Journal of Theoretical and Applied Information Technology
10th May 2014. Vol. 63 No.1

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

MINING FAQ FROM FORUM THREADS: THEORETICAL FRAMEWORK

[1]ADEKUNLE ISIAKA OBASA, [2]NAOMIE SALIM

[1]SCRG Lab, Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

[2]Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

E-mail: [1]iaobasa@yahoo.com , [2]naomie@utm.my

ABSTRACT

Frequently Asked Questions (FAQ)'s tag is becoming more popular on websites. Research activities have been concentrated on its retrieval rather than construction. FAQ construction can be achieved using a number of sources. Presently, it is mostly done manually by help desk staff and this tends to make it static in nature. In this paper, a comprehensive review of various components that can guarantee effective mining of FAQ from forum threads is presented. The components encompass pre-processing, mining of questions, mining of answers and mining of the FAQ. Besides the general idea and concept, we discuss the strengths and limitations of the various techniques used in these components. In fact, the following questions are addressed in the review. What kind of pre-processing technique is needed for mining FAQ from forum? What are the recent techniques for mining questions from forum threads? What approaches are currently dominating answer retrieval from forum threads? How can we cluster out FAQ from question and answer database?

Keywords: *FAQ, Forum thread, Question answering, mining FAQ, Internet forum*

1. INTRODUCTION

FAQ stands for "frequently asked questions." As the name implies, it is a type of web page (or group of web pages) that lists questions frequently asked by users, usually about different aspects of the website or its services. The answers are typically shown with the questions. FAQ pages have become commonplace on many websites for many reasons. The main reason is that, they offer a way to provide support, most commonly, customer support, without having to repeat solutions to common problems. For large companies, a good FAQ page can have effect on the number of supporting staff that are needed to be hired.

A preliminary investigation carried out by [1, 2] revealed that about 30 to 40 per cent of inquiries sent to help desk operator are related to FAQ. Therefore, over 30 per cent of inquiries are solved by simply showing FAQ. Building up FAQ is an important task of help desk. Companies build up and post FAQ on their Web site to reduce the number of inquiries from users. Users can browse the FAQ pages and clear up their unfamiliar matters before they send emails to help desk. Help desks also analyses inquiry records, constructs question and answer sets, and adds them to their FAQ pages. The task of analysing great deal of inquiries, however, needs a lot of time, hence, the need for automatic FAQ. With the aim of overcoming the costs of call centres, companies usually try to anticipate typical customer's questions and answer them in advance. Those users questions and expert answers about specific domains are often collected and organized in FAQ's lists [3].

According to Pareto 80/20 Principle as cited in [4] discovered that 20% of users' questions are frequently asked, and the rate of occurrence of these frequently asked questions amounted to about 80% of the frequency of all questions asked. The 20% of users' questions that are asked frequently are called Frequently Asked Questions. In the real sense of it, human activities are repetitive, we repeatedly eat certain kinds of foods, repeatedly wear certain types of clothes, likewise many other things. The Frequently asked questions are now being used by some QA systems to build FAQ databases. The FAQ database is used to search for questions similar to the question asked by user. Whenever similar question is found, the similar FAQ's answer is returned to the user as a candidate answer. FAQ question answering system can conveniently answer about 80% of the questions

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

asked by user if properly constructed and effectively managed.

There are a number of problems that need to be tackled for FAQ to achieve this ground but the most prevailing ones are Sentence Similarity Computing and FAQ Construction [4]. FAQ can be constructed from a number of sources; some of these sources are shown in table 1. In this paper, we shall focus on FAQ construction from Internet forums. Our consideration for forum lies in the fact that there is less segregation in forum when compared with Community-based Question Answering (CQA) and Mailing list. For this, any member can contribute anything so the restriction is less.

2. INTERNET FORUMS

An Internet forum, which is also known as discussion board, can be considered as a topic-based document set that has a definite boundary separated by members and non-members. Forums are either public or private. The private forums usually have stronger participant boundaries. Almost all forums have hierarchical structures. A forum comprises of sub-forums depending on the broad topic categories. A sub-forum is made up of threads. A thread is the minimal topical unit that addresses a specific topic. A thread is usually initiated by an author's post (usually called initial post), which constitute the topic of discussion. Members who are interested in the topic send reply posts. The reply posts relations establish what is known as conversational structure which can be classified as the entire thread, a post, a pair of posts or a dialogue that starts with an initial post and terminate in a leaf node [5]. Hence, forums typically have both hierarchical structures and conversational structures. A schematic diagram of Internet forum is shown in figure 1.

A framework for the mining of human generated contents of forums into FAQ is the focus of this paper. Previous studies on FAQ have focused mostly on the retrieval [6-11]. The work is non-trivial due to the nature of content generated in forum. Forum contents contain short sentences that pose difficulty for similarity measurement models. Also, much of the contents are delivered like spoken speech and as a result a lot of noise and redundant text used to be created. The characteristics of human contents generated in forums along with their associated problems are given in table 2. Much work have been recorded in the areas of FAQ retrieval and maintenance as seen in the literature but little efforts have been put on automatically mining FAQ. In fact, to the best of our knowledge, only two works so far attempt generating FAQ from forums: [12] and [13]. The scanty work in the area may be attributed to the challenges summarised in table 2.

3. OVERVIEW OF THE FRAMEWORK

FAQ is domain dependent. Research activities in restricted domain question answering addresses problems concerning the incorporation of domain-specific information into current state-of-the-art QA technology. The aim of doing this is to achieve deep reasoning capabilities and reliable accuracy performance in real world applications. In this framework, we shall consider four different components that can guarantee successful mining of FAQ, namely, pre-processing, mining of questions, mining of answers and FAQ generation.

3.1 Pre-processing

The noisy nature of forum calls for an extensive noise removal concepts. Standard questions and answers are expected to be considered as FAQ. Redundant phrases need to be removed. These phrases are often used to express appreciation or emotion of the writer. Examples of such text are shown in table 3. Detailed discussion of these phrases is contained in [14]. Regular expression heuristic for removing redundant text is discussed by [13]. Error types among others include spelling and out of vocabulary. Frequency of spelling errors has been found to be higher than other types [15]. A good number of queries given to forums contain words which are legitimate but not present in Standard English dictionaries. Examples of these words are pics, hotmail, ebay, etc.

Preliminary investigation of forum threads revealed the four classes of noise shown in table 4. In fact, about 70% of text inspected contains the four classes of errors. In order to be able to mine meaningful questions and answers from forum threads adequate attention needs to be paid to noise level. The pre-processing steps shown in figure 2 are suggested for mining FAQ from forum threads.

3.2 Mining of Questions

Mining of questions is a major phase in the development of a FAQ system from forum threads. The process is not a trivial one given the nature of forum threads as depicted in table 2 above. Mining of questions from forums started with the work of Cong et al [16] and since then a host of other works have been carried out. Review of these works is

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

shown in table 5. However, question detection from forums has not been investigated thoroughly. The importance of this process in FAQ generation should be well noted. It will be difficult getting right answer for an ill-formed question. Question detection approaches generally can be considered to be a method based on *learning* or *non-learning* as pointed out by [17]. To the best of our knowledge only the work of [18] centres around non-learning heuristics. As we can see in table 5, the performance of the approach is impressive. In the email domain, [17] used non-learning approach called regex to achieve a result close to learning method. Looking at the demands of learning approach – *time consuming*, *too expensive* and *complexity*, the non-learning appears to be promising. A good tool that can be used in this regard is regular expression. Regular expressions can be built around question detection heuristics to mine questions from forum. Question mining schematic diagram is shown in figure 3.

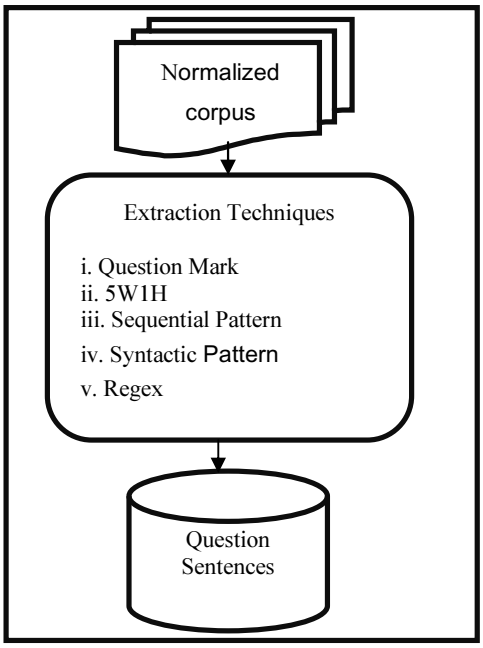

Figure 3: Mining of Question Sentences

3.3 Mining of Answers

Mining of answers is another important process in FAQ generation. In fact, it is good not to have an answer than to have a wrong answer as this may have an adverse effect on the user. Here attempt is made to obtain answers to the questions generated in 3.2. A popular approach to this is to establish semantic similarity between the question and its answer candidate. It has been observed that lexical similarity that performs fairly well in information retrieval generally suffers lexical chasm problem raised in table 2 in forums. A popular lexical similarity is the cosine similarity. Different authors that use this approach have to find a way of bridging the lexical chasm. The problem can be attributed to different ways of writing that calls for the use of synonymy (same word with different meanings, such as "book" as in the following examples: "The book is on the table" and "I will book my flight tomorrow"), polysemy (different words with the same or similar meanings, such as "agree" and "approve" as in "I agree with his going to London" and "I approve his going to London") and paraphrasing. A common method for tackling this problem is query expansion. At times it is used with lexical resources such as WordNet. One of the major constraints with query expansion is how to assign appropriate weights to expanded terms [19]. Details of how query expansion can be used to tackle lexical chasm can be found in [20].

3.3.1 Answer Mining Features

The various features used in mining answers from forums can be broadly classified into two, namely, Textual features and Non-textual features. These features utilize post textual content and forum meta data to extract good answers from forums. Textual features are usually used to measure degree of relevance between a document and a query while non-textual features can be employed to estimate the quality of the document [21]. Several authors have combined the two to mine good answers from forums as either of the two approaches has been found less effective. Figure 4 shows a schematic diagram of mining answers and pairing them up with questions.

a) Textual Features: These features are at times referred to as content based features. The text of a post should be a very good parameter for classifying a post as an answer post, remark post, question clarification post, answer clarification post or a mere junk. For example, one will expect a post that answers the question raised in the initial post to have a higher similarity scores with the topic of the thread and the initial post. On the other hand an off topic post will have relatively low scores. Specific examples of these features are shown in table 6.

b) Non-Textual Features: These features are at times referred to as structural features. Forum Meta

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

data such as authorship, answer length, normalized position of post, etc. are used in determining answers. Descriptions of these features are shown in table 7.

3.4 Mining FAQ from Question and Answer Pairs

Generating FAQ from the question and answer pairs will require sentence similarity and clustering. Clustering without specific labelled information can be considered as a concise model of the data which can be interpreted in a sense as either a summary or a generative model. It should be noted that there are a number of alternatives for getting the FAQ generated. It is possible to generate questions from forum and get these questions clustered to determine the FAQ. Thereafter, the answers to these FAQ questions can then be obtained. This approach will reduce the volume of data to be processed but stands the risk of generating questions that have no answers as FAQ as it has been shown empirically by [22] that about 14% of community-based questions do not get answered by the community. In this paper, we propose generation of question and answer pair prior to clustering and use only the questions for generating the clusters.

3.4.1 Sentence Similarity

Sentence similarity is needed during clustering to establish closeness between question objects. Sentence similarity between two questions can be established using either statistic similarity or semantic similarity or both. It is a common practice in traditional document similarity measures to represent document with high dimensional space in which each word in the document is considered as a dimension, which is the TFIDF. This practice does yield good result in large document retrieval applications due to the fact that the documents share a lot of words. Similarity calculation between questions cannot toll this line since question sentence contain just few words. Using high dimension representation with questions will make the vector space to be very sparse and will lead to a poor performance. An approach used by [23] is recommended. Similarity between two questions is performed using a low dimension vector space composed of the words of the two questions instead of vector space of all question words.

3.4.2 Clustering

Clustering and classification are two concepts that are often used interchangeably because both place data objects into groups of similar objects. There exists a clear cut distinction between the two. Classification arranges data objects into pre-defined classes, that is, prior to grouping of objects the classes must have been defined. On the other hand, clustering uses similarity between data objects to place them into groups. Here, no pre-defined classes, one only needs to define number of clusters into which the data objects will be placed. Clustering passes through an iterative process to refine members of the clusters until when members of the same cluster are close to each other as much as possible and object of other clusters are as distinct as possible. The various clustering techniques can be found in [24], a commonly used technique, the k-means will be discussed in this paper.

K-means: This is a partitioning method in which each cluster is represented by the centroid of the cluster. The main steps of this method are as follows:

a. First determine the number of clusters k you wish to have
b. Randomly select k objects as the initial centroids
c. For each of the remaining objects, calculate the distance between the object and the k centroids (that is, the difference between the two values) and assign it to the cluster that it is much close to its centroid.
d. Recalculate the centroid of each cluster by finding the mean of the members. Assign the nearest object to the calculated mean of the cluster as the new centroid.

Repeat steps c and d, until the centroids of all the clusters do not change again

In order to mine FAQ from the question and answer pairs, each question of the question set will be considered as an object. Question similarities are performed as discussed in section 3.4.1. and the k-means steps discussed above can be followed to determine the clusters. The main issue here is the choice of k as there is no standing rule for picking k. If k is too large, large number of clusters will be obtained after clustering and some of the questions which supposed to be in one cluster can split into two or more clusters and this will lead to a problem of having similar questions in the final

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

FAQ. Also, number of questions in some clusters may be too small to guarantee being considered for FAQ thereby reducing the number of standard FAQ questions. If k is too small, the number of clusters also will be small and this may lead to assigning questions that supposed to be in different clusters to the same cluster. Consequently, the number of FAQ questions could be reduced. In view of this, we will suggest the approach used by [4] to determine the value of k. In their approach, value of k is determined by measuring the degree of similarity of the questions in the set of questions. A schematic diagram showing the FAQ mining process is shown in figure 5.

4. CONCLUSION

This paper provides a general survey on mining questions, answers and FAQ from forums. The study is done in a way that new researchers in the area will quickly see the various techniques, their strengths and limitations. Some recommendations have also been made as regards arrears of exploration that will yield better performance. To the best of our knowledge it is the first paper to address the topic holistically. Different sources of mining FAQ have been outlined. Inherent challenges of forum corporal are discussed. In the future, we hope to confirm empirically the most optimal techniques for mining both questions and answers from forum threads. In particular, we wish to confirm the efficacy of the non-learning technique using regular expression.

REFERENCES:

[1] C.-H. Hsu, S. Guo, R.-C. Chen, S.-K. Dai. Using domain ontology to implement a frequently asked questions system, *in: Computer Science and Information Engineering, 2009* WRI World Congress on: IEEE, 2009, pp. 714-718.

[2] K. Iwai, K. Iida, M. Akiyoshi, N. Komoda. A help desk support system with filtering and reusing e-mails, *in: Industrial Informatics (INDIN), 2010 8th IEEE International Conference on: IEEE, 2010, pp. 321-325.*

[3] A. Moreo, M. Romero, J.L. Castro, J.M. Zurita. FAQtory: A framework to provide high-quality FAQ retrieval systems, *Expert Systems with Applications, 39 (2012) 11525-11534.*

[4] J. Mao, J. Zhu. FAQ Auto Constructing Based on Clustering, *in: Computer Science and Electronics Engineering (ICCSEE), 2012* International Conference on: IEEE, 2012, pp. 468-472.

[5] J. Seo, W. Bruce Croft, D.A. Smith. Online community search using conversational structures, Information retrieval, 14 (2011) 547-571.

[6] H. Kim, J. Seo. Cluster-Based FAQ Retrieval Using Latent Term Weights, *IEEE Intelligent Systems, 23 (2008) 58-65.*

[7] H. Kim, H. Lee, J. Seo. A reliable FAQ retrieval system using a query log classification technique based on latent semantic analysis, *Information processing & management, 43 (2007) 420-430.*

[8] G. Kothari, S. Negi, T.A. Faruquie, V.T. Chakaravarthy, L.V. Subramaniam. SMS based interface for FAQ retrieval, *in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 852-860.*

[9] A. Moreo, E.M. Eisman, J.L. Castro, J.M. Zurita. Learning regular expressions to template-based FAQ retrieval systems, *Knowledge-Based Systems, (2013).*

[10] A. Moreo, M. Navarro, J.L. Castro, J.M. Zurita. A high-performance FAQ retrieval method using minimal differentiator expressions, *Knowledge-Based Systems, 36 (2012) 9-20.*

[11] M. Romero, A. Moreo, J.L. Castro. A cloud of FAQ: A highly-precise FAQ retrieval system for the Web 2.0, *Knowledge-Based Systems, 49 (2013) 81-96.*

[12] W.-C. Hu, D.-F. Yu, H.C. Jiau. A FAQ Finding Process in Open Source Project Forums, *Fifth International Conference on Software Engineering (2010) 259-264.*

[13] S. Henβ, M. Monperrus, M. Mezini. Semi-automatically extracting FAQs to improve accessibility of software development knowledge, *in: Proceedings of the 2012 International Conference on Software Engineering: IEEE Press, 2012, pp. 793-803.*

[14] C.-J. Lin, C.-H. Cho. Question pre-processing in a QA system on internet discussion groups, in: Proceedings of the Workshop on Task-Focused Summarization and Question Answering: Association for Computational Linguistics, 2006, pp. 16-23.

[15] S. Cucerzan, E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users, in: Proceedings of EMNLP, 2004, pp. 293-300.

[16] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, Y. Sun. Finding question-answer pairs from online forums, in: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2008, pp. 467-474.*

[17] H. Kwong, N. Yorke-Smith. Detection of imperative and declarative question–answer pairs in email conversations, *AI Communications, 25 (2009) 271-283.*

[18] B. Wang, B. Liu, C. Sun, X. Wang, L. Sun. Extracting Chinese question-answer pairs from online forums, in: *Systems, Man and Cybernetics, 2009 SMC 2009 IEEE International Conference on: IEEE, 2009, pp. 1159-1164.*

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

[19] H. Fang. A re-examination of query expansion using lexical resources, *Proceedings of ACL-08: HLT, (2008) 139-147.*

[20] C. Carpineto, G. Romano. A survey of automatic query expansion in information retrieval, *ACM Computing Surveys (CSUR), 44 (2012) 1.*

[21] J. Jeon, W.B. Croft, J.H. Lee, S. Park. A framework to predict the quality of answers with non-textual features, in: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval: ACM, 2006, pp. 228-235.*

[22] L. Yang, S. Bao, Q. Lin, X. Wu, D. Han, Z. Su, Y. Yu. Analyzing and Predicting Not-Answered Questions in Community-based Question Answering Services, in: *AAAI, 2011.*

[23] W. Song, M. Feng, N. Gu, L. Wenyin. Question Similarity Calculation for FAQ Answering, (2007) 298-301.

[24] C.C. Aggarwal, C.K. Reddy. Data Clustering: Algorithms and Applications: *CRC Press, 2013.*

[25] L. Sun, B. Liu, B. Wang, D. Zhang, X. Wang. A study of features on Primary Question detection in Chinese online forums, in: *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on: IEEE, 2010, pp. 2422-2427.*

[26] K. Muthmann, A. Löser. Detecting near-duplicate relations in user generated forum content, in: *On the Move to Meaningful Internet Systems: OTM 2010 Workshops: Springer, 2010, pp. 698-707.*

[27] K. Pattabiraman, P. Sondhi, C. Zhai. Exploiting Forum Thread Structures to Improve Thread Clustering, in: *Proceedings of the 2013 Conference on the Theory of Information Retrieval: ACM, 2013, pp. 15.*

[28] Wang, Bao-Xun. Thread Segmentation Based Answer Detection in Chinese Online Forums, *in: Liu, Bing-Quan (Eds.), 2013.*

[29] P. Raghavan, R. Catherine, S. Ikbal, N. Kambhatla, D. Majumdar. Extracting problem and resolution information from online discussion forums, *Management of Data, 77 (2010).*

[30] L. Hong, 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: ACM, 2009, pp. 171-178.*

[31] B. Sumit, B. Prakhar, M. Prasenjit. Classifying User Messages For Managing Web Forum Data, *Fifteenth International Workshop on the Web and Databases (WebDB 2012), Scottsdale, AZ, USA, (2012).*

[32] B.-X. Wang, B.-Q. Liu, C.-J. Sun, X.-L. Wang, L. Sun. Thread Segmentation Based Answer Detection in Chinese Online Forums, *Acta Automatica Sinica, 39 (2013) 11-20.*

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

Table 1: FAQ Generation- Existing Practice

	Approach/Source	Limitation	Advantage	Usability
1	Information supplier tries to anticipate typical questions that users might have and answer them in advance.	▪ The candidate questions don't always satisfy users' needs ▪ A FAQ is sometimes scattered over several sentences that include small pieces of information ▪ The FAQ are mostly static in nature	It gives starting FAQ even before the commencement of the operation	Very common
2	Compose FAQ from web log, online search and some documentation	▪ It requires a lot of man power and time ▪ It is too subjective ▪ The Q/A pair may not be FAQ	It gives starting FAQ even before the commencement of the operation	Very common
3	Internet forums	▪ There must be active members for the forums ▪ Level of expertise must be high ▪ Too Noisy	▪ FAQ in its real sense can be generated ▪ Unanticipated Q/A pairs can emerge	[12]
4	Community-based Question Answering (CQA)	▪ Duplicates are highly controlled ▪ Extra efforts needed to generate FAQ	▪ High level of experts ▪ Contains standard Q/A pairs ▪ Large volume of Q/A pairs ▪ Mostly publicly available	[4]
5	Mailing List	May be difficult to obtain (confidentiality issue)	High level of experts	[13]

Journal of Theoretical and Applied Information Technology
10th May 2014. Vol. 63 No.1

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

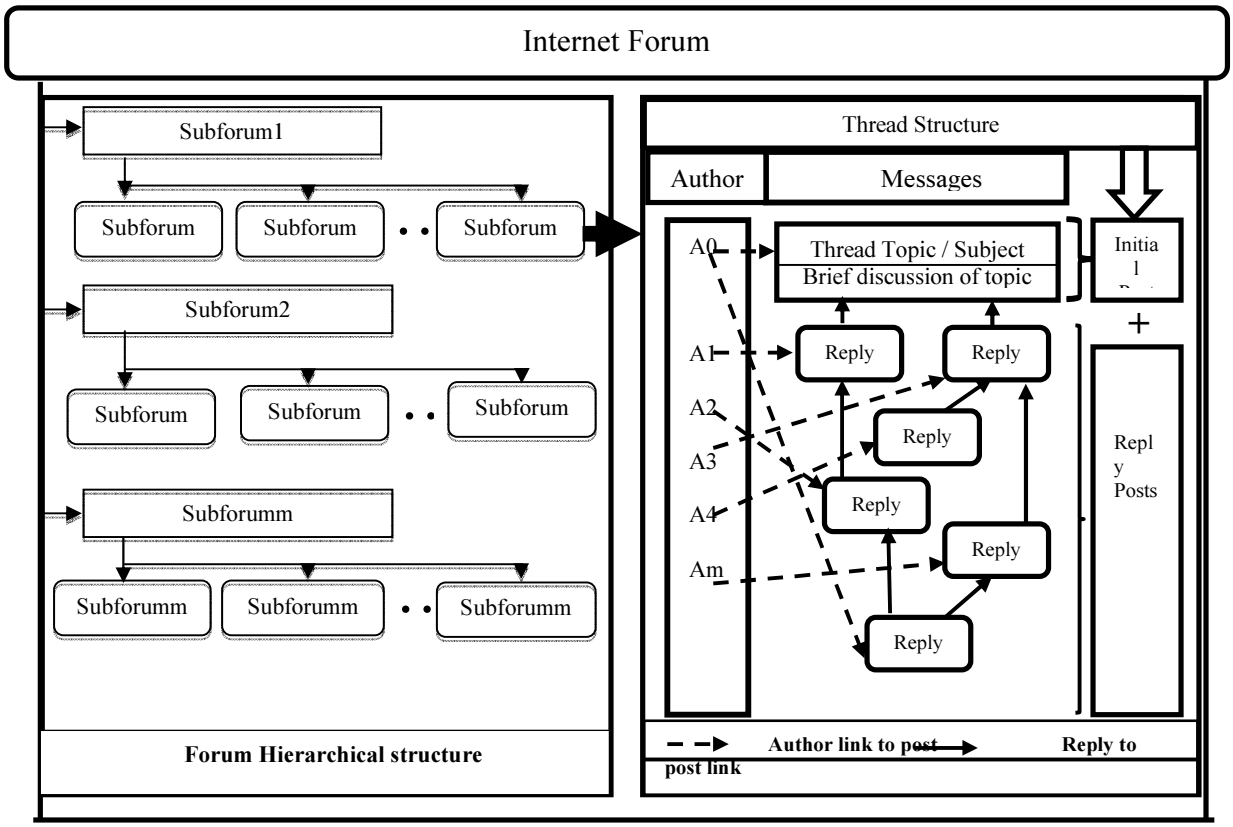

Figure 1: Hierarchical and Conversational Structures of Internet Forum

Table 2: Challenges of Human Generated Content of Forums

Problem	Forum Content characteristics	Effect on Data Mining	Source
Lexical chasm	Posts generated in forum usually include very short content that contains much fewer sentences than that of web pages.	▪ This makes similarity computing model to be less powerful. ▪ Also, the short contents cannot provide enough semantic or logical information for deep language processing	[16, 18]
Non- structure	Use of an informal tone that is close to oral statement	This creates a lot of noise in forum corpus and make the text to be less structured	[16, 18, 25-27]
Asynchronous	Multiple questions and answers may be discussed in parallel and are often interwoven together.	▪ This makes it difficult to establish reply relationship between posts ▪ One post may contain answers to multiple questions and one question may have multiple replies	[28]
Textual Mismatch between Q & A	Question words are not necessarily used in answers. Also, there are user-generated spam or flippant answers	These make it difficult to establish similarity between questions and answers.	
Multiple Authors	Discussion forum threads involve many contributors.	This makes them less coherent and more vulnerable to abrupt jumps in topics.	[29]

Table 3: Examples of Redundant Text

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

Forum Text	Redundant Text	Standard FAQ Statement
Hi Gurus, can you explain how I can protect a cell in excel?	Hi Gurus, can you explain	How can I protect a cell in excel?
About MS-Word, **I will like to know**, how to underline text.	I will like to know	About MS-Word, how to underline text
Where can I get pet friendly hotel, **Can anybody suggest some?**	Can anybody suggest some	Where can I get pet friendly hotel?

Table 4: Classes of Noise with Examples

Class of Noise	Example
Orthographic	Msg= Message, befour =before Positon=position
Phonetic	Rite=right, gooood= good Smokin= smoking
Contextual	In other to = in order to I can here you= I can hear you
Acronym	Asap = as soon as possible Lol = laughs out loudly

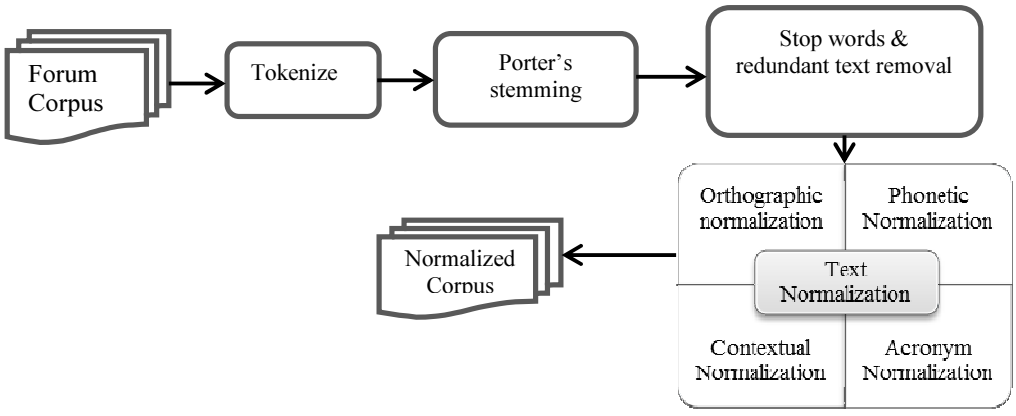

Figure 2: Pre-processing Steps

Table 5: Techniques for Mining Questions from Forums

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

Approach	Description	Strength	Limitation	Author
Question mark (?)	Considering sentences that end with question mark as question	Very simple to apply	Not all questions end with "?" mark in forum.	Popular method used by authors
Simple question words (5W1H)	Considering sentences containing What, Where, Why, When, Who and How as question	Very simple to apply	There are sentences that contain question words but are not questions e.g. Everyone knows how to locate his office	Popular method used by authors
Labelled Sequential pattern (LSP)	Classifying question sentences using extracted patterns	i. It has the potential to find more questions ii. It gives high precision	To learn a pattern for every question may be practically impossible if the thread is large	[16, 29]
Sequential Rule Based (SRB)	Developed 7 rules using combination of beginning word, question word, Mood word and Singleton word	i. It is able to detect 97% of questions in the dataset. ii. It gives high precision	The rules were manually coded, this may be tedious for large dataset	[18]
LSP + POS	Tagging labelled sequential pattern with part-of-speech	It has the potential to detect large number of questions	POS is a time consuming process.	[16]
N-gram + Non-textual features	N-gram combined with total number of posts and the authorship	Has ability to reveal hidden questions through language patterns	May lead to generation of high number of redundant language patterns especially if higher N-grams are considered.	[25, 30]

Table 6: Textual Features for Mining Answers from Forums

Label	Textual Feature	Description	Data Type	Author
TF1	Number of answer words shared with question	Shared words are considered to be of great value in determining answers. Notional words are at times used.	Integer	[18]
TF2	Similarity with question based on HowNet	This is a semantic similarity measure. It is often used to bridge lexical chasm. It computes semantic similarity between two corresponding words using HowNet	Float	[18]
TF3	Number of words shared with thread topic	It has been noted that thread topic often share the same words with the initial post which is often the question post.	Integer	[18]
TF4	Similarity with thread topic based on HowNet	Used to establish semantic similarity between the candidate answer and the thread topic.	Float	[18]
TF5	Cosine similarity between the post and thread topic	Measures lexical similarity between the post and thread topic	Float	[18, 31]
TF6	Cosine similarity between the post and the question	Measures lexical similarity between the post and the question	Float	[16, 18, 31, 32]
TF7	KL-divergence of the post and the question	Measures the relevance of the question and is candidate answer. It introduces the concept of entropy into relevance quantification.	Float	[16, 32]
TF8	Query likelihood Model Score	It is a basic language model that calculates the likelihood that a reply post is relevant to the original question post	Float	[16, 30]
TF9	Quote presence in the post	Confirms presence of quote in the post.	Binary	[31]

Table 7: Non-Textual Features for Mining Answers from Forums

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

Label	Non-Textual Feature	Description	Data Type	Author
NTF1	Answer length	The length of the answer. It measures the comprehensiveness and completeness of the answer. It could be considered as the number of characters or words forming the answer.	Integer	[18, 32]
NTF2	Distance between question and answer	It is expected that the post that contains the answer to a question is not far from the question post	Integer	[18, 32]
NTF3	Normalized position of the post in the thread.	It is calculated as the absolute position of the post divided by the number of posts in the thread	Float	
NTF4	Total number of words in the post after stopwords removal	It is believed that an authority will compose answers with lesser number of stopwords.	Integer	[31]
NTF5	Unique number of words in the post after stopwords removal	Measures the domain vocabulary skill of the answerer. The higher the value, the more skilled the answerer.	Integer	[31]
NTF6	Total number of words in the post after stopwords removal and stemming	It measures inflexion contribution. It has ability to enhance lexical similarity	Integer	[31]
NTF7	Unique number of words in the post after stopwords removal and stemming	It combines the effects of NTF5 and NTF6	Integer	[31]
NTF8	Presence of acknowledgement	Politeness is a requirement of forums. Questioners always express acknowledgement for getting help from answerers by using words like *Thank you, I appreciate.*	Binary	[18, 32]
NTF9	Answerer's activity	Forum users can be classified as answerers and questioners based on the number of posts answered and number of question posts raised.	Integer	[18, 32]
NTF10	Quote counts	It has been observed that an answer contained in a post is often being quoted by other posts because of its usefulness.	Integer	[18, 32]

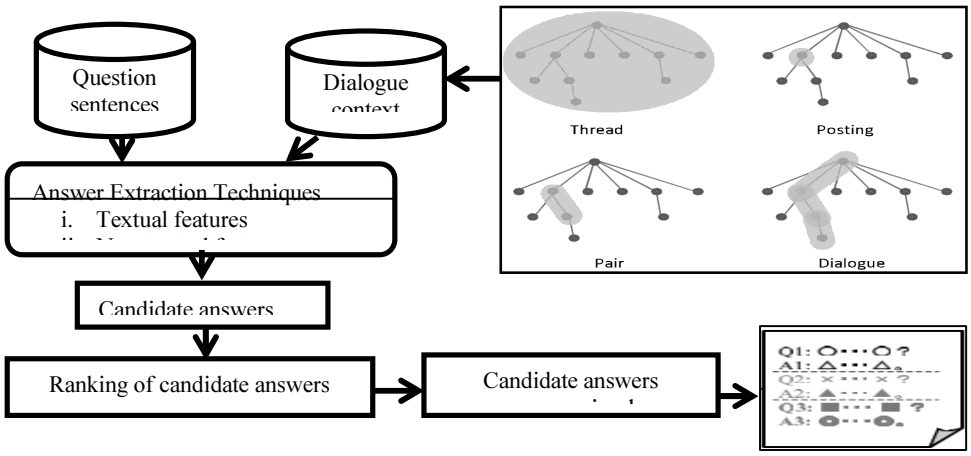

Figure 4: Mining of Answers and Pairing them up with Questions

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

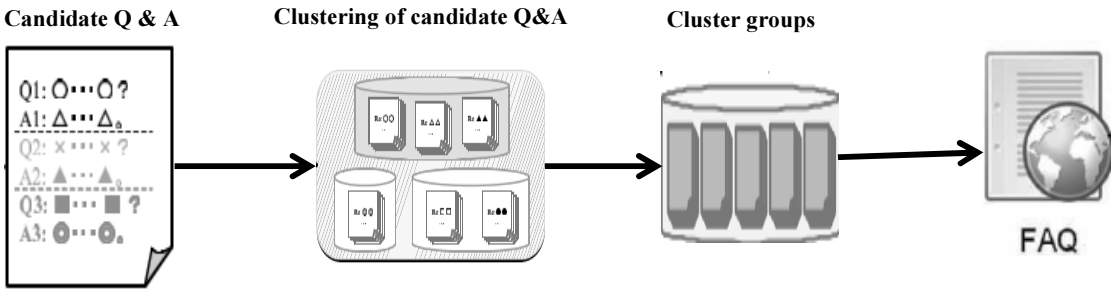

Figure 5: FAQ Mining Process from Q & A Pairs