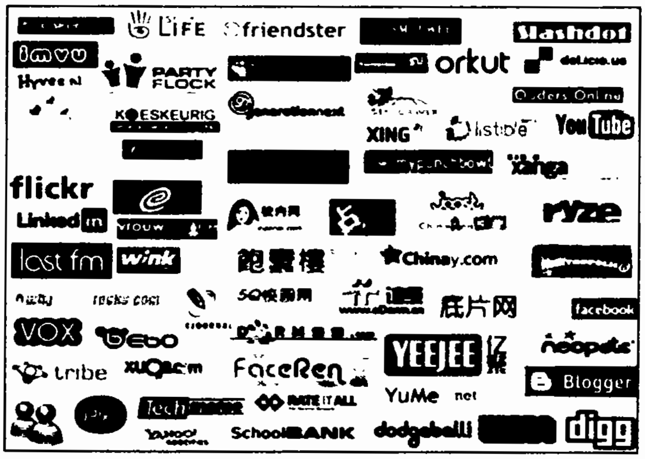


Figure 1.1 : Social Web applications (*sources: Lifewire, socialnetworking, Technorati*)

1.2. Conversational media

Blogs, Online discussion forums, wikis and community question answering sites are known services of conversational media, providing useful information services to users. Facebook provides a convenient way to chat whereas LinkedIn[4] provides a discussion platform for professionals. Micro-blogging services such as Twitter revolutionized the communication ways. In Twitter, users may follow to each other or to whom they influenced. In this way, information is propagated on large scale. Users may share image

[1] http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/
[2] https://www.facebook.com/
[3] http://www.twitter.com/
[4] https://www.linkedin.com

tweets along with text description. Twitter improved the information spread mechanism for dissemination of news in a short time. In conversational media, users may initiate conversation through number of channels. For example, community question answering sites like Yahoo answers provide a general discussion platform. Nevertheless, more specific discussions can be found in online forums. BBC discussion forum[5], Digital Spy[6], BestoftheWeb[7], NewsForum[8], MailOnline[9], Stackoverflow[10] and Ubuntu[11] are the well-known examples of online discussion forums. Details of online forums are provided in appendix A. These forums provide the easiest ways to obtain answers for difficult questions and allow users to interact with domain experts. Access to online forums is easier than conventional web because information is organized in static web pages. Online forums are of two type, rated and non-rated forums. Rated forums are those where best-answer selection mechanism or rating criteria is provided. For example, StackOverflow, Reddit, Yahoo! answers and Ubuntu forum are examples of rated-forums while BBC discussion forum[12]and Digital Spy[13] are examples of non-rated forums where rating criteria regarding user and answer is provided. In online forums, the organized information in categories can be easily browsed to find required question-answer pairs on several topics. There are several categories in an online forum such as world news, UK news, sports, religion and entertainment. Each category contains several threads. People can easily search answers in forums because discussions are grouped together and all replies/answers to a single question follow a strict hierarchal structure. The hierarchal structure helps in finding a chain of relevant discussions on any topic. A typical online discussion forum has a thread/post structure. Thread is an asked-question or a topic shared by a user and answer-posts/posts are a replies to that thread. A thread may consist of head post (initial post) along with number of answering posts as shown in figure 1.2 & 1.3. The common criterion of all of these forums is to judge users' expertise by their answers frequency. A user is assumed to be an expert if he answers frequently to multiple questions. In a study of Yahoo Answers, a small proportion of 25% users provided answers [1] while others asked questions. To conclude, online forums are better choice for straightforwardly getting detailed, productive and topic specific discussions which is not easy to find on conventional web. Therefore, presence of domain experts made online discussion forums more reliable. The best answer selection and expert finding is an essential requirement in rated and non-rated forums. In the following

[5]http://www.bbc.co.uk/blogs/internet

[6] http://forums.digitalspy.co.uk/

[7] http://botw.org.uk/News/Discussion_Forums/

[8] http://www.newsforum.com

[9] http://boards.dailymail.co.uk/news-board-moderated

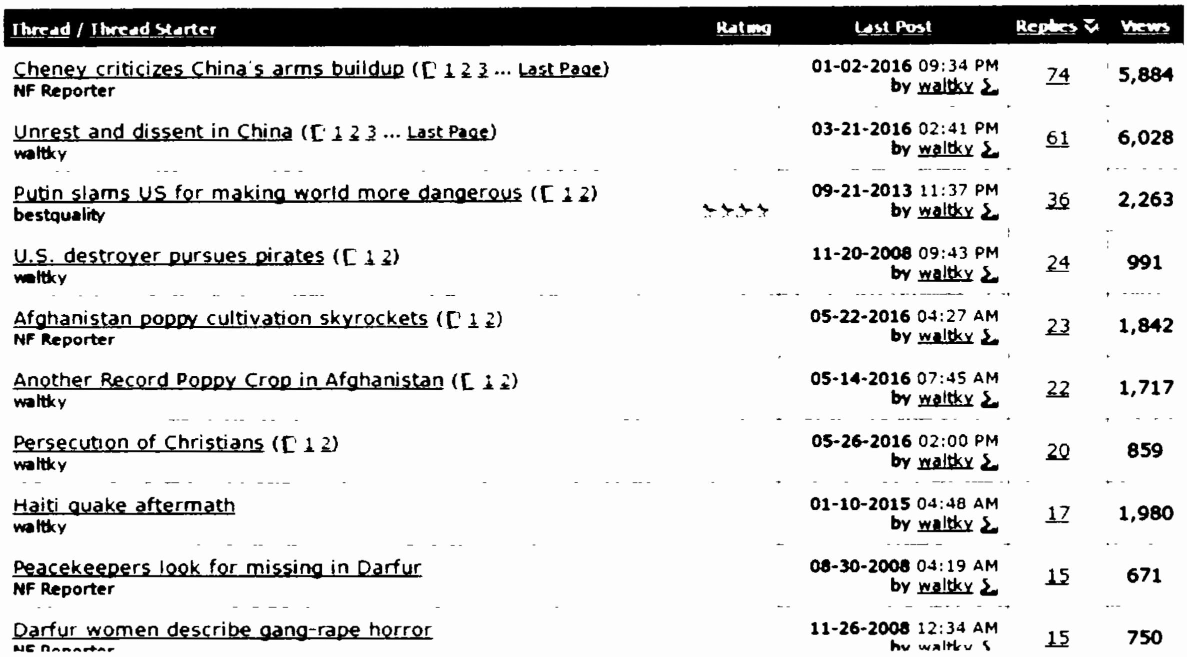
[10] http://stackoverflow.com/

[11] http://www.ubuntu.com/

[12]http://www.bbc.co.uk/blogs/internet

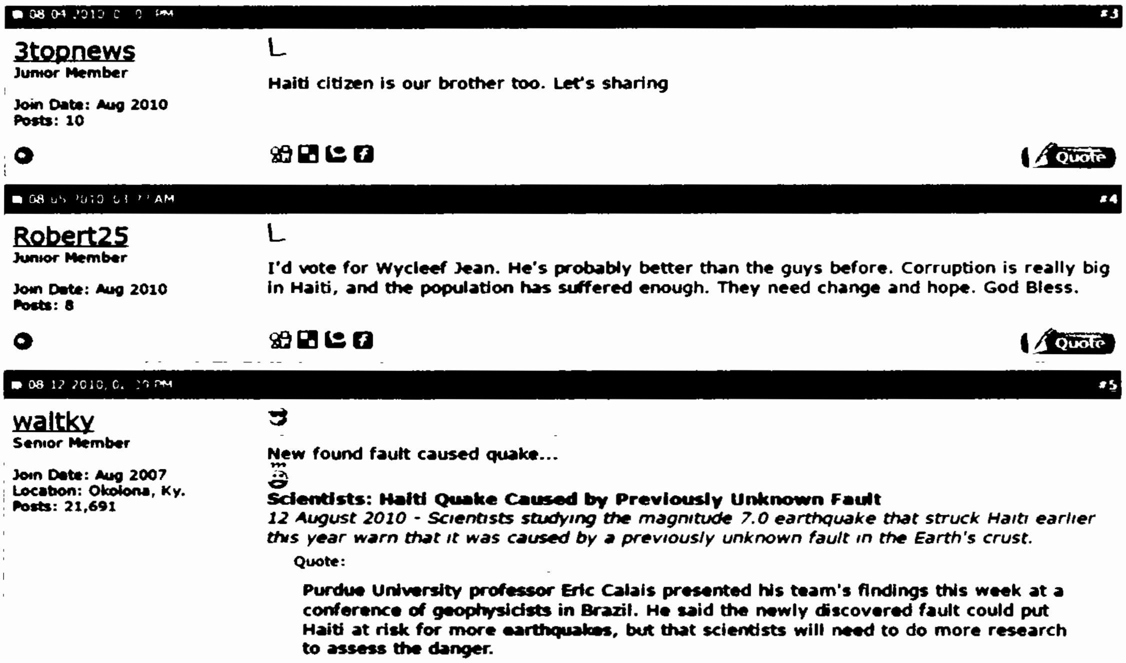
[13] http.//forums.digitalspy.co.uk/

subsections, we describe more particular aspects of expert ranking and thread ranking problems that we consider in the present thesis including the techniques that we use for each problem.



Thread / Thread Starter	Rating	Last Post	Replies	Views
Cheney criticizes China's arms buildup (1 2 3 ... Last Page) NF Reporter		01-02-2016 09:34 PM by waltky	74	5,884
Unrest and dissent in China (1 2 3 ... Last Page) waltky		03-21-2016 02:41 PM by waltky	61	6,028
Putin slams US for making world more dangerous (1 2) bestquality	★★★★	09-21-2013 11:37 PM by waltky	36	2,263
U.S. destroyer pursues pirates (1 2) waltky		11-20-2008 09:43 PM by waltky	24	991
Afghanistan poppy cultivation skyrockets (1 2) NF Reporter		05-22-2016 04:27 AM by waltky	23	1,842
Another Record Poppy Crop in Afghanistan (1 2) waltky		05-14-2016 07:45 AM by waltky	22	1,717
Persecution of Christians (1 2) waltky		05-26-2016 02:00 PM by waltky	20	859
Haiti quake aftermath waltky		01-10-2015 04:48 AM by waltky	17	1,980
Peacekeepers look for missing in Darfur NF Reporter		08-30-2008 04:19 AM by waltky	15	671
Darfur women describe gang-rape horror NF Reporter		11-26-2008 12:34 AM by waltky	15	750

Figure 1.2: Threads and posts in an online discussion forum (*Sources: phpBB, MacRumors*)



08 04 2010 [illegible] PM #3

3topnews
Junior Member

Join Date: Aug 2010
Posts: 10

Haiti citizen is our brother too. Let's sharing

08 05 2010 [illegible] AM #4

Robert25
Junior Member

Join Date: Aug 2010
Posts: 8

I'd vote for Wycleef Jean. He's probably better than the guys before. Corruption is really big in Haiti, and the population has suffered enough. They need change and hope. God Bless.

08 12 2010, [illegible] PM #5

waltky
Senior Member

Join Date: Aug 2007
Location: Okolona, Ky.
Posts: 21,691

New found fault caused quake...

Scientists: Haiti Quake Caused by Previously Unknown Fault
12 August 2010 - Scientists studying the magnitude 7.0 earthquake that struck Haiti earlier this year warn that it was caused by a previously unknown fault in the Earth's crust.

Quote:

> **Purdue University professor Eric Calais presented his team's findings this week at a conference of geophysicists in Brazil. He said the newly discovered fault could put Haiti at risk for more earthquakes, but that scientists will need to do more research to assess the danger.**

Figure 1.3: Forum's comments/Posts hierarchy (*Sources: phpBB, MacRumors*)

1.3. Expert ranking

Allowing users to have a list of real experts in online forums is an increasingly important research problem in social web mining since the current online forums and blogs do not possess effective mechanism for finding domain experts. Normally expert finding techniques consider user's social network authority score to measure their reputation [3, 20, 21]. In a social network, user authority score represent their influence or popularity which they gain through their social connections strength [22-24]. However, these techniques do not consider the answer quality of users, which is considered as a primary reputation measure [11, 12, 25]. Co-occurrence is a well-known concept in social web domain, provides the relationships between people, concepts and other entities [26-28]. Social web applications have been modeled and visualized through graphs in several domains such as expert finding, mining influential users and sentence similarity etc. Therefore, importance of co-occurrence concept has increased with the rapid growth of social web because it visualizes the relationships among entities in a better way. Co-occurrence phenomenon can be expressed in many ways, for example, terms A and B may be said to "co-occur" or "co-exist" if they both appear in a particular article. In another scenario of product reviews, two people Ahmed and Ali may be said to "co-occur" or "co-reviewers" if they both write reviews for a product. Co-occurrence phenomenon has been widely and effectively used in several social web domains such as review spam detection[29-32], review recommendation [33, 34], plagiarism detection[35-37], product recommendation[38, 39], URL clustering[40, 41], passage similarity computation[42, 43]and social tagging systems[44, 45]. Being motivated from successful applications of co-occurrence concepts in several domains, we have proposed the concept of "co-occurred users" or "co-existing users" in online forum. Co-existing users are those who answer/reply together in a thread. It is assumed that users who co-exists or discuss topics in groupings are more reputed than those who do not participate in groupings. We proposed that users' reputation can be better measure through their participation strength and co-existing users' reputation. Participation strength of a user indicates that user is actively participating in discussions. Moreover, presence of reputed co-existing users indicates the presence of real experts with quality discussions. Answer quality features are proposed because answer quality has been widely used to measure users' expertise in online forums [4, 7, 46-48]. We believe that our assumption is practically reasonable and consequently our techniques give promising results.

For rated-forums such as StackOverflow, features such as vote ratio, badge count and accepted answers count etc are used to measure user expertise. However, these features are not fully capable for finding real experts because they do not consider the element of consistent performance for the users. Moreover important features such as tags quality and voter reputation are not considered for measuring users' reputation. We argued that link-based expert finding techniques are not well suited for rated forums as they don't consider the answer quality, performance consistent and forum based features. We have adapted a popular bibliometric g-index for finding StackOverflow users who consistently perform well. Moreover, we have derive novel features from StackOverflow dataset which gives promising results.

1.4. Thread ranking

Thread ranking/retrieval is an application of information retrieval for online forums. Thread retrieval, especially automatically, is a challenging research problem as current online discussion forums and community question answering (CQA) sites do not provide effective procedures for finding quality threads. Classic thread retrieval is normally done by exploiting threads' content [11-14, 25, 49-52] and structure [17, 53] as thread content is considered as primary element in measuring thread's quality. A thread may have several posts/answers, thus post quality reflects overall thread quality. In our thesis, we have used semantic information of the thread's content for thread ranking. The approach of exploiting semantics in a thread's content is particularly motivated by an application of increasing importance, namely the discovery of useful and relevant threads for a given query in online discussion forums. Our approach is based on semantic similarity measure [49, 54], widely used in information retrieval for the study of computing overlap between query and documents [18, 55, 56]. Semantic overlap among terms indicates the relevance of document to a given query [57, 58]. More importantly, semantic techniques consider the meaning of terms and their relationships, which is not considered by conventional techniques such as simple overlap, Jaccard coefficient, dice coefficient and tf-idf [59-61].We have used WordNet database for computing semantic similarity scores between thread's content and query. WordNet is a lexical semantic database of English language. Different variations of WordNet are used to compute semantic similarity using approaches like node-based, edge-based and hybrid methods for similarity computation [19].In addition, we have enhanced ranking through post quality and participant reputation features. In order to evaluate the relevance of whole thread's content, semantic similarity of its substructures such as, thread title/question, initial post, and central post have been computed. Finally, by computing similarity between a query and thread's sub-structures contents, it is possible to define an effective thread ranking scheme. For thread ranking, we experimentally validate our approach and achieve promising results on BBC forum dataset.

1.5. Research goals and Contributions

To provide a viable solution, goals of ranking users according to their expertise and ranking threads according to their quality respectively are presented in this section by highlighting the limitations. The, major contributions in these domains are discussed on our suggested techniques such as expert ranking and thread ranking.

Firstly, we have discussed the techniuqes like expert ranking problem in non-rated online discussion forums such as BBC discussion forum. Online discussion forums provide the knowledge sharing facilities to online communities. Usage of online discussion forums has increased tremendously due to the variety of services and their ability of common users to ask question and provide answers. With the passage of time, these forums can accumulate huge contents. Several of these posted discussions may not contain quality contents and may reflect users' personal opinions about topic which contradicts with a relevant answer. These low quality discussions indicate the existence of the unprofessional users. In addition, there is no rating mechanism for users and answers in discussion forums like BBC complicates the task of expert finding. The expert finding in the online forums is an important research problem. Most of the existing expert finding techniques consider only user's social network authority as a parameter of evaluating the user expertise that are not much effective because they don't consider the quality of the answer provided by the user. From the literature study, it has been observed that users' expertise have not been measured through their participation frequency and their neighbors' reputation. In first module (chapter no. 3), our goal is to find real experts in non-rated online discussion forums. To achieve our goal, following research questions have been identified in literature [2-7].

1. What is an effective mechanism of finding real domain experts in non-rated online discussion forums?
2. How to measure user reputation according to participation frequency, content relevancy and Co-existing users' reputation in multiple threads?
3. How to predict the user reputation effectively through the quality and frequency of the answer in online forums?

The above research questions are addressed by expert ranking technique to achieve our goals. In this context, our major contributions are as below:

- Co-occurrence concepts is proposed and applied to identify reputed co-existing/co-participants in online discussion forums.
- A reputation measure computes a users' expertise based on their participation frequency and reputation of their co-existing users in multiple threads.
- Answer quality and user activity features are proposed.

- We have extended a link-based expert ranking algorithm [3] with our techniques to assess its performance.

We also present an expert ranking problem for rated-forums like StackOverflow forum, a useful medium to exchange programming problems. Users from beginner to expert levels, participate in multiple discussions. With the passage of time a huge amount of information is accumulated in the forum which improves forum quality. Moreover, existence of expert users made StackOverflow a reputed forum. StackOverflow provides a built-in incentivization mechanism for computing users' reputation score. The mechanism includes accumulated up-vote, accepted answer, and accepted suggested edit scores etc. However, the built-in incentivization mechanism misclassify very active users for knowledgeable ones, and misjudge activeness for expertise. StackOverflow measure users' expertise based on their up-votes, favorite count, accepted answers, and down-votes, but ignore the user's answer quality and consistency. In this module (chapter no.4), based on a bibliometric, g-index and a set of novel user reputation features, expert-ranking techniques are proposed, applied to a StackOverflow forum dataset. We measure user reputation and expertise according to their answer quality and the consistency in providing quality answers. Our goal is to find real experts in StackOverflow forum. To achieve our goal, following research questions have been identified in literature [8-14].

1. What is an effective mechanism of finding real experts in StackOverflow discussion forum?
2. How to measure user reputation according to consistent performance and answer quality in the StackOverflow forum?

The above research questions are addressed by expert ranking technique to achieve our goals. The important contributions of this work for programming problems in the selected CQA site, i.e., StackOverflow, are given below:

- Novel user reputation features are proposed for the StackOverflow dataset.
- Expert-ranking techniques Exp-PC, Weighted Exp-PC are proposed.
- Our proposed, baseline and feature-specific expert-ranking techniques are evaluated against StackOverflow's reputation score using the standard performance-ranking measures: O_{sim}, Spearman's Rank Correlation, and Kendall's Rank Correlation.

In, in last module (Chapter no. 5), we present a thread ranking problem for non-rated online forums. In online forums, users may share or exchange ideas by posting the contents in the form of questions and answers. With the increasing volume of online forums content, finding relevant information is not only a challenging task, but also knowledge management and quality assurance of online forum's content get an importance for further investigation. In most of the cases, online discussion forums offer search services based on keyword search. In the beginning, research was focused on improving the performance of thread retrieval using cosine-similarity. Cosine-similarity based techniques were only based on lexical overlap between documents rather than semantic similarity. Our goal is to analyze the impact of semantic similarity of content and user reputation features on thread ranking in online forum. To achieve our goal, following research questions have been identified in literature Lee, Yang [12], [13-19] according to the highlighted limitations.

- Does semantic similarity techniques better than cosine similarity techniques?
- Which of the thread's elements are important and contribute in measuring its overall quality?
- Are user reputation and post quality features helpful in finding relevant threads for a query?
- Can we achieve better thread ranking by combining semantic and cosine similarity techniques?

The above research questions are addressed by thread ranking technique to achieve our goals. In this context, our major contributions are as below:

- Computing semantic similarity (cosine similarity vs. WordNet-based similarity) and aggregating post similarity up to the thread level.
- Characterization of the involvement of thread features for an overall efficacy of the thread.

1.6. Thesis Organization

- In second chapter, we describe the related work regarding our three modules, expert ranking for non-rated forums like BBC discussion forums, expert ranking in rated-forums like StackOverflow forum and thread ranking.

- In third chapter, we explain our first module which are expert ranking techniques for non-rated forums such as BBC discussion forum. Proposed and baseline algorithms are explained. Dataset are described. Results are discussed in last section of this chapter.

- In fourth chapter, we explained our second module which are expert ranking techniques for rated-forums like StackOverflow forum. Proposed and baseline algorithms are explained. Dataset are described. Results are discussed in last section of this chapter.

- In fifth chapter, we explained our third module which are thread ranking techniques in online forums. Ranking techniques and the dataset are explained and results are discussed at the end of this chapter.
- In sixth chapter, we give conclusions and future works for all abovementioned three modules.

Chapter 2
Related Work

We have thoroughly investigated literature on expert ranking/finding and thread ranking. In this chapter, we describe related works on each of the aforesaid domains.

2.1. Expert ranking techniques

Based on literature review, expert ranking approaches for online discussion forums can be classified into link-based, content-based and Bibliometric based techniques as depicted in figure 2.1 with detailed discussion in the subsequent subsections.

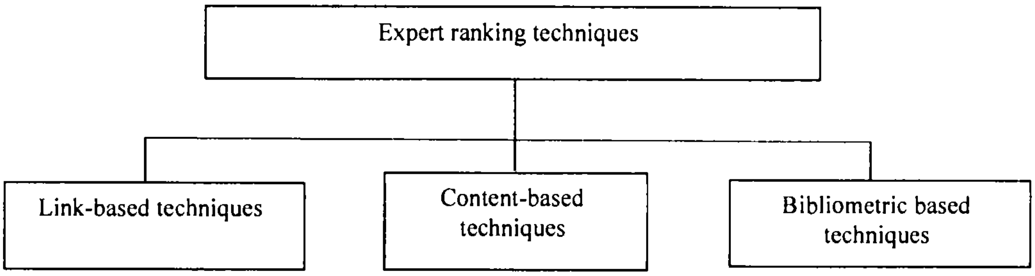


Figure 2.1 Classification of Expert ranking techniques

2.1.1. Link based expert ranking techniques

The link-based expert finding techniques focus on analysing the link structure among individuals rather than their document's content. Links analysis techniques were used for finding experts according to questions-answering relationships[3-5, 20-27], email communications[28, 29], citation or co-citation networks[30-33]. Given a specific expertise query, existing link-based expert finding techniques consist of three steps.

1. Identify query related documents (question or answers posts in a community question answering site),
2. Construct communication graphs such as asker-replier graphs (based on the connection among users, who post question and those who provide their answers), 3. Analyze the communication graphs to measure users' expertise score on the graph using link analysis algorithms.

Conventional web page ranking algorithms such as PageRank [34] and HITS [35] are commonly used algorithms for link analysis tasks in social networks. For example, PageRank [34] and HITs algorithm [35]have been used to measure user authority score [3-5, 29, 34, 36, 37] in question answering sties and citation networks. PageRank [34] counted the number and quality of links to a page to determine a rough estimate the importance of website. The assumption is that more important websites are likely to receive more links from other websites. HITS [35] identifies good authorities and hubs for a topic by assigning two numbers to a web page: an authority and a hub weight. These weights are defined recursively. A higher authority weight occurs if the page is pointed to by pages with high hub weights. A higher hub weight occurs if the page points to many pages with high authority weights. HITs algorithm[35] is used to find experts by formulating a graph structure in Yahoo Answers network [21]. They consider question-askers as hub and answer-providers as authorities. For each user, hub and authority score is computed. Users posted quality questions and received high hub scores. Similarly, users posted poor quality questions, received low hub scores. However, HITS only performed well for few categories and it only achieved performance on local graph structures.

Automatic expert finding mechanism[25] is presented for online help seeking communities. Users' profiles are constructed through their social network authority score and posts terms. Social network authority is computed through algorithms such as Z-score, in-degree, PageRank and HITS adaptations. Posts terms are extracted from posts and reflected the users' domain expertise. For each query, users' profiles are matched to find relevant and high ranked experts. However, [25] did not address the cold start problem and thus new users couldn't achieve high social network authority scores with few answers. Moreover, similarity between users' post content and other posts for the same thread is not considered for indicating the users posts relevance to the thread/posted question. A hybrid approach for expert finding is presented[22]. Features such as user subject relevance, user reputation and authority of a category are used for finding experts. However, the content similarity between users provided answers and questions is measured through cosine similarity techniques without giving attention to semantics, thus lacked to capture the answer relevancy to the question. Link structure of email networks [28, 29] has been analyzed and performance of link analysis algorithms is studied on email datasets. An adaptation of HITS algorithm[35] is presented to address expert finding problem[28] for email communication dataset. Expert ranking is achieved through email content quality and communication patterns. By combining email content and link structure, the approach presented better results on a too small dataset. A network of only 15 people and the formulation of HITS algorithm for expert finding problem is also unclear. DSARank [38], an extension of PageRank measures the relative importance of users in collaboration networks through a link intensity based ranking model. Expertise are measured through interaction metrics and contextual link information. In addition to profile information of users, their interactions in social contexts are considered to measure their expertise. Asking or replying questions are examples of social context in collaboration environments. As the authors focused on service oriented crowd sourcing systems, but interaction logs and email contents for measuring expertise of users were overlooked including answer content quality as an important feature..

Extended-category link graph and topic analysis approach is developed to measure user authority in community question answering websites [27]. Relevance between categories is measured through their content similarity using LDA model and KL-divergence. After grouping similar categories, users ranking is performed through topical random surfer approach LeaderRank [24], an extension of PageRank algorithm quantifies influential people in a popular social network, Delicious. The impact of emotion in online debates is also analyzed with the primary focus on transfer of emotions between participants. Superedge algorithm [26] an adaptation of PageRank [34], identify opinion leaders through influence score during information dissemination and lexical overlap between terms. Content and link features can be used to recommend potential experts in online forums by constructing threads, profile and cluster based language models [39]. In context based expert finding, WordNet is used to measure users' posts' similarity and users ranking is measured through their social network authority score [36]. Context based expert finding approach (CEF) [20], based on PageRank [34], measures user expertise based on their social network authority. In graph representation, nodes represents users and links represent communication between them in the form of asking-questions and replying-questions. Answer providers are assumed as experts as they have more knowledge than question-askers. In CEF [1], link weights are assigned based on (1) the number of answers and the level of communication between users, (2) semantic similarity of posts and context. However, the concept of post similarity and internet is unclear. Moreover, the authors did not consider the similarity between questions and answers as a measure. Document relevance and user's social network authority features [4] are used to measure the user reputation in knowledge communities. For computing document relevance, expertise profiles are created for each user by combining the content of his all given replies for questions in the forum. Finding an expert for a given query, users' expertise profiles are matched with the query terms and most relevant users are returned for a given topic. Similarity between query and profile terms is computed using cosine similarity. For computing users' social network authority score, ExpertiseRank algorithm [3] is used. Finally, overall expert rank is computed by assigning weights to users' ExpertiseRank score with their document relevancy scores. Nevertheless, profiles are constructed using cosine similarity that did not consider the content semantics. Moreover, users' neighbor reputation is not considered. Experts can be identified based on their behavioral properties [9, 39, 40].

Link-based features are used to find experts in a Java programming forum [3]. This technique is an adaptation of the PageRank algorithm [34], which is used to rank web pages. In [3], user reputation is measured based on the total number of answers provided by the users, taking into account the reputation of users to whom they answer. If a user A provide answer to a user B's question who is a domain expert, then it means that user A has more expertise than user B because it answered an expert's question. Assume User X has answered questions for users $U_1 \ldots U_n$, then the ExpertiseRank of User X is given by equation 2.1.

$$ER(X) = (1-d) + d\,(ER\,(U_1)/L\,(U_1) + \ldots + ER\,(U_n)/L\,(U_{n-i})) \qquad (2.1)$$

Where, ER(X) is ExpertiseRank for user X, U_n is the user who is answered by X, d is a damping factor and $L(U_{n-i})$ is defined as the total number of users who helped U_i, according to this idea, a user had more expertise if he replied to questions posted by expert users. User rank can be decreased if he posted several questions in online forum. However, answer's content quality is not considered for reflecting the users' competency and knowledge level in a proper way.

2.1.2. Content quality based expert finding techniques

Content quality is considered as an important aspect to measure user reputation in online forums [41]. For measuring user expertise in online forums, various aspects of answer's content quality [42, 43] , question quality [44] , social network [3, 4, 20, 34, 36, 45, 46] and user profile [3, 47] were analyzed. Numerous features such as asked-questions frequency, provided- answers frequency, votes, reputation, content quality and semantics are used in measuring user expertise [43, 44]. Content based techniques handled expert ranking as an information retrieval problem [4, 20, 22, 26, 27, 40, 48, 49], where user expertise-profiles are constructed from their answer contents. Expertise-profiles contain the topics on which user provides answers. Leveraging theses profiles, online forums present a list of relevant experts for user queries [3, 27, 47, 50, 51]. Profile based approaches [3, 47, 50], assume that answers provided by an individual are reliable indication of his expertise. For example, users' answer content provides strong evidences for their expertise in some specific category or domain such as news, sports or education and thus can be used to construct their profiles [27, 51]. However, this assumption may not always be reasonable in online forums because users' skills varies from novice to professional. For example, sometimes novice or beginner users do not provide relevant answers and their answers' content may contains spam or abusive material, thus their profiles do not reflect correct expertise. Using neighbour and self-preferences, relevant posts are recommended to forum users, content relatedness is measured through semantic similarity techniques [42].

Technical programming terms and tags associated with each query are used to mine user expertise [52]. In [52], technical terms such as class names, methods, built-in functions and tags were selected as features. These features did not reflect the user expertise with accuracy. For example, programming terms are generic and commonly used in programs, therefore their usage do not reflect an individual's expertise. However, tags may be the better indicator of user expertise because they are applied according to the posted problem and indicated user domain. Exiting value (EV) of a question is measured through its answers quality [53], answer quality is computed through its number of answers, votes received, answer status, author reputation and content quality. A quality answer is beneficial to reader and user reputation can be measured through answer. Although, EV-model gave better results than baselines, without considering the answer's content quality in detail. For example, they only considered answer length, number of citations and number of hyperlinks, however failed to consider post content's similarity. Moreover, number of citations or inlinks did not indicate content quality and reflected user's social network authority.

In context of expert finding , most of the previous works investigated the application development by analyzing source code, configuration management and activity of a developers within the developmental environment [9]. To the best of our knowledge, the quality of user contributions in the StackOverflow forum is not addressed in literature. User contributions have been evaluated through question debatableness and the utility of the answer [9]. Debatableness referred to the total number of answers for a question, while the utility of an answer is measured according to its relative rank in the list of answers. Although they considered answer frequency for a question, however, the up-vote ratio and content quality of source code is ignored which are considered as an important aspect in expert mining. A user-activity model [8] is proposed for classifying real experts in the StackOverflow forum. To measure user reputation, this model considered the ratios of questions to answers, accepted answers to total answers, and up-voted answers to total answers. The model used basic features for measuring reputation and ignores user consistency. Moreover, tag quality is not considered, which could lead to more accurate expert recommendations [54, 55]. Hybrid expert-finding approaches [56, 57] are effective in discovering experts in social forums. According to these techniques, similar users are first identified by association rule mining, and then content and link based techniques are applied to measure expertise. Similarity between the applied tags and the question title have also been used to find experts in the StackOverflow forum [9]. State of the art expert finding methods such as document-based model [58] and candidate-based model [58] along with non-textual features have been applied on StackOverflow dataset. Users expertise are measured by combining content/textual and social network features in online QA services [39, 59]. Co-occurrence is a well-known and widely used concept in social web domain, provided the relationships between people, concepts and other entities [45, 60, 61]. User reputation is measured through the co-occurrence of user name in multiple web pages for expert search problem [62]. In a spammer detection method, the co-occurrence between two reviewers for the same review indicated a spamming activity [63]. Co-occurrence can be used for discovering web communities and URL clustering. Web communities are discovered through co-occurred hyperlinks, a webpage pointed by the most frequent hyperlinks was considered as a new member to the community [64]. Analyzing aforesaid works, to the best of our knowledge, co-existing users' reputation measure for expert ranking has not yet been considered for online conversational media particularly for online discussion forums.

2.1.3. Bibliometrics

Bibliometrics such as g-index and h-index are used to compare the scientific output of researchers, research groups, and institutions [65]. These indexes measure the authors' consistency in their performances. Both h and g-indexes are applied to a range of problems such as influence mining in social media[66-70] and finding interesting topics in social networks[65]. The computation method of G-index is given in following sub sections.

2.1.3.1. g-index

G-Index is proposed for evaluating scientific productivity of research scholars by Egghe [71]. It was proposed because of the criticism on h-index that h-index deals extraordinary cited paper in ordinary manner. We have adapted g-index metric for expert ranking problem by applying it on posts scores of StackOverflow users. g-index is defined as "Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number g such that the top g articles received a total of at least g^2 citations" [67]. For the StackOverflow forum, it may be rewritten as "*the unique largest number such that the top-g answers received together at least g^2 score*". Although the complexity of g-index is higher than h-index but it produce better results for top-posts. We can notice that total number of top *g* answers is used in g-index calculation. Hence, answers of higher number of scores contributed more weight to the index than a smaller one. g-index is further explained as follows,

Suppose user A has 9, 13, 10, 15, 12, 6, 8, 3 and 1 score in their answer-posts. After arranging it in descending order score becomes15, 13,12, 10, 9, 8, 6, 3 and 1. Next step is to give them numbers in ascending order as shown in table 2.1.

Table 2.1: Computation of Exp-PC (Step-1)

Post-serial numbers	Answer-score
1	15
2	13
3	12
4	10
5	9
6	8
7	6
8	3
9	1

After giving serial numbers, square of serial numbers is then calculated, as shown in table 2.2.

Table 2.2: Computation of Exp-PC (Step-2)

Post-serial numbers	**(Serial Numbers)^2**	**Answer-score**
1	1	15
2	4	13
3	9	12
4	16	10
5	25	9
6	36	8
7	49	6
8	64	3
9	81	1

Cumulative sum of Score of answer posts of user 'A' is then computed as shown in table 2.3.

Table 2.3: Computation of Exp-PC (Step 3)

Post-serial number	**(Serial Numbers)^2**	**Answer-score**	**Cumulative Score**
1	1	15	15
2	4	13	28
3	9	12	40
4	16	10	50
5	25	9	59
6	36	8	67
7	49	6	73
8	64	3	76
9	81	1	77

Serial number after which the square becomes greater than cumulative score is the Exp-PC (user reputation score). Exp-PC of a user is 8, as shown in table 2.3.

2.1.3.2. g-index applications in social web domain

Applications of bibliometrics such as, h-index and g-index are used to quantify the user reputation and productivity in domains like social web and academic social networks [72]. G and H-indexes can be applied in a social web context for purposes such as finding influential bloggers [73, 74] and to determine the recognition of a researcher among his peers in open source software repositories [75]. Influence of link based techniques can be assessed on retrieval process using h-index and in-degree[76]. Blogs quality is measured through their posts scores. Posts scores are boosted through their in-degree and h-index scores [76]. H-index performed better than in-degree approach for measuring blog quality. Similarly, G-index is used to discover the influential members of a community [67]. G-index can also be applied to rank video-content creators on YouTube [77]. iFinder is another example to calculate the top *k* influential bloggers, proposed by Agarwal et al. [64]. This model scored each post in a community based on its quality. Mean of posts is used to measure the user influence score. The model in [64] is not only dependent on mean score of all posts, but also considers influential and non-influential posts. In a statistical analysis of Slashdot network, h-index is used to measure the controversy in thread discussions [78, 79]. Thread's semantic and structure information is used to measure the degree of controversy. H-index gave quality results due to its simplicity and robustness. Metrics such as MEIBI and MEIBIX are presented for ranking influential bloggers in community blogs [80]. First technique, MEIBI considers the frequency of post's inlinks, comments and publication date, Second, MEIBIX considers the number and age of the post's inlinks and comments. These metrics considered productivity and temporal aspects of blogging behavior. Moreover, interlinkages between blogs posts are also considered. h-index family is used to measure the bloggers influence in community blogs. G-index is used to measure the change in blogger's influence[67]. Post score is estimated through inbound, outbound links, comments and post content.

Summary of related works

In this section, we summarized the existing expert ranking techniques as given in table 2.4, for both rated and non-rated forums such as BBC discussion forum and StackOverflow. In our proposed techniques, we argued that link-structure or link-analysis based expert ranking techniques are not effective as they do not consider the aspects such as content quality, consistency element in user performances and impact of co-existing users on user reputation and discussion quality. Therefore in this literature review, we have reviewed link, content and hybrid features based expert ranking techniques. The purpose is to highlight the issues in link, content and hybrid features based expert finding techniques, and to highlight the aspects which may overcome their deficiencies. As described earlier, rated forums are those platforms where public can give ratings to answers and may select the best answer from the given ones. While in non-rated forums, there is no such mechanism of rating answers and answers frequency is considered as the measure of experts' discovery. Several link-graph, content based and bibliometric techniques have been used for expert ranking problem. Purpose of link-graph or structure based techniques is to rank experts based on their link quality. Link quality may be measured through social network features such as closeness, relatedness and centrality. Moreover, link analysis algorithms such as PageRank and HITS are used to rank experts based on their in-degree and out-degree features. In content based techniques, answer's content quality is considered as important measure in expert finding task. Content quality may be measured through various features such as text similarity, answer length, punctuation, syntactic and semantic complexity, grammar and URL or quotes existence etc. Link-graph based techniques do not consider content quality which is considered as an important aspect in expert recommendation. Similarly, content based techniques do not consider the link structure. Few hybrid techniques exist which consider both link-structure and content quality features for expert finding in social web domain. However, consistency in users' performances, a significant measure in expertise mining, has not been considered for online discussion forums. In some hybrid cases, rated-forum features are not utilized which could be effective in mining user expertise. Majority of the techniques consider individual user performance and do not consider their co-existing/co-participants performance to measure their authority, particularly in case of online forums where users' participates in multiple threads and share topics of their common interests. Characteristics of co-existing users such as, activity and productivity can better predict co-existing users expertise and can be exploited to enhance expert ranking in online forums. Similarly, answer quality and post content similarity among co-existing users which is an effective measure, has not been considered before for expert ranking problem. Expert finding in programming forums such as StackOverflow, has been addressed through features like debatableness of question and utility of answer while ignoring the consistency element in user performance. In StackOverflow forum (SO) case, user reputation is measured through features like up-votes to down-votes ratio, approved-edit score and frequency of accepted answers[14]. However, these features are not fully capable of measuring expertise, for example, SO gives high reputation score to users who receive high up-votes for their provided answers but SO do not consider the reputation of user who provided those up-votes. It is necessary to consider the reputation of user who provides up-vote because users are of different levels for example, from novice to professional, thus receiving up-vote from a professional user has more impact than receiving a vote from a novice or beginner. Similarly, SO present a list of tags to users to

[14] http://meta.stackoverflow.com/questions/269653/why-did-i-gain-lose-reputation-can-i-audit-my-reputation-history

enhance searching and browsing, however, tags quality have not been exploited to infer user expertise. The SO does not measure the consistency in users' performance over the time which is significant in enhancing expert ranking. Expert ranking problem in online programming forums has not been fully addressed through bibliometrics like h-index and g-index, which can be better used with SO features.

Table 2.4: Summary of expert ranking techniques in online forums

Source	**Problem addressed**	**Characteristics**		**Techniques**	**Findings**	**Dataset**
		Link	**Content**			
Zhang, Ackerman [3]	Identification of expert users	✓	×	PageRank [34], HITS, ExpertiseRank, Z_number, Z_degree	Z_number and ExpertiseRank methods performed better.	Java discussion forum
Wang, Jiao [4]	expert finding for online knowledge communities	✓	✓	PageRank, Cosine simailrity	Document-based relevance and user authority, best performance is achieved.	Microsoft Discussion Groups
Bouguessa, Dumoulin [5]	Identifying Authoritative Actors in Question-Answering Forums	✓	×	In-degree, Bayesian Information Criterion, Expectation-Maximization	Normalized In-Degree is found as the best feature for finding authoritative actors.	Yahoo Answers
Schall [38]	Expert ranking in collaboration networks	✓	×	Dynamic social aware PageRank (DSARank), interaction context, IIL (Interaction intensity level) , RA (Relative availability)	Interaction intensity level and Relative availability techniques performed well to recommend important users.	Reality Mining and Enron email archives

Li, Zhang [40]	Ranking of potential experts	✓	✓	Concept and users networks are constructed and then query-concept mapping is performed to search experts.	Semantic information is significant in searching Experts.	Java forum, Eclipse forum
Yang, Qiu [59]	Expert Finding	×	✓	KL-divergence and LDA topic model are used for measuring category relevancy. TSR model is extended through category based link analysis approach.	Category relevancy based authority ranking approach performed well in identifying experts.	Yahoo Answers
Kardan, Omidvar [20]	Expert Finding in online communities	✓	×	Social network analysis techniques, PageRank, total number of answers provided	CEP method gained higher accuracy in comparison with other methods in all contexts like social network and WordNet.	MetaFilter Forum
Ma and Liu [26]	Opinion leaders identification	✓	✓	Extended PageRank through network topology analysis and text mining features such as influential degree of information dissemination and similarity between keywords.	Supernetwork analysis method and Superedge-Rank algorithm are reliable.	Japan's Nuclear Leak Crisis and MATLAB programming tool

Xu and Ramanathan [81]	Expertise finding in microblogs	✗	✓	Probabilistic rules and linguistic methods, thread data, local evidence quality, thread evidence quality, thread filtering	Candidate models significantly outperformed document models. Thread weight is beneficial, candidate centric models and candidate associations do better.	Microblog dataset
Yang, Tao [9]	Characterisation of Expert Behaviour in StackOverflow	✗	✓	Debatableness of a question (number of answers) and answer utility (relative rank in the list) are used as reputation measures	Mean Expertise Contribution	StackOverflow dataset
Movshovitz-Attias, Movshovitz-Attias [8]	Reputation System and User Contributions analysis in StackOverflow	✗	✓	User activity model (Answers, question, accepted, up-voted, comments, QA ratio), PageRank, Singular Value Decomposition, Random Forest Classifier	Analysis of singular value decomposition has successfully detect the extreme user cases who have been influential in the network	StackOverflow dataset
MacLeod [82]	Tagging based reputation measurement	✗	✓	exploratory analysis of StackOverflow data to determine the correlation between user reputation scores and the tags diversity	Blondel's community detection algorithm	StackOverflow dataset

2.2. Thread ranking techniques

We describe two general approaches to thread ranking. Firstly, Content based thread ranking techniques are described. Secondly, link-structure based thread ranking techniques are presented, which considers post structure, content and user reciprocal relationship aspects as depicted in figure 2.2.

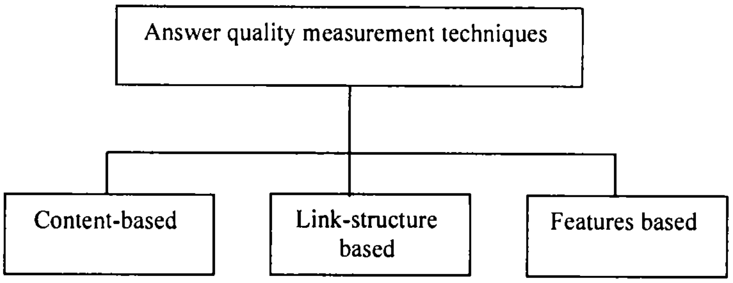


Figure 2.2: Classification of Thread ranking techniques

2.2.1. Content based thread retrieval techniques

Textual representation features are commonly used to identify best answers in CQA [83], like question length, answer length and question-answer content similarity. Social network authority, answer count, ratings and answer acceptance ratio are also used to analyze answer quality [13, 83-85] as additional features. Thread's participants' popularity is utilized as a quality indicator in recent studies on CQA sites. Moreover, interaction among users, in the form of question-answers, can be measured the users interest on several topics [14]. Yahoo Answers and Ubuntu, voting techniques are used to find the best answers. Voter's reputation is used to measure the answer quality [86, 87]. Best answer selection in a CQA can be addressed through finding similar resolved-questions to a newly posted question [15, 88, 89]. An autonomous agent is presented [90] to recommend relevant threads for a given query in medical forums. In this regard, two schemes are presented, first one is Monotonic Post Weighting and other is Parabolic Post Weighting scheme. Both schemes assign weights to posts based on their relative position in a thread. Forum categories assign weights based on their topic relevancy to the queries. Medical entity extraction and shallow medical extraction information can be used as semantic weighting approaches. However, only first post [90] is selected for measuring thread quality which does not represent entire thread's quality, for example combinations such as thread title with head post, top-k posts and bag of posts are not evaluated. Furthermore, a small number (i.e. 20) of selected questions are considered. BM25 model evaluated content quality without considering semantic aspects. For a given thread, finding similar threads from a threads' collection is addressed by exploiting thread structures [17]. Non-textual features are used to predict answer quality in community question answering sites [84]. Features include user reputation, answer length, answer frequency and editor's recommendation. The study was based on non-textual features and textual information and lacked in answer content. Moreover, user participation is measured through their answer count rather than the quality of the provided answers. For finding answers of high quality content in rated-forums, various features [14] are extracted from questions and answers, user to user relations, content quality and online forum usage. The highest results are achieved by combining text, intrinsic and relation features. Although answer length produced better results than other features and considered as a quality indicator by showing the answer depth. However, answer position within a thread is not considered for better thread structure representation. Moreover, no content similarity technique is applied. Question and answer pairs detection is presented as a binary classification task in forums [15]. Questions are detected through sequential patterns while answers are discovered through graph based propagation technique. Features such as, labeled sequential patterns are used to classify questions as they are helpful to identify comparison and erroneous sentences. Answers are detected using cosine similarity, Query likelihood language model (QLM) and classification based re-ranking techniques. All of above methods achieved 90% MAP while QLM and KL-divergence performed better than cosine similarity. Content semantics are not taken into consideration for measurement of answer quality that caused the lower performance of cosine similarity. Automatic methods are useful in differentiating high vs low quality conversation in online question-answering sites [12]. Thread quality is measured through post ratings and without post rating features. Post rating based thread quality is measured through four metrics such as rating ratio, average rating, absolute rating and Bayesian average rating. Thread quality measurement without post ratings involved thread surface, thread initiator authority and temporal features. Although various features have been used to measure the thread quality, however, content quality is not considered. Moreover, participant reputation is measured only through their answer count. Total number of asked questions by the participant and their answer's content relevancy to the question is not considered. Identification of thread length is carried through a classification framework [91]. Thread length indicates the topic specificity of a thread as well as its productivity. Features such as sentiments, topic modeling, semantics, text and link are used in thread length prediction task. According to [92], the longer threads contain viral topics in internet and may have interesting debate. Latent and semantic knowledge is used to recommend similar resolved-questions for a new query in online forums because resolved questions have relevant answers for a given thread. In this framework [92], initially similar resolved-questions for a thread are detected. Secondly, word translation probabilities are learned from questions. Thirdly, semantic relatedness between candidate questions and targeted questions is measured through translation based language model. A hybrid thread recommender systems considered [93] both collaborative and content features. User interest and topic-discovery models are used to assign threads to clusters and then to discover users' interest on each cluster. Users' preferences are used to cluster similar threads. Moreover, based on the content similarity between users' and their neighbor's, threads are clustered. Question and answer pairs identification in discussion forums is carried out using 5W1H, question marks presence and posts frequency feature [94]. In an unsupervised solution for post identification in discussion forums, the lexical correlation between threads and their respective posts has been examined [95]. Mutual reinforcement label propagation is used to predict questions' and question-asker's quality [11]. Question quality is measured through features such as, title's subject length, questions' punctuation, typos, spaces and part of speech entropy. Asker's quality is measured through points earned, answers count and received stars. Thread retrieval problem is addressed for web forums [96]. LDA model is used to rank answers in web forum. LDA model is found effective in terms of running time and space complexity. Authority and content quality features have been used [97] for the best answer selection. Hierarchal classifiers are used to detect question-type and answer-quality in community question answering sites [98]. Internet statistics and WordNet structure information is used to develop hybrid semantic similarity methods [19]. Various methods have been proposed [99, 100] to measure semantic relatedness between terms, as implemented in WordNet. WordNet[15] is an on-line lexical reference database system, attempts to model the lexical knowledge of an English speaker. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also organized into senses (i.e., corresponding to different meanings of the same term or concept). The synsets (or concepts) are related to other synsets higher or lower in the hierarchy defined by different types of relationships. The most common relationships are the Hyponym/Hypernym (i.e., Is-A relationships), and the Meronym/Holonym (i.e., Part-Of relationships) [99]. Figure. 2.3 illustrates a fragment of the WordNet Is-A hierarchy.

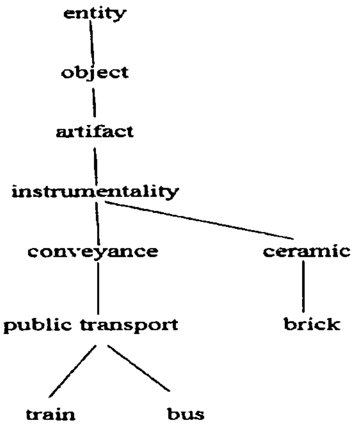


Figure 2.3 : A fragment of the WordNet Is-A hierarchy

[15] https://wordnet.princeton.edu/

WordNet is a collection of implemented semantic techniques like node-base, edge-based and hybrid methods for similarity computation [19, 99, 100]. Several semantic similarity approaches such as named-entity recognition [101], WordNet [102], LDA [101] and word representation measures have been used for measuring semantic textual similarity for English language content [101]. Intelligent question routing and expert finding problems [56, 103] are addressed through WordNet based on semantic similarity. The question dependencies are mapped to hierarchies of WordNet answers for the clarity of the semantic class [104]. For instance, a query: "*How much could you rent a house in London for a year?*" provides the dependency between rent and the amount, WordNet derives the Money as the answer type. WordNet's answer-type taxonomy (ATT) influences the performance of open question-answering (QA) systems due to the category of answer-type. Semantic similarity methods [19, 105] performed better than the cosine-similarity techniques in multiple domains. The use of ATT for QA systems showed 10% improvement using TREC(8-10) collection [104]. Query expansion or more precisely question and answer analysis task is performed using WordNet [102], ConceptNet [106] and SemNet [107]. In an abusive user analytics framework, content degeneration is the relevance of posts in a given thread, the mutual information between a post and the thread is computed using WordNet [107]. A maximum-weight-matching algorithm [108] based on WordNet is found effective in finding the similarities between two texts. In an effort [109] of extracting medication information from veterinary discussion forums, semantic features and WordNet is used for finding similar words. Short answer grading problem is addressed through several graph based features and semantic similarity methods using machine learning techniques [110]. In another case [42], semantic similarity among posts for co-existing users in multiple threads is measured to identify relationships among users and topics of their interest. In a thread summarization task [111], semantic similarities between words is added to post propagation model to reduce the effect of text sparseness. Semantic technique is applied for sentiment analysis for web forum topics [112].

2.2.2. Link-structure based thread retrieval techniques

Thread quality is analyzed through its sub-structures such as whole thread, head-post, top-k posts and reciprocal structures. Using cosine similarity technique [113], threads sub-structures similarity is computed to measure the jointly contained information. Semantics in contents are considered by Wan [13] despite the fact of comparison with the several baseline methods [14-16]. Besides this, Wan [13] ignored participant reputation aspect and only relied on content of thread. Thread structure is more effective rather than considering the entire document for thread retrieval. In this context, various techniques such as PCS, MAX, START, small and large documents are used to compare thread retrieval performance [114]. Thread structure are discovered through posts' intrinsic and extrinsic features [115]. Thread structure may be represented through post relations, reply structure, post position in a thread. In intrinsic features, a reply relation between posts is measured through their contents similarity which shows that both posts discusses same topic. Extrinsic features include time gap, same author and author reference. Methods for annotating post-post discourse structure are presented for online forums [116]. Post label set is developed to capture the post level interactions. Label set consisted of various Question-Answer categories. Post characteristics are captured through lexical, contextual, structural and semantic features. Interaction links between posts are classified in the same feature set. Reply structure within a thread is predicted through conditional random fields [117]. Two kinds of features are used to capture the dependencies among the posts, replying patterns and user interactions. Edge and node features are used to capture structural dependency among threads' posts. Node features include, post similarity, first-post-first, last-post-last, time recency and author reference. Edge features include repeat-reply, jumping-reply, author response, author preference and content propagation. However, well-known techniques such as WordNet have not been applied for content analysis, matching syntax-parsing structures.

Summary of related works

In this section, we summarized the various approaches for thread ranking in online discussion forums as given in table 2.5. Several link-structure and content based approaches have been proposed for thread ranking problem. The purpose of link-structure techniques is to measure the threads quality based on their structure. Link/structure strength reflects the importance of post for a given thread. Thread structure may be extracted through post to post reciprocities or through thread to post hierarchies. Content based techniques consider answer's content quality as a primary measure for thread relevance. For measuring the relevance of a post in a thread, text similarity between posts and thread is used. As our proposed work is based on thread's content quality and post structure, therefore we have discussed thread structure and content quality techniques in literature. Importantly, we have raised the issue of cosine similarity techniques and discussed their limitations to thread ranking problem. Often, posts terms may have multiple meaning (polysemy), however cosine similarity technique only consider the lexical overlap of terms, thus ignore semantics. By ignoring semantics, context of the term in that thread is not captured. In addition to other issues, cosine similarity techniques do not consider phrase structure, proximity information and word order as provided in literature. It is noted that few techniques consider semantic aspects but in some of these methods, semantic similarity has only been computed at a low level, between thread title and user query for example, and detailed semantic similarity measures have not been applied to thread's sub-structures. Answer quality indicators are missing in existing techniques which may produce better thread ranking when combined with content similarity scores.

Table 2.5: Summary of thread ranking techniques in online discussion forums

Source	problem-addressed	Threads' features				Techniques	Findings	Domain/Data set
		Link/Structure	Content/Textual	User-reputation	Non-textual / Forum based			
Adamic, Zhang [1]	Analysis of knowledge sharing activities	✓	×	✓	×	Best answer prediction, cluster analysis of categories, question answers network structural analysis, expertise depth through user entropy	Lower entropy correlates with receiving higher answer ratings in categories where factual expertise are desired.	Yahoo! Answers
Lee, Yang [12]	Discover high quality threads	✓	✓	✓	✓	Features are based on thread surface, temporal, user reputation	Surface and user reputation features are the most influential feature types.	Slashdot forum
Singh and Raghu [17]	Retrieval of similar Discussion Forum Threads	×	✓	×	×	Thread structure is exploited to measure similarity between threads. Cosine similarity technique is applied on following	bop, head-post and bop + head-Post outperformed	Apple Discussion Forums

						substructures, Bag of words, bag of posts, head post, bop-head post and central post.	baseline techniques	
Wang, Tu [88]	Answers ranking in community question answering sites	✓	✓	✓	×	Answer to question latent links are modeled by Bayesian logistic regression. Textual, statistical and user social interaction features are used. Cosine similarity is used.	Answer structure and community intelligence are beneficial in finding quality answers.	Yahoo! Answers
Elsas and Carbonell [114]	Thread retrieval in online forums	✓	✓	×	×	Language models are used, comparison of thread retrieval methods is performed by exploiting thread structures.	Thread's structure, message selection are useful features in finding quality threads. MAX and PCS method outperformed others.	MacRumors Forum
Seo, Croft [115]	Thread retrieval in online forums			×	✓	Cosine technique is used for n-gram text similarity, other features include location prior, time gap, same author and author reference	Unigram and n-gram gave same results, location prior and time gap features are most helpful.	Cancun dataset

Kim, Wang [116]	Post and link classification in online forums			×	×	Structural, post context and semantic features are used. Classifiers included maximum entropy, support vector machines and conditional random fields.	Structural features gave better results when combined with the context feature “previous post from the same author”.	CNET forums
Albaham, Salim [118]	Post quality based thread ranking in online forums	×		×	✓	Quality features include words, sentence, characters and URL frequency in a post. CombSUM and Okapi BM25 method is used to generate a list of ranked threads.	all-all method is always better than the all-sum method	Ubuntu forum
Jeon, Croft [84]	Prediction of answer quality	×	×	×	✓	Non-textual features, maximum entropy approach is used to evaluate answer quality	Answerer acceptance ratio and answer length are better	Naver Q&A service

							indicators to predict the answer quality.	
Agichtein, Castillo [14]	Identify high quality content in social media				×	Content quality features includes punctuation and typos, semantic, syntactic. User authority is measured using HITS and PageRank. Usage statistics features. Several classifiers have been used like SVM, log-linear and decision tree.	The best performance is given by stochastic gradient boosted trees.	Yahoo! Answers
Hong and Davison [94]	Identification of threads and their potential answers in discussion boards	×			✓	Question and answer detection features includes. Support vector machine is used as classifier.	Features (Position + Stop-word) in combination performed reasonably well on both datasets. Content and non-content based methods also performed well.	Ubuntu dataset, DC dataset
Lee, Yang [12]	Discovering high quality answers in CQA	×		×	✓	Question classification is carried out through lexical, syntactic, semantics features	SVM and logistic regression achieve the best accuracy	Yahoo! Answers.

Chapter 3:
Expert Ranking techniques for Non-rated discussion forums

3.1. Introduction

Online discussion forums are rich source of knowledge to find useful information. In such forums, users can easily ask questions or reply to a posted question. Several users participate in online discussions with different skills. however, few users provide quality answers while the majority of the public usually ask questions on multiple topics [1]. It is easy to find answers or a little discussed topic content in online forum because question-answer pairs are provided in categories such as education, news, sports, religion and entertainment. With the passage of time, these forums can accumulate huge contents with unsatisfactory quality of posted discussions. The low quality discussions indicate the existence of unprofessional users. Moreover, for a given problem, it is difficult for public to find an expert. Therefore, expert finding in online forums is an important research problem. Expert ranking problem has been widely addressed through several features such as, social network authority [3, 38, 45, 46, 119, 120], content relevancy [1, 4, 43, 94, 120, 121], combination of links and content [27, 48, 49, 122, 123] features. Structural analysis of social network communities yields a better understanding of user authorities such as finding influential people and opinion leader [67, 74]. Systematic analysis and visualization of social web applications give insight of their structural aspects [124]. Link-based algorithms such as PageRank [34] and HITS [35] have been successfully used for ranking web pages and experts in social web domain [3]. Content quality evaluation is an essential task for performing information retrieval on the community question answering tasks [41]. As mentioned in literature review, several intrinsic content quality features based on semantical, syntactical and grammatical features have been proposed [14].

Co-occurrence is a well-known concept in social web domain, provides the relationships between people, concepts and other entities [45, 60, 61]. In [62-64, 125-133], theoretical ways to incorporate co-occurrence approach for discovering frequent objects are suggested. To the best of our knowledge, the full utility of users' participation in the activity and their co-existing users' reputation has not been considered in online discussion forums. According to co-occurrence concept, the co-occurrence of two or more reputed users in large number of threads with provided quality answers are considered as domain experts. In the subsequent sections, we have proposed users' co-occurrence based reputation measures for expert ranking problem such as users' participation activity, answer quality and their co-existing users' reputation. In order to evaluate the efficacy of our proposed techniques against link-graph based techniques, we have extended a link-graph based expert ranking technique [3]. The experimental study based on real BBC forum dataset shows that our expert ranking techniques achieve promising results.

3.2. Problem Definition

In this section, we first describe basic terms used in online discussion forums and then formally define the expert ranking problem.

Terminologies:

Thread: Thread is a question, asked by a user or it may be a topic initiated by a user in an online forum.

Subject: Each thread has a subject which is usually indicated by a title that appears on top of the thread.

Post: Post is a reply or an answer provided by a user to a question/thread. A thread may consist of many posts.

Forum user: Forum user is a person who ask question or provide replies or answers to the questions.

Definition

Let NF be the non-rated forum containing a set of threads $T=\{t_1, t_2, t_3 \ldots\ldots t_m\}$, where t_i contains the number of posts $P_i=\{pi_1, pi_2, pi_3 \ldots\ldots\ldots\ldots pi_n\}$ where Pi be the collection of posts by the user u_j. Let U is the group of co-existing users. We have to find the expert users from U who's self and neighbours reputation should be high. Reputation of the users is computed through their thread count, and posts content similarity scores.

3.3. Baseline technique

ExpertiseRank [3] an expert ranking algorithm, selected as a baseline technique which is based on PageRank [34] algorithm. PageRank [34] is a popular and well-known algorithm for ranking web pages. ExpertiseRank identify expertise in online help seeking communities using social network features. According to ExpertiseRank, a user is considered as professional, if he gives answer to experts' questions. ExpertiseRank, as a baseline approach, is a better choice for further extension because of its scalability. Besides this, it allows us to analyze the impact of answer's content quality and co-existing users' reputation in addition to link and structure aspects.

Based on PageRank algorithm [34], ExpertiseRank [3] is presented for finding experts in online community forum. According to ExpertiseRank [3], if a user *A* provide answer to an expert user *B's* question, then it means that user *A* has more expertise than user *B* because user *A* answered an expert's question. Assume *User X* has answered questions for users $U_1 \ldots U_n$, then the ExpertiseRank of *User X* is given in equation 3.1.

$$ER(X) = d\left(\frac{ER(U_1)}{L(U_1)} + \ldots\ldots\ldots + \frac{ER(U_n)}{L(U_n)}\right) \qquad (3.1)$$

ER(X) is ExpertiseRank for user *X*, U_l is the user who is answered by *X*, *d* is a damping factor which is set to *0.85* because we tested multiple values of *d* such as *0.25* and *0.65* without any remarkable difference. $L(U_l)$ is defined as the total number of users who helped U_l. Generally, a user will have more expertise if he replies to the questions posted by expert users. User rank will be decreased if he asks too many questions in online forum.

ExpertiseRank [3] measure the users' expertise based on the total number of expert users to whom they answered. However, important aspects such as, answer content quality [12-14], users' participation frequency and co-existing user reputation [46, 48, 49] are not considered. In addition to our proposed features, ExpertiseRank [3] is extended with the use of answer quality and user reputation features. Experiments on real dataset showed that our techniques are effective in ranking expert. As discussed above, for extended methods (ExpRank-COM and ExpRank-AQCS*)*, damping factor *d's* value is set to 0.85.

For expert ranking problem, first stage of our task is to extract co-existing users from BBC forum dataset. In the following section, we describe user extraction process.

3.4. Co-existing Users Extraction

Based on co-occurrence concept in social web applications, the co-existing users can be defined as "*the users who co-exist in multiple threads as answer providers*", as shown in figure 3.1. In Apriori algorithm [134], support is considered as a primary measure in extracting co-existing users. The support value of an item X with respect to transaction T is defined as the proportion of transactions in the database which contains the item-set X, illustrated by equation 3.2. Aprori algorithm's support measure [134] is considered in extracting co-existing users from BBC forum dataset. For extracting users in BBC discussion forum, Apriori algorithm is applied on 10,000 threads' data which contains threads and posts of users. As a result, 450 co-existing users were obtained. Frequent users obtained in BBC forum dataset have minimum support of 2 and maximum-support of 22 which indicates that small proportion of users in BBC forum dataset have co-participated/co-existed.

$$Suppor(X) = \frac{|X \cup Y|}{n} \qquad (3.2)$$

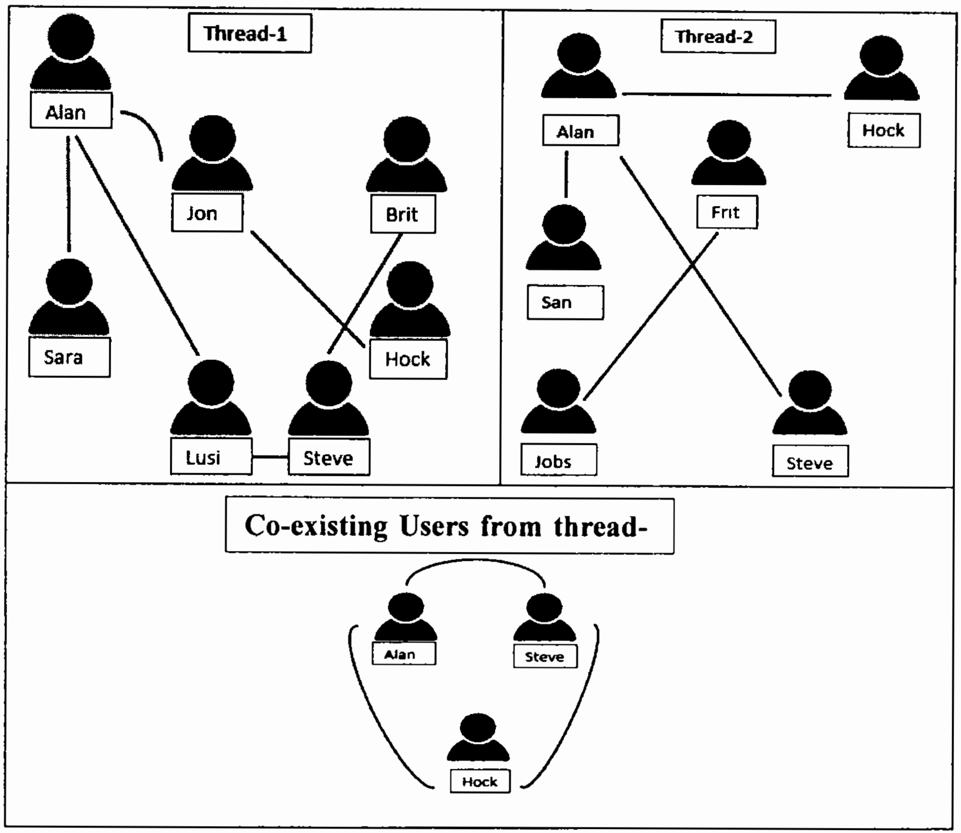


Figure 3.1: Co-existing users in multiple threads

3.5. Proposed Expert Ranking Techniques

Expert ranking techniques ExpRank-CRF and ExpRank-FB and their extensions with ExpertiseRank [3] have been presented in following sections.

3.5.1. ExpRank-CRF

Our first proposed technique is ExpRank-CRF which ranks the user based on features such as, thread support, posts similarity and co-existing user's reputation. Aforesaid features are described in following sub sections.

3.5.1.1. Threads support count for user

The motivation behind this feature is that, more the user co-exists in different threads, the higher the chance of remaining active and provides several answers. Thread support count of a user is defined as the total threads where user co-exists with other users as given in equation 3.3.

$$U_{CT_k} = \sum_{k=1}^{n} Sup(u_k, u_{k+1}) \qquad (3.3)$$

Where U_{CTk} is the thread count of user *k*, *Sup* is the support count. Let ψ be the support threshold i.e. ψ = 2. If U_{CTi} >= ψ for a user $U= \{u_1, u_2....u_k\}$ then u_k is the active participant in thread $T = \{t_1, t_2........t_m\}$.

3.5.1.2. Semantic Similarity among posts of co-existing users for a given thread

Post's content similarity represents the answer quality [12, 13, 42, 43, 98] in community question answering sites. Relevant answers show a user's knowledge depth and understanding of a topic and expected that user provide similar and relevant post content to his co-existing user. To conclude the provision of the relevant answers by the users on a topic, they are discussing in the scope of the topic and can be expressed as follows: Let S be the set of semantic similarity scores of co-existing users C_e's post content in their respective threads. i.e. S = {s_1ce_1, s_2ce_2, s_3ce_3 ...., s_nce_m}. If S_nCE_m>= β then the co-existing users have same area of expertise and have highly relevant content for a given question or topic, where β>=0.75, because average similarity score among posts of reputed users is 0.60.

Cosine similarity technique [113] has widely been used [11, 12, 135] for computing content similarity. However, it only considers lexical overlap between documents and ignores words' semantics. Due to this limitation, the context in discussion's content is totally ignored that gives rise to polysemy problem [136]. On the other hand, semantic similarity techniques considers the context/meaning of discussion [19, 100] and preferred for evaluating content overlap in posts. Semantic similarity is computed for different post contents of co-existing users using WordNet [102]. Given the posts of co-existing users, WordNet determines the similarity of the posts in terms of sense (synset) and relationship. A high semantic score for given posts infers that the context of the posts is the same. Computation of semantic similarity involves processes including the tokenization of posts content, part of speech tagging, stemming, determining the sense of every word in a post and computing the similarity of posts based on pairs of words. Leacock and Chodorow [100, 102] took the maximum depth of taxonomy into account and computed semantic similarity between posts by equation 3.4.

$$Sim_{sem}(p_i, p_j) = -\log \frac{length(p_i, p_j)}{2 * deep\_\max} \qquad (3.4)$$

Here, *length (p*p_i*, p*$_j$*)* is the shortest path between p_i and p_j and *deep_max* is the maximum depth of the taxonomy. Sim_{sem} (p_i, p_j) is the semantic similarity between posts which lies between 0 and log *(2deep_max+1)*. Similarity is computed based on the shortest path between the synsets associated with posts terms [100, 102].We computed similarity of posts (both which consist of multiple terms) by taking the average shortest synset distance between all the terms in both posts. If terms of p_i and p_j have the same sense, then *length* (p_i, p_j) *=0*. In practice, we add 1 to both *length* (p_i, p_j) and 2*deep_max* to avoid *log (0)*.

3.5.1.3. Co-existing user reputation

Rank of users' is increased due to the high reputation of their co-existing users. Users' self-reputation score is computed by combining the thread support count (equation 3.3) and semantic similarity score of their posts (equation 3.4), is illustrated in equation 3.5.

$$UR_i = U_{CT_i} + Sim(Post) \qquad (3.5)$$

Where, UR_i represents the user's self-reputation score. U_{CT_i} is the thread count and *Sim* is semantic similarity score between user and his co-existing users' posts.

Overall reputation score (ExpRank-CRF*)* of each user is computed by adding his co-existing users reputation CE_{rep} score to his self-reputation score UR_i, as illustrated in equation 3.6.

$$ExpRank-CRF = UR_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} CE_{rep}(UR_i, UR_j) \qquad (3.6)$$

Where, UR_i is self-reputation score for each user computed in (equation 3.5) and $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} CE_{rep}(UR_i, UR_j)$ is the reputation score of all other users who co-exist with user U_i. Co-existing user reputation CE_{rep} is computed using equation 3.5.

3.5.2. ExpRank-COM

We have extended our technique ExpRank-CRF (section 3.5.1) through a baseline approach ExpertiseRank [3] and named it as ExpRank-COM. In ExpRank-COM, user's expertise are measured based on the reputation of their question-askers *(ER (U_n))*. The ExpRank-COM score of a user is computed by multiplying his question-asker's reputation score (rep_n) with ExpertiseRank [3] score and can be expressed by equation 3.7.

$$CR(A) = d\left(\frac{ER(U_1)}{C(U_1)} * rep_1 + \dots\dots\dots + \frac{ER(U_n)}{C(U_n)} * rep_n\right) \qquad (3.7)$$

Where, *CR(A)* is ExpRank-COM score for user *A*, *ER* is ExpertiseRank score of user U_i answered by user *A*, rep_n is ExpRank-CRF score of user U_i computed in (equation 7), $C(U_n)$ is the total number of users who helped user *(ER (U_n))* and *d* is damping factor whose value is set to 0.85.

3.5.3. ExpRank-FB

According to ExpRank-FB, A user is an expert if he provides quality answers in specific categories. Following features have been proposed:

-f1: Count User's highly similar replies for each thread: Highly similar replies to a thread's title/question indicates answer quality [14, 104, 118, 137]. Similarity between user post (answer) and thread title (question) is computed using WordNet's dictionary [102].

-f2: Mention links: Presence of URL references in the posts indicates post quality [15, 84].

-f3: Answer count in each category: Frequent replies by a user in a specific category shows his domain specificity [51].

-f4: Mention quotes: Existence of quotes in user's post contents indicates answer quality [121].

-f5: Answer count: Reputed users are expected to give frequent answers rather than asking questions [14, 84].

-f6: Answer length: Detailed written replies indicates an answer's quality [13].

In order to rank experts, each feature scores (f_1-f_6) have been summed for all users, given in equation 3.8.

$$ExpRank-FB = \sum_{j=1}^{m} \sum_{i=1}^{n} (F_i U_i) \qquad (3.8)$$

ExpRank-FB is based on above features for expert ranking and F_i is the feature score for user U_i.

3.5.4. ExpRank-AQCS

We have extended our ExpRank-FB technique (section 3.5.3) through a baseline approach ExpertiseRank [3] and named it ExpRank-AQCS. The notion behind the ExpRank-AQCS is to enhance users rankings and to assess the combined impact of answer quality, category speciality features and ExpertiseRank technique [3]. Users' expertise are measured not only through the total number of their question-askers, but also included the question-asker's answer quality and category speciality score. As a result, user rank is computed by multiplying question-asker's (*ER* (U_n)) answer quality and category speciality score with their ExpertiseRank score [3], named as ExpRank-AQCS and can be illustrated by equation 3.9.

$$AQCS(A) = d\left(\frac{ER(U_1)}{C(U_1)} * f_1 + \ldots\ldots\ldots + \frac{ER(U_n)}{C(U_n)} * f_n \right) \qquad (3.9)$$

Where, AQCS is ExpRank-AQCS score for user *A*, *ER* is ExpertiseRank of user U_1 answered by user *A*, *f* is a summed features (answer quality, category speciality) score for each user computed as ExpRank-FB in (equation 3.8), $C(U_n)$ is the total number of users who helped user (*ER* (U_n)) and *d* is damping factor whose value is set to 0.85. Algorithm for expert ranking in non-rated forums is given below,

```
ALGORITHM: Expert ranking for non-rated forums
Input: List of all Co-existing $CE_i$ users participated in threads
Output: Ranked list of expert users

For each $CE_i$ do
    Compute ExpRank-CRF_i (by equation 3.6);
    Compute ExpRank-COM_i (by equation 3.7);
    Compute ExpRank-FB_i (by equation 3.8);
    Compute ExpRank-AQCS_i (by equation 3.9);
End
```

The time complexity of the proposed expert ranking algorithm for Non-rated forums is quadratic. The big oh asymptomatic notation for the algorithm is O (n2). The time complexity of the proposed algorithm is same as that of the baseline algorithms.

3.6. Experimental Setup

In this section, we give details on dataset, performance measures and results. We have used a public BBC Message board's discussions dataset obtained from cyberemotions[16]. This dataset consisted of four years threads and posts/comments from several categories as given in table 3.1.

Table 3.1: BBC dataset statistics

Threads	97,946
Posts	2,592,745
Users	18,000
Categories	News, Sports, world news, religious, entertainment

Statistics of dataset shows that large number of discussions has been taken place in four years which indicates the significance of online discussion forums. For accomplishing expert ranking task, 10,000 threads with their participants from several categories have been selected. There were 450 co-existing participants who co-exist in multiple threads. Each participant in BBC forum is either a question-asker, answer provider or has both roles. Co-existing users' participation statistics are given in table 3.2.

Table 3.2: Co-existing users' participation statistics

Group-members count	Thread count	Category
186	19039	World News

[16] http://www.cyberemotions.eu/data.html

46	7016	UK News
31	5617	TV and Radio
29	4942	Jewish topic
88	9875	Ethics and free thought
23	4375	Eastern Religions
47	7439	Christian topic

As there was no explicit user-supplied expertise ranking/rating data in BBC discussion boards, therefore we used human experts to label the users' expertise based on their answer quality. Majority of BBC dataset is related to world and UK news domains. Therefore, two domain experts are selected to rate/annotate the users. In order to rate users, Zhang et al. [3] categorized users into three expertise levels. By adapting criteria [3], threads and their respective posts have been selected for each user for rating/judgement. Rating levels have been assigned to users on the basis of their answer quality (relevancy to the topic) [3, 138]. Rating level [3] details are provided in table 3.3.

Table 3.3: Expertise rating levels

Level	**Category**	**Description**
3	Expert user	Highly informative and can answer critical and domain specific questions regarding domestic and international issues
2	Average user	Can answer general questions and have some basic knowledge
1	Beginner	Just starting to know about general issues or want to gain insight on some hot issue

Our techniques produce continuous values for users ranking. To normalize continuous values the min-max normalization technique [139, 140] is applied.

3.6.1. Performance measures

Spearman's rho [3] and Kendall's Tau [3, 141] are the commonly used measures and have been used in several works [3, 4, 74]. Upon receiving 450 users' ratings from human raters, the human rater's judgement reliability has been checked by intra-rater correlation. The Kendall's Tau distance between the two human raters was found *0.773*, and the Spearman's rho correlation coefficient was *0.791*, which are sufficiently a high rate of inter-rater correlation. Spearman's rank order correlation, represented by ρ, is a technique to compute association between the ranking orders of scores on two variables [74] . In our case, correlations between our techniques and baseline expert ranking techniques are computed by using the equation 3.10.

$$\rho = 1 - \frac{6\sum df_i^2}{l(l^2 - 1)} \qquad (3.10)$$

Where, *df* shows the difference of ranks and, *l* is the number of items in each case and equal to 50 and 100 for top-50 and top-100 users respectively. Kendall's rank correlation [141] is a measure that determine the strength of dependence between two variables. The variation among two different ranking results is reported in figure 1 to 4. Kendall's rank correlation is represented by τ ranging from -1 to +1, and computed using the equation 3.11.

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)} \qquad (3.11)$$

3.6.2. Results and Discussion

In this section, we describe the performances of expert ranking techniques. First, in figures 3.2 & 3.3, we compare the performances of Co-existing (CE) and Non-Coexisting (NCE) users, using popular features such as answer count [3, 12, 14, 20, 142] , support count [3, 62, 143] , question count [12, 14, 142], mention link and quotes [13, 14] and ExpertiseRank (baseline technique) [3]. It is evident from the figures 3.2 & 3.3, CE's performed better than NCE's validating our notion that CE's provide quality discussions and their activity and participation levels are higher than NCE's. Almost, nearly in all features, CE's achieve higher scores than NCE's which indicates their reputation.

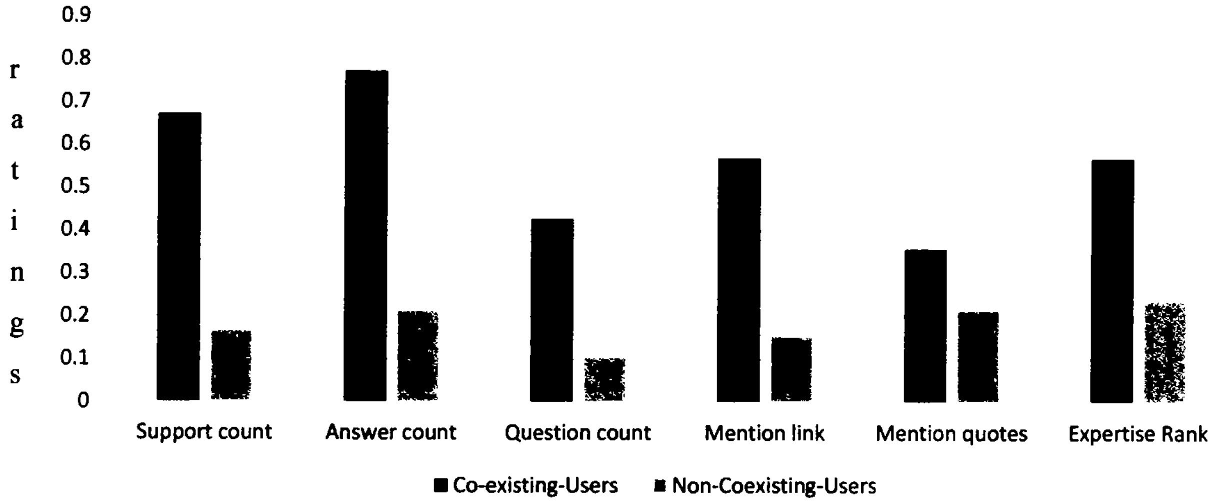


Figure 3.2: Performance comparison of top-10 Co-existing and Non-coexisting users

Regarding answer count feature, CE's give more answers than NCE's, reflects that CE's are real experts and can answer frequently on multiple topics. Moreover, CE's provide quality answers as their answers contain references/links to external sources. Interestingly, the CE's achieved high score in baseline technique [3] than NCE's which reflect that CE's give answers to those users who are experts and their reputation is higher than those answered by NCE's. Support count shows the number of threads in which users' participated. CE's support count is higher than NCE's which indicates CE's participation strength. Although CE's have higher question count but it is expected because an expert, while answering a question, may ask clarification questions from question-asker.

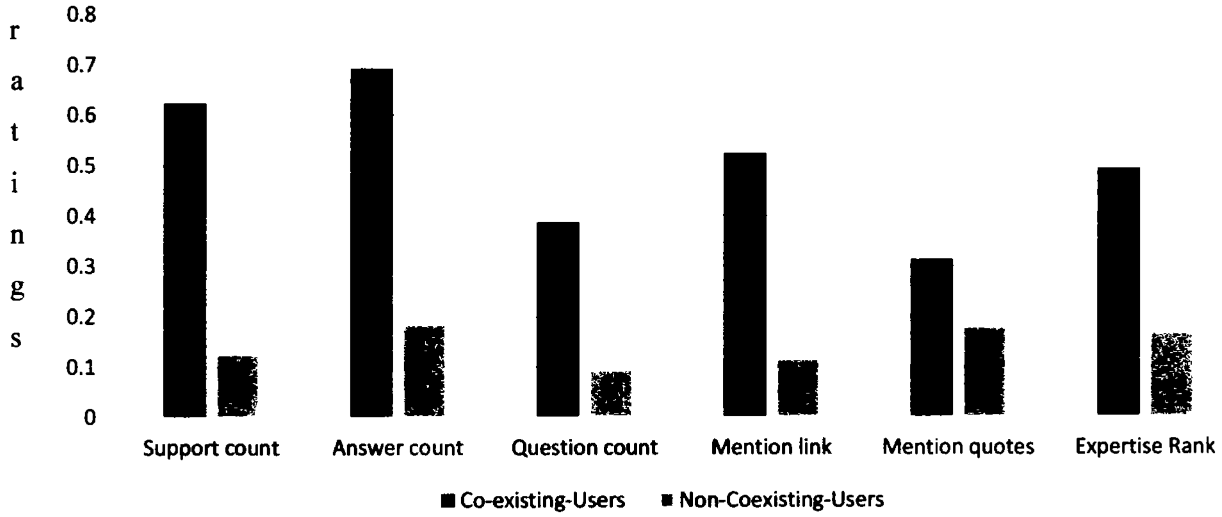


Figure 3.3: Performance comparison of top-20 Co-existing and Non-coexisting users

Now we compare the performances of proposed (ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS) and baseline technique ExpertiseRank [3] techniques against human ratings through correlation and overlap similarity measures. Figures 3.4 & 3.5, shows the statistical correlations between various expert ranking techniques and human ratings scores. A high correlation can be used to evaluate users' expertise in online discussion forums.

All of the ranking techniques give a relatively high correlation with the human-assigned ratings. This tells us that, indeed, co-existing users' information and answer quality could be used to help evaluate users' expertise in online community networks. We have observed that BBC forum network structure is different from the Web, and we have also seen that some algorithms, such as PageRank, which excel at ranking Web pages, do not outperform the proposed techniques in this network. The key to understanding the performance of the ExpertiseRank algorithm is in understanding the human dynamics that shape an online community. From these figures 3.4 & 3.5, it can be seen that the proposed techniques achieves relatively high correlations scores with human assigned ratings which support our notion that features like users' participation frequency, answer quality and their co-existed user reputation can significantly improve expert rankings. A high O_{sim} score obtained by proposed ExpRank-FB, ExpRank-COM and ExpRank-Hybrid indicates the quality of rankings they produced. In comparison to O_{sim}, the spearman and Kendall correlations have obtained low score which indicates that proposed techniques produce different ranking orders. It is found that in most of the cases, the baseline technique ExpRank [3] doesn't perform better than proposed techniques, indicates that only the link-structure features are not effective in identifying real experts in online forums.

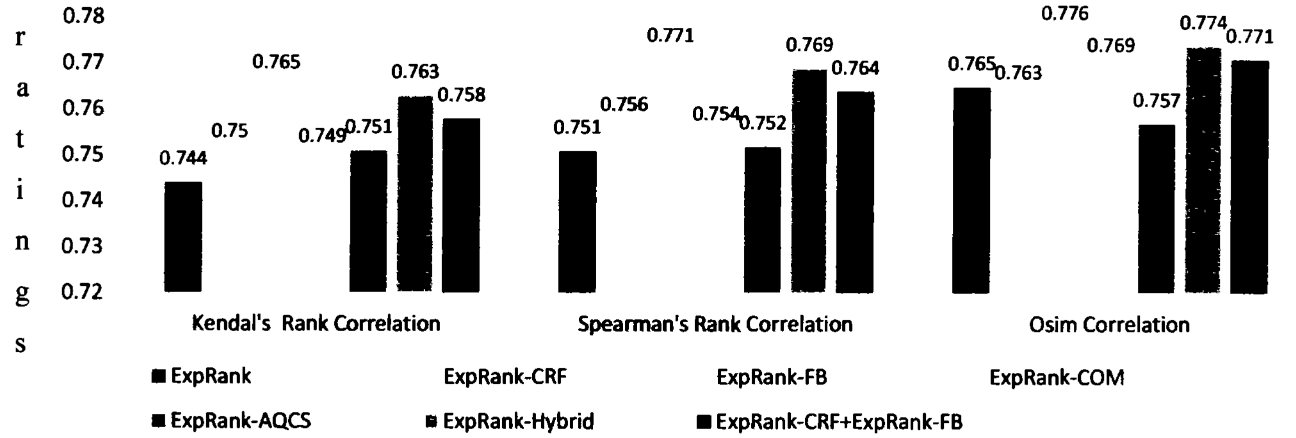


Figure 3.4: Correlation scores for Top-50 users

In figure 3.4 & 3.5, the performance of hybrid techniques such as (ExpertiseRank + Hybrid) indicate that better rankings may be achieved by combing features such as co-existing users' reputation and link-structure. In figure 10, for top-50 users, ExpRank-FB based ranks tends to produce slightly better among all techniques and indicate that features such as, answer quality, category specialty and content similarity are significant in expert finding. For top-50 users in figure 10, second dominant rankings are achieved by combining features such as, co-existing users' reputation and their answer quality (ExpRank-CRF + ExpRank-FB). For top-20 and top-30 users, some of the techniques obtained relatively low correlation scores against human rating scores except ExpRank-FB and ExpRank-hybrid.

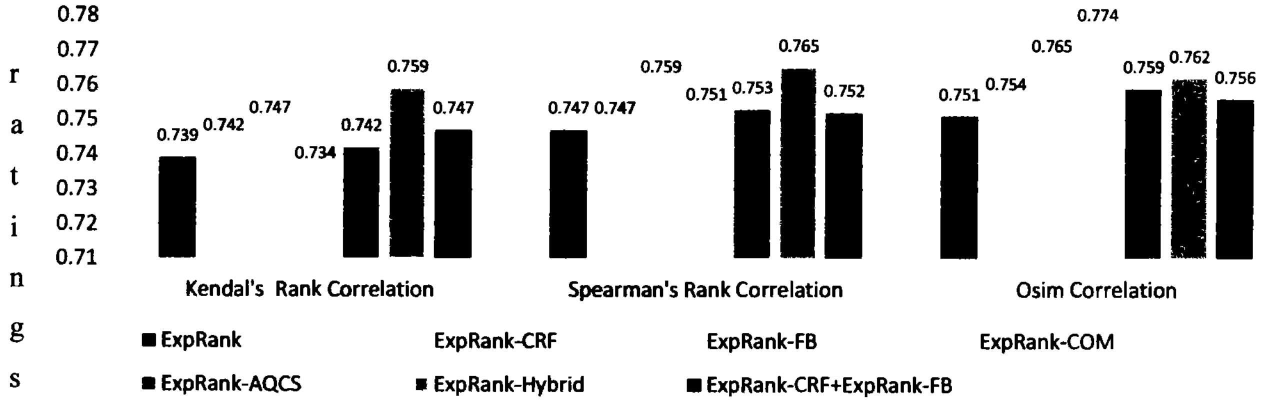


Figure 3.5: Correlation scores for Top-100 users

Chapter Summary

We studied the problem of expert ranking in online discussion forums and proposed techniques for such problem like ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS. Basically we rank forum participants through their self-reputation, co-existing members' reputation and answer quality. In contrast to classical expert finding techniques, which are based on users social network authority and provided-answers frequency features, our techniques (ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS) considers both answer quality and co-existing members reputation to identify experts in online discussion forum. We provided some experimental evidences that ranking produced by our techniques such as, ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS have better quality than those of traditional ranking algorithms such as ExpertiseRank [3]. We believe that proposed techniques find expert communities in online discussion forums in a natural and productive way.

Chapter 4
Expert-ranking techniques for the online rated discussion forums

4.1. Introduction

Community question answering (CQA) sites provide valuable knowledge services to online users. Yahoo! Answers, StackOverflow, Ubuntu, wikis, and other online discussion forums are popular examples of CQA services. These online forums provide an easy and interactive place for knowledge sharing and the exchange of ideas to take place, and users prefer specific discussion forums to conventional web pages for finding topic-specific and useful information. StackOverflow and Yahoo! Answers are popular CQA sites, and they typically present a list of experts based on their reputation score. The quality of information provided by CQA sites has greatly improved in recent years [6]. Answer providers are the main drivers of online discussion forums [6, 144], and expert finding in these forums is an extensively considered problem. In online rated-forums such as StackOverflow forum (SO-forum) which is a technical forum for software developers with knowledge sharing facilities, users may 1) find solutions to programming issues, 2) post technical questions, 3) provide answers to posted questions, and 4) vote on existing answers and questions. StackOverflow provides browsing facilities for the posted questions, tags, and users.

Link-analysis techniques are used for finding experts according to questions-answering relationships[3-5, 20-27], email communications[28, 29], citation or co-citation networks [30-33]. The link-analysis techniques focus on analyzing the link structure among individuals rather than their answer's content. Therefore these techniques are not suitable for finding experts in online rated forums such as StackOverflow-forum[17], Ubuntu[18] and Yahoo Answers![19] etc. StackOverflow-forum employ a user reputation mechanism, through which, users obtain reputation score based on their post quality, up-voted answers and up-voted-questions [8]. It is observed that in StackOverflow-forum, active users respond to many programming problems. However, it may still be a challenge for the StackOverflow-forum community to produce creative or novel knowledge [9]. Users may gain greater reputation by providing frequent answers to simple questions rather than answering complex questions [144]. A better characterization of a users' expertise is required using their contribution quality rather than their reputation scores received through built-in incentivization mechanism. Currently online rated-forums such as StackOverflow-forum have no mechanism to measure the performance consistency of participants. The

[17] http://stackoverflow.com/

[18] http://www.ubuntu.com/

[19] https://answers.yahoo.com/

expert-finding technique proposed here rely on G-index [71] which is a popular bibliometric technique. G-index is applied in the social web domain to find influential bloggers [73, 74], and to determine the recognition of a researcher among her peers in open source software repositories [75]. Similarly, G-index is used to discover influential members of online communities [145] and is applied to rank video-content creators on YouTube [77]. Link-analysis techniques such as [3, 36] did not consider content quality, but combined content quality a social network authority score to improve a user's rank in online forums. Considering [3] the baseline approach, we came up with the novel features such as votes ratio, voter's reputation and tag quality using the SO-forum dataset to measure user reputation. Moreover, used the bibliometrics, such as G-index [65, 71, 146] to measure user consistency in providing quality answers.

In this module, we propose two expert-ranking techniques for the StackOverflow forum: Exp-PC and Weighted Exp-PC. Firstly, Exp-PC, an adaptation of G-index, considers the user's consistency in receiving a high reputation score on each of the provided answers. Moreover, we propose Rep-FS (user reputation features) including vote's ratio, voter's reputation, tag quality and participant reputation. Secondly, Weighted Exp-PC, an adaptation of Exp-PC with user reputation features is proposed. The experimental results of the proposed expert-ranking techniques, Exp-PC and Weighted Exp-PC, validate that these methods identify genuine experts in a more effective way.

4.2. Problem definition

In this section, we first describe basic terms used in online rated-forums and then formally define the expert ranking problem.

Terminologies:

Question: A query posted by the user in an online forum.

Answer: An answer provided by a user to a question. A question may have many answers.

Forum user: A person who may initiate or answer the questions.

Definition

Le RF be the rated-forum containing a set of questions $Q= \{q_1, q_2, q_3 \ldots \ldots q_m\}$, where q_i contains the number of answers $A_i= \{ai_1, ai_2, ai_3 \ldots \ldots \ldots \ldots ai_n\}$ by a particular user. Let U be the group of users. We have to find the expert users from U whose performance is consistent and whose reputation should be high. Performance consistency is measured through a bibliometric g-index. Reputation of the users is computed through their voters' reputation, neighbours' reputation and tag quality.

4.3. Baseline technique

We have chosen a link analysis based expert-ranking approach, ExpertiseRank [3], as a baseline technique. According to [3], if user A provides an answer to a question posed by user B, who is a domain expert, then it means that user A has more expertise than user B, because they answered an expert's question. Assume that user *X* has answered questions for users $U_1 \ldots U_n$, then the ExpertiseRank of user *X* is given in equation 4.1.

$$ER(X) = d\left(\frac{ER(U_1)}{L(U_1)} + \ldots\ldots\ldots + \frac{ER(U_n)}{L(U_n)}\right) \qquad (4.1)$$

ER(X) is ExpertiseRank for user *X*, U_l is the user who is answered by *X*, *d* is a damping factor that is set to .85, and $L(U_l)$ is defined as the total number of users who responded to U_l. According to this technique, a user will have more expertise if replying to questions posted by expert users. User rank will be decreased if he poses too many questions in the online forum.

ExpertiseRank [3], considers link-structure of the forum users where links are formed on the basis of questioning-answering relationships between users. This technique does not consider users' answer quality and their consistency in providing quality answers. As the social media applications such as forums and blogs, support users to generate their favorite content, therefore utilization of content in information retrieval tasks is of prime importance. In the social web domain, particularly in CQA sites such as Yahoo! Answers, a question receives multiple answers and the best answer is selected by the user posing the question [1, 11, 37, 88]. StackOverflow-forum has a reputation measuring mechanism, according to which users are listed according to their reputation scores. The reputation score is computed based on their cumulative past performance [8].

4.4. Proposed Expert ranking techniques

The expert-ranking technique consists of the application of G-index approaches to various reputation features to identify experts on CQA sites for programming language problems, in this case, SO-forum. In this section, expert-ranking techniques, namely *Exp-PC* and *Weighted Exp-PC* are discussed in details.

4.4.1. Exp-PC

Exp-PC is an adaptation of G-index as explained in section 2.1.3.1, a popular author-productivity measure in academic social networks. Exp-PC can be used to measure user reputation in online programming forums such a StackOverflow.

Formally, Exp-PC (an adaptation of G-index) can be written as: *given a set of posts ranked in decreasing order of the number of post scores that a user received, the Exp-PC rank is the largest number, n, such that the top-n posts received a total of at least n^2 post scores.*

4.4.1.1. User Reputation features (Rep-FS)

Expert finding is an acknowledged research problem for the social web domain, and it has been addressed using several features in different datasets. For the most part, the features are based on content quality and social network authority. For the StackOverflow forum dataset, we presented reputation features such as voter reputation, tag quality and participant's reputation in the following way:

Voter reputation (Rep-f1): In StackOverflow forum, high up-votes count of an answer shows its quality. However, voter reputation is not considered by StackOverflow which is an important aspect to determine the users' different skills, from beginner to professional. Therefore, for an answer, receiving up-votes from less reputed or beginner user is less effective than receiving up-votes from professional or real experts. To handle this problem, voter's reputation feature is proposed. Voter reputation is computed by Exp-PC technique (Section 4.4.1).

Up-vote to down-vote ratio (Rep-f2): In StackOverflow, up-votes indicate the usefulness of an answer, while down-votes show its irrelevance. This feature measures the user's consistency in gaining high numbers of up-votes and low number of down-votes [147]. This ratio is computed by dividing a user's up-votes by the down-votes. High number of up-votes indicate user expertise.

Participant-based reputation (Rep-f3): The number of answers provided for a question has been proven to be a simple but effective value in evaluating answer quality [12, 13]. In each StackOverflow conversation, multiple users may participate in discussion. In a conversation, the participation of reputed users (who are consistent in performance), made it productive. Moreover their existence boosts the rank of neighbor users [16]. Participants' reputation is computed through Exp-PC (Section 4.4.1).

Popular tags (Rep-f4): StackOverflow presented the list of available tags and their usage frequency to facilitate the browsing of questions and answers. Tag quality is an important aspect in mining user expertise in collaborative systems [54, 55]. Assigning popular tags (frequently-used tags in StackOverflow) to the question indicates a user's domain knowledge and his capability to better present his question to community [148]. Tags similarity is computed through cosine similarity technique [113].

Chapter Summary

We studied the problem of expert ranking in online discussion forums and proposed techniques for such problem like ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS. Basically we rank forum participants through their self-reputation, co-existing members' reputation and answer quality. In contrast to classical expert finding techniques, which are based on users social network authority and provided-answers frequency features, our techniques (ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS) considers both answer quality and co-existing members reputation to identify experts in online discussion forum. We provided some experimental evidences that ranking produced by our techniques such as, ExpRank-CRF, ExpRank-COM, ExpRank-FB and ExpRank-AQCS have better quality than those of traditional ranking algorithms such as ExpertiseRank [3]. We believe that proposed techniques find expert communities in online discussion forums in a natural and productive way.

Chapter 4
Expert-ranking techniques for the online rated discussion forums

4.1. Introduction

Community question answering (CQA) sites provide valuable knowledge services to online users. Yahoo! Answers, StackOverflow, Ubuntu, wikis, and other online discussion forums are popular examples of CQA services. These online forums provide an easy and interactive place for knowledge sharing and the exchange of ideas to take place, and users prefer specific discussion forums to conventional web pages for finding topic-specific and useful information. StackOverflow and Yahoo! Answers are popular CQA sites, and they typically present a list of experts based on their reputation score. The quality of information provided by CQA sites has greatly improved in recent years [6]. Answer providers are the main drivers of online discussion forums [6, 144], and expert finding in these forums is an extensively considered problem. In online rated-forums such as StackOverflow forum (SO-forum) which is a technical forum for software developers with knowledge sharing facilities, users may 1) find solutions to programming issues, 2) post technical questions, 3) provide answers to posted questions, and 4) vote on existing answers and questions. StackOverflow provides browsing facilities for the posted questions, tags, and users.

Link-analysis techniques are used for finding experts according to questions-answering relationships[3-5, 20-27], email communications[28, 29], citation or co-citation networks [30-33]. The link-analysis techniques focus on analyzing the link structure among individuals rather than their answer's content. Therefore these techniques are not suitable for finding experts in online rated forums such as StackOverflow-forum[17], Ubuntu[18] and Yahoo Answers![19] etc. StackOverflow-forum employ a user reputation mechanism, through which, users obtain reputation score based on their post quality, up-voted answers and up-voted-questions [8]. It is observed that in StackOverflow-forum, active users respond to many programming problems. However, it may still be a challenge for the StackOverflow-forum community to produce creative or novel knowledge [9]. Users may gain greater reputation by providing frequent answers to simple questions rather than answering complex questions [144]. A better characterization of a users' expertise is required using their contribution quality rather than their reputation scores received through built-in incentivization mechanism. Currently online rated-forums such as StackOverflow-forum have no mechanism to measure the performance consistency of participants. The

[17] http://stackoverflow.com/

[18] http://www.ubuntu.com/

[19] https://answers.yahoo.com/

expert-finding technique proposed here rely on G-index [71] which is a popular bibliometric technique. G-index is applied in the social web domain to find influential bloggers [73, 74], and to determine the recognition of a researcher among her peers in open source software repositories [75]. Similarly, G-index is used to discover influential members of online communities [145] and is applied to rank video-content creators on YouTube [77]. Link-analysis techniques such as [3, 36] did not consider content quality, but combined content quality a social network authority score to improve a user's rank in online forums. Considering [3] the baseline approach, we came up with the novel features such as votes ratio, voter's reputation and tag quality using the SO-forum dataset to measure user reputation. Moreover, used the bibliometrics, such as G-index [65, 71, 146] to measure user consistency in providing quality answers.

In this module, we propose two expert-ranking techniques for the StackOverflow forum: Exp-PC and Weighted Exp-PC. Firstly, Exp-PC, an adaptation of G-index, considers the user's consistency in receiving a high reputation score on each of the provided answers. Moreover, we propose Rep-FS (user reputation features) including vote's ratio, voter's reputation, tag quality and participant reputation. Secondly, Weighted Exp-PC, an adaptation of Exp-PC with user reputation features is proposed. The experimental results of the proposed expert-ranking techniques, Exp-PC and Weighted Exp-PC, validate that these methods identify genuine experts in a more effective way.

4.2. Problem definition

In this section, we first describe basic terms used in online rated-forums and then formally define the expert ranking problem.

Terminologies:

Question: A query posted by the user in an online forum.

Answer: An answer provided by a user to a question. A question may have many answers.

Forum user: A person who may initiate or answer the questions.

Definition

Le RF be the rated-forum containing a set of questions $Q= \{q_1, q_2, q_3\ldots\ldots q_m\}$, where q_i contains the number of answers $A_i= \{ai_1, ai_2, ai_3\ldots\ldots\ldots\ldots ai_n\}$ by a particular user. Let U be the group of users. We have to find the expert users from U whose performance is consistent and whose reputation should be high. Performance consistency is measured through a bibliometric g-index. Reputation of the users is computed through their voters' reputation, neighbours' reputation and tag quality.

4.3. Baseline technique

We have chosen a link analysis based expert-ranking approach, ExpertiseRank [3], as a baseline technique. According to [3], if user A provides an answer to a question posed by user B, who is a domain expert, then it means that user A has more expertise than user B, because they answered an expert's question. Assume that user X has answered questions for users $U_1 \dots U_n$, then the ExpertiseRank of user X is given in equation 4.1.

$$ER(X) = d\left(\frac{ER(U_1)}{L(U_1)} + \dots\dots\dots + \frac{ER(U_n)}{L(U_n)}\right) \qquad (4.1)$$

ER(X) is ExpertiseRank for user X, U_l is the user who is answered by X, d is a damping factor that is set to .85, and $L(U_l)$ is defined as the total number of users who responded to U_l. According to this technique, a user will have more expertise if replying to questions posted by expert users. User rank will be decreased if he poses too many questions in the online forum.

ExpertiseRank [3], considers link-structure of the forum users where links are formed on the basis of questioning-answering relationships between users. This technique does not consider users' answer quality and their consistency in providing quality answers. As the social media applications such as forums and blogs, support users to generate their favorite content, therefore utilization of content in information retrieval tasks is of prime importance. In the social web domain, particularly in CQA sites such as Yahoo! Answers, a question receives multiple answers and the best answer is selected by the user posing the question [1, 11, 37, 88]. StackOverflow-forum has a reputation measuring mechanism, according to which users are listed according to their reputation scores. The reputation score is computed based on their cumulative past performance [8].

4.4. Proposed Expert ranking techniques

The expert-ranking technique consists of the application of G-index approaches to various reputation features to identify experts on CQA sites for programming language problems, in this case, SO-forum. In this section, expert-ranking techniques, namely *Exp-PC* and *Weighted Exp-PC* are discussed in details.

4.4.1. Exp-PC

Exp-PC is an adaptation of G-index as explained in section 2.1.3.1, a popular author-productivity measure in academic social networks. Exp-PC can be used to measure user reputation in online programming forums such a StackOverflow.

Formally, Exp-PC (an adaptation of G-index) can be written as: *given a set of posts ranked in decreasing order of the number of post scores that a user received, the Exp-PC rank is the largest number, n, such that the top-n posts received a total of at least n^2 post scores.*

4.4.1.1. User Reputation features (Rep-FS)

Expert finding is an acknowledged research problem for the social web domain, and it has been addressed using several features in different datasets. For the most part, the features are based on content quality and social network authority. For the StackOverflow forum dataset, we presented reputation features such as voter reputation, tag quality and participant's reputation in the following way:

Voter reputation (Rep-f1): In StackOverflow forum, high up-votes count of an answer shows its quality. However, voter reputation is not considered by StackOverflow which is an important aspect to determine the users' different skills, from beginner to professional. Therefore, for an answer, receiving up-votes from less reputed or beginner user is less effective than receiving up-votes from professional or real experts. To handle this problem, voter's reputation feature is proposed. Voter reputation is computed by Exp-PC technique (Section 4.4.1).

Up-vote to down-vote ratio (Rep-f2): In StackOverflow, up-votes indicate the usefulness of an answer, while down-votes show its irrelevance. This feature measures the user's consistency in gaining high numbers of up-votes and low number of down-votes [147]. This ratio is computed by dividing a user's up-votes by the down-votes. High number of up-votes indicate user expertise.

Participant-based reputation (Rep-f3): The number of answers provided for a question has been proven to be a simple but effective value in evaluating answer quality [12, 13]. In each StackOverflow conversation, multiple users may participate in discussion. In a conversation, the participation of reputed users (who are consistent in performance), made it productive. Moreover their existence boosts the rank of neighbor users [16]. Participants' reputation is computed through Exp-PC (Section 4.4.1).

Popular tags (Rep-f4): StackOverflow presented the list of available tags and their usage frequency to facilitate the browsing of questions and answers. Tag quality is an important aspect in mining user expertise in collaborative systems [54, 55]. Assigning popular tags (frequently-used tags in StackOverflow) to the question indicates a user's domain knowledge and his capability to better present his question to community [148]. Tags similarity is computed through cosine similarity technique [113].

Features' descriptions are given in table 4.1. The Rep-FS (Reputation features score) for each user is computed by combining the scores for each of the above-mentioned features (*Rep-f_1* to *Rep-f_4*), using equation 4.2.

$$Rep\text{-}FS = \sum_{j=1}^{m}\sum_{i=1}^{n}(F_i U_i) \qquad (4.2)$$

Where F_i is the feature score for user U_i.

Table 4.1: Rep-FS features descriptions

Feature No	**Feature Name**	**Feature description**
Rep-f1	Voter reputation	Reputation of user who gives Up-vote to an answer
Rep-f2	Up-vote to down-vote ratio	Ratio of up-votes to down-votes received
Rep-f3	Participant-based reputation	A discussion with reputed participants
Rep-f4	Popular tags	Total popular tags applied by the user

4.4.2. Weighted Exp-PC

In this technique, Exp-PC is extended with user reputation features described in section 4.4.1.1. The notion behind this technique is to enhance Exp-PC through StackOverflow features. The reputation of a user is considered who gives vote to some answer. This feature is proved as more logical and effective than the vote ratio feature. Weighted-Exp-PC is presented to measure user expertise in realistic way and it is computed by multiplying users' Rep-FS features scores to their Exp-PC score.

Algorithm for expert ranking in rated forums is given below,

```
ALGORITHM: Expert ranking for rated forums
Input: List of users participated in discussions
Output: Ranked list of expert users

For each user do
    Compute Exp-PC, (in section 4.4.1);
    Compute Weighted-Exp-PC, (in section 4.4.2);
End
```

The time complexity of the proposed expert ranking algorithm for rated-forums is quadratic. The big oh asymptomatic notation for the algorithm is O (n2). The time complexity of the proposed algorithm is same as that of the baseline algorithms.

4.5. Experimental Setup

4.5.1. StackOverflow dataset

StackOverflow is one of the leading online programming forums where users can post and respond to questions, and find information from previously answered questions. StackOverflow has a rich repository of previously solved questions and a moderation policy that closes or removes duplicate questions. Almost all of the popular programming languages are discussed, including Java, MATLAB, php, C#, HTML, Java Script, etc. To perform experiments for ranking experts, we used the StackOverflow benchmark dataset used in [9]. The dataset[20] is freely available for research and has been used for finding experts and quality-answer [9, 149]. The dataset is relatively large with a considerable number of users and their discussions. Statistics for the StackOverflow dataset are given in table 4.2.

Table 4.2: StackOverflow dataset statistics

Users	120148
Answers	717118
Comments	1289175
Comments per post	1.79
Average answer length	1513.5
Hyperlinks	857694

4.5.2 Performance evaluation measures

In StackOverflow forum, reputation score is assigned to each user based on their performance. The expert-ranking techniques, namely Exp-PC, Weighted-Exp-PC and baseline technique ExpertiseRank [3], have been evaluated against the StackOverflow reputation score which we considered as true value for users expertise. The performance of baseline and Exp-PC, Weighted-Exp-PC is compared through standard evaluation measures. For this purpose O_{sim} [150], Spearman's Rank Correlation [3], and Kendall's Rank Correlation [3, 141] are used. O_{sim} is used to measure the similarity between two lists or the results of two ranking methods, and it is computed by taking the users common to both lists, normalized by the number of records under consideration. We used O_{sim} to analyze the similarity of the results for the common users in our expert-ranking techniques. For two ranked lists, L_1 and L_2, O_{sim} can be computed for the top 10 results using equation 4.3.

[20] https://archive.org/download/stackexchange

$$O_{Sim} = (L_1 \cup L_2)/n \qquad (4.3)$$

Spearman's Rank Correlation, represented by ρ, is a technique to compute the association between the ranking orders on two variables [3]. In our case, correlations between the Exp-PC, Weighted-Exp-PC and the baseline expert-ranking technique are computed using equation 4.4.

$$\rho = 1 - \frac{6\sum d_i^2}{l(l^2-1)} \qquad (4.4)$$

Where d is the difference between ranks, and l is the number of items in each case. For clarity of results, we considered the top 20 and top 30 users: l is equal to 20 and 30, respectively.

Kendall's Rank Correlation is a measure that determines the strength of the dependence between two variables [3, 141]. The variation between two different ranking results is reported. Kendall's Rank Correlation is represented by τ, ranging from -1 to +1, and computed using equation 4.5.

$$\tau = \frac{(number\ of\ concordant\ pairs)-(number\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)} \qquad (4.5)$$

4.5.3 Results and discussion

In this section, we describe the results; the top-20 and top-30 users have been selected for evaluation. Correlations and similarity scores have been computed for the baseline and proposed expert-ranking techniques against the benchmark StackOverflow reputation scores. Results are presented in figures 4.1 & 4.2. It is evident from figures 4.1 & 4.2, the proposed techniques consistently outperform the baseline technique and achieve better correlation scores.

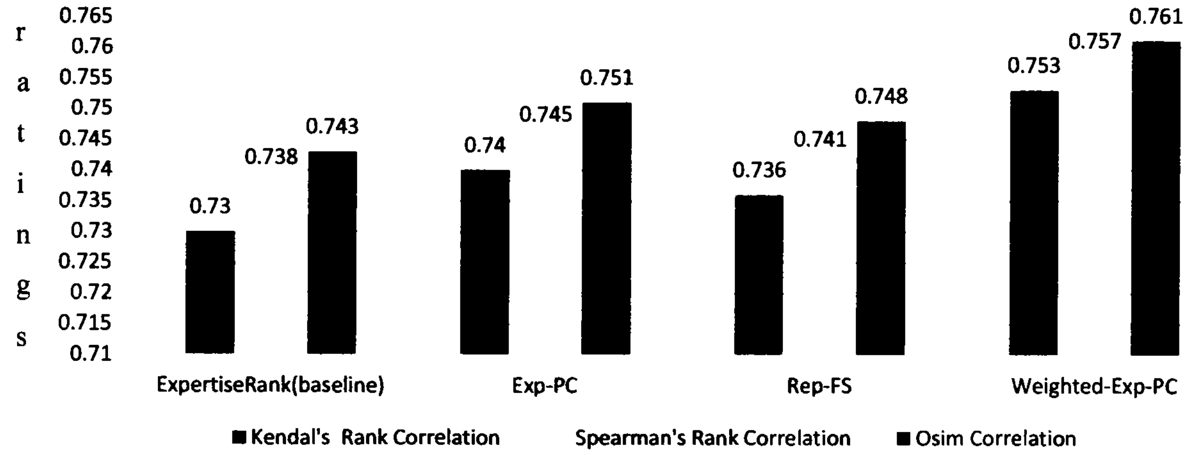


Figure 4.1: Correlations and Overlap similarity scores for the top-20 users

Figure.4.1 shows the correlation analysis of the various approaches. For example, Exp-PC-Weighted achieved high quality results with the highest Spearman's Correlation and O_{sim} scores because it combines the characteristics of users' consistence performances and their reputation features. Exp-PC achieves high correlation with StackOverflow reputation score which indicates the effectiveness of consistent performance feature.

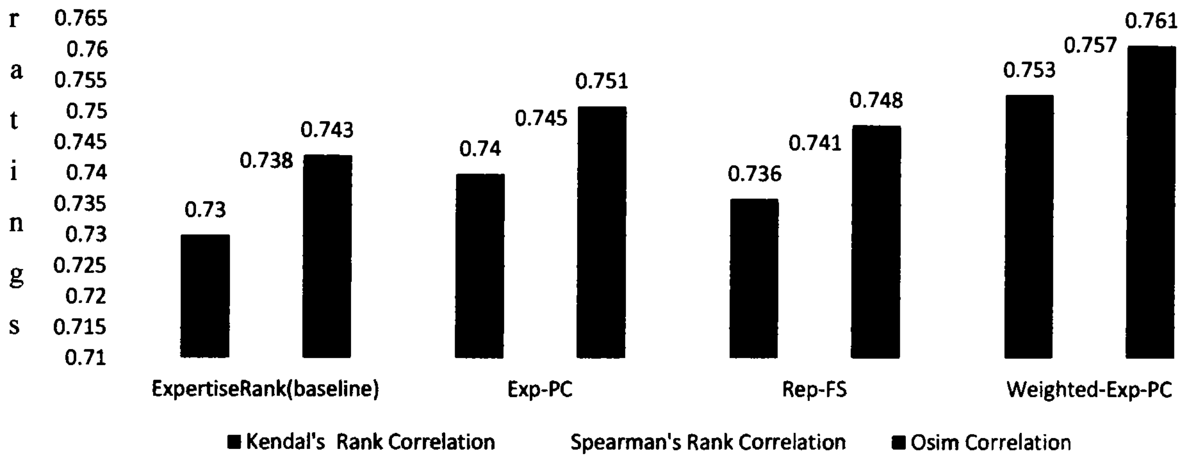


Figure 4.1: Correlations and Overlap similarity scores for the top-20 users

Exp-PC results indicate that highly ranked users are consistent in providing quality posts and their subsequent posts quality is high which is remarkable in case of online forum where users are not always expected to produce quality subsequent posts. Exp-PC is good for users because it takes into account the cumulative sum of their previous post scores which other indexes such as h-index don't. In StackOverflow case, high Exp-PC score indicates that top users have excellent skills in multiple programming languages therefore their subsequent posts are of high quality.

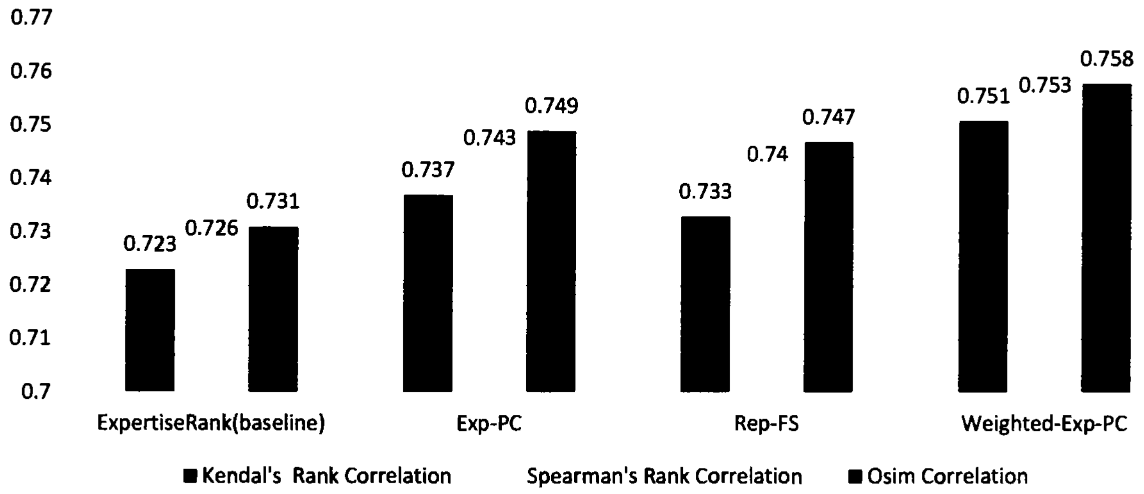


Figure 4.2: Correlations and overlap similarity scores for the top-30 users

It is observed that for both top-20 and top-30 rankings, weighted-Exp-PC technique gives better correlation scores than Rep-FS and Exp-PC technique, indicates the effect of combining users reputation features score to their consistent performance score. Rep-FS gives relatively low correlation scores from other techniques because its features are realistic than StackOverflow reputation mechanism and it improved users' rankings. Moreover, overall the ExpertiseRank (the baseline technique) did not perform well because this technique is based on the link-structure features and don't consider answer quality and users reputation and performance features.

4.5.4. User performance analysis

In this section, the users' performances are analyzed for the proposed (Exp-PC, Rep-FS and Weighted-Exp-PC) and the baseline technique (ExpertiseRank). For each technique, the top-20 users were selected for evaluation and are listed in table 4.3. We present cases for proposed techniques in which the user's ranks are improved.

From table 4.3, Jon Skeet and Marc Gravell are ranked as top users by nearly all expert-ranking techniques, which show that they are reputed users and give consistent performances. In comparison of Exp-PC and ExpertiseRank, we can see that the proposed technique Exp-PC ranked the user SLott at 7th position while the he is ranked at 12th position by the ExpertiseRank. This indicates that SLott consistently gives quality answers, while baseline technique is failed to rank him high because it only considers the link-structure features of users. Now we discuss a user Adam Rosenfeild, ranked at 19th position by the Exp-PC technique, indicates that Adam was not performing consistently and his post scores are low on subsequent answers. Although he has answered many people, therefore, he is ranked at 11th position by the baseline technique. John is another case whose rank is improved from 13th to 11th position by Exp-PC.

Table 4.3: Top 20- users ranking for all techniques

	Stack overflow- Reputation	ExpRank(Baseline)	Exp-PC	Rep-FS	Weighted-Exp-PC
1	Jon Skeet	Jon Skeet	Jon Skeet	Van Fosson	Jon Skeet
2	Marc Gravell	Marc Gravell	Marc Gravell	Jared Par	Marc Gravell
3	Van Fosson	Jared Par	Jared Par	Marc Gravell	Van Fosson
4	Jared Par	Van Fosson	Van Fosson	Jon Skeet	Jared Par
5	Cletus	Cletus	Cletus	Cletus	Cletus
6	Greg Hewgill	Greg Hewgill	Greg Hewgill	Greg Hewgill	Greg Hewgill
7	SLott	Joel Coehoorn	SLott	SLott	SLott
8	Joel Coehoorn	Mehrdad	Joel Coehoorn	Joel Coehoorn	Joel Coehoorn
9	Mehrdad	Pax	Mehrdad	Mehrdad	Mehrdad
10	Pax	Konrad Rudolph	Pax	Pax	Pax
11	John	Adam Rosenfield	John	John	John
12	Konrad Rudolph	SLott	Konrad Rudolph	Konrad Rudolph	Konrad Rudolph
13	Von C	John	Von C	Von C	Von C
14	Bill the Lizard	Von C	Bill the Lizard	Alex Martelli	Bill the Lizard
15	Scott Anderson	Bill the Lizard	Scott Anderson	Bill the Lizard	Scott Anderson
16	Andrew Hare	Scott Anderson	Andrew Hare	Scott Anderson	Andrew Hare
17	Alex Martelli	Andrew Hare	Alex Martelli	Andrew Hare	Adam Rosenfield
18	Paolo Bergantino	Alex Martelli	Paolo Bergantino	Mitch Wheat	Alex Martelli
19	Adam Rosenfield	Paolo Bergantino	Adam Rosenfield	Paolo Bergantino	Paolo Bergantino
20	Mitch Wheat	Mitch Wheat	Mitch Wheat	Adam Rosenfield	Mitch Wheat

Moreover, we have separately analyzed user performances in Rep-FS, presented in table 4.4. Now we compare Rep-FS and ExpertiseRank, the proposed Rep-FS computes users' reputation based on their voters' reputation, up-vote to down-vote ratio, participants' reputation and tags popularity scores. Major ranking differences are noted in Rep-FS rankings. For example, it can be seen in table 4.3 that, Jon Skeet, ranked at top-most position by all techniques, however, Rep-FS ranked him at 4th position. This decrease in rank is caused by Jon Skeet's voters' reputation and tags quality which can be seen in table 4.4.

Table 4.4: Rep-FS features based ranking of top-20 users

User-id	Participant Rep	Up-vote to Down-vote ratio	Popular Tags	Voter Rep
1	Jon Skeet	Jon Skeet	Marc Gravell	Van Fosson
2	Marc Gravell	Marc Gravell	Jon Skeet	Jon Skeet
3	Van Fosson	Cletus	Van Fosson	Marc Gravell
4	Cletus	Jared Par	Jared Par	SLott

5	Jared Par	Van Fosson	Cletus	Cletus
6	Greg Hewgill	Greg Hewgill	Greg Hewgill	Greg Hewgill
7	SLott	Mehrdad	SLott	Jared Par
8	Pax	Joel Coehoorn	Joel Coehoorn	Joel Coehoorn
9	Mehrdad	SLott	Mehrdad	Mehrdad
10	Joel Coehoorn	Pax	Pax	Pax
11	John	John	John	John
12	Konrad Rudolph	Konrad Rudolph	Konrad Rudolph	Konrad Rudolph
13	Von C	Von C	Von C	Von C
14	Alex Martelli	Bill the Lizard	Alex Martelli	Bill the Lizard
15	Bill the Lizard	Scott Anderson	Bill the Lizard	Alex Martelli
16	Scott Anderson	Andrew Hare	Scott Anderson	Scott Anderson
17	Andrew Hare	Adam Rosenfield	Andrew Hare	Andrew Hare
18	Mitch Wheat	Alex Martelli	Mitch Wheat	Mitch Wheat
19	Paolo Bergantino	Paolo Bergantino	Paolo Bergantino	Paolo Bergantino
20	Adam Rosenfield	Mitch Wheat	Adam Rosenfield	Adam Rosenfield

From table 4.4, Jon Skeet performance in Rep-FS supports our notion that voter's reputation must be considered while ranking him because all the voters are not reputed users and may be a beginner or novice users who may not judge the answer quality in true sense. While a real expert or professional user's vote matters. Therefore a user with up-votes from reputed users is a real expert. StackOverflow does not consider voters reputation due to which its rankings are not much reliable. In voter rep feature, shown in table 4.4 & 4.4, VanFosson is another case who is ranked at 1st position by Rep-FS while he is ranked at 3rd and 4th positions by other techniques. This again validates our notion that voters' reputation of the user effects user's ranking. It is shown in table 4.3 that, ExpertiseRank ranked Alex Martelli at 18th position but it is shown in table 4.4 that Rep-FS brought him at 14th position which is a significant change in rank. Jared Par and Marc Gravell ranks are also affected by Rep-FS.

In comparison of Rep-FS with StackOverflow reputation score, we can see in table 4.3 that VanFosson's rank is improved by Rep-FS while Jon Skeet's rank is decreased. Alex Mirtelli rank is improved by Rep-FS. Above three cases shows that benchmark StackOverflow reputation mechanism is not fully capable of finding real experts. By incorporating Rep-FS features, StackOverflow mechanism may be improved. We have seen techniques such as Exp-PC and weighted Exp-PC have given quality rankings, however the Rep-FS is found to be the most effective technique.

The overall results support the fact that consistently providing better answers is an important factor in finding real experts. Moreover, voter reputation feature is found to be a significant feature in discovering experts. Moreover, the link-structure features, when combined with consistence performance element and voter reputation features, give better user rankings.

Chapter Summary

We studied the expert-ranking problem for the StackOverflow forum and propose Exp-PC, Rep-FS and Weighted-Exp-PC for such problem. Basically, we considered the element of users' consistency in providing quality answers. In contrast to classical link-structure based expert ranking techniques, which are normally based on users' social network authority or answer quality score, our techniques such as, Exp-PC, Rep-FS and Weighted-Exp-PC performed well and improved user ranking. To evaluate our techniques, standard performance evaluation measures of a benchmark StackOverflow dataset were used for comparative analysis. Experimental results confirmed that our techniques identify the experts in a more effective manner and show better performance after extensive testing. The experts identified in the results are plausible based on their consistent performance in productive discussions in an online programming forum.

Chapter 5
Thread ranking techniques for Non-rated discussion forums

5.1. Introduction

Online discussion forums are a valuable source of knowledge. These platforms can be particularly useful for users who have a keen interest in a particular subject and are searching in depth for specific or expert information [1]. Forum threads may contain factual or opinionative knowledge and can be an important source of information where technical or domain specific topics are discussed. With the growing volume of content in online forums, finding quality answers and relevant information turn out to be a challenging task.

Online discussion forums such as BBC, provides keyword-based search services, where users may search or browse the topics of their interests. In keyword-based search, on searching some topic, the forum search engine returns a list of similar threads to a given query. However, the quality of retrieved threads is not up to standard because keyword-based techniques don't consider contents semantics like synsets. In most cases the keyword-search is accomplished through cosine similarity technique [113]. Earlier research efforts for improving the performance of thread retrieval were primarily based on content similarity and thread structure features. In most of the thread ranking techniques [11, 14, 15, 17, 151], content similarity is measured through cosine-similarity techniques [152] . A major limitation of the cosine-similarity technique is that it is based on lexical overlap between documents. However, semantics for similarity are not considered [153-155]. Documents are semantically relevant to each other due to common terms and thus related. Semantically similar concepts may be expressed in different words in the documents and the queries. For that reason, direct word to word comparison by cosine similarity techniques [113, 152] is not effective [155]. For instance, cosine similarity technique cannot recognize synonyms or semantically similar terms (e.g., "tennis-ball", "sports"). In case of online forums, often, posts terms may have multiple meaning (polysemy), however cosine similarity technique only consider the lexical overlap of terms, thus ignore semantics. By ignoring semantics, context of the term in that post is not captured. For example, two threads having different terms with similar meaning cannot be interpreted as common threads by cosine similarity techniques. Moreover, a word may be described in many ways (synonym) and different terminologies are used by the documents to describe these words. A query based on the terminology of one document will not retrieve other related documents. Due to these issues, even the threads contain the relevant terms for a query but they are not retrieved as a relevant thread for a given query. Another problem of cosine similarity technique is that it involves many computations like term frequency count, document frequency, inverse document frequency and term weights. Computations like these take significant processing time and space when applied to big data. Moreover, cosine similarity techniques do not consider phrase structure, proximity information and word order

Increasing use of online discussion forums for finding details on specific topics, thread ranking problem has gained importance. Relevant threads for a given query can be better retrieved through semantic similarity techniques as they consider the meaning and relation of terms rather than simple lexical overlap between them. As semantic similarity techniques are successfully applied in information retrieval domain, therefore these techniques can also retrieve more relevant threads for a given query. Computing semantic similarity between a query and thread's posts can lead to better posts ranking. WordNet[21] is a lexical database for English language with controlled vocabulary and thesaurus. This database offered a taxonomic hierarchy of natural language terms, developed by Princeton University [155]. WordNet can frequently be used for computing semantic relatedness among documents [18, 19, 155, 156]. Several semantic similarity techniques [18, 19, 155] are implemented in WordNet database.

In this thesis, we moved, one step forward, on thread ranking problem by considering structure, semantics of the thread's content and participants' reputation. Considering aspects such as structure, content semantics and reputation, better thread are retrieved for specific topics. In [114, 157-159], various possible ways on thread ranking/retrieval are presented using cosine similarity techniques for ranking threads. To the best of our knowledge, the maximum efficacy of thread ranking has not been taken into consideration. To provide a solution, we proposed ways of improving thread ranking by using semantic similarity, thread sub-structures and user reputation features. An edge counting based semantic similarity method [100, 102] is selected for thread ranking which is a widely used technique for information retrieval and document clustering tasks [19, 99, 103-105, 110, 153, 155, 156, 160, 161]. In this perspective, our major contributions are as follows:

1. Computing semantic similarity (cosine similarity vs. WordNet-based similarity) for ranking threads.
2. Different ways of aggregating post similarity up to the thread level.
3. The involvement of thread features characterizes the overall usefulness of the thread.
4. Comparison between Semantic similarity and cosine similarity techniques for thread ranking over real BBC discussion forum dataset

5.2. Problem definition

In this section, we first describe the basic terms used in online discussion forums and then formally define the thread ranking problem.

[21] https://wordnet.princeton.edu/

Terminologies:

Thread: Thread is a question or it may be a topic initiated by a user in an online forum.

Subject: Each thread has a subject usually represented by a title that appears on top of the thread.

Post: Post is a reply or an answer provided by a user to a thread. A thread may consist of many posts.

Definition

Let NF be the non-rated forum containing a set of threads $T= \{t_1, t_2, t_3 \ldots\ldots t_m\}$, where t_i contains the number of posts $P_i= \{pi_1, pi_2, pi_3 \ldots\ldots\ldots\ldots pi_n\}$ where Pi be the collection of posts by a particular user. We have to find the high quality threads for this user. Quality of a thread are measured through semantic similarity score between thread subject and posts content. Moreover, positive sentiments, url-references and reputed participants indicate threads quality.

5.3. Baseline techniques

Cosine similarity is the most popular measure [113] for estimating document similarity based on Vector Space Model. The similarity between a query and thread q_i and t_j can be defined as the normalized inner product of the two corresponding vectors v_i and v_j as is given in equation 5.1,

$$Sim(q_i,t_j) = \frac{v_i.v_j}{|v_i|*|v_j|} = \frac{\sum t_{x\in u}\left(wt_x(q_i)*wt_x(t_j)\right)}{\sqrt{\sum t_{x\in d_i} w^2 t_x(q_i)*\sum t_{x\in d_j} w^2 t_x(t_j)}} \qquad (5.1)$$

Where $u= (q_i \cap t_j)$ i.e., common terms of query q_i and thread t_j; $wt_x\,(q_i)$ and $wt_x\,(t_j)$ are the weights of term t_x in query q_i and thread t_j respectively.

The problem of recognizing similarity among threads has been addressed computationally using cosine similarity [113], applied to baseline methods including bag-of-posts [114], head-post [157] , maximum similarity between posts [158] and central post similarity [159]. These techniques [114, 157-159] have been used by [17] to compute thread-thread, thread-post and post-post content similarity. We have extended the baseline techniques [114, 157-159] , as used by [17], to address the problem of query-based thread ranking using a WordNet based semantic similarity [102]. Likewise, thread post content quality and participants' reputation are also used to enhance thread retrieval.

6. Proposed Framework

Thread ranking in online forums can be classified in two categories. First, *T-SimRank* is a semantic based similarity for improving thread ranking by considering the meanings of terms. Semantic relationship between words can be measured by using WordNet's edge counting [102]. WordNet is a lexical database for the English language [162]. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members [99]. WordNet can be seen as a combination of dictionary and thesaurus. Different variations of WordNet[22] are used to compute semantic similarity using approaches like node-based, edge-based and hybrid methods for similarity computation [19]. Given a query and a thread, WordNet determines the similarity of the two documents in terms of sense (synset) and relationship. A high semantic score for a given query and post infers that the context of the two documents is the same. Computation of semantic similarity involves processes including the tokenization of sentences, part of speech tagging, stemming, determining the sense of every word in a sentence and computing the similarity of queries to threads based on pairs of words. Leacock and Chodorow [100, 102] took the maximum depth of taxonomy into account and computed semantic similarity between query q and thread by equation 5.2.

$$Sim_{sem}(q,t) = -\log \frac{length(q,t)}{2 * deep\_\max} \qquad (5.2)$$

Here, *length (q, t)* is the shortest path between q and t and *deep_max* is the maximum depth of the taxonomy. *Sim_{sem} (q, t)* is the semantic similarity between a query terms q and a thread terms t which lies between 0 and log *(2deep_max+1)*. Similarity is computed based on the shortest path between the synsets associated with query-term q and thread-term t [100, 102].We computed similarity of a query and a thread (both which consist of multiple terms) by taking the average shortest synset distance between all the terms in the query and the thread. If terms of q and t have the same sense, then *length (q, t) =0.* In practice, we add 1 to both *length (q, t)* and 2*deep_max* to avoid *log (0).*

Second, *T-CRRank* anticipated for better content quality and the existence of reputable participants in a thread for ensuring relevancy in depth. Therefore we used several content quality and participant reputation features to evaluate thread's usefulness.

Architectural depiction of thread ranking technique's including baselines (cosine similarity), proposed technique (Semantic similarity through WordNet database) and participant reputation, post quality features is given in fig.5.1.

[22] http://wordnet.princeton.edu

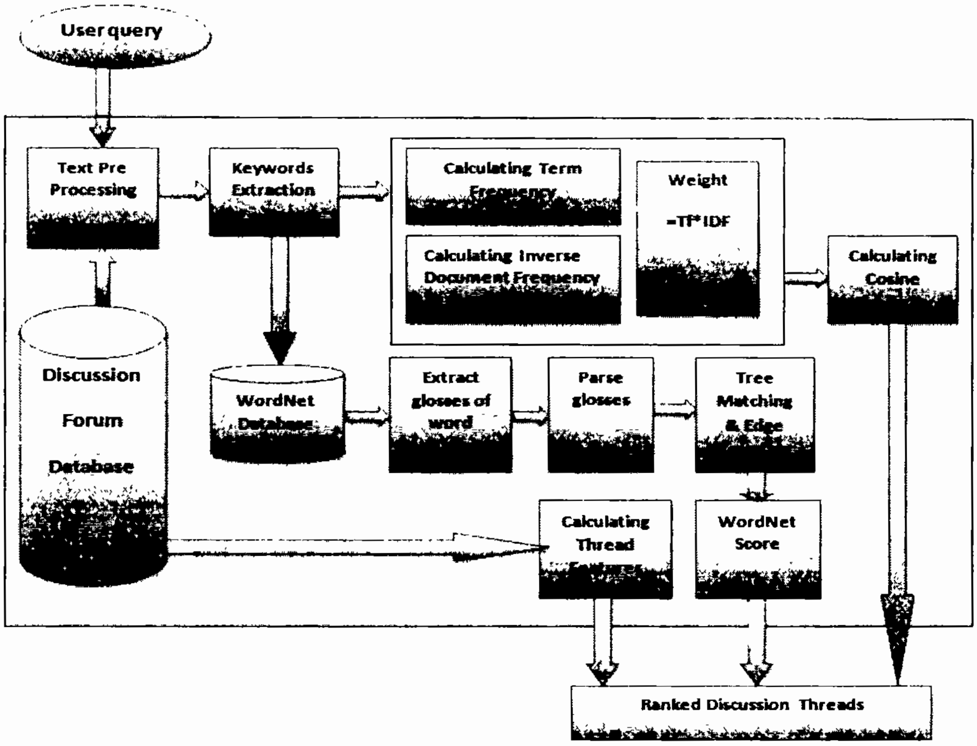


Figure 5.1: Proposed thread ranking system architecture

In following sections, we describe T-SimRank and T-CRRank techniques respectively.

6.1. T-SimRank

T-SimRank approach is anticipated for ranking threads based on their semantic similarity score in relation to a query. Threads sub-structures such as, initial/starting post, middle post, combination of initial post and title has been considered. A thread is considered relevant and will be ranked high if its contents are semantically similar to a given query. Our extended techniques with semantic similarity technique are discussed in following sub sections.

6.1.1. Bag of Words

In information retrieval, documents can be thought of as a collection of words, or 'Bag of Words' (BOW) [163]. In this work we consider a thread as a document and the associated collection of posts as words [114]. The similarities between query and threads can be estimated as the similarities between documents as computed by equation 5.3.

$$BOW(q, t) = Sim_{sem}(q, t) \qquad (5.3)$$

Where *BOW (q, t)* considers all posts within the thread collectively and Sim_{sem} represents the semantic similarity scores of the query with respect to the combined contents of the thread.

6.1.2. Bag of posts (BOP)

BOP technique considers threads as a collection of elements and posts [114].To calculate query-thread semantic similarity, *BOW* method is applied. Post-wise *BOW* similarities are then aggregated by computing the arithmetic mean to obtain a single score for each thread, which is used for thread ranking and given by equation 5.4,

$$BOP\ (q,\ t) = Mean\ \{BOW\ (q_i,\ t_j)) \mid q_i \epsilon Q,\ t_j \epsilon T\} \qquad (5.4)$$

Where Q and T represents query and thread respectively.

6.1.3. Head-Post

Head-Post combines the significance of a thread's title and its first post. The thread title is important as it draws the user's attention and influences them to participate in discussion. Generally, the initial post of a thread gives an insight about thread's topic/question and succeeding posts normally hold answers or clarifications. We assumed that the thread's title and body of the initial post of a thread can be considered as the outline of the discussion and is therefore highly representative of the thread's content [157]. The estimation of the degree of relation between a query and thread can be expressed as a function of the relevancy between the query and the head-post (thread's title and first post) content expressed in equation 5.5.

$$Head\text{-}Post\ (q,\ t) = Sim_{sem}\ (q,\ head\text{-}post) \qquad (5.5)$$

Where q represents user query t represents thread and *head-post* combines thread title and its first post. Sim_{sem} is the semantic similarity function used to compute similarity between query and head-post.

6.1.4. BOP-HPost

In BOP-HPost, we linearly combine the Bag of Posts and Head-post methods as in [17] to analyze their impact in combination, given in equation 5.6.

$$BOP\text{-}HPost\ (q,\ t) = BOP\ (q,\ t) + Head\text{-}Post\ (q,\ t) \qquad (5.6)$$

6.1.5. Maximum Similarity between Posts

Online forums contain content from users with varying skill levels, content quality is a major issue. BOP method is based on the arithmetic mean of the similarities of all the posts, building on an implicit assumption that the thread is free of noisy or irrelevant posts. In fact, noisy posts and even spam posts often exist. As these noisy or spam posts will not have a high similarity score with the user's query, therefore their presence in a thread may decrease the thread ranking score and relevant information can be missed. To avoid this problem, the post with the highest semantic similarity score in a given thread is selected to produce the overall score, given in equation 5.7.

$$MAX\ (q,\ t) = max\ \{BOW\ (q_i,\ t_j)\ |\ q_i \in Q,\ t_j \in T\} \quad (5.7)$$

6.1.6. Top-k Similarity

Top-k method computes the mean score of the top-5 similar posts for a given query [17] as given in equation 5.8.

$$Top\text{-}k\ (Q,\ T) = mean\ (top\text{-}k\ \{BOW\ (q_i,\ t_j)\ |\ q_i \in Q,\ t_j \in T\}) \quad (5.8)$$

We have selected top-5 posts because most of the useful discussions are occurred in top posts [164].

6.1.7. Central Post Similarity

This method estimates the relevancy between the middle post of a thread ($t_{central}$) and a query (Q) [17]. The use of the central post of a discussion thread is analogous to the K-medoid clustering approach [159] as given in equation 5.9.

$$Central\ (q,\ t) = BOW\ (q,\ t_{central}) \qquad (5.9)$$

6.1.8. Sum-of-All

We proposed two hybrid techniques to sum the scores though CombSum method [165-167] of the baseline methods, discussed in Section 6.1.1 to 6.1.7. Firstly, for ranking threads, we used a *tf-idf* based cosine similarity measure [113], the vector space model (VSM). Secondly, we applied a WordNet (W-Net) based semantic similarity method [100, 102].

6.2. T-CRRank

T-CRRank is based on ranking threads according to their post quality and participants' reputation. *T-CRRank* is to favor threads with: (a) posts written by "experts", (b) posts expressing a positive sentiment (possibly evidence of successful answers), and (c) posts count for each thread.

6.2.1. Post Quality Features

A thread can be regarded as significant or productive if it maintains a certain level of quality in the posts. Following features have been identified;

Post count: Answer length is proven to be a simple and effective indicator of answer quality [14, 168]. The post count indicates the degree of participation in a thread and it is expected that a thread with several posts carries sufficient knowledge and has productive discussion material.

Total clarification posts: Lesser the number of clarification posts in a thread, corresponds to a higher level of thread productivity because existence of clarification posts in a thread indicates uncertainty in answer [168].

Count of 5W-1H words: This indicates how many rhetorical terms such as, "what", "why", "who", "where", "which" and "how" are used in a thread. The existence of such words suggests that the question has not been fully resolved or is not understandable and further clarification is required [56, 121]. Thus, the thread is unlikely to contain the correct answer to a given user query.

Sentiment Score: Users may express their opinions in online forums by reflecting their level of agreement with answers [168]. Hence, a high level of positivity can indicate that a question has been successfully answered and a consensus has been reached. Ranking threads based on content quality features is computed by equation 5.10,

$$\textit{Thread-Rank}_{cq} = \textit{Post-count} + \textit{Clarific-posts-count} + \textit{5W-1H-count} + \textit{Pos-sentiment} \qquad (5.10)$$

Where, *Post-count* represents total posts in a thread, *Clarific-posts* represents a count of clarification posts within the thread, 5W-1H represents Five Ws and *Pos-sentiment* represents the positive sentiment score of a thread.

6.2.2. Participant Reputation Features

Thread participants can be categorized as follows:

- *Expert participants*: Counting how many members one helps could be a better indicator than enumerating the count of replies [1, 56]. An expert is a user who provides several quality answers on rarely ask questions in a forum. Thus the productivity and reliability of a thread is expected to increase with the number of expert participants. Answer count feature is used to measure frequent answers by a user and answer quality is measured through its content quality. For a user, if the semantic similarity score between a query and his average posts is high (75% in this case) then the answer is considered as quality answer.

- *Askers-only*: Askers-only are the novice users. They always post questions with few or without any answer [13]. They are not considered experts, so are expected to contribute at a basic level and their existence within a thread is assumed to effect the overall quality of discussion.

- *Unique Users*: A greater number of unique users [12, 168] contributing to a thread indicates a greater source of opinions and ideas, increasing the quality of the thread. Ranking threads based on participant's reputation features is computed in equation 5.11,

$$\textit{Thread-Rank}_{pr}= \textit{Exp-count} + \textit{Asker-count} + \textit{Unique-users} \qquad (5.11)$$

Where, *Exp-count* represents the total number of expert participants in a thread, *Asker-count* represents the count of users who only ask questions or provide few answers and *Unique-users* represents the total number of distinct users in a thread.

The *T-CRRank* score is computed in equation 5.12, by combining a thread's content quality (equation 5.10) and participant reputation (equation 5.11) scores.

$$\textit{T-CRRank}= \textit{Thread-Rank}_{cq} + \textit{Thread-Rank}_{pr} \qquad (5.12)$$

Proposed aalgorithm for thread ranking techniques in non-rated forums is given in table 5.1,

Table 5.1: Algorithm for thread ranking for non-rated forums

ALGORITHM: Thread ranking for non-rated forums

Input: List of threads and a query set

Output: Ranked list of threads for each query

Q: Query set

T: List of threads

Posts-count (p): Total posts within a thread

Clarification-posts (cp): Total clarification posts/questions within a thread

5W1H (wh): Rhetorical words count in a thread

```
Sentiment Score (sem): Sentiment words count within a thread
Experts (exp): Number of experts who answered in a thread
Askers (ask): Number of users who only asked questions
Answerers (ans): Users count who answered the questions
Unique-users (u): unique users count within a thread

For each query ($q$) in Q
Begin
    For each thread (t) in Threads (T)
    Begin
        Score1 = $Sim(q_i, t_j) = \frac{v_i . v_j}{|v_i| * |v_j|}$   // for techniques described in section 6.1.1 to 6.1.8, calculating cosine similarity between a query and thread (using equ. 5.1)

        Score2 = $Sim(t, q) = max[-\log \frac{length(t,q)}{2D}]$   // for techniques described in section 6.1.1 to 6.1.8, Calculating semantic similarity using WordNet (using equ. 5.2)

        Score3 = $\sum(p, cp, wh, sem, exp, ask, ans, u)$  // calculating sum of post and participants features (using equ. 5.12)

        Thread-rank= score1 + score2 + score3
    End
End
```

The time complexity of the proposed thread ranking algorithm is quadratic. The big oh asymptomatic notation for the algorithm is O (n2). The time complexity of the proposed algorithm is same as that of the baseline algorithms.

7. Experimental setup

7.1. Dataset and Performance measure

We have used the BBC Message board's discussions public dataset from cyberemotions[23]. The BBC's discussion forum is a platform for the discussion of current affairs around the world. It enables users to distribute information and provide answers to the questions posted by other users. Several users participate in discussion with each other by posting questions and answers. BBC forum data is organized into several categories such as news, religion, ethics and tv/radio. Each category has several threads. All the threads and posts in the dataset are well processed; every thread contains primary information such as the title of the thread which may be a question or a shared topic/idea by the user, initializing post, number of replies, replying posts, user information, and categories of threads. The BBC forum dataset encompasses threads collected over a time span of four years, with 80,000 threads, 1,925,200 posts and 19,500 users.

To generate a query set that reflects the scenario of BBC forum search, firstly, we have extracted a set of keywords from BBC's thread's title using n-gram techniques [169-172]. Secondly, using these keywords,

[23] http://www.cyberemotions.eu/data.html

we crawled Yahoo! Answers categories such as UK news, World news, religion, ethics and tv/radio to find similar questions/queries to these extracted keywords. Randomly 30 posted questions have been selected whose words are appeared in BBC forum's threads and posts content. For example, in BBC forum, we found a thread's keywords such as "coverage, political parties and UK newspapers". We found an equivalent posted question in UK news category in Yahoo Answer. Sample queries from Yahoo! Answers are given in table 5.2.

Table 5.2: Yahoo! Answers sample queries

Category: News
Question 1: What is the real American Death toll in Iraq????
Question 2: Does The Situation in Iraq Qualify as a Civil War?
Question 3: Can England Really Win World Cup 2010?

Category: Ethics
Question 1: Capital Punishment: Still a good idea?
Question 2: Why is religion concerned with ethics and ethical behavior?
Question 3: Is Christian ethics of no use when addressing a modern moral dilemma?

Category: Religion
Question 1: Why is there sharia law in the UK?
Question 2: Jesus in History Books?
Query 3: Male Atheists: Why don't you join the Reorganized LDS Church?

Category: TV/Radio
Question 1: In future elections, should all media coverage be banned?
Question 2: Has the BBC shown courage by removing restriction regarding depictions of Mohammed, and airing controversial cartoons?
Question 3: Do you enjoy watching sporting events on TV?

Using techniques (section 6.1.1 to 6.1.8), for each question/query we have retrieved top-30 threads from BBC dataset. BBC discussions forum is a non-rated and does not contain any criteria to rate threads and posts. As there was no labeled data for the evaluation of retrieved thread quality and relevance in BBC forum dataset, therefore we adopt human judgement criteria [135, 173] for labeling retrieved threads and their posts. For human judgement (annotation) of threads, we avail the services of two experts for thread labeling task. The annotator's task was to label the threads as relevant and non-relevant for a given query. In past, several attempts have been made to label online forum/community question answering sites threads based on their quality [13, 18, 135, 173] . A thread may be considered as less useful due to several reasons such as, it doesn't address a specific problem or the answer description may unclear [135]. Moreover it may contain spam or abusive text. While a thread is assumed to be helpful and productive if it contains relevant information. In our BBC dataset case, following the existing approaches [135, 173], we have defined a criteria to identify relevant threads based on task orientation, relevance, spam, problem type and solvedness [135].

(1). Task orientation: A thread is task oriented if it focuses on solving a specific problem.

(2). Relevance: A thread is significant if its post's contents are relevant to the question.

(3). Solvedness: If a useful and complete solution is provided for the question or topic.

(4). Spam: A spam thread contains irrelevant or abusive content, sometime it is blank.

(5). Problem type: Keywords provision in thread's content describe the thread subject and gives topic overview.

Following the judgment criteria of [173], we ask the annotators to judge the relevance of each thread. The annotators were asked to rate the "Task orientation", "Completeness" and "Solvedness" of each thread based on a five point scale, with a score of 5 indicating a high degree of fit, and a score of 1 indicating a low degree of fit. The mean numeric value across the 2 annotators was used to derive the gold-standard value for three tasks. A thread with an average rating of 3.5 or above is considered as relevant thread for a given query. The correlation between the annotators is measured using Kendall's Tau [174]. Kendall's tau coefficient is a statistic used to measure the association between two values, as given in equation 5.13.

$$\tau_A = \frac{n_c - n_d}{n_0} \qquad (5.13)$$

Where n_c is the number of concordant pairs, n_d is the number of discordant pairs and $n_0 = n\ (n\text{-}1)/2$. The Kendall's Tau distance is 0.823 which is sufficiently high rate of intra-annotator correlation.

7.2. Results and Discussion

We have compared several methods of thread quality evaluation based on cosine and semantic similarity techniques, comparing techniques both separately and in combination. A comparison of methods is shown in figures 5.2 to 5.4, where cosine-similarity techniques are labelled as cosine-sim, semantic similarity techniques are labelled as W-Net (WordNet) and the thread's content quality and participant reputation features are labelled as thread features. We have evaluated three different components: (1) cosine vs. WordNet similarity, (2) different ways of combining the similarities between the query and the posts in the thread (e.g., the similarity with the first post, all the posts, the most similar post, the top-k most similar posts, etc.), and (3) the contribution from thread quality evidence. Plus symbol is used in several diagrams to represent the hybrid techniques which combined (summed) scores of multiple techniques into single one. A standard evaluation measure, mean average precision (MAP) [175] is used for evaluation of the top-k (30 in our case) retrieved threads against each query. The mean average precision for a set of queries is the mean of the average precision scores for each query, given in equation 5.14.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \qquad (5.14)$$

Where Q is the number of queries. In our case higher MAP score means better performance. We have discussed the performance of the cosine similarity method, followed by the results of the semantic similarity method and a comparison of the two. In most of the cases, the proposed method (section 6.1 &6.2), which utilizes semantic similarity technique (WordNet) and various thread features, achieves higher MAP value than the baselines.

Results in figures 5.2 to 5.4, reflect that the Top-k (cosine-sim) and (MAX (cosine-sim) + thread features) methods outperformed other cosine-sim techniques, indicating a high lexical overlap between query and top posts. However improved results were achieved by combining the sum of the cosine-sim, W-Net techniques and thread features, indicating the importance of semantic aspects and participant's features. In figure 5.2, we can see a performance gain from using both (Top-k (cosine-sim) and MAX (cosine-sim)) when tested in combination with thread features. The results of the semantic similarity techniques are illustrated in figure 5.2, showing that the Top-k (WordNet) method combined with thread features outperformed all other techniques. This is expected as top-k only consider similarity between query and top posts content of a thread whereas thread features considers post quality and participant reputation features, by combining them , better rankings are achieved. Moreover, by considering top-k posts for each thread, low similar posts are discarded leading to improving thread retrieval. It can be seen in figure 5.2, that (MAX (cosine-sim) + thread features) and Max (WordNet) do not show significant differences in performance which indicates that cosine similarity decreases the effect of thread features.

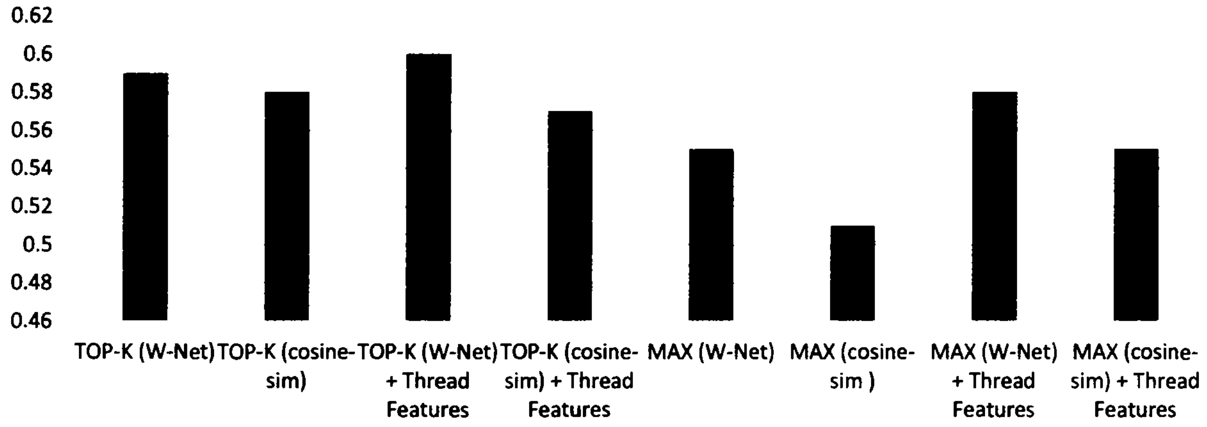


Figure 5.2: Mean Average Precision score for Top-k and MAX

In figure 5.3, the similar results for BOP (WordNet) and (BOP + HeadPost (WordNet)) indicates that the overall post content quality is comparable. Nevertheless, HeadPost (cosine-sim) performed below average which might be due to the content quality of initial post or inappropriate thread title. HeadPost represents the Initial post text and thread title. It can be seen in figure 5.3 that HeadPost (WordNet) improves the performance slightly over the (HeadPost + thread features) setting which means that reputed participants replied in lower posts. BOP with thread features have given similar performances for both cosine and WordNet techniques.

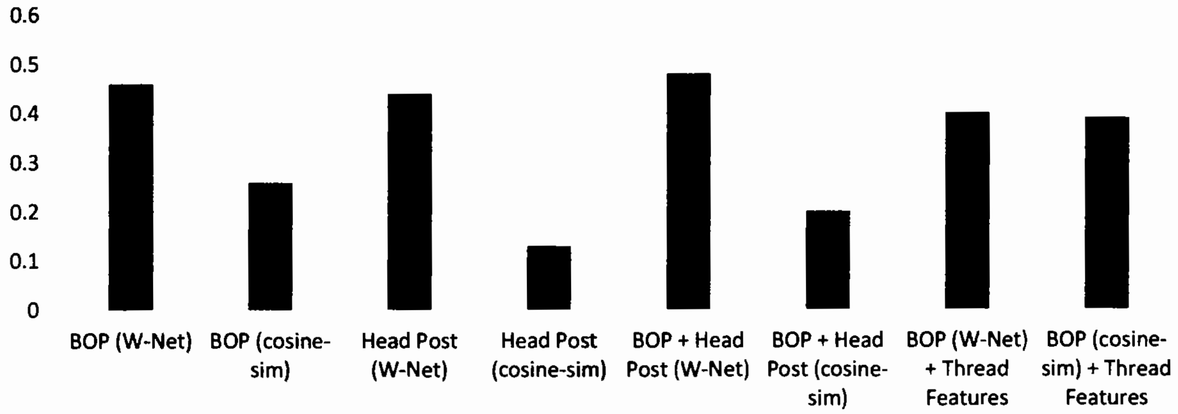


Figure 5.3: Mean Average Precision score for Bag of Posts and HeadPost

In figure 5.4, it was found that considering thread features with the sum of cosine-sim and WordNet actually decrease the MAP due to the existence of participants with poor reputations and low content quality. Techniques described in section 6.1.8, such as Sum (All-W-Net-sim) and Sum (All-cosine-sim) represent the sum of all methods using WordNet and cosine similarity techniques. Sum (All-W-Net-sim) performed well than Sum (All-cosine-sim) which validates our notion that semantic similarity techniques are significant in finding relevant threads.

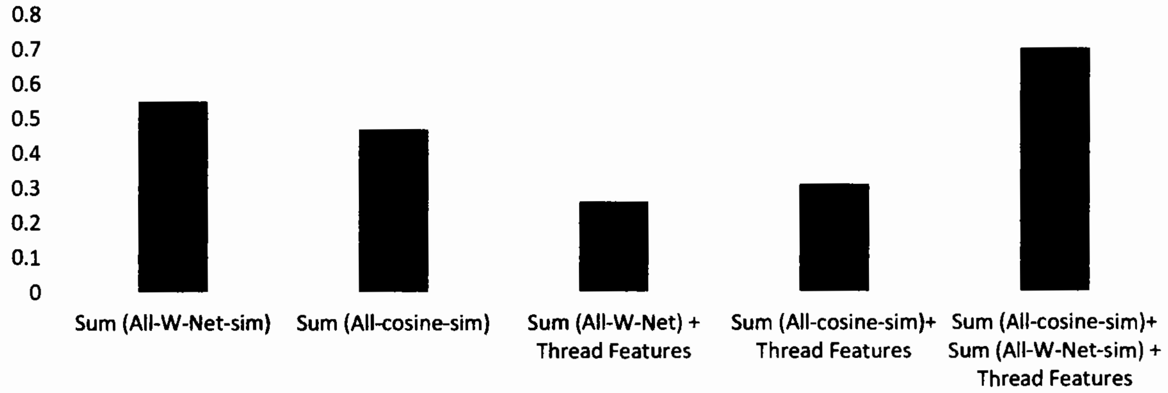


Figure 5.4: Mean Average Precision score for W-Net, Cosine similarity and Thread features (Combined)

Comparing the baseline methods, the semantic approach consistently outperforms the cosine-sim equivalent. It was also found that the threads with more reputed participants and posts, tended to be more productive and have a greater chance of being of a high quality.

Chapter Summary

In this chapter, we have investigated the impact of semantic similarity technique, answer quality and participant's reputation on thread ranking in online forums. Finding relevant threads for a query has various applications such as providing links to similar threads and providing a clustered interface of related threads for a given query. The primary objective was to improve the thread retrieval process by considering the context of a thread's content. In contrast to cosine similarity techniques, which consider lexical overlap of content, our techniques T-SimRank and T-CRRank considers both threads' semantics and participant's reputation in evaluating threads' quality. By demonstrating the utility of our approach using a real BBC discussion forum, we found that the rankings produced through T-SimRank and T-CRRank have better quality than methods based solely on lexical overlap (e.g., cosine similarity technique).

Chapter 6:
Conclusions and Future directions

6.1. Conclusions

In this thesis, we studied several social web application such as expert ranking in online discussion forums (non-rated and rated) and thread ranking in forums. We propose novel ways of using such features which have not fully exploited by social web domain for information retrieval and recommendation applications. For expert ranking in online non-rated discussion forums such as BBC message boards, we proposed expert ranking techniques. Firstly, for BBC discussion forum we proposed features such as, user participation activity, content relevancy and Co-existing users' reputation for expert ranking. Our techniques performed better than link-graph based techniques [3, 4], indicates that co-existing users' reputation and answer quality are better indicators of measuring user reputation. Experiments showed that proposed techniques ExpRank-CRF, ExpRank-COM, ExpRank-AQCS and ExpRank-FB based on aforesaid features have improved expert rankings.

We have proposed expert ranking techniques for rated-forums like StackOverflow. We argued that link-structure based techniques are not well suited for expert ranking problem in rated-forums. Moreover, current StackOverflow reputation mechanism is not fully capable of finding real experts. We have applied bibliometrics like g-index on users' posts score to compute users' consistency in providing quality answers. Novel reputation features have been extracted from StackOverflow dataset. Experiments showed that our techniques Exp-PC, Rep-FS and Weighted-Exp-PC outperformed to both link-structure and StackOverflow, reputation mechanisms.

In the last, we have evaluated the impact of semantic similarity technique, user participant and post quality features on thread ranking in BBC discussion forum. In contrast to cosine similarity techniques which only considers the lexical overlap between terms, semantic similarity techniques consider the context/meaning of terms which is helpful in retrieving relevant threads for a given query. By applying WordNet based semantic similarity function [102], participant reputation and answer quality features, better thread rankings are achieved. By combining cosine similarity and semantic similarity techniques we achieved significant results.

6.2. Future lines of research

There are several directions in which our work can evolve. In this section, we present some possible future directions.

6.2.1. Expert ranking

- *Applications of Evolutionary algorithms*: For BBC and StackOverflow discussion forums, our features combination strategies such as ExpRank-CRF and ExpRank-AQCS are simple and straightforward. We are currently reviewing to model our expert finding problem as optimization problem through more powerful fusion strategies such as particle swarm optimization and genetic algorithm to automatically design and fine tune the fusion strategies.
- *Community detection in StackOverflow forum*: We are investigating to address the community detection problem through co-existing users' data. It can be achieved in two ways. Firstly interactions between forum users (as an answer provider and askers) will be used to construct a social network. Social network strength (relationships between users) can suggest the users who collaborate most of the time on specific topics. Secondly the topic similarity will be computed among co-existing users posts. Expert communities may be formed by combining social network strength and content similarity of users for a given topic.
- *Expert databases for Organizations*: Currently expert ranking techniques are proposed for online discussion forums. These techniques can be extendable to organizations for building expert databases. Moreover we also plan to apply our techniques to some related areas such as mining influential persons and opinion leader identification.

6.2.2. Thread ranking

- *Question quality*: We have recently started to work on assessment of question quality because a topic-specific and well written question attracts more people to answer and it may receive quality answers. Therefore we plan to measure the correlation between question quality and answers quality through salient features such as question content quality, author reputation, punctuation density, word length, entropy of part of speech tags and votes received etc.

- *Identifying topic drift*: Currently we have investigated thread's content quality but some threads becomes longer with several posts because threads participants sometime get away from the actual topic which decrease the threads' quality. Our thread ranking techniques can be further enhanced by measuring the threads' topic drift. By discarding lengthy and irrelevant threads, better thread ranking may be achieved.

- *Subjectivity analysis*: Applying subjectivity analysis techniques may lead to better thread ranking. Supervised and semi-supervised techniques may be used to classify subjective threads in online forums.

- *Temporal features*: Thread retrieval may be enhanced through temporal features such as replies latency, duration of replies. These features indicates users' continuous interest towards topic as well their activeness in the forum.

which might be due to the content quality of initial post or inappropriate thread title. HeadPost represents the Initial post text and thread title. It can be seen in figure 5.3 that HeadPost (WordNet) improves the performance slightly over the (HeadPost + thread features) setting which means that reputed participants replied in lower posts. BOP with thread features have given similar performances for both cosine and WordNet techniques.

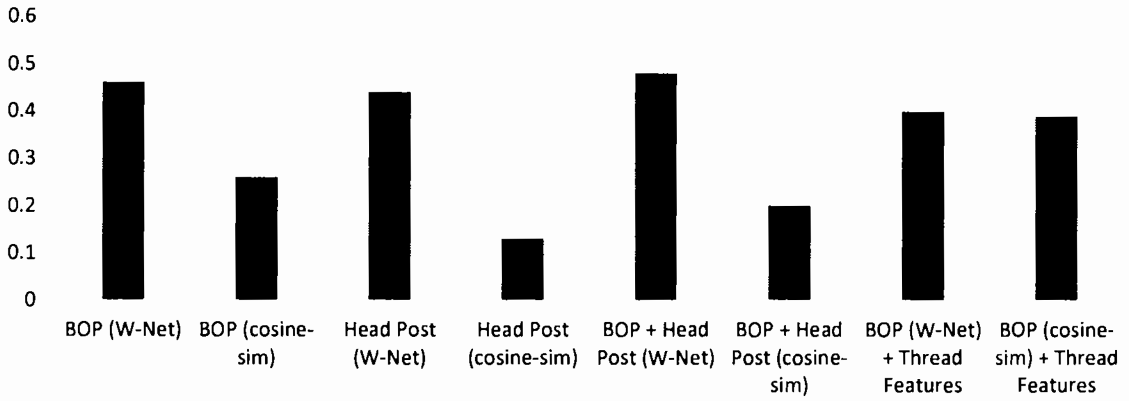


Figure 5.3: Mean Average Precision score for Bag of Posts and HeadPost

In figure 5.4, it was found that considering thread features with the sum of cosine-sim and WordNet actually decrease the MAP due to the existence of participants with poor reputations and low content quality. Techniques described in section 6.1.8, such as Sum (All-W-Net-sim) and Sum (All-cosine-sim) represent the sum of all methods using WordNet and cosine similarity techniques. Sum (All-W-Net-sim) performed well than Sum (All-cosine-sim) which validates our notion that semantic similarity techniques are significant in finding relevant threads.

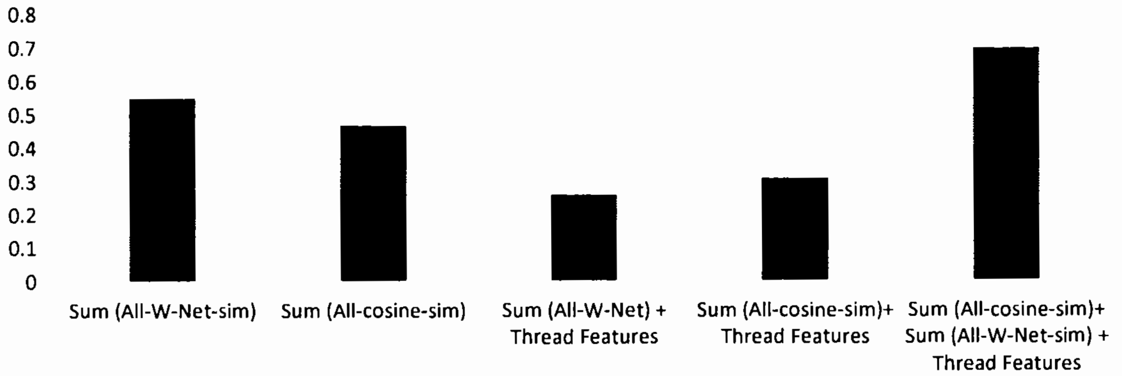


Figure 5.4: Mean Average Precision score for W-Net, Cosine similarity and Thread features (Combined)

Comparing the baseline methods, the semantic approach consistently outperforms the cosine-sim equivalent. It was also found that the threads with more reputed participants and posts, tended to be more productive and have a greater chance of being of a high quality.

Chapter Summary

In this chapter, we have investigated the impact of semantic similarity technique, answer quality and participant's reputation on thread ranking in online forums. Finding relevant threads for a query has various applications such as providing links to similar threads and providing a clustered interface of related threads for a given query. The primary objective was to improve the thread retrieval process by considering the context of a thread's content. In contrast to cosine similarity techniques, which consider lexical overlap of content, our techniques T-SimRank and T-CRRank considers both threads' semantics and participant's reputation in evaluating threads' quality. By demonstrating the utility of our approach using a real BBC discussion forum, we found that the rankings produced through T-SimRank and T-CRRank have better quality than methods based solely on lexical overlap (e.g., cosine similarity technique).

Chapter 6:
Conclusions and Future directions

6.1. Conclusions

In this thesis, we studied several social web application such as expert ranking in online discussion forums (non-rated and rated) and thread ranking in forums. We propose novel ways of using such features which have not fully exploited by social web domain for information retrieval and recommendation applications. For expert ranking in online non-rated discussion forums such as BBC message boards, we proposed expert ranking techniques. Firstly, for BBC discussion forum we proposed features such as, user participation activity, content relevancy and Co-existing users' reputation for expert ranking. Our techniques performed better than link-graph based techniques [3, 4], indicates that co-existing users' reputation and answer quality are better indicators of measuring user reputation. Experiments showed that proposed techniques ExpRank-CRF, ExpRank-COM, ExpRank-AQCS and ExpRank-FB based on aforesaid features have improved expert rankings.

We have proposed expert ranking techniques for rated-forums like StackOverflow. We argued that link-structure based techniques are not well suited for expert ranking problem in rated-forums. Moreover, current StackOverflow reputation mechanism is not fully capable of finding real experts. We have applied bibliometrics like g-index on users' posts score to compute users' consistency in providing quality answers. Novel reputation features have been extracted from StackOverflow dataset. Experiments showed that our techniques Exp-PC, Rep-FS and Weighted-Exp-PC outperformed to both link-structure and StackOverflow, reputation mechanisms.

In the last, we have evaluated the impact of semantic similarity technique, user participant and post quality features on thread ranking in BBC discussion forum. In contrast to cosine similarity techniques which only considers the lexical overlap between terms, semantic similarity techniques consider the context/meaning of terms which is helpful in retrieving relevant threads for a given query. By applying WordNet based semantic similarity function [102], participant reputation and answer quality features, better thread rankings are achieved. By combining cosine similarity and semantic similarity techniques we achieved significant results.

6.2. Future lines of research

There are several directions in which our work can evolve. In this section, we present some possible future directions.

6.2.1. Expert ranking

- *Applications of Evolutionary algorithms*: For BBC and StackOverflow discussion forums, our features combination strategies such as ExpRank-CRF and ExpRank-AQCS are simple and straightforward. We are currently reviewing to model our expert finding problem as optimization problem through more powerful fusion strategies such as particle swarm optimization and genetic algorithm to automatically design and fine tune the fusion strategies.
- *Community detection in StackOverflow forum*: We are investigating to address the community detection problem through co-existing users' data. It can be achieved in two ways. Firstly interactions between forum users (as an answer provider and askers) will be used to construct a social network. Social network strength (relationships between users) can suggest the users who collaborate most of the time on specific topics. Secondly the topic similarity will be computed among co-existing users posts. Expert communities may be formed by combining social network strength and content similarity of users for a given topic.
- *Expert databases for Organizations*: Currently expert ranking techniques are proposed for online discussion forums. These techniques can be extendable to organizations for building expert databases. Moreover we also plan to apply our techniques to some related areas such as mining influential persons and opinion leader identification.

6.2.2. Thread ranking

- *Question quality*: We have recently started to work on assessment of question quality because a topic-specific and well written question attracts more people to answer and it may receive quality answers. Therefore we plan to measure the correlation between question quality and answers quality through salient features such as question content quality, author reputation, punctuation density, word length, entropy of part of speech tags and votes received etc.

- *Identifying topic drift*: Currently we have investigated thread's content quality but some threads becomes longer with several posts because threads participants sometime get away from the actual topic which decrease the threads' quality. Our thread ranking techniques can be further enhanced by measuring the threads' topic drift. By discarding lengthy and irrelevant threads, better thread ranking may be achieved.

- *Subjectivity analysis*: Applying subjectivity analysis techniques may lead to better thread ranking. Supervised and semi-supervised techniques may be used to classify subjective threads in online forums.

- *Temporal features*: Thread retrieval may be enhanced through temporal features such as replies latency, duration of replies. These features indicates users' continuous interest towards topic as well their activeness in the forum.

References

1. Adamic, L.A., et al. *Knowledge sharing and yahoo answers: everyone knows something.* in *Proceedings of the 17th international conference on World Wide Web.* 2008. ACM.
2. Budanitsky, A. and G. Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness.* Computational Linguistics, 2006. **32**(1): pp. 13-47.
3. Zhang, J., M.S. Ackerman, and L. Adamic. *Expertise networks in online communities: structure and algorithms.* in *Proceedings of the 16th international conference on World Wide Web.* 2007. ACM.
4. Wang, G.A., et al., *ExpertRank: A topic-aware expert finding algorithm for online knowledge communities.* Decision Support Systems, 2013. **54**(3): pp. 1442-1451.
5. Bouguessa, M., B. Dumoulin, and S. 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 and data mining.* 2008. ACM.
6. Harper, F.M., et al. *Predictors of answer quality in online Q&A sites.* in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2008. ACM.
7. Seo, J. and W. Croft. *Thread-based expert finding.* in *SIGIR 2009 Workshop on Search in Socal Media. New York: ACM.* 2009. Citeseer.
8. Movshovitz-Attias, D., et al. *Analysis of the reputation system and user contributions on a question answering website: Stackoverflow.* in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on.* 2013. IEEE.
9. Yang, J., et al., *Sparrows and owls: Characterisation of expert behaviour in stackoverflow*, in *User Modeling, Adaptation, and Personalization*2014, Springer. pp. 266-277.
10. Vasilescu, B., V. Filkov, and A. Serebrenik. *StackOverflow and GitHub: Associations between software development and crowdsourced knowledge.* in *Social Computing (SocialCom), 2013 International Conference on.* 2013. IEEE.
11. Li, B., et al. *Analyzing and predicting question quality in community question answering services.* in *Proceedings of the 21st international conference companion on World Wide Web.* 2012. ACM.
12. Lee, J.-T., M.-C. Yang, and H.-C. Rim, *Discovering High-Quality Threaded Discussions in Online Forums.* Journal of Computer Science and Technology, 2014. **29**(3): pp. 519-531.
13. Shah, C. and J. 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.* 2010. ACM.
14. Agichtein, E., et al. *Finding high-quality content in social media.* in *Proceedings of the 2008 International Conference on Web Search and Data Mining.* 2008. ACM.
15. Cong, G., et al. *Finding question-answer pairs from online forums.* in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* 2008. ACM.
16. Arai, K. and A.N. Handayani, *Predicting quality of answer in collaborative Q/A community.* International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2013. **2**(3).

17. Singh, A. and D. Raghu. *Retrieving similar discussion forum threads: a structure based approach.* in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* 2012. ACM.

18. Mohler, M. and R. Mihalcea. *Text-to-text semantic similarity for automatic short answer grading.* in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* 2009. Association for Computational Linguistics.

19. Liu, G., et al. *A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge.* in *SEKE.* 2011.

20. Kardan, A., A. Omidvar, and M. Behzadi, *Context based Expert Finding in Online Communities using Social Network Analysis.* International J of Computer Science Research and Application, 2012. **2**(1): pp. 79-88.

21. Jurczyk, P. and E. Agichtein. *Discovering authorities in question answer communities by using link analysis.* in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* 2007. ACM.

22. Kao, W.-C., D.-R. Liu, and S.-W. Wang. *Expert finding in question-answering websites: a novel hybrid approach.* in *Proceedings of the 2010 ACM Symposium on Applied Computing.* 2010. ACM.

23. Bouguessa, M. and L.B. Romdhane, *Identifying authorities in online communities.* ACM Transactions on Intelligent Systems and Technology (TIST), 2015. **6**(3): p. 30.

24. Lü, L., et al., *Leaders in social networks, the delicious case.* PloS one, 2011. **6**(6): p. e21202.

25. Zhang, J., et al. *QuME: a mechanism to support expertise finding in online help-seeking communities.* in *Proceedings of the 20th annual ACM symposium on User interface software and technology.* 2007. ACM.

26. Ma, N. and Y. Liu, *SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork.* Expert Systems with Applications, 2014. **41**(4): pp. 1357-1368.

27. Zhu, H., et al. *Towards expert finding by leveraging relevant categories in authority ranking.* in *Proceedings of the 20th ACM international conference on Information and knowledge management.* 2011. ACM.

28. Campbell, C.S., et al. *Expertise identification using email communications.* in *Proceedings of the twelfth international conference on Information and knowledge management.* 2003. ACM.

29. Dom, B., et al. *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.* 2003. ACM.

30. Ding, Y., *Topic-based PageRank on author cocitation networks.* Journal of the American Society for Information Science and technology, 2011. **62**(3): pp. 449-466.

31. Yan, E., Y. Ding, and C.R. Sugimoto, *P-Rank: An indicator measuring prestige in heterogeneous scholarly networks.* Journal of the American Society for Information Science and technology, 2011. **62**(3): pp. 467-477.

32. Zhou, J., et al., *Ranking scientific publications with similarity-preferential mechanism.* Scientometrics, 2015: pp. 1-12.

33. Amjad, T., et al., *Topic-based heterogeneous rank.* Scientometrics, 2015. **104**(1): pp. 313-334.

34. Brin, S. and L. Page, *Reprint of: The anatomy of a large-scale hypertextual web search engine.* Computer networks, 2012. **56**(18): pp. 3825-3833.

35. Kleinberg, J.M., *Authoritative sources in a hyperlinked environment.* Journal of the ACM (JACM), 1999. **46**(5): pp. 604-632.

36. Omidvar, A., M. Garakani, and H.R. Safarpour, *Context based user ranking in forums for expert finding using WordNet dictionary and social network analysis.* Information Technology and Management, 2014. **15**(1): pp. 51-63.

37. Jurczyk, P. and E. Agichtein. *Hits on question answer portals: exploration of link analysis for author ranking.* in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* 2007. ACM.

38. Schall, D., *Expertise ranking using activity and contextual link measures.* Data & Knowledge Engineering, 2012. **71**(1): pp. 92-113.

39. Zhou, Y., et al. *Routing questions to the right users in online communities.* in *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on.* 2009. IEEE.

40. Li, W., C. Zhang, and S. Hu. *G-Finder: routing programming questions closer to the experts.* in *ACM Sigplan Notices.* 2010. ACM.

41. Chai, K., V. Potdar, and T. Dillon, *Content quality assessment related frameworks for social media*, in *Computational Science and Its Applications–ICCSA 2009*2009, Springer. pp. 791-805.

42. Kardan, A.A. and M. Ebrahimi, *A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups.* Information Sciences, 2013. **219**: pp. 93-110.

43. Li, Y.-M., T.-F. Liao, and C.-Y. Lai, *A social recommender mechanism for improving knowledge sharing in online forums.* Information Processing & Management, 2012. **48**(5): pp. 978-994.

44. Pal, A., F.M. Harper, and J.A. Konstan, *Exploring question selection bias to identify experts and potential experts in community question answering.* ACM Transactions on Information Systems (TOIS), 2012. **30**(2): pp. 10.

45. Fu, Y., et al. *Finding experts using social network analysis.* in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence.* 2007. IEEE Computer Society.

46. Davoodi, E., M. Afsharchi, and K. Kianmehr, *A social network-based approach to expert recommendation system*, in *Hybrid Artificial Intelligent Systems*2012, Springer. pp. 91-102.

47. Yang, K.-H., et al. *EFS: Expert finding system based on Wikipedia link pattern analysis.* in *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on.* 2008. IEEE.

48. Reihanian, A., B. Minaei-Bidgoli, and H. Alizadeh, *Topic-oriented community detection of rating-based social networks.* Journal of King Saud University-Computer and Information Sciences, 2015.

49. Zhao, Z., et al., *Topic oriented community detection through social objects and link analysis in social networks.* Knowledge-Based Systems, 2012. **26**: pp. 164-173.

50. Fu, Y., et al. *A CDD-based formal model for expert finding.* in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* 2007. ACM.

51. Zhu, H., et al., *Ranking user authority with relevant knowledge categories for expert finding.* World Wide Web, 2014. **17**(5): pp. 1081-1107.

52. Venkataramani, R., et al. *Discovery of technical expertise from open source code repositories*. in *Proceedings of the 22nd international conference on World Wide Web companion*. 2013. International World Wide Web Conferences Steering Committee.

53. Pal, A. and J.A. Konstan. *Expert identification in community question answering: exploring question selection bias*. in *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010. ACM.

54. Xu, Z., et al. *Towards the semantic web: Collaborative tag suggestions*. in *Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*. 2006.

55. Sen, S., J. Vig, and J. Riedl. *Tagommenders: connecting users to items through tags*. in *Proceedings of the 18th international conference on World wide web*. 2009. ACM.

56. Kardan, A.A., A. Omidvar, and M. Behzadi, *Context based Expert Finding in Online Communities using Social Network Analysis*. International Journal of Computer Science Research and Application, 2012. **2**(01): pp. 79-88.

57. Kardan, A.A. and A. Zeinab. *Design and implementation of a recommender system to introducing experts in an online forums*. in *E-Learning and E-Teaching (ICELET), 2013 Fourth International Conference on*. 2013. IEEE.

58. Brandão, L. and P. Diviacco, *Expert finding in question-and-answer web services*.

59. Yang, L., et al. *Cqarank: jointly model topics and expertise in community question answering*. in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. 2013. ACM.

60. Leydesdorff, L. and L. Vaughan, *Co-occurrence matrices and their applications in information science: extending ACA to the web environment*. Journal of the American Society for Information Science and technology, 2006. **57**(12): pp. 1616-1628.

61. Mika, P., *Ontologies are us: A unified model of social networks and semantics*, in *The Semantic Web–ISWC 2005*2005, Springer. pp. 522-536.

62. Guan, Z., et al., *Co-occurrence-based diffusion for expert search on the web*. Knowledge and Data Engineering, IEEE Transactions on, 2013. **25**(5): pp. 1001-1014.

63. Fei, G., et al., *Exploiting Burstiness in Reviews for Review Spammer Detection*. ICWSM, 2013. **13**: pp. 175-184.

64. Murata, T. *Discovery of Web communities based on the co-occurrence of references*. in *Discovery Science*. 2000. Springer.

65. Alonso, S., et al., *h-Index: A review focused in its variants, computation and standardization for different scientific fields*. Journal of Informetrics, 2009. **3**(4): pp. 273-289.

66. Romero, D.M., et al., *Influence and passivity in social media*, in *Machine learning and knowledge discovery in databases*2011, Springer. pp. 18-33.

67. Bui, D.-L., T.-T. Nguyen, and Q.-T. Ha, *Measuring the influence of bloggers in their community based on the h-index family*, in *Advanced Computational Methods for Knowledge Engineering*2014, Springer. pp. 313-324.

68. Young, A.L. and A. Quan-Haase. *Information revelation and internet privacy concerns on social network sites: a case study of facebook*. in *Proceedings of the fourth international conference on Communities and technologies*. 2009. ACM.

69. Razis, G. and I. Anagnostopoulos, *InfluenceTracker: Rating the impact of a Twitter account*, in *Artificial Intelligence Applications and Innovations*2014, Springer. pp. 184-195.

70. Li, N. and D. Gillet. *Identifying influential scholars in academic social media platforms.* in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* 2013. ACM.

71. Egghe, L., *Theory and practise of the g-index.* Scientometrics, 2006. **69**(1): pp. 131-152.

72. Wang, X., Z. Wang, and S. Xu, *Tracing scientist's research trends realtimely.* Scientometrics, 2013. **95**(2): pp. 717-729.

73. Agarwal, N., et al., *Modeling blogger influence in a community.* Social Network Analysis and Mining, 2012. **2**(2): pp. 139-162.

74. Akritidis, L., D. Katsaros, and P. Bozanis, *Identifying the productive and influential bloggers in a community.* Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2011. **41**(5): pp. 759-764.

75. Capiluppi, A., A. Serebrenik, and A. Youssef. *Developing an h-index for OSS developers.* in *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories.* 2012. IEEE Press.

76. Devezas, J.L., S. Nunes, and C. Ribeiro. *FEUP at TREC 2010 Blog Track: Using h-index for blog ranking.* in *TREC.* 2010.

77. Hovden, R., *Bibliometrics for Internet media: Applying the h-index to YouTube.* Journal of the American Society for Information Science and technology, 2013. **64**(11): pp. 2326-2331.

78. Wang, G.A., et al., *Examining micro-level knowledge sharing discussions in online communities.* Information Systems Frontiers, 2015. **17**(6): pp. 1227-1238.

79. Gómez, V., A. Kaltenbrunner, and V. López. *Statistical analysis of the social network and discussion threads in slashdot.* in *Proceedings of the 17th international conference on World Wide Web.* 2008. ACM.

80. Akritidis, L., D. Katsaros, and P. Bozanis. *Identifying influential bloggers: Time does matter.* in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on.* 2009. IET.

81. Xu, Z. and J. Ramanathan, *Thread-based probabilistic models for expert finding in enterprise Microblogs.* Expert Systems with Applications, 2016. **43**: pp. 286-297.

82. MacLeod, L. *Reputation on Stack Exchange: Tag, You're It!* in *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on.* 2014. IEEE.

83. Bian, J., et al. *Finding the right facts in the crowd: factoid question answering over social media.* in *Proceedings of the 17th international conference on World Wide Web.* 2008. ACM.

84. Jeon, J., et al. *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.* 2006. ACM.

85. Lou, J., et al., *Contributing high quantity and quality knowledge to online Q&A communities.* Journal of the American Society for Information Science and technology, 2013. **64**(2): pp. 356-371.

86. Blooma, M.J., A.Y. Chua, and D.H.-L. Goh. *A predictive framework for retrieving the best answer.* in *Proceedings of the 2008 ACM symposium on Applied computing.* 2008. ACM.

87. Surdeanu, M., M. Ciaramita, and H. Zaragoza, *Learning to rank answers to non-factoid questions from web collections.* Computational Linguistics, 2011. **37**(2): pp. 351-383.

88. Wang, X.-J., et al. *Ranking community answers by modeling question-answer relationships via analogical reasoning*. in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009. ACM.

89. Zhou, T.C., M.R. Lyu, and I. King. *A classification-based approach to question routing in community question answering*. in *Proceedings of the 21st international conference companion on World Wide Web*. 2012. ACM.

90. Cho, J.H., et al. *Resolving healthcare forum posts via similar thread retrieval*. in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2014. ACM.

91. Sharif, H., et al., *A Classification Based Framework to Predict Viral Threads*. 2015.

92. Zhou, T.C., et al., *Learning to suggest questions in social media*. Knowledge and Information Systems, 2015. **43**(2): pp. 389-416.

93. Tang, X., M. Zhang, and C.C. Yang. *Leveraging user interest to improve thread recommendation in online forum*. in *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*. 2013. IEEE.

94. Hong, L. and B.D. Davison. *A classification-based approach to question answering in discussion boards*. in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009. ACM.

95. Deepak, P. and K. Visweswariah, *Unsupervised Solution Post Identification from Discussion Forums*.

96. Zhang, Y.Y. *A Retrieval Sorting Approach for Online Forums Based on Domain Topics*. in *Advanced Materials Research*. 2013. Trans Tech Publ.

97. John, B.M., A.Y.-K. Chua, and D.H.-L. Goh, *What makes a high-quality user-generated answer?* Internet Computing, IEEE, 2011. **15**(1): pp. 66-71.

98. Toba, H., et al., *Discovering high quality answers in community question answering archives using a hierarchy of classifiers*. Information Sciences, 2014. **261**: pp. 101-115.

99. Hliaoutakis, A., et al., *Information retrieval by semantic similarity*. International Journal on Semantic Web and Information Systems, 2006. **2**(3): pp. 55-73.

100. Meng, L., R. Huang, and J. Gu, *A review of semantic similarity measures in wordnet*. International Journal of Hybrid Information Technology, 2013. **6**(1): pp. 1-12.

101. Vu, T.T., Q.H. Tran, and S.B. Pham, *TATO: Leveraging on Multiple Strategies for Semantic Textual Similarity*.

102. Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification*. WordNet: An electronic lexical database, 1998. **49**(2): pp. 265-283.

103. kralja Aleksandra, B., *The Role of Semantic Similarity for Intelligent Question Routing*.

104. Pasca, M. and S. Harabagiu. *The informative role of WordNet in open-domain question answering*. in *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*. 2001.

105. Corley, C. and R. Mihalcea. *Measuring the semantic similarity of texts*. in *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*. 2005. Association for Computational Linguistics.

106. Speer, R. and C. Havasi. *Representing General Relational Knowledge in ConceptNet 5*. in *LREC*. 2012.

107. Agt, H. and R.-D. Kutsche. *Automated construction of a large semantic network of related terms for domain-specific modeling*. in *Advanced Information Systems Engineering*. 2013. Springer.

108. Proisl, T., et al., *SemantiKLUE: Robust Semantic Similarity at Multiple Levels Using Maximum Weight Matching.* SemEval 2014, 2014: pp. 532.

109. Ding, H. and E. Riloff, *Extracting Information about Medication Use from Veterinary Discussions.*

110. Mohler, M., R. Bunescu, and R. Mihalcea. *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments.* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* 2011. Association for Computational Linguistics.

111. Ren, Z., et al. *Summarizing web forum threads based on a latent topic propagation process.* in *Proceedings of the 20th ACM international conference on Information and knowledge management.* 2011. ACM.

112. Shi, L., et al. *Web forum Sentiment analysis based on topics.* in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on.* 2009. IEEE.

113. Wan, X., *A novel document similarity measure based on earth mover's distance.* Information Sciences, 2007. **177**(18): pp. 3718-3730.

114. Elsas, J.L. and J.G. Carbonell. *It pays to be picky: an evaluation of thread retrieval in online forums.* in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 2009. ACM.

115. Seo, J., W.B. Croft, and D.A. Smith, *Online community search using conversational structures.* Information Retrieval, 2011. **14**(6): pp. 547-571.

116. Kim, S.N., L. Wang, and T. Baldwin. *Tagging and linking web forum posts.* in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning.* 2010. Association for Computational Linguistics.

117. Wang, H., et al. *Learning online discussion structures by conditional random fields.* in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* 2011. ACM.

118. Albaham, A.T., N. Salim, and O.I. Adekunle. *Leveraging Post Level Quality Indicators in Online Forum Thread Retrieval.* in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).* 2014. Springer.

119. Kardan, A., A. Omidvar, and F. Farahmandnia. *Expert finding on social network with link analysis approach.* in *Electrical Engineering (ICEE), 2011 19th Iranian Conference on.* 2011. IEEE.

120. Li, Y., S. Ma, and R. Huang, *Social Context Analysis for Topic-Specific Expert Finding in Online Learning Communities*, in *Smart Learning Environments*2015, Springer. pp. 57-74.

121. Bhatia, S., P. Biyani, and P. Mitra, *Classifying user messages for managing web forum data.* 2012.

122. Breslin, J.G., et al. *Finding experts using Internet-based discussions in online communities and associated social networks.* in *First International ExpertFinder Workshop.* 2007.

123. Ehrlich, K., C.-Y. Lin, and V. Griffiths-Fisher. *Searching for experts in the enterprise: combining text and social network analysis.* in *Proceedings of the 2007 international ACM conference on Supporting group work.* 2007. ACM.

124. Hua, G. and D. Haughton, *A network analysis of an online expertise sharing community.* Social Network Analysis and Mining, 2012. **2**(4): pp. 291-303.

125. Gomaa, W.H. and A.A. Fahmy, *A survey of text similarity approaches.* International Journal of Computer Applications, 2013. **68**(13): pp. 13-18.

126. Mobasher, B., R. Cooley, and J. Srivastava. *Creating adaptive web sites through usage-based clustering of URLs.* in *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on.* 1999. IEEE.

127. Gustafson, N., M.S. Pera, and Y.-K. Ng. *Nowhere to hide: Finding plagiarized documents based on sentence similarity.* in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01.* 2008. IEEE Computer Society.

128. Lee, T.Y. *Needs-based analysis of online customer reviews.* in *Proceedings of the ninth international conference on Electronic commerce.* 2007. ACM.

129. Patterson, A.L., *Detecting spam documents in a phrase based information retrieval system*, 2009, Google Patents.

130. Krestel, R. and L. Chen. *Using co-occurrence of tags and resources to identify spammers.* in *Proceedings of 2008 ECML/PKDD Discovery Challenge Workshop.* 2008.

131. Glance, N., M. Hurst, and T. Tomokiyo. *Blogpulse: Automated trend discovery for weblogs.* in *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics.* 2004. New York.

132. Craswell, N., A.P. de Vries, and I. Soboroff. *Overview of the TREC 2005 Enterprise Track.* in *Trec.* 2005.

133. Sood, S., et al. *TagAssist: Automatic Tag Suggestion for Blog Posts.* in *ICWSM.* 2007.

134. Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules.* in *Proc. 20th int. conf. very large data bases, VLDB.* 1994.

135. Baldwin, T., D. Martinez, and R.B. Penman. *Automatic thread classification for Linux user forum information access.* in *Proceedings of the Twelfth Australasian Document Computing Symposium (ADCS 2007).* 2007.

136. Sussna, M. *Word sense disambiguation for free-text indexing using a massive semantic network.* in *Proceedings of the second international conference on Information and knowledge management.* 1993. ACM.

137. Gottipati, S., D. Lo, and J. Jiang. *Finding relevant answers in software forums.* in *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering.* 2011. IEEE Computer Society.

138. Wenwen, H., et al., *Ranking potential reply-providers in community question answering system.* Communications, China, 2013. **10**(10): pp. 125-136.

139. Lee, D.H., P. Brusilovsky, and T. Schleyer, *Recommending collaborators using social features and mesh terms.* Proceedings of the American Society for Information Science and Technology, 2011. **48**(1): pp. 1-10.

140. Weerkamp, W. and M. de Rijke, *Credibility-inspired ranking for blog post retrieval.* Information Retrieval, 2012. **15**(3-4): pp. 243-277.

141. Abdi, H., *The Kendall rank correlation coefficient.* Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, 2007: p. 508-510.

142. Handayani, A.N. and K. Arai, *Predicting Quality of Answer in Collaborative Q/A Community.* culture, 2013. **3779**(93): pp. 37799.

143. Mukherjee, A., B. Liu, and N. Glance. *Spotting fake reviewer groups in consumer reviews.* in *Proceedings of the 21st international conference on World Wide Web.* 2012. ACM.

144. Fisher, D., M. Smith, and H.T. Welser. *You are who you talk to: Detecting roles in usenet newsgroups.* in *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on.* 2006. IEEE.

145. Lanagan, J., N. Anokhin, and J. Velcin, *Early Stage Conversation Catalysts on Entertainment-Based Web Forums*, in *State of the Art Applications of Social Network Analysis*2014, Springer. pp. 97-118.

146. Hirsch, J.E., *An index to quantify an individual's scientific research output.* Proceedings of the National academy of Sciences of the United States of America, 2005. **102**(46): pp. 16569-16572.

147. Wingkvist, A. and M. Ericsson. *Asked and Answered: Communication Patterns of Experts on an Online Forum.* in *36th Information Systems Research Seminar in Scandinavia.* 2013.

148. Bhat, V., et al. *Min (e) d your tags: Analysis of question response time in stackoverflow.* in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on.* 2014. IEEE.

149. Burel, G., Y. He, and H. Alani, *Automatic identification of best answers in online enquiry communities*, in *The Semantic Web: Research and Applications*2012, Springer. pp. 514-529.

150. Haveliwala, T.H. *Topic-sensitive pagerank.* in *Proceedings of the 11th international conference on World Wide Web.* 2002. ACM.

151. Li, C., J. Yin, and J. Zhao, *Using improved ICA method for hyperspectral data classification.* Arabian Journal for Science and Engineering, 2014. **39**(1): pp. 181-189.

152. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 1988. **24**(5): pp. 513-523.

153. Mihalcea, R., C. Corley, and C. Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity.* in *AAAI.* 2006.

154. Vallet, D., I. Cantador, and J.M. Jose, *Personalizing web search with folksonomy-based user and document profiles*, in *Advances in Information Retrieval*2010, Springer. pp. 420-431.

155. Varelas, G., et al. *Semantic similarity methods in wordNet and their application to information retrieval on the web.* in *Proceedings of the 7th annual ACM international workshop on Web information and data management.* 2005. ACM.

156. Kannan, V. and G. Srinivasan, *Yet another way of Ranking web Documents Based On Semantic Similarity.*

157. Bhatia, S. and P. Mitra. *Adopting Inference Networks for Online Thread Retrieval.* in *AAAI.* 2010.

158. Jain, A.K. and R.C. Dubes, *Algorithms for clustering data.* Vol. 6. 1988: Prentice hall Englewood Cliffs.

159. Park, H.-S. and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering.* Expert Systems with Applications, 2009. **36**(2): pp. 3336-3341.

160. Tari, L., et al. *Passage Relevancy Through Semantic Relatedness.* in *TREC.* 2007. Citeseer.

161. Chahal, P., M. Singh, and S. Kumar. *Ranking of web documents using semantic similarity.* in *Information Systems and Computer Networks (ISCON), 2013 International Conference on.* 2013. IEEE.

162. Miller, G.A., et al., *Introduction to wordnet: An on-line lexical database**. International journal of lexicography, 1990. **3**(4): pp. 235-244.

163. Xu, Z., et al. *An alternative text representation to TF-IDF and Bag-of-Words.* in *Proceedings of 21st ACM Conf. of Information and Knowledge Management (CIKM).* 2012.

164. Grozin, V.A., N.F. Gusarova, and N.V. Dobrenko, *Feature Selection for Language Independent Text Forum Summarization*, in *Knowledge Engineering and Semantic Web*2015, Springer. pp. 63-71.

165. Montague, M. and J.A. Aslam. *Relevance score normalization for metasearch.* in *Proceedings of the tenth international conference on Information and knowledge management*. 2001. ACM.

166. Gopalan, N. and K. Batri, *Adaptive Selection of Top-m Retrieval Strategies for Data Fusion in Information Retrieval.* International Journal of Soft Computing, 2007. **2**(1): pp. 11-16.

167. Fox, E.A. and J.A. Shaw, *Combination of multiple searches.* NIST SPECIAL PUBLICATION SP, 1994: pp. 243-243.

168. Biyani, P., et al. *Thread Specific Features are Helpful for Identifying Subjectivity Orientation of Online Forum Threads.* in *COLING*. 2012.

169. Cavnar, W.B. and J.M. Trenkle, *N-gram-based text categorization.* Ann Arbor MI, 1994. **48113**(2): pp. 161-175.

170. Kumar, N. and K. Srinathan. *Automatic keyphrase extraction from scientific documents using N-gram filtration technique*. in *Proceedings of the eighth ACM symposium on Document engineering*. 2008. ACM.

171. Shah, U., et al. *Information retrieval on the semantic web.* in *Proceedings of the eleventh international conference on Information and knowledge management*. 2002. ACM.

172. Wang, X., A. McCallum, and X. Wei. *Topical n-grams: Phrase and topic discovery, with an application to information retrieval.* in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. 2007. IEEE.

173. Duan, H. and C. Zhai, *Exploiting thread structures to improve smoothing of language models for forum post retrieval*, in *Advances in Information Retrieval*2011, Springer. pp. 350-361.

174. Lapata, M., *Automatic evaluation of information ordering: Kendall's tau.* Computational Linguistics, 2006. **32**(4): pp. 471-484.

175. Rijsbergen, C.J.V., *Information Retrieval*1979, MA, USA: Butterworth-Heinemann Newton.

88. Wang, X.-J., et al. *Ranking community answers by modeling question-answer relationships via analogical reasoning*. in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009. ACM.

89. Zhou, T.C., M.R. Lyu, and I. King. *A classification-based approach to question routing in community question answering*. in *Proceedings of the 21st international conference companion on World Wide Web*. 2012. ACM.

90. Cho, J.H., et al. *Resolving healthcare forum posts via similar thread retrieval*. in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2014. ACM.

91. Sharif, H., et al., *A Classification Based Framework to Predict Viral Threads*. 2015.

92. Zhou, T.C., et al., *Learning to suggest questions in social media*. Knowledge and Information Systems, 2015. **43**(2): pp. 389-416.

93. Tang, X., M. Zhang, and C.C. Yang. *Leveraging user interest to improve thread recommendation in online forum*. in *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*. 2013. IEEE.

94. Hong, L. and B.D. Davison. *A classification-based approach to question answering in discussion boards*. in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009. ACM.

95. Deepak, P. and K. Visweswariah, *Unsupervised Solution Post Identification from Discussion Forums*.

96. Zhang, Y.Y. *A Retrieval Sorting Approach for Online Forums Based on Domain Topics*. in *Advanced Materials Research*. 2013. Trans Tech Publ.

97. John, B.M., A.Y.-K. Chua, and D.H.-L. Goh, *What makes a high-quality user-generated answer?* Internet Computing, IEEE, 2011. **15**(1): pp. 66-71.

98. Toba, H., et al., *Discovering high quality answers in community question answering archives using a hierarchy of classifiers*. Information Sciences, 2014. **261**: pp. 101-115.

99. Hliaoutakis, A., et al., *Information retrieval by semantic similarity*. International Journal on Semantic Web and Information Systems, 2006. **2**(3): pp. 55-73.

100. Meng, L., R. Huang, and J. Gu, *A review of semantic similarity measures in wordnet*. International Journal of Hybrid Information Technology, 2013. **6**(1): pp. 1-12.

101. Vu, T.T., Q.H. Tran, and S.B. Pham, *TATO: Leveraging on Multiple Strategies for Semantic Textual Similarity*.

102. Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification*. WordNet: An electronic lexical database, 1998. **49**(2): pp. 265-283.

103. kralja Aleksandra, B., *The Role of Semantic Similarity for Intelligent Question Routing*.

104. Pasca, M. and S. Harabagiu. *The informative role of WordNet in open-domain question answering*. in *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*. 2001.

105. Corley, C. and R. Mihalcea. *Measuring the semantic similarity of texts*. in *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*. 2005. Association for Computational Linguistics.

106. Speer, R. and C. Havasi. *Representing General Relational Knowledge in ConceptNet 5*. in *LREC*. 2012.

107. Agt, H. and R.-D. Kutsche. *Automated construction of a large semantic network of related terms for domain-specific modeling*. in *Advanced Information Systems Engineering*. 2013. Springer.

108. Proisl, T., et al., *SemantiKLUE: Robust Semantic Similarity at Multiple Levels Using Maximum Weight Matching.* SemEval 2014, 2014: pp. 532.

109. Ding, H. and E. Riloff, *Extracting Information about Medication Use from Veterinary Discussions.*

110. Mohler, M., R. Bunescu, and R. Mihalcea. *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments.* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* 2011. Association for Computational Linguistics.

111. Ren, Z., et al. *Summarizing web forum threads based on a latent topic propagation process.* in *Proceedings of the 20th ACM international conference on Information and knowledge management.* 2011. ACM.

112. Shi, L., et al. *Web forum Sentiment analysis based on topics.* in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on.* 2009. IEEE.

113. Wan, X., *A novel document similarity measure based on earth mover's distance.* Information Sciences, 2007. **177**(18): pp. 3718-3730.

114. Elsas, J.L. and J.G. Carbonell. *It pays to be picky: an evaluation of thread retrieval in online forums.* in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* 2009. ACM.

115. Seo, J., W.B. Croft, and D.A. Smith, *Online community search using conversational structures.* Information Retrieval, 2011. **14**(6): pp. 547-571.

116. Kim, S.N., L. Wang, and T. Baldwin. *Tagging and linking web forum posts.* in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning.* 2010. Association for Computational Linguistics.

117. Wang, H., et al. *Learning online discussion structures by conditional random fields.* in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* 2011. ACM.

118. Albaham, A.T., N. Salim, and O.I. Adekunle. *Leveraging Post Level Quality Indicators in Online Forum Thread Retrieval.* in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).* 2014. Springer.

119. Kardan, A., A. Omidvar, and F. Farahmandnia. *Expert finding on social network with link analysis approach.* in *Electrical Engineering (ICEE), 2011 19th Iranian Conference on.* 2011. IEEE.

120. Li, Y., S. Ma, and R. Huang, *Social Context Analysis for Topic-Specific Expert Finding in Online Learning Communities*, in *Smart Learning Environments*2015, Springer. pp. 57-74.

121. Bhatia, S., P. Biyani, and P. Mitra, *Classifying user messages for managing web forum data.* 2012.

122. Breslin, J.G., et al. *Finding experts using Internet-based discussions in online communities and associated social networks.* in *First International ExpertFinder Workshop.* 2007.

123. Ehrlich, K., C.-Y. Lin, and V. Griffiths-Fisher. *Searching for experts in the enterprise: combining text and social network analysis.* in *Proceedings of the 2007 international ACM conference on Supporting group work.* 2007. ACM.

124. Hua, G. and D. Haughton, *A network analysis of an online expertise sharing community.* Social Network Analysis and Mining, 2012. **2**(4): pp. 291-303.

125. Gomaa, W.H. and A.A. Fahmy, *A survey of text similarity approaches.* International Journal of Computer Applications, 2013. **68**(13): pp. 13-18.

126. Mobasher, B., R. Cooley, and J. Srivastava. *Creating adaptive web sites through usage-based clustering of URLs*. in *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on*. 1999. IEEE.

127. Gustafson, N., M.S. Pera, and Y.-K. Ng. *Nowhere to hide: Finding plagiarized documents based on sentence similarity*. in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. 2008. IEEE Computer Society.

128. Lee, T.Y. *Needs-based analysis of online customer reviews*. in *Proceedings of the ninth international conference on Electronic commerce*. 2007. ACM.

129. Patterson, A.L., *Detecting spam documents in a phrase based information retrieval system*, 2009, Google Patents.

130. Krestel, R. and L. Chen. *Using co-occurrence of tags and resources to identify spammers*. in *Proceedings of 2008 ECML/PKDD Discovery Challenge Workshop*. 2008.

131. Glance, N., M. Hurst, and T. Tomokiyo. *Blogpulse: Automated trend discovery for weblogs*. in *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*. 2004. New York.

132. Craswell, N., A.P. de Vries, and I. Soboroff. *Overview of the TREC 2005 Enterprise Track*. in *Trec*. 2005.

133. Sood, S., et al. *TagAssist: Automatic Tag Suggestion for Blog Posts*. in *ICWSM*. 2007.

134. Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules*. in *Proc. 20th int. conf. very large data bases, VLDB*. 1994.

135. Baldwin, T., D. Martinez, and R.B. Penman. *Automatic thread classification for Linux user forum information access*. in *Proceedings of the Twelfth Australasian Document Computing Symposium (ADCS 2007)*. 2007.

136. Sussna, M. *Word sense disambiguation for free-text indexing using a massive semantic network*. in *Proceedings of the second international conference on Information and knowledge management*. 1993. ACM.

137. Gottipati, S., D. Lo, and J. Jiang. *Finding relevant answers in software forums*. in *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. 2011. IEEE Computer Society.

138. Wenwen, H., et al., *Ranking potential reply-providers in community question answering system*. Communications, China, 2013. **10**(10): pp. 125-136.

139. Lee, D.H., P. Brusilovsky, and T. Schleyer, *Recommending collaborators using social features and mesh terms*. Proceedings of the American Society for Information Science and Technology, 2011. **48**(1): pp. 1-10.

140. Weerkamp, W. and M. de Rijke, *Credibility-inspired ranking for blog post retrieval*. Information Retrieval, 2012. **15**(3-4): pp. 243-277.

141. Abdi, H., *The Kendall rank correlation coefficient*. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, 2007: p. 508-510.

142. Handayani, A.N. and K. Arai, *Predicting Quality of Answer in Collaborative Q/A Community*. culture, 2013. **3779**(93): pp. 37799.

143. Mukherjee, A., B. Liu, and N. Glance. *Spotting fake reviewer groups in consumer reviews*. in *Proceedings of the 21st international conference on World Wide Web*. 2012. ACM.

144. Fisher, D., M. Smith, and H.T. Welser. *You are who you talk to: Detecting roles in usenet newsgroups*. in *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*. 2006. IEEE.

145. Lanagan, J., N. Anokhin, and J. Velcin, *Early Stage Conversation Catalysts on Entertainment-Based Web Forums*, in *State of the Art Applications of Social Network Analysis*2014, Springer. pp. 97-118.

146. Hirsch, J.E., *An index to quantify an individual's scientific research output.* Proceedings of the National academy of Sciences of the United States of America, 2005. **102**(46): pp. 16569-16572.

147. Wingkvist, A. and M. Ericsson. *Asked and Answered: Communication Patterns of Experts on an Online Forum.* in *36th Information Systems Research Seminar in Scandinavia.* 2013.

148. Bhat, V., et al. *Min (e) d your tags: Analysis of question response time in stackoverflow.* in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on.* 2014. IEEE.

149. Burel, G., Y. He, and H. Alani, *Automatic identification of best answers in online enquiry communities*, in *The Semantic Web: Research and Applications*2012, Springer. pp. 514-529.

150. Haveliwala, T.H. *Topic-sensitive pagerank.* in *Proceedings of the 11th international conference on World Wide Web.* 2002. ACM.

151. Li, C., J. Yin, and J. Zhao, *Using improved ICA method for hyperspectral data classification.* Arabian Journal for Science and Engineering, 2014. **39**(1): pp. 181-189.

152. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 1988. **24**(5): pp. 513-523.

153. Mihalcea, R., C. Corley, and C. Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity.* in *AAAI.* 2006.

154. Vallet, D., I. Cantador, and J.M. Jose, *Personalizing web search with folksonomy-based user and document profiles*, in *Advances in Information Retrieval*2010, Springer. pp. 420-431.

155. Varelas, G., et al. *Semantic similarity methods in wordNet and their application to information retrieval on the web.* in *Proceedings of the 7th annual ACM international workshop on Web information and data management.* 2005. ACM.

156. Kannan, V. and G. Srinivasan, *Yet another way of Ranking web Documents Based On Semantic Similarity.*

157. Bhatia, S. and P. Mitra. *Adopting Inference Networks for Online Thread Retrieval.* in *AAAI.* 2010.

158. Jain, A.K. and R.C. Dubes, *Algorithms for clustering data.* Vol. 6. 1988: Prentice hall Englewood Cliffs.

159. Park, H.-S. and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering.* Expert Systems with Applications, 2009. **36**(2): pp. 3336-3341.

160. Tari, L., et al. *Passage Relevancy Through Semantic Relatedness.* in *TREC.* 2007. Citeseer.

161. Chahal, P., M. Singh, and S. Kumar. *Ranking of web documents using semantic similarity.* in *Information Systems and Computer Networks (ISCON), 2013 International Conference on.* 2013. IEEE.

162. Miller, G.A., et al., *Introduction to wordnet: An on-line lexical database*.* International journal of lexicography, 1990. **3**(4): pp. 235-244.

163. Xu, Z., et al. *An alternative text representation to TF-IDF and Bag-of-Words.* in *Proceedings of 21st ACM Conf. of Information and Knowledge Management (CIKM).* 2012.

164. Grozin, V.A., N.F. Gusarova, and N.V. Dobrenko, *Feature Selection for Language Independent Text Forum Summarization*, in *Knowledge Engineering and Semantic Web*2015, Springer. pp. 63-71.

165. Montague, M. and J.A. Aslam. *Relevance score normalization for metasearch.* in *Proceedings of the tenth international conference on Information and knowledge management*. 2001. ACM.

166. Gopalan, N. and K. Batri, *Adaptive Selection of Top-m Retrieval Strategies for Data Fusion in Information Retrieval.* International Journal of Soft Computing, 2007. **2**(1): pp. 11-16.

167. Fox, E.A. and J.A. Shaw, *Combination of multiple searches.* NIST SPECIAL PUBLICATION SP, 1994: pp. 243-243.

168. Biyani, P., et al. *Thread Specific Features are Helpful for Identifying Subjectivity Orientation of Online Forum Threads.* in *COLING*. 2012.

169. Cavnar, W.B. and J.M. Trenkle, *N-gram-based text categorization.* Ann Arbor MI, 1994. **48113**(2): pp. 161-175.

170. Kumar, N. and K. Srinathan. *Automatic keyphrase extraction from scientific documents using N-gram filtration technique.* in *Proceedings of the eighth ACM symposium on Document engineering*. 2008. ACM.

171. Shah, U., et al. *Information retrieval on the semantic web.* in *Proceedings of the eleventh international conference on Information and knowledge management*. 2002. ACM.

172. Wang, X., A. McCallum, and X. Wei. *Topical n-grams: Phrase and topic discovery, with an application to information retrieval.* in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. 2007. IEEE.

173. Duan, H. and C. Zhai, *Exploiting thread structures to improve smoothing of language models for forum post retrieval*, in *Advances in Information Retrieval*2011, Springer. pp. 350-361.

174. Lapata, M., *Automatic evaluation of information ordering: Kendall's tau.* Computational Linguistics, 2006. **32**(4): pp. 471-484.

175. Rijsbergen, C.J.V., *Information Retrieval*1979, MA, USA: Butterworth-Heinemann Newton.

Appendices

Appendix A:

List of popular Community Question Answering Sites

General QA Forums

1. Yahoo! Answers

While Yahoo! Answers[24] may sometimes vary in quality, it makes up for in quantity. Statistically, it's quite probable that you'll find a good answer to a question — although you may have to through a few series and threads of questions and discussions before you run into something you find credible. No disrespect meant: Users have found many great leads and sources to answers that they have been able to confirm with other sources over the internet with Yahoo! Answers.

2. Quora

Quora[25] is a site where people post answers to your questions. It also allows you to follow Topics, People, and specific Questions, which is great for keeping up with trends and questions that you never ran into yet. Its advantage lies in its community of reputable experts. Quora covers a wide variety of topics and interests, and is bound to have something for your curiosity. User can also choose to search for answers to specific questions right in the search bar on every page.

3. WikiAnswers

WikiAnswers[26] is designed to help users gather information and provide knowledge. The website allows its users to ask questions and post responses to those posed by others. Users can also toggle between those submitted by the community and references topics.

4. Qhub

Qhub[27] lets users create their own, professional looking question and answer area to go with a website or application. Qhub integrates with Facebook and Twitter so users can bring their answers to social networks.

[24] https://answers.yahoo.com/

[25] https://www.quora.com/

[26] http://answers.wikia.com/wiki/Wikianswers

[27] http://qhub.com/

A moderator control panel gives the user control so they can approve or delete entries or make edits as necessary. Filters are included so users can catch questionable submissions before they appear on their Q&A page. Along with post filters, language filters are also available so users can prevent abusive language issues

5. Answers.com

Answers.com[28] is another wiki-styled source for information.

6. O'Reilly Answers

O'Reilly Answers[29] is a community site for sharing knowledge, asking questions, and providing answers.

7. Ask.fm

Ask.fm[30] is a questions and answers platform.

8. Ask Q&A

Ask[31] offers users a way to get answers for their questions through a free, easy to use platform

News forums

1. MailOnline

MailOnline carries News, sport, celebrity, science and health stories.

2. NewsForum.com

The friendly general discussion community for all types of news! At News Forum[32] we encourage intelligent, non-biased, discussion and friendly debates on everything from entertainment to politics.

3. Rednews UK forum

Rednews[33] contains news, sport, celebrity, science and health stories.

[28] http://www.answers.com/

[29] http://support.oreilly.com/oreilly

[30] http://ask.fm/

[31] https://www.reference.com/

[32] http://www.newsforum.com/

[33] http://www.rednews.co.uk/forum/forumdisplay.php/9-RED-NEWS-FORUMS

4. Neowin

It is for non-technical discussion and light hearted news[34]. Political news, questions and debates must be posted in Real World Issues.

5. Orange Power

A world news and political discussion forum[35].

Political forums

1. Political Forum

Political Forum[36] message board contains current events, polls, debate, and humor for all US politics, world politics & political issues

2. Digital Point Forum

The Digital Point[37] Forum is the latest forum. It has sections for everything about internet marketing, including: search Engines: with subcategories Google, Yahoo, Microsoft, and Directories. Marketing: with subcategories General Marketing, Search Engine Optimization, Social Networks, Link Development, PPC Advertising, and Affiliate Programs.

3. Canadian content

Canadian content[38] is the place where Canadians can post and discuss current events. If you want to discuss something more of a political nature, please browse to our Political forums and start a discussion there. Please refrain from posting one liners and make sure you title your threads on-topic with the current event description. News articles, links and discussion about current events.

4. DebatePolitics

[34] http://www.neowin.net/forum/forum/58-real-world-news/

[35] https://www.orangepower.com/forum/world-news-politics.22/

[36] http://www.politicalforum.com/forum.php

[37] http://www.smartpassiveincome.com/5-of-the-best-internet-marketing-and-blogging-forums/

[38] http://forums.canadiancontent.net/news/

This is a political forum[39] that is non-biased/non-partisan and treats every person's position on topics equally. This debate forum is not aligned to any political party. In today's politics, many ideas are split between and even within all the political parties. Often we find ourselves agreeing on one platform but some topics break our mold. We are here to discuss them in a civil political debate

5. US Politics online

This is a non-biased political forum[40] to discuss USA politics.

Programming forums

1. Stack Overflow

Stack Overflow[41] is a Q&A site dedicated to answering inquiries about programming. There are specific questions about chunks of code, or mechanisms and how they function. Users can have their questions voted up or down, and that determines how much visibility each one gets.

2. Super User

Super User[42] is a community that collaborates and gives advice on how to help out computer enthusiasts with their questions. It is geared more towards the power user, hence you'll find geeky questions and their more geeky answers abound on the site.

3. Ask Ubuntu

Ask Ubuntu[43] is a question and answer site for Ubuntu users and developers. Join them; it only takes a minute.

4. CFD Online Discussion Forums

CFD[44] is a free community for everyone interested in Computational Fluid Dynamics.

5. Coding Forums.com

[39] http://www.debatepolitics.com/
[40] https://www.uspoliticsonline.com/
[41] http://stackoverflow.com/
[42] http://superuser.com/
[43] http://askubuntu.com/
[44] http://www.cfd-online.com/Forums/

Coding Forum[45] is a C/C++ programming forum. Provide solutions to a variety of questions.

6. HOTSCRIPTS

Hot scripts[46] is a C and C++ Forum and Online Communities. Visit Hot Scripts today for the largest collection of C and C++ Scripts that free or commercial to all web developers.

[45] http://www.codingforums.com/

[46] http://www.hotscripts.com/category/scripts/c-cpp/forums-online-communities/