*International Journal of Advanced Information in Engineering Technology (IJAIET) ISSN:* 2454-6933
*Vol.3, No.8, August 2016*

# Extraction and Summarization of Relevant Forum Posts

Shahad P.
Dept. of Computer Science and Engineering
MEA Engineering College
Perinthalmanna, Kerala, India

Saani H.
Dept. of Computer Science and Engineering
MEA Engineering College
Perinthalmanna, Kerala, India

***Abstract*—**A forum is a web application for holding discussions and posting user generated content. Messages within these forums or sub-forums are displayed either in chronological order or as threaded discussions. There are forums based on topics which acts like a big virtual information center. Information on wide range of topics is discussed through the forum. Here, we extract recent and relevant posts coming under a specific thread and summarizing it. Pattern matching and other data mining techniques are applied for extracting these forum posts. Existing Google indexing is utilized for extracting posts in threads. Posts are summarized by Latent Semantic Analysis and are further enhanced by Latent Dirichlet Allocation method.

***Keywords—Forums, Crawler, Latent Semantic Analysis***

## 1.INTRODUCTION

Information Extraction refers to the automatic extraction of structured informationsuch as entities, relationships between entities, and attributes describing entities fromunstructured sources. This enables much richer forms of queries on the abundant un-structured sources than what is possible with keyword searches alone. The extractionof structure[1] from noisy, unstructured sources is a challenging task that has engaged averitable community of researchers for over two decades now. With roots in the NaturalLanguage Processing (NLP) community, the topic of structure extraction now engages many different communities spanning machine learning, information retrieval, database,web, and document analysis. Early extraction tasks were concentrated around the identification of named entities, like people and company names and relationship amongthem from natural language text. Applications such as comparison shopping, and otherautomatic portal creation applications, lead to a frenzy of research and commercial activity on the topic. As society became more data oriented with easy online access toboth structured and unstructured data, new applications of structure extraction camearound.

Online communities are valuable information sources where knowledge is accumulatedby interactions between people. Forums are one of such communities which is a goodsource of information. Online forum pages are not the same as general web pages. Mygoal is to design effective retrieval models that incorporate online forum propertiesso that the effectiveness of forum search and hence posts extraction can be improved.The forum pages have unique textual or structural features that distinguish them fromgeneral web pages. Generally, a forum has several sub-forums covering high-level topiccategories. Each sub-forum has many threads. A thread is a more focused topic-centricdiscussion unit and is composed of posts created by community members. This featureencourages in depth discussion, compared to general web pages.

Web forum has become an important resource on the Web due to its rich informationcontributed by millions of Internet users every day. In the forum sites people communicate and discuss with each other through a thread. A typical bulletin board is verypopular almost all over the world for opening discussion. Every day, there are innumerable new posts created by millions of Internet users to talk about any conceivable topicsand issues. Thus, forum data is actually a tremendous collection of human knowledge,and therefore is highly valuable. It is also noticed that some recent research efforts havetried to mine forum data to find out useful information such as business intelligence andexpertise[2]. Whatever the application, the fundamental step is to fetch data pages fromvarious forum sites distributed on the whole Internet.

In forum sites people can hold conversations in the form of posted messages. Thesemessages are often longer than single line and are at least temporarily achieved. Also,depending on the access level of a user or the forum set-up, a posted message might needto be approved by a moderator before it becomes visible. Depending on the forum's settings, users can be anonymous or have to register with the forum and then subsequentlylog in to post messages. On most forums, users do not have to log in to read existingmessages. Millions of posts are posting per day in the forum sites.

A post is a user-submitted message enclosed into a block containing the user's detailsand the date and time it was submitted. Posts have an internal limit usually measuredin characters. Often one is required to have a message of minimum length of 10 characters. There is always an upper limit but it is rarely reached most boards have 10,000,20,000, 30,000, or 50,000 characters. Most forums keep track of a user's post count.The post count is a measurement of how many posts a certain user has made. Userswith higher post counts are often considered more reputable than users with lower postcounts, but not always. In this project I extract such posts related to the query andmakes summary.

Summarization is the way of expressing ideas in fewer words. The introduced summarization technique summarizes forum posts. The system extracts posts and provides asummary about the discussion. Two different techniques are used for

summarization.Latent semantic analysis (LSA)[3] is a technique in natural language processing, specifically in victoria semantics, of analyzing relationships between a set of documents andthe terms they contain by producing a set of concepts related to the documents andterms. LSA assumes that words that are close in meaning will occur in similar piecesof text. A matrix containing word counts per paragraph (rows represent unique wordsand columns represent each paragraph) is constructed from a large piece of text anda mathematical technique called singular value decomposition (SVD) is used to reducethe number of columns while preserving the similarity structure among rows. Words arethen compared by taking the cosine of the angle between the two vectors formed by anytwo rows. Values close to 1 represent very similar words while values close to 0 representvery dissimilar words.
Latent Dirichlet Allocation (LDA)[4] is used to discover the topics that contained inthe given set of sentences. LDA represents documents as mixture of topics that split outwords with certain probabilities. Here, the individual posts are treated as documentsand are having some fixed number of sentences and words. The summary produced bythis generic model is more reduced than LSA.

## 2. RELATED WORK

iRobot[5] is one of the most important attempt to identify the Forum sites from theirURLs. IRobot's first step is learning the sitemap of forum site. This is achieved bycollecting pre-sampled pages from the forum sites and then decides how to select an optimal traversal path to avoid duplicates and invalids. Forums sitemap is Re-constructinghere. This paper builds a prototype of an intelligent forum crawler, iRobot, which hasintelligence to understand the content and the structure of a forum site, and then decidehow to choose traversal paths among different kinds of pages. After that, we selectan optimal crawling path which only traverses informative pages and skips invalid andduplicate ones. One of the important advantages of this paper is the division of longthreads into multiple pages and this can be reconstructed and concatenated, which is ofgreat help for further indexing and data mining.
It is also noticed that some recent research efforts have tried to mine forum data[6] to find out useful information such as business intelligence and expertise. Forum pages are generated by pre-defined templates, and different templates are usually adopt topresent different content such as list-of-board, listof-thread, post-of-thread, user profile,etc. Moreover, valuable information in forum pages is mostly shown in some repetitive manners, as it is essentially data records stored in a database. Thus, a forum page canbe well characterized by what kinds of repetitive regions it contains, and where theseregions are located. Location information should be integrated with the URL patterns to decide which links should be followed and how to follow them.
Jingtian Jiang designed and developed a supervised web-scale forum crawler, FoCUS[7].The FoCUS aims to crawl relevant forum content with minimal overhead. This paperexplains about the common pattern existing among forums and their threads even withdifferent supporting software packages. FoCUS reduced the forum crawling problem intoURL type recognition problem. Here scholars suggest some regular expressions as patterns for identifying URLs of forum threads. A single pattern can be used for crawlinga large set of similar forum sites. Supervised method is applied here so the training setcreation is tedious task. Special interest had been given by FoCUS to identify forumsentry page by proposing an algorithm.
Similar work to crawl forum sites includes the work by Vidal et al. they proposed amethod for learning regular expression patterns of URLs. This method compares DOMtree[8] structure and is effective for the sampled forum alone. Similarly FOCUS architecture has two major parts: Learning part to learn ITF regexes of a given forum fromautomatically constructed URL training examples. Second part is online crawling part to crawl all threads. The classifier used in FOCUS is very effective. The method introducedby Koppula et al[9]. is used here to identify the different URL patterns and are refinedrecursively. This refined pattern is retained only if its matching URLs are greater thanan empirically determined threshold. Online crawling is done by depth-first-strategy.

### *A. Feature Identification Methods*

Summarization technique is employed to reduce the size of information. The systemwill summarize the posts and provide the user an overview about the posts. A LatentSemantic Analysis (LSA) based feature-identification approach works best to identifyfeatures. Features and opinion word identification are essential in feature-based summarization. The effectiveness of Latent Dirichlet Allocation (LDA) for feature identificationis examined in this paper.
Latent Semantic Analysis (LSA) projects an original vector space or term-documentmatrix into a small factor space. The dimensional reduction of a matrix is accomplishedusing singular value decomposition which decomposes an original matrix into three matrixes, a document eigenvector matrix, an eigen value matrix, and a term eigen vectormatrix. In turn, an original matrix can be approximated by multiplying these threematrixes with only high eigenvalues. Because of orthogonal characteristic of factors,words in a factor have little relations with words in other factors, but words in a factorhave high relations with words in that factor.
Latent Dirichlet Allocation (LDA) incorporates the generative process of documentswith Dirichlet distribution. According to LDA process, each document is generated inthe following three steps. First, the number of words used in a document is determinedby sampling with the Poisson distribution. Second, a distribution over topics for a document is elicited from the Dirichlet distribution. Third, based on the document-specificdistribution, topics are generated, and then words for each topic are generated. LDAalso provides topics in which words have probability values. Assuming this generativemodel for a collection of documents, LDA then tries to backtrack from the documentsto find a set of topics that are likely to have generated the collection.

## 3. PROPOSED SYSTEM

A detailed description on the implementation aspects of the proposed work is discussedin this chapter.

### *A. System Architecture*

The collection of valuable user generated posts and their summary generation process is done by dividing the task into three different modules. These modules areimplemented[10] in the order as they are given in the following section. Details ofimplementation process are given in next section of this chapter.

*1) Document collection Module*

Before implementing A Forum crawler some forum sites are deeply studied. Based onthese studies, document collection module was started [11].This step includes finding the best forums and threads for the posts that is required. Forums selection is done after fixing the domain of the discussions carrying out.Most of the data mining projects utilize the data warehouse to collect the required data.Here we need to extractfrom online forums and we need online analytical processing tocollect required data. For this purpose pattern matching technique is used. Pattern iswritten in regular expression and is made possible by continuous analysis of the forumsites.

*2) Pre-Processing Module*

The extracted posts[12] are selectively discarded before making their contents as a single file for making summary. Discarding posts by the forum etiquettes and by considering sentence. All these elimination is done by constructing regular expression corresponding to each rule[13] in the forum etiquette.

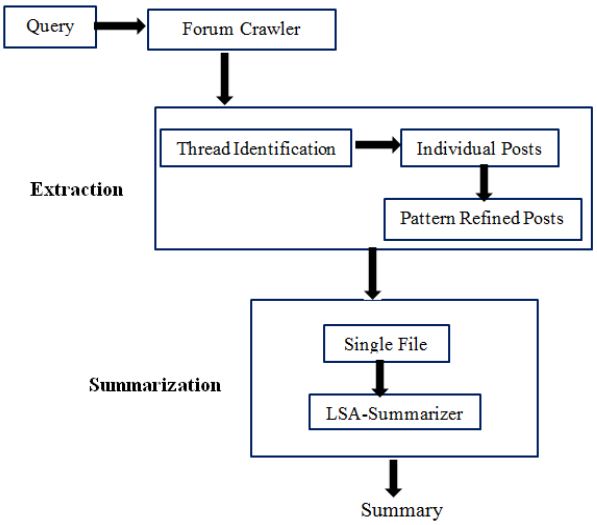


Figure 5.1 System Architecture

*3) Information Extraction Module*

The posts are extracted and saved in temporary file[14] because the queries are different each time. This temporary file is used by the summarization Module

*4) Summarization Module*

Fuzzy logic is a form of knowledge representation suitable for notions that cannot bedefined precisely, but which depend upon their contexts. It deals with reasoning that is approximate rather than fixed and exact. Fuzzy logic has been extended to handle theconcept of partial truth, where the truth value may range between completely true andcompletely false. It can also use with linguistics. Linguistics is the scientific study oflanguage. There are broadly three aspects to the study, which include language form,language meaning, and language in context. Fuzzy logic provides an alternative wayto represent linguistic and subjective attributes of the real world in computing. Herefor the summarization purpose I considered the two existing methods Latent SemanticAnalysis and Latent Dirchilet Allocation.

## 4. IMPLIMENTATION DETAILS

All experiments were performed on a Windows 8 machine with two 2.3 GHz core i3processors and 2GB memory. Forums are well structured and are organized as Indexpage, Thread Page and post pages respectively.

### *A. Extraction of Relevant Posts*

Inorder to extract the forum posts it is necessary to know the depth of forums and theamount of contents, the user posts, in them. It is difficult and not necessary to crawl allthe posts. Few of the forums are analyzed before implementing a post extractor. Thishas been generalized to fix the number of relevant posts to be extracted from forums.The basic forum structure includes three different types of pages they are the following.Board page is the index page of the Forum site it consist of some sub forums which isspecific to the domain of the Forums. This is looking like:

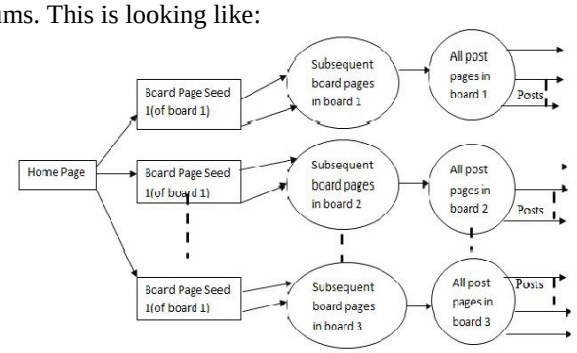


Figure 4.1: Basic Forum Structure


Figure 4.2: Board Forum Page

In thread page there are different thread titles of each sub-Forums. Each ofthem have many number of post pages.


Figure 4.3: Thread Page

Post page holds the posts or opinions of members temporarily. Each page is having acapacity of keeping around 30 posts. This can be viewed as;


Figure 4.4: Post Page

The proposed system accepts questions or issues from the user as the search query.Thesequeries are related to different domains. The same domain has different topics and sohas to face many queries. Forums are generally domain specific and have subforums tobe more specific in that domain. Structure of forums is different but in specific format,there will have a lot of forums. This makes the task of creating template for specificpackage and structure. This template is used for identifying the thread and throughthat relevant posts are extracted. Output of the extraction phase is the posts relatedto the search query. The query is compared with thread headings in the forum sitesand most appropriate and relevant posts are extracted with the help of a good regularexpression. This system utilizes the existing index of the Google[15] and string matchingalgorithm used in this. The requirement for proposed system is now available. To reachthe defined goal it is necessary to follow the below phases.

### *B. Feature Based Summarization*

Feature identification algorithms are employed to find out related feature terms of specified seed features, and these related terms could be regarded as being semanticallyrelated to the specified features. These related terms can be employed to select summary sentences. The two feature identification algorithms examined in this work aredescribed below.

#### *1) Algorithm 1: LSA for feature Identification*

*Input:*
A nxm term-document matrix M, feature seed set S, n number of extracted features foreach seed f in S

*Output:*
An association array F, where each key represents a feature seed f and its correspondingvalue is f0s related features

*Steps:*
1. Initialize associated array F.
2. Call SingularValueDecomposition(M).
3. Assign the product of term eigen vector and eigen value to M.
4. For each f in S :
5. Assign getTermVectorFromTermDocMatrix(f,M) to wf .
6. Initialise similarity list sim.
7. For each column vector w of M :
8. Assign wf :w to sim[i].
9. Sort(sim).
10. Assign getTopRelatedFeatures(sim,n) to F.
11. Return F.

In algorithm 1, line 2 calls the Singular Value Decomposition function. It decomposesthe M matrix into its corresponding term eigen vector, eigen values and document eigen vector. Line 3 assigns the product of term eigen vector and eigen value to N matrix.In lines 4-8, for each seed feature get the term vector wf from term document matrixand compute its dot product with each element in the column vector of N. Then sortthem in increasing order inorder to get the top related features.

#### *2) Algorithm 2: LDA for feature identification*

*Input:*
A set of documents, fixed number of K topics, n number of extracted features for eachseed topic f of K

*Output:*
An association array F, where each key represents a topic f of K and its correspondingvalue is f0s related features

*Steps:*
1. Initialize associated array F.
2. Initialize each word in the document to one of the K topics.
3. For each document d, repeat until no change:
4. For each word w in d:
5. For each topic t :
6. Compute p(topic t j document d).
7. Compute p(word w j topic t).
8. Reassign w a new topic, where we choose topic t with probability p(topic t jdocument d) * p(word w j topic t).
9. Update the n words assigned to F under each aspect.
10. Return F.

In algorithm 2, line 2 randomly assigns each word in the document to one of theK topics. Lines 3-7 computes the product of the proportion of words assigned to atopic in each document and the proportion of assignments to a topic throughout allthe documents coming from this word w. Line 8 assigns w a new topic based on therepetition of the previous steps, until no change in the above computed value. Line 9updates assignment of words under each topic.

## 5. RESULTS AND OBSERVATIONS

Results of the two important phases of this thesis work can be given as;

### *A. Posts Extraction*

Many different measures for evaluating the performance of information retrieval[16]systems have been proposed. The measures require a collection of documents and aquery. All common measures described here assume a ground truth notion of relevancy:every document is known to be either relevant or non-relevant to a particular query. Inpractice queries may be ill-posed and there may be different shades of relevancy. Thischapter discusses about the results obtained from the proposed system. The proposedsystem is an attempt to implement information extraction from discussion forums.
Data set contains n number of posts downloaded from discussion thread.
To evaluate this system, Precision and Recall are used as evaluation metrics. Precision-is the fraction of the documents retrieved that are relevant to the user's informationneed. Recall-is the fraction of the documents that are relevant to the query that aresuccessfully retrieved. We provide the Precision of downloaded posts and the Recall ofposts to evaluate the effectiveness of this crawler. We suppose that, for crawling a

| Precision Calculated | | | |
|---|---|---|---|
| Query | Extracted Post | Effective post | Precision |
| Body fat | 100 | 64 | 64 |
| Protein Powder | 100 | 62 | 62 |
| Vitamins | 100 | 78 | 78 |
| Minerals | 100 | 62 | 62 |

Table 5.1: Precision table
Precision=Post-effective/Posts-downloaded*100

Recall=Post-effective / Post-all * 100

Webforum site,Posts-downloaded is the count of all posts that could be download,and Posts-effective is the count of all post that are chosen in summary generation, and Post-all is the
count of all posts in the related topic. Then the Precision and Recall are computed asfollows:

### *B. Summarization*

Two different feature identification algorithms, namely LSA and LDA are compared[4].The results of each of which are then used in generating the summary. Precision, recalland f-value are the matrices used for reporting the performance of feature identificationalgorithms. Precision is the ratio of the number of correctly extracted features to thetotal number of features extracted. Recall is the number of correctly extracted featuresto the total number of standard features. F-value gives the ratio of twice the productof precision and recall to their sum.
Precision values are higher for LSA and as the number of terms increases a significantimprovement can be seen for LDA based feature identification. As far as recall is concerned, again LSA shows better results when compared with LDA but its performancebegins to degrade when the number of terms increases and LDA seems to show goodrecall values at this point. Therefore LSA based feature identification gives betterresults than LDA based feature identification.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we design and develop an intelligent Forum crawler. This Forum crawlerextracts posts from relevant discussion threads of the forum. Pattern matching anddata mining techniques are applied for extracting the forum posts. Before exacting the posts into separate files for summarization, specific rules are applied to removecertain words, sentences or posts. This makes posts more effective for applying machinelearning algorithms and natural languageprocessing technique. As the forums arerich in posts or relevant contents this technique should be enhanced by extractinggrammatically correct sentences or posts. Here, these extracted posts are summarized by Latent Semantic Analysis and enhanced by Latent Dirichlet Allocation method.
It would be interesting to extract high quality posts for decision making. Extractgrammatically correct posts and give score to users for understanding their value indecision making. Forum users were unaware of the forum etiquettes and importance oftheir posts. This area also needs improvement. Moderators of most of the forum sites are inefficient so enhancement in this side is also necessary. Pattern matching by usingregular expression is used in this system to extract the posts. An obvious future workwould be using natural language processing for extracting large training data. Thissystem is domain and structure specific. Another future work is Scaling the system toa generic domain. The summary generation for the posts will be more accurate if we
make the above enhancement. The behavior of forum members can be predicted byenhancing this System.

## REFERENCES

[1] A. Arasu and H. Garcia-Molina, “Extracting structured data fromweb pages," in proceedings of the 2003 ACM SIGMOD international conference on management

[2] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, “Deriving marketing intelligence from online discussion," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005, pp. 419-428.

[3] S. T. Dumais, \Latent semantic analysis," Annual review of information science and technology, vol. 38, no. 1, pp. 188{230, 2004.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan,”Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[5] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, “iRobot: An intelligent crawler for web forums," in Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp. 447-456.

[6] Y. Guo, K. Li, K. Zhang, and G. Zhang, “Board forum crawling: a web crawling method for web forum," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, 2006, pp. 745-748.

[7] J. Jiang, X. Song, N. Yu, and C.-Y. Lin, “Focus: learning to crawl web forums," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 6, pp. 1293-1306, 2013.

[8] Y. Zhai and B. Liu, “Structured data extraction from the web based on partial tree alignment," Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 12, pp. 1614{1628, 2006.

[9] H. S. Koppula, K. P. Leela, A. Agarwal, K. P. Chitrapura, S. Garg, and A. Sas-turkar, “Learning url patterns for webpage de-duplication," in Proceedings of the hird ACM international conference on Web search and data mining. ACM, 2010, pp. 381-390.

[10] G. Gupta, Web Data Mining,in Introduction to Data Mining with Case Studies. New Delhi, India: Prentice-Hall, 2006.

[11] Forum Matrix," [Online] Available :http://www.forummatrix.org/index.php, accessed Feb 2014.

[12] Cordeiro and P. Brazdil, \Learning text extraction rules, without ignoring stop words." in PRIS. Citeseer, 2004, pp. 128-138.

[13] C.-H. Chang, C.-N. Hsu, and S.-C. Lui, "Automatic information extraction from semi-structured web pages by pattern discovery," Decision Support Systems, vol. 35, no. 1, pp. 129-147, 2003.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in IJCAI, vol. 7, 2007, pp. 2670-2676.

[15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Computer networks and ISDN systems, vol. 30, no. 1, pp. 107-117, 1998.

[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, pp. 168-177.