Journal of Theoretical and Applied Information Technology
15th April 2012. Vol. 38 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

QUALITY-BIASED RETRIEVAL IN ONLINE FORUMS

[1]AMEER TAWFIK ALBAHAM, [2]NAOMIE SALIM

[1]Faculty of Computer Science and Information System, UTM, Johor, Malaysia-81310

[2]Prof., Facultyof Computer Science and Information System, UTM, Johor, Malaysia-81310

E-mail: ameer.tawfik@gmail.com, naomie@utm.my

ABSTRACT

This paper presents an ongoing research in utilizing content quality to enhance thread and post retrieval tasks in support oriented forums. First, we adapt the established information quality approach as a theoretical way to understand quality. Then, we outline research areas that help in quantifying quality. Lastly, we propose to use learning to rank method to weight quality parameters while ranking.

Keywords: *Online Forums, Information Quality (IQ), Information Retrieval (IR), Learning To Rank (LTOR).*

1. INTRODUCTION

Online discussion forums or message boards are web-based software that enables people with the same backgrounds, interests or geographic locations to build virtual communities. Usually, a user starts a discussion through posting a post (Initial post); asking for help or opening a dialog for discussion. Then, the other users read the initial post and reply to it. Each initial post and its replies are grouped into threads. The complete thread posts list provides a cohesive view of the discussion. Threads are grouped into sub-forums according to their theme which in turn builds up the entire community.

Online forums' asynchronous nature and accessibility enableusersto share and seek knowledge by contributing thousands of posts. As a result, support oriented forums have become repositories of hidden, valuable and huge volume of archived knowledge [1, 2].

Nevertheless, users need search tools to fully utilize such knowledge [1, 3-6]. By default, forums provide three methods to search: keywords matching, database backend full text search and customized web search. However, these methods are not effective. The first method is limited because it is just a keyword matching[4, 7]. The second method causes confusion. Database engine indexes posts rather than threads. Therefore, when the database search engine matches a user's query against posts, it calculates the similarity score between all posts in the post table and user's query. Later, it returns the highest scoring posts to users. Returning a single post might confuse users as it will be taken out of the discussion context especially if the post addresses other posts in the thread. Lastly, some forums allow users to search using commercial webs search. However, this method is not adequate due to the different structure and nature of retrieval in online forums [1, 3, 8].

As a result, many studies have paid attention to forum unique characteristics through introducing special link-based algorithms[8, 9] or leveraging forums structure to improve thread and post retrieval[1, 3, 4, 10, 11].

Nevertheless, the previous proposed methods assumed that all content have equal quality. This assumption is inherited from traditional information retrieval evaluation context. In ad hoc evaluation experiments, documents quality was assumed be constant or ignored[12]. That is because of the nature of evaluation corpus used to assess ranking function performance[13]: most of them were collection of newswire and government documents. As a result, the quality of document was assumed to be distributed uniformly[12]. However, the characteristic of collection has changed especially web and user generated content hence this assumption is not valid anymore[12, 13]. Additionally, studies in web and user generated content retrieval found that quality scores not only contribute a significant improvement in search performance [6, 12-17], but also enable users to make an informed decision about search result[18].

Therefore, we hypothesize that users in online forums are more interested in content that have more quality. To test the hypothesis, the following research questions are to be answered:

- What are the parameters of forums thread and post quality?
- How to make search engines biased toward good content?

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Does utilizing quality improve search performance?

In the next section, we discuss related works to information quality, forum ad hoc retrieval tasks and approaches in incorporating quality into ranking. In section 3, we detail our candidate proposed method. Lastly, section 4 outlines our future works.

2. RELATED WORK

2.1 Information quality

In the context of information quality assessment, researchers and practitioners do not distinguish between information and data[19]. Putting that in mind, a common definition of data quality is data's fitness for use[19]. Following this definition, several information quality frameworks have been proposed[20, 21]. Knight and Burn [22]studied several frameworks and extracted common used dimensions. Some of those dimensions are accuracy, consistency, timeliness, completeness, accessibility, objectiveness and relevance.

Nevertheless, in the context of this research, two issues are to be resolved. The first issue is the understanding of what contributes to threads or posts quality. The second issue is how to automate and quantify content quality assessment.

As pertaining to the first issue, there has been no attempt to understand quality of post or thread in online forums as perceived by forums users.

In contrast, one way to automatically estimate quality is to utilize user feedback such rating. However, manual rating is not reliable in online forums [23-26]. As a result, some studies tried to solve the problem of manual rating by classifying forums posts as low, good or high quality posts [25-28]. Nevertheless, some researches utilized different class of evidences to measure quality of content. You et al [28] developed a wavelet based method to identify high quality topics in forums. They found that threads that attract many authors and replies while lasting for long period contain high quality content. Sun et al[29] proposed a real time forums crawler that utilizes quality. They used thread's number of views, number of replies and temporal features. However, all previous studies failed to explain why those posts or threads are good or useful. Is it because of their informative value, readability or objectivity?

Therefore, in this paper, we propose to derive content quality dimensions using information quality approach, and then we quantify these dimensions.

2.2 Ad hoc retrieval in online forums

In the literature of forums ad hoc retrieval, there two groups of researches. The first group focuses in finding posts [1, 10, 11, 30], whereas the other group treats thread as unit of retrieval [1, 3, 4, 6, 8, 9, 31]. The following subsections discuss them respectively.

2.2.1 Post retrieval

Finding relevant posts in online forums has been applied in several applications. Although they all rely in the fundamental concept that is given a query return a list of potential relevant posts, they differ in the ranking strategy and query nature. These applications are chatbot extraction[32, 33], answers detection[34, 35], thread structure discovery[1, 10] and post ad hoc retrieval in threaded discussion[1, 10, 11, 30]. This study addressesonly post ad hoc retrieval in which the query is a natural language query.

Recent studies in post retrieval focused in how to incorporate thread structure to smooth and improve post retrieval[1, 11]. Both [1] and [10] first recovered thread structure, and then they used thread structure to improve performance. In contrast, [11] used raw thread structure to improve performance. Using raw thread structure appeals prosing as it does not require training. All recent studies used language model framework[36] to estimate post relevance. Additionally, [11] found that language model approaches have equal or better performance than BM25.

2.2.2 Thread ad hoc retrieval

The other task of interest is thread retrieval. Nevertheless, the thread retrieval is not a trivial task. That is because thread is not the unit of contribution. The unit of contribution is posts. As a result, researchers in thread retrieval have leveraged various combinations of ranking methods and document representations to estimate thread relevance.

First attempt was to build a special PageRank algorithm for online forum[8, 9]. Xu and Ma [8] built implicit links between threads by extracting each thread topic and constructing a topic hierarchy from extracted topics. After experimenting with the new algorithm (Fing-graindRank), Fine-grainedRank was found to perform better than

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

normal Page Rank. Another link-based algorithm is Posting Rank [9]. The authors argued that users' interactions can be used to improve ranking in forums. Base on the assumption that the more common repliers between threads, the more the threads are related. The co-existence of users implies a mutual recommendation. They used Posting Rank to measure authority of thread pages and BM25 to measure relevance. Their study showed that Posting Rank outperformed Page Rank. Although, both studies confirm the inferiority of web search algorithm with respect to a special Page Rank for online forums, they suffer from a similar problem. In both studies, the unit of retrieval is thread pages. Ranking pages instead of threads might confuse users: If the relevant information is in the third or fourth page, returning these pages might be out of context[3]. Additionally, users, when they search forums expect either thread or posts. Therefore, returning web pages is not consistent with what users want.

Recent studies have given more attention to thread structure by considering thread as unit of retrieval [1, 3, 4, 6, 31]. However, the challenge as mentioned before is how to calculate thread relevance since posts are the textual representation of threads. One approach is to concatenate all posts content into one large virtual document [1, 4, 6, 31]. However, this approach suffers from problem of low relevant content swapping away high relevant information[1]. Additionally, other approaches have been proven to be superior [1, 4].

Another approach is to estimate thread relevance by fusing individual posts relevance scores [1, 3, 4]. This approach can be divided into three techniques. The first technique is a plain aggregation of posts' scores [4]. However, it has been found to produce no significance improvement over the one virtual document representation. The second technique is to consider only a subset of the thread's posts [1, 4]. Both [1, 4] adapted pseudo cluster selection method proposed by [37] to thread retrieval. The applied method consistently outperformed other retrieval methods. The third technique is to utilize inference network[3]. [3] adopted inference network retrieval[38] to online forums thread search. In addition to thread posts,the authorsleveraged other structure information such title. In their approach, they divided thread posts into initial post and reply posts. Then, each initial post and its replies form a network. The final score is estimated using inference network framework[38]. Once again, this approach produced better result than the one virtual document.

Some far, all discussed techniques tackle the problem of query to thread relevance estimation. Nevertheless, some studies do utilize other relevance evidence such document usefulness [3, 6, 31] and authority [3, 8, 9]. Raghavan et al. [31] gives more weight to threads that has at least one solution. However, there was no performance evaluation. Wang et al.[39] used BM25 to get an initial result of threads, and then they re-ranked threads base on their credibility and argument quality. Although, re-ranking using credibility and argument quality proved superior to BM25, their first step is not optimal. They considered thread as one whole document resulted from concatenation of posts. As discussed previously, this method is not effective. Zhang [6] extended [39] work by training a genetic algorithm to weight the importance of each quality indicators. Recently, [3] used thread number of replies, number of links and authority of users as prior probabilities . When applying each prior into the query language model, it was found to outperform baseline methods. However, when combine all priors, the performance was the worst. One analogy of that counter intuitive result is that these priors are useful but need an appropriate method to fuse their scores.

2.2.3 Discussion

Researches in post ad hoc retrieval have been focusing in leveraging thread structure. Additionally, most of them used the query language model. However, no attempt has been made to incorporate quality of post content into ranking.

In contrast, attempts have been made to utilize the notion of quality in thread retrieval[3, 31]. However, those attempts address only some aspects of quality heuristically. For instance,[31] considered one aspect of quality which is whether thread contains solution or not. Surely, the existence of solution is an aspect of quality but it is not the whole story. Similarly, in [3], user's authority, threads linkages, and length are indicators of some quality dimensions such as thread source credibility, thread reputation, informativeness. In fact, [6] used more dimensions such thread completeness, relevance , richness and timeliness to improve ranking. Therefore, one could see that each study uses different set of dimensions base on the researchers understanding of quality. However, it is more adequate to ask the users of online forums to outline what is quality and then automate quality quantification.

That is exactly what this study is trying to accomplish. Our quality framework is based on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

users' feedback, forums guidelines and researches works in this area. Therefore, our quality models not only are more comprehensive, but they are also complimentary to previous works.

2.3 Quality Based Retrieval

In literature there have been two main streams in incorporating quality indicators into ranking. The first approach casts the problem of utilizing quality parameters into ranking as two independent tasks. The first task is quality estimation using machine learning algorithms and the second task is incorporating the quality score into ranking. We call this approach the two stage approach. The second approach utilizes quality parameters alongside the retrieval experiment which we call the one stage approach. In this approach, the machine learning techniques are used to learn quality parameters importance from information retrieval objective.

2.3.1 Two stages approach

Researches following this approach operate in the notion that quality of document is independent from retrieval and that quality has its application to information filtering. Therefore, they first train a machine learning classifier to predicate the quality of document then the quality score is incorporated into ranking. The classification is conducted using entropy based classifier [13, 15, 17] or support vector machine classifier [14]. As for the second task, it is either prior document of query language model [13, 15, 17] or weighted sum of relevance and quality scores[14]. Intuitively, this approach is the easiest but it has two limitations. First, if the objective is to increase search performance through quality, there is an extra cost of building an appropriate data collection for classification. Second, from information retrieval perspective, it is not possible to identify which feature or dimension has an impact in retrieval performance. That is because that the final output of the classifier is a number that indicates the overall quality of the document.

2.3.2 One stage approach

In this approach the quality features are utilized directly to improve retrieval performance. Works following this approach utilize the concept of prior document probability. Tow techniques are used in this approach: parameter tuning[16] and learning to rank[12]. In this study, learning to rank will be used as it is more conceptually and theoretically sound to information retrieval[40].

3. PROPOSED METHOD

Our proposed method consists of two main components. The first component is the quality models for thread and post content. The second component is the quality biased retrievals in online forums. The rest of this section describes our methodology to develop these components.

3.1 Multi-dimensional quality models in online forums

The development of the quality models consists of two steps: dimension identification and quantification.

3.1.1 Dimension identification

The primary aim of this stage is to elicit the different dimensions of what makes useful information. To achieve that, three sources will be investigated. The first source is a pilot study in four support oriented forums. The second source is an analysis of 10 forums guidelines. The third source is current literature in content quality estimation.

PILOT STUDY

The pilot study is to elicit initial quality dimensions as perceived by users. Ubuntu[1], tripadvisor[2] and Cnet[3] forums are chosen because they are the sources of this study dataset[3, 11]. The VBCity[4] forum is selected because it was used in the most similar work to this study[6]. To collect users' opinions, a question was posted in these forums. The question was "What makes a good post or good thread?" Base on users' replies, dimensions and metrics are extracted. Examples below show some users responses and potential dimensions.

Example 1 :
Response: "#1 a descriptive title #2 one that stays on topic "
Dimension: Descriptive title, cohesiveness
Level : Thread

Example 2:

[1]http://ubuntuforums.org/
[2]http://www.tripadvisor.com/ForumHome
[3]http://forums.cnet.com/
[4]http://vbcity.com/forums/default.aspx

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Response:

"I judge the initial thread posts to support forums based on whether the information provided allows me to have some idea of what the problem is and where to start investigating a solution to the problem. I judge responses to the threads based on whether the advice will work or not."

Dimension: Informativeness, accuracy

Level : Message

FORUMS GUIDELINES

Every forum has guidelines that govern user contribution or provide tips in how to utilize the forum content. The reason to consult these guidelines is they are the de facto standard of contribution. Users are expected to abide these guidelines. Furthermore, forums moderators use them to monitor thread discussion.

In addition to the study dataset forums, another 7 forums are selected from the literature. Several information retrieval related studies are considered. These areas are ad hoc retrieval, expert finding, question answer detection and question routing. The reason to choose these areas is that most of these studies focus in support oriented forums. In extracting dimensions, the same method applied in pilot study method is applied.

CONTENT QUALITY LITERATURE

The third source to collect information quality parameters is literature [23, 25, 26, 28, 29]. These studies provide features that can be mapped to dimensions. In other words, at least dimensions extracted from them are guaranteed to be measureable.

3.1.2 Dimension quantification

After finding the quality dimensions, three types of works are to be examined. The first area is work in quality estimation in online forums [24-26].The second area is work in question and answers detection methods[34, 35]. The question –answer methods provide tools to locate answers to first thread post. In other words, these tools provide ways to measure the relevance of reply. The third area is social network[41, 42] and expert finding[43] as they will help in measuring source credibility.

3.2 Quality biased retrieval

Query Language model[36] based methods will be used to measure the relevance of threads or post with user query. Query language model is a suitable approach for four reasons. Firstly, query language model's approach toward information retrieval strongly exists in online forums. The model tries to simulate how users come up with search query. Furthermore, online communities originate around users sharing common dominators. Therefore, searching base on how users could write documents and words is logically sound: the knowledge seeker and provider share the same backgrounds and thoughts hence there is a higher possibility that both of them could express their ideas using similar words and terminologies. Therefore, the proposed model fits online forums environments. Secondly, the prior document probability component of the model could be used to plug-in document quality. That is has been successfully applied in several studies [13, 15, 17, 44]. Thirdly, most researches in forums ad-hoc retrieval used query language model [1, 3, 4, 11, 45]. Lastly, query language model is known to perform better than other information retrieval models [36, 46, 47]. The rest of this subsection explains the query language model and how quality is plugged into it.

According to Ponte and Croft [36], under maximum likelihood estimation, the probability of generating the term t from document model M_d is given by equation 1:

$$P_{(t|M_d)} = \frac{tf(t,d)}{dl} \quad (1)$$

Where tf is the frequency of term t in document d and dl is the number of tokens in document d. Then, under the assumption of term independency, the probability of generating a query $Q = q_1, q_2, q_2 \ldots q_m$ from document d is given by equation 2:

$$P_{(Q|M_d)} = \prod_{q \in Q} \frac{tf(t,d)}{dl} \quad (2)$$

According to the pioneers of this model, this equation has one problem. If one term of the query terms is not present in the document, then the whole probability will be zero. To rectify this problem, several techniques have been proposed. These techniques are called smoothing techniques. According to Zhai and Lafferty [48], the primary common smoothing methods are Jelinek-Mercer, Dirichlet and Absolute discount.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In the context of information retrieval, there are set of document $D = d_1, d_2, d_3 \ldots d_3$. The probability of the likelihood of relevance of document d_i given the observed query set Q is given by Bayes's theorem in equation 3:

$$P(d_i|Q) \propto P(d_i)P(Q|d_i) \quad (3)$$

$P(Q|d_i)$ is the likelihood of generating Q from document d_i calculated using equation 2. $P(d_i)$ is the probability of observing d_i. This is called the prior probability of the document being relevant. Most of time, it is assumed to be constant hence it does not affect the ranking[47]. However, in some cases, it has been used to include other features such as authority or styles of documents[44].

Since the hypothesis of this study is that a user is more interested in high quality content, we could replace prior probability $P(d_i)$ in equation 3 with the quality score of thread and post respectively. The same approach has been applied in [13, 17]. The result will be as in equation 4:

$$P(d_i|Q) = P(d_{iQlty})P(Q|d_i) \quad (4)$$

Where d_i is either thread or post and $P(d_{iQlty})$ is the estimated thread or post quality prior. To estimate the $P(d_{iQlty})$, we adapt the learning to rank method proposed by [12]. $P(Q|d_i)$ is estimated using query language models . In the case of post retrieval task, standard query language model will be used. However, for thread retrieval, Pseudo-cluster selection model[1, 4] will be used. Additionally, this study assumes term independence and uses Dirichlet smoothing[49].

4. CONCLUSION AND FUTURE WORKS

In this paper, we discussed the need to utilize information quality in online forums to enhance search performance. We also outlined current approaches to leverage quality in ranking functions. Consequently, we proposed a method that leverages content quality to enhance post and thread retrieval tasks. Our proposed method consistent of two components; a multi-dimension quality models and quality biased retrieval methods.

Future work will focus in the multi-dimensional quality models development for thread and posts. Afterward, the proposed quality biased retrieval methods will be test in two datasets; post[11] retrieval and thread[3] retrieval test collection.

Acknowledgment

This research is sponsored by Ministry of Science Technology and Innovation under the university research grant vote number 01H74, Universiti Teknologi Malaysia.

REFERENCES

[1] Seo, J., W. Bruce Croft, and D. Smith, *Online community search using conversational structures*. Information Retrieval, 2011: p. 1-25.

[2] Lin, C., et al., *Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications*, in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, ACM: Boston, MA, USA. p. 131-138.

[3] Bhatia, S. and P. Mitra. *Adopting Inference Networks for Online Thread Retrieval*. in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010. Atlanta, Georgia, USA, .

[4] 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: Boston, MA, USA. p. 714-715.

[5] Elsas, J.L. *Search in Conversational Social Media Collections*. in *Proceedings of the Third Annual Workshop on Search in Social Media (SSM 2010)*. 2010. New York City, USA.

[6] Zhang, X., *Effective Search in Online Knowledge Communities: A Genetic Algorithm Approach*. 2009, Virginia Polytechnic Institute and State University: Blacksburg, Virginia, USA.

[7] Baldwin, T., et al., *Intelligent linux information access by data mining: the ILIAD project*, in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. 2010, Association for Computational Linguistics: Los Angeles, California. p. 15-16.

[8] Xu, G. and W.-Y. Ma, *Building implicit links from content for forum search*, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, ACM: Seattle, Washington, USA. p. 300-307.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[9] Chen, Z., L. Zhang, and W. Wang, *PostingRank: Bringing Order to Web Forum Postings*, in *Information Retrieval Technology*, H. Li, et al., Editors. 2008, Springer Berlin / Heidelberg. p. 377-384.

[10] Wang, H., et al. *Learning Online Discussion Structures by Conditional Random Fields*. in *The 34th Annual International ACM SIGIR Conference (SIGIR'2011)*. 2011.

[11] Duan, H. and C. Zhai, *Exploiting Thread Structure to Improve Smoothing of Language Models for Forum Post Retrieval*, in *To appear in the 33rd European Conference on Information Retrieval (ECIR 2011)*. 2011: Dublin, Ireland.

[12] Bendersky, M., W.B. Croft, and Y. Diao, *Quality-biased ranking of web documents*, in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, ACM: Hong Kong, China. p. 95-104.

[13] Zhou, Y. and W.B. Croft, *Document quality models for web ad hoc retrieval*, in *Proceedings of the 14th ACM international conference on Information and knowledge management*. 2005, ACM: Bremen, Germany. p. 331-332.

[14] Chen, C.C. and Y.-D. Tseng, *Quality evaluation of product reviews using an information quality framework.* Decision Support Systems, 2011. **50**(4): p. 755-768.

[15] Suryanto, M.A., et al., *Quality-aware collaborative question answering: methods and evaluation*, in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. 2009, ACM: Barcelona, Spain. p. 142-151.

[16] Weerkamp, W. and M.d. Rijke. *Credibility Improves Topical Blog Post Retrieval*. in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistic: Human Language Technologies (ACL-08: HLT)*. 2008. Columbus, Ohio, USA: Association for Computational Linguistics.

[17] 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: Seattle, Washington, USA. p. 228-235.

[18] Yamamoto, Y. and K. Tanaka, *Enhancing Credibility Judgment of Web Search Results*, in *CHI 2011*. 2011, ACM: Vancouver, BC, Canada.

[19] Strong, D.M., Y.W. Lee, and R.Y. Wang, *Data quality in context.* Commun. ACM, 1997. **40**(5): p. 103-110.

[20] Lee, Y.W., et al., *AIMQ: a methodology for information quality assessment.* Inf. Manage., 2002. **40**(2): p. 133-146.

[21] Kahn, B.K., D.M. Strong, and R.Y. Wang, *Information quality benchmarks: product and service performance.* Commun. ACM, 2002. **45**(4): p. 184-192.

[22] Knight, S.-a. and J. Burn, *Developing a Framework for Assessing Information Quality on the World Wide Web.* Informing Science Journal, 2005. **8**.

[23] Chai, K., et al. *Assessing post usage for measuring the quality of forum posts*. in *4th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. 2010.

[24] Brennan, M.R., S. Wrazien, and R. Greenstadt, *Learning to Extract Quality Discourse in Online Communities*. 2010. 2010.

[25] Wanas, N., et al., *Automatic scoring of online discussion posts*, in *Proceeding of the 2nd ACM workshop on Information credibility on the web*. 2008, ACM: Napa Valley, California, USA. p. 19-26.

[26] Weimer, M. and I. Gurevych, *Predicting the Perceived Quality of Web Forum Posts*, in *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*. 2007.

[27] Kim, J., et al., *Modeling and Assessing Student Activities in On-Line Discussions*, in *In Proceedings of the AAAI Workshop on Educational Data Mining*. 2006.

[28] You, C., C. Xue-Qi, and H. Yu-Lan. *A Wavelet-Based Model to Recognize High-Quality Topics on Web Forum*. in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*. 2008.

[29] Sun, J., H. Gao, and X. Yang, *Towards a quality-oriented real-time web crawler*, in *Proceedings of the 2010 international conference on Web information systems and mining*. 2010, Springer-Verlag: Sanya, China. p. 67-76.

[30] Xi, W., J. Lind, and E. Brill, *Learning effective ranking functions for newsgroup search*, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, ACM: Sheffield, United Kingdom. p. 394-401.

[31] Raghavan, P., et al., *Classification and Retrieval from Mailing Lists and Forums*, in *Forum for Information Retrieval Evaluation*. 2010: DAIICT, Gandhinagar, India.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[32] Huang, J., M. Zhou, and D. Yang, *Extracting Chatbot Knowledge from Online Discussion Forums.* 2006.

[33] Feng, D., et al., *An intelligent discussion-bot for answering student queries in threaded discussions*, in *Proceedings of the 11th international conference on Intelligent user interfaces*. 2006, ACM: Sydney, Australia. p. 171-177.

[34] 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: Boston, MA, USA. p. 171-178.

[35] 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: Singapore, Singapore. p. 467-474.

[36] Ponte, J.M. and W.B. Croft, *A language modeling approach to information retrieval*, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, ACM: Melbourne, Australia. p. 275-281.

[37] Seo, J. and W.B. Croft, *Blog site search using resource selection*, in *Proceeding of the 17th ACM conference on Information and knowledge management*. 2008, ACM: Napa Valley, California, USA. p. 1053-1062.

[38] Turtle, H. and W.B. Croft, *Evaluation of an inference network-based retrieval model.* ACM Trans. Inf. Syst., 1991. **9**(3): p. 187-222.

[39] Wang, G.A., J. Jiao, and W. Fan, *Searching for Authoritative Documents in Knowledge-Base Communities*, in *Proceeding of International Conference on Information Systems (ICIS)*. 2009, Association for Information Systems Electronic Library (AISeL): Phoenix.

[40] Liu, T.-Y., *Learning to Rank for Information Retrieval.* Found. Trends Inf. Retr., 2009. **3**(3): p. 225-331.

[41] Skopik, F., H.-L. Truong, and S. Dustdar, *Trust and Reputation Mining in Professional Virtual Communities*, in *Proceedings of the 9th International Conference on Web Engineering*. 2009, Springer-Verlag: San Sebasti\&\#225;n, Spain. p. 76-90.

[42] Petrovčič, A., V. Vehovar, and A. Žiberna, *Posting, quoting, and replying: a comparison of methodological approaches to measure communication ties in web forums.* Quality & Quantity, 2011: p. 1-26.

[43] Seo, J. and W.B. Croft, *Thread-based Expert Finding*, in *SIGIR'09 SSM Workshop*. 2009, ACM: Boston, Massachusetts.

[44] Peng, J. and I. Ounis, *Combination of document priors in web information retrieval*, in *Proceedings of the 29th European conference on IR research*. 2007, Springer-Verlag: Rome, Italy. p. 732-736.

[45] Elsas, J.L. and N. Glance, *Shopping for Top Forums: Discovering Online Discussion for Product Research*, in *ACM KDD SOMA 2010 Workshop on Social Media Analytics.* 2010, ACM: Washington, DC, USA.

[46] Zhai, C., *Statistical Language Models for Information Retrieval A Critical Review.* Found. Trends Inf. Retr., 2008. **2**(3): p. 137-213.

[47] Manning, C.D., P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. 2008, New York, NY, USA: Cambridge University Press.

[48] Zhai, C. and J. Lafferty, *A study of smoothing methods for language models applied to Ad Hoc information retrieval*, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001, ACM: New Orleans, Louisiana, United States. p. 334-342.

[49] Zhai, C. and J. Lafferty, *A study of smoothing methods for language models applied to information retrieval.* ACM Trans. Inf. Syst., 2004. 22(2): p. 179-214.