International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 4 Issue: 1 09-Apr-2015,ISSN_NO: 2321-3337

Empirical Forum Mining

BhosalePranita.B,BuchkulPriyanka.S,BankarPriyanka.B,Borawake Priyanka. M

Prof. Kumbhar H.R.

SVPM's COE Malegaon

ABSTRACT— *Empirical forum mining is the online discussion board where user can request and exchange information. The forum contains the lot of data and information can be handle by using threads in the forum crawler. Crawler is nothing but the linking between the pages which we traverse during our searching of any content. There are many existing system which provide the facility to get content but there is a problem in getting the appropriate data. But by using our proposed system we have to obtain the most appropriate answer to the posted question out of thousand responses. We have to avoid wasting of resources from some responses which may not yield desired result. Here in this system we have three pages as index page, entry page and thread page. In entry page we can ask or see the question, in index page there is information on URL pointing to the board. Thread page contains the post to the question. We are using index/thread URL detection, page flipping URL detection and entry URL discovery algorithms to do this. After this FAQ generation technique is applied. In this we are mining the most frequently asked questions and finding the most suitable answer to the question by reducing uninformative data. The conversion of similar URL into the regular expression is done. This is known as index thread flipping (ITF). The clustering of data is done with the help of k-means algorithm.*

KEYWORDS: threads, crawler, entry page, index page, thread page, summarization, FAQ

1. INTRODUCTION

In recent years web application allow users to interact or collaborate with each other in a social media dialogue. Website is familiar to everyone and from many days we are using it to access data. But sometimes we didn't get the desired data because of duplicate link's means there are some URL that point to the next page but that page again contain the same data. It also has many uninformative pages such as login control, advertisements [9]. There is wastage of time due to navigation for long time. This is not efficient due to loss of time.This navigation of page for long time is known as crawling. Following these links, a crawler will crawl many uninformative pages thus making it inconsistent for web crawling. To crawl data more efficiently and effectively, we

propose an approach to exploring an appropriate traversal strategy to direct crawling of a given target data. This strategy is known as forum crawler. Forum mining means to mine all data that is useful or not useful information but empirical forum mining can mine exact information related to topic.

Empirical is a way of gaining knowledge by mean of direct and indirect observation and experience. Forum is used for requesting and exchanging information with each other [1]. Forum mining means combination of forum crawling and mining forum. Crawler is nothing but the linking between the pages which we traverse during our searching of any content.

Mining helps us to extract new information and streaming data and also the process of discovering hidden patterns in large sets. Forum sites allow user to view forum posting and to post message in it. In this system we are implementing the URL classifier and web page classifier to identify entry page, index page and thread page as well as page flipping URL.

In entry page we can ask or see the question, in index page there is information on URL pointing to the board and thread page contains the post to the question. By using Regular Expression we can find the correct pattern that should cover many URLs [4] and finally we will generate FAQs and summarize it[10]. FAQ stands for frequently asked question. It is type of web page that lists questions frequently asked by user. The answers are typically shown with the questions. The first section of this paper tells us about the existing and proposed system. Next section gives architectural diagram and forum crawler details. Last section provides features of our system.

2. SYSTEM ANALYSIS

2.1 Existing System

Some techniques are present like “Forum Mining” but they have some limitation. Few of them are as follows:

2.1.1 Dom Tree

It is very efficient but it only works for the specific sites from which the samplepage is drawn. The same process has to be repeated every time therefore it is notsuitable for large-scale crawling[1].

2.1.2 IRobot

This approach automatically understand the content and structure of each forumsites and then decides how to traverse to different pages in forum sites to find outSuch traversal path, it first automatically rebuild the site map of the target web forumand then it select the optimal traversal path which only traverses informative pagesand skip invalid and duplicate pages[3]. iRobot system consist of two measure part:-

(a) Off-line site map reconstructing and traversal path selection.

(b) On-line crawling.

2.1.3 Web forum

Automated traversal of web to collect all the useful informative pages, effectively and efficiently gather information about link structure interconnecting the informative Pages [2]. Web application designed to manage user created content. Pre-samples few Pages to discover the repetitive regions.

2.1.4 Near Duplicate Detection

Another related work is near duplicate detection. Forum crawling also needs to remove duplicates. But content-based duplicate detection is not bandwidth efficient, because it only be carried out when pages have been downloaded [10]. URL based duplicate detection is not helpful. It tries to mine rules of different URLs with similar text, but such methods still need to analyze logs from target sites or results of a previous crawler.

2.2 Proposed System

In the proposed system a new approach that is empirical forum mining in which we are mining the data with minimized question and answer having minimal overhead. In this system we are giving facility for user to see the post and to give is opinion about the post or question. Then the most appropriate answer is send to user so that he will get desired answer in a minimum time. The main feature of this model is simplicity.

2.2.1 Page classification:

In page classification forum page is classified into three types as follows: Entry page: The home page of the forum, which contains lowest common ancestor of all threads, Index page: Major link on the home page and contains information on URL's pointing to a board or a thread, Thread page: A page of a thread in a forum that contains a list of post with user generated contents belonging to the same discussion.

2.2.2 URL classification:

URL type is classified for each page type as follows:

Index URL: Index URL is a URL on entry page or index page and points to the index page. Its anchor text shows the title of its destination board.

Thread URL: It is a URL on index page and points to a thread page. Its anchor text is the title of its destination thread.

Page flipping URL: that leads users to another page of the same board or the same thread.

2.2.3 Learning ITF regex:

In regular expression we get positive and negative URL. After this by using the threshold formula the minimization of the positive and negative URL takes place.

2.2.4 Clustering:

In clustering, we place data object into groups of similar objects called as cluster that share common characteristics. We can cluster question and answer by using Fuzzy C-means algorithm for generating FAQ that is asked questions frequently

3. SYSTEM ARCHITECTURE

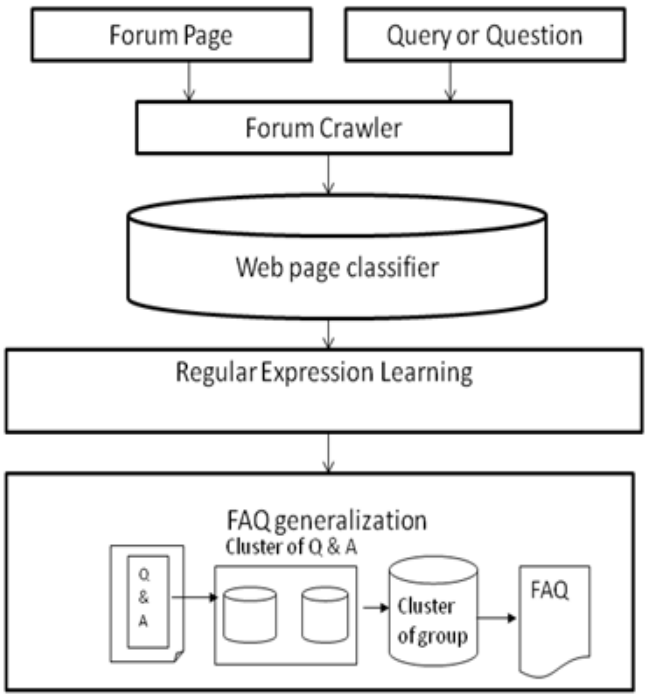

Fig: 3.1System Architecture

The figure shows the architecture of Empirical forum mining. In this first home page of forum is provided to the user so that user can fire query. When query is fired or user enter the URL forum crawler can crawl the relevant information with minimal overhead. Web page classifier can classifies the different pages in three group. Pages such as Entry page, Index page, Thread page. Entry page is a page that is the lowest common ancestor of all thread pages in a forum. Index page is a page that contains information on URLs pointing to a board or a thread. List of board & thread page, board page are all index pages. Thread page is a page that contain a list of post with user generated content. Other page is a page is not an entry, index, thread page. After that URLs are classified with the help of URL classifier. There are four types of URL such as index URL, thread URL, page flipping URL, other URL. Index URL is a URL that is on an entry page or index page and points to an index page. Its anchor text shows the title of its destination board. Thread URL is

a URL that is on index page and points to a thread page. Its anchor text shows the title of its destination thread. Page flipping URL is a URL that leads to user to another page of a same board or same thread. Other URL is a URL is not index, thread or page flipping URL. Regular Expressions are used next to minimize the URLs. With the help of threshold formula we will reduce the URL pattern. We get different questions from URLs. With the basis of question we get answers that is threads. Here one answer is one document.so we get number of documents. Cosine formula and K-mean algorithm are applied on the documents to cluster all the documents. Clustering uses similarity between data objects to place them into groups. By applying preprocessing to get summarized topic that is frequently asked questions.

ALGORITHM:-

1. Index/Thread URL Detection:-

- Extract the all URL's from the page that user requested.
- Find maximum anchor text length of that URL's.
- If anchor text is long and short plain text then that page is index page and type of that URL is index URL.
- Otherwise anchor text is small and post has a very long text block then that page is Thread page and type of that URL is Thread URL.

2. Page Flipping URL Detection:-

- Extract all outgoing URL's from Index page or Thread page.
- Find out anchor text length of that URL's.
- If anchor text is digits or special text such as "Last" or the URL's are appear at same location or pages have the similar layout then this URL's are the Page Flipping URL's.

3. Entry URL Discovery:-

- Almost every page contains a link to lead users back to the entry page of a forum.
- Extract the all URL's from the page that user requested that are candidate URL's.
- After that find out the common URL's of that candidate URL'sand the outgoing URL's of that page.
- An entry page has most index URL's.
- First assign the count be 0.
- Then detect the index URL's of that pages.
- If the count is less than that index URL's then that page is Entry page and type of that URL is Entry URL.

4. k-means Algorithm:-

- First determine the no of clusters k you wish to have.
- Randomly select k objects as the initial centroids.
- For each of the remaining objects, calculate the distance between the objects and the k centroids (that is difference between the two values) and assign it to the cluster that it is much close to its centroid.
- 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.
- This process continues until the centroids of allthe clusters do not change again.

4. MAIN FEATURES OF EMPIRICAL FORUM MINING

4.1 User can ask any question

User can ask the question in the forum, and post the answers to the question which are askedby others. Then there is mining of answers carried out with the help of FAQ generalization technique. This is beneficial for the user to get the actual content. This have the high priority to get the actual content instead of getting the large data.

4.2 User can select any topic present on the entry page.

User not only select or ask question in the forum but also he can select the topics which are already discussed in the forum and get information about it. One user can also see the detail about the question that are asked by other user. By this feature the user has not to search again if the topic is already present on the page, he can get the information directly on the main page.

4.3 FAQ mining technique

There are many questions asked by the user and the posted answer to them. So we are in confusion to know the actual one. So we are doing the FAQ mining. In FAQ we are calculating the mostly asked question and the mostly matched answer to the question. Then we are mining the answer by removing stop words, doing stemming and lowercase conversion. After this clustering of this all words is done to obtain the information related to our desired topic.

EXPECTED RESULT:

User select or fire the query or his question to the forum then there are many result that user get. Our forum summarizes the topics and gives the appropriate answer so that user can get desired result in minimum time.

User can also only see the post on the different topics that are already discussed and also post his opinion to any topic that he want.

RESULT ANALYSIS:

INPUT	OUTPUT
1.Fetching URL http://www.aspforums.net...	Links: (210) <> (Dismiss) <http://www.aspforums.net/Home.aspx> () <> () <http://www.aspforums.net/ForgotPassword/> (Forgot Password)<http://www.aspforums.net/SignUp/> (Create Account)

	<http://www.aspforums.net/Tutorials/AddImage> (How to add image or screenshot to .) <http://www.aspforums.net/Login/?NewQuestion=1&RedirectUrl=http%3a%2f%2fwww.aspforums.net%2fForums%2f> (Ask Question) <http://www.aspforums.net/Login/?NewQuestion=1&RedirectUrl=http%3a%2f%2fwww.aspforums.net%2fForums%2f> () <http://www.aspforums.net/Net-Framework/176/Forums> (.Net Framework) <http://www.aspforums.net/Handlers/RSS.ashx?ForumId=233> ()
2.Afterremoving unnecessary links or space	http://www.aspforums.net/Threads/766763/How-to-publish-AspNet-website-on-IIS-using-IP-address/Replies/3#Replies http://www.aspforums.net/Forums/Data-Controls/210/Threads http://www.aspforums.net/Threads/141646/Copy-all-rows-records-from-one-Table-to-another-of-MySql-database-in-ASPNet/Replies/4#Replies http://www.aspforums.net/Threads/478487/crystal-report-log-in-problem/Replies/1#Replies
3.Index/Thread URL detection I. For Index URL <http://www.aspforums.net/Tutorials/NewThread> (Ask Question)	<http://www.aspforums.net/Tutorials/NewThread> (Ask Question) Anchor Text: (Ask Question) Length:17 Destination Page Type is:Index page URL Type is:Index URL Count:5
II. For Thread URL <http://www.aspforums.net/Threads/978670/Radio-Buttons-Not-working-in-IE11/Replies/6#Replies> (RE: Radio	<http://www.aspforums.net/Threads/978670/Radio-Buttons-Not-working-in-IE11/Replies/6#Replies> (RE: Radio Buttons Not working in I.) Anchor Text: (RE: Radio Buttons Not working in I.) Length:40

Buttons Not working in I.)	Destination Pagetype is:Thread page URL type is:Thread URL Count is:3
3.Page flipping URL http://www.aspforums.net/Threads/179469/count-occurance-of-specified-character-in-string/Replies/2#Replies RE: count occurance of specified character in string	1 7 9 4 6 9 http://www.aspforums.net/Threads/179469/count-occurance-of-specified-character-in-string/Replies/2#Replies RE: count occurance of specified character in string Url is Page flipping
4.Entry URL Discovery Algorithm **4a. Fetch URL**	Output of first fetching URL (point 1).
4b. Remove URL	Output of Remove URL (point 2).
4d. Intersection Of URL	http://www.aspforums.net/Forums/

4e. Fetching Common URL	<http://www.aspforums.net/Forums/Net-Basics/233/Threads> (391) <http://www.aspforums.net/Threads/502872/Validating-two-date-fields-when-1st-must-be-within-7days-before-today-2nd-is-today-or-yesterday/Replies/4#Replies> (RE: Validating two date fields whe.) <http://www.aspforums.net/Forums/Net-Classes/224/Threads> (.Net Classes) <http://www.aspforums.net/Forums/Net-Classes/224/Threads> (26) <http://www.aspforums.net/Forums/Net-Classes/224/Threads> (26) <http://www.aspforums.net/Threads/179469/count-occurance-of-specified-character-in-string/Replies/2#Replies> (RE: count occurance of specified c.) <http://www.aspforums.net/ASPNet/171/Forums> (ASP.Net) <http://www.aspforums.net/Forums/ASPNet-AJAX/212/Threads> (ASP.Net AJAX) <http://www.aspforums.net/Forums/Net-Basics/233/Threads> (.Net Basics) Anchor Text: (.Net Basics) length:16 Dstpagetype is:Index page URL type is:Index URL Link is: <http://www.aspforums.net/Forums/Net-Basics/233/Threads> (.Net Basics) Count:5
	<http://www.aspforums.net/Forums/ASPNet-AJAX/212/Threads> (ASP.Net AJAX) Anchor Text: (ASP.Net AJAX) length:17 Dstpagetype is:Index page URL type is:Index URL Link is: <http://www.aspforums.net/Forums/ASPNet-AJAX/212/Threads> (ASP.Net AJAX) Count:12
4f. Get Entry URL	http://www.aspforums.net/Forums/

5.cosine similarity Q1:http://www.aspforums.net/Threads/172879/Convert-PDF-to-Image-In-ASPnet-with-C/Replies/3#Replies Q2: http://www.aspforums.net/Threads/830856/How-to-save-images-locally-on-Users-machine-in-ASPNet/Replies/1#Replies Q3: http://www.aspforums.net/Threads/178172/How-to-fit-an-image-responsive-to-fit-in-every-resolution-using-Bootstrap/Replies/11#Replies	11.157177565710489,80.47189562170502,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, 11.157177565710489,0.0,80.47189562170502,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, 11.157177565710489,0.0,0.0,80.47189562170502,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, 0.0,0.0,0.0,0.0,160.94379124341003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, 0.6563045626888522,0.0,0.0,0.0,0.0,4.733640918923824,4.733640918923824,4.733640918923824,9.467281837847649,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,9.467281837847649,9.467281837847649,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,4.733640918923824,
Clustering Questions:	1.Get Connection String from App.Config in Windows Forms Application using C# .Net and VB.Net 2. How to get value of HTML Input TextBox in ASP.Net code behind using C# and VB.Net
Clustering Answers:	- How to add connection string in app.config using C# in Windows Application? And How to Open Connection in C# windows Application? Reply Voted Answer Selected Reply Posted on Jan 04, 2015 10:35 PM Mudassar Expert Refer: How to add and use connection string from App.Config file in Windows Application using C# and VB.Net ASP.Net • If this valid is a valid duplicate/abuse/broken link reply you will earn 5 bonus points. • But if this reply is not a valid duplicate/abuse/broken link reply you loose 10 points. Abuse Reply Duplicate Broken Link Report Cancel Once you mark this reply as Not Satisfactory, it will get deleted and you will not be able to view this reply. Confirm Cancel Submit Cancel

	- Hi, How to retrieve HTML input value from code behind (VB) Regards, Tempalli Reply Voted Answer Selected Reply Posted on Oct 31, 2014 02:04 AM Mudassar Expert Refer Get value of HTML Input TextBox in ASP.Net code behind using C# and VB.Net ASP.Net • If this valid is a valid duplicate/abuse/broken link reply you will earn 5 bonus points. • But if this reply is not a valid duplicate/abuse/broken link reply you loose 10 points. Abuse Reply Duplicate Broken Link Report Cancel Once you mark this reply as Not Satisfactory, it will get deleted and you will not be able to view this reply. Confirm Cancel Submit Cancel

CONCULSION

In this paper we propose the Empirical Forum Mining which is important programming model for minimizing the time wastage to large search. We presented a solution empirical Forum Mining system for effectively mining the forum to get FAQ.

We have studied that how the forum can crawl data and how to get information. By using FAQ we are mining the mostly asked question and finding appropriate data hence we feels that this system will be useful in future.

REFERENCES

[1] J. JIANG, X. SONG, AND N. YU, "FOCUS: LEARNING TO CRAWL WEB FORUMS", IEEE TRANS. Knowledge and Data Engg, pp. 1293-1306, 2013.

[2] "Web Forum Crawling Techniques", by Namrata H.S Bamrah, B.S Satpute, Pramod Patil, International journal of computer Applications, No 17, Vol 85, January 2014.

[3] "iRobot: An Intelligent Crawler for Web Forums" by R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, Proc. 17th Int"l Conf. World Wide Web, pp. 447-456, 2008.

[4] "Learning URL Patterns for Webpage De-Duplication", by H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg and A. Sasturkar, Proc. Third ACM Conf. WebSearch and Data Mining, pp. 381-390, 2010.

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

[6] J. Mao, J. Zhu. FAQ Auto Constructing Based on Clustering, in: Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on: IEEE, 2012

[7] Gao,C.Lin,C-Y. andSong,Y-I. Wang,L.(2008) , "Finding Question- Answer Pairsfrom Online Forums", Proc. 31st Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval.

[8] "Automatic Extraction of Web Data Records Containing User-Generated Content ",IEEE Trans. Knowledge and Data Engg, pp. 1293-1306, 2013.

[9] "Web Crawler", by Raja Iswary, KeshabNath, October 2013, International Journalof Advanced Research in Computer and Communication Engineering, Vol. 2

[10]G. S. Manku, A. Jain, and A. D. Sarma. Detecting near duplicates for Web crawling. In Proc.of 16thWWW, pages 141-150, 2007.

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

[12] Ran Vijay Singh and M.P.S Bhatia , "Data Clustering with Modified K-means Algorithm", IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp 717-721.

[13]D. Napoleon and P. Ganga Lakshmi, "An Efficient k-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE 2010.

[14] Neha Aggarwal and KritiAggarwal, "A Mid- point based k –mean Clustering Algorithm.For Data Mining". International Journal on Computer Science and Engineering (IJCSE) 2012.

[15]KoheiArai, AliRidhoBarakbah, Hierarchical K-means: an algorithm for centroids initialization for k-means.