WEB FORUMS CRAWLER FOR USER RESPONSE ANALYSIS

B.Muthulakshmi[1], **Dr.V.Thiagarasu**[2]

[1]M.Phil Research Scholar, PG & Research Department of Computer Science,
Gobi Arts & Science College (Autonomous), Gobichettipalayam, India.

[2]Associate Professor of Computer Science, PG & Research Department of Computer Science,
Gobi Arts & Science College (Autonomous), Gobichettipalayam, India.

ABSTRACT

The goal of FoCUS is to crawl relevant forum content from the web with minimal overhead. The forums have different layouts and are powered by different forum software packages; they always have similar implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages. The web forum crawling problem is reduced to a URL-type recognition problem and classifies them as Index Page, Thread Page and Page-Flipping page. To address the scalability issue, the research proposes an edge-centric clustering scheme is to extract sparse social dimensions approach can efficiently handle networks of millions of actors while representing a comparable prediction performance to other non-scalable methods. In addition, the research includes a new concept called sentiment analysis which transforms the cases into a standard model of features and classes. This is developed in two stages: emotional polarity computation and integrated sentiment analysis based on K-means clustering. The proposed unsupervised text-mining approach is used to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span. As a result, the behavior of individuals is collected through their posts in a forum and then they are classified as positive/negative posts. The positive and negative value is assign to each word and to classify the word in the document.

KEYWORDS: Entry index thread path, Forum crawling, Page classification, Page type, URL pattern learning, URL type.

I.INTRODUCTION

Web forums are important services where users can request and exchange information with others to collect information from forums, their content must be downloaded first. Though, forum crawling is not a trivial problem. Generic crawlers which adopt a breadth-first traversal strategy are usually ineffective and inefficient for forum crawling. This is mainly due to two non crawler friendly characteristics of forums: duplicate links and uninformative pages and page-flipping links. Besides duplicate links and uninformative pages, a long forum board or thread is usually divided into multiple pages which are linked by page-flipping links.Web Crawler was originally a separate search engine with its own database, and displayed advertising results in separate areas of the page. The number of possible crawl able URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Generic crawlers process each page individually and ignore the relationships between such pages. These relationships should be preserved while crawling to facilitate downstream tasks such as page wrapping and content indexing. Forums exist in many different layouts or styles and are powered by a variety of forum software packages, but they always have implicit navigation paths to lead users from entry pages to thread pages. The proposed approaches K-means clustering is a method of vector quantization originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

ISSN: 2278 – 1323

II.RELATED WORK

De-duping URLs [1], [10] is an extremely important problem for search engines, since all the principal functions of a search engine, including crawling, indexing, ranking, and presentation, are adversely impacted by the presence of duplicate URLs. The de-duping problem has been addressed by fetching and examining the content of the URL, their approach here is different. Given a set of URLs partitioned into equivalence classes based on the content to address the problem of mining this set and learning URL rewrite rules that transform all URLs of an equivalence class to the same canonical form.Vidal et al [6] proposed a method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. Target pages were found through comparing DOM trees of pages with a preselected sample target page. In the research "Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms" the author Monika Hen zinger[2], [12] stated that Broder et al.'s Shingling algorithm and Charikar's random projection based approach are considered state-of-the-art" algorithms for finding near-duplicate web pages. In the paper "Learning URL Patterns for Webpage De-duplication" the authors Hema Swetha Koppula, Krishna P. Leela and Amit Agarwal [3] stated that Presence of duplicate documents in the World Wide Web adversely affects crawling, indexing and relevance, which are the core building blocks of web search. Their technique is composed of mining the crawl logs and utilizing clusters of similar pages to extract transformation rules, which are used to normalize URLs belonging to each cluster. They compare the precision and scalability of our approach with recent efforts in using URLs for de-duplication. In this research "Extracting and Ranking Product Features in Opinion Documents" [4] the authors Lei Zhang and Bing Liu stated that an important task of opinion mining is to extract people's opinions on features of an entity. The extracted opinion words and features are utilized to identify new opinion words and new features, which are used again to extract more opinion words and features. They rank feature candidates by feature importance which is determined by two factors: feature relevance and feature frequency. The problem is formulated as a bipartite graph and the well-known web page ranking algorithm HITS is used to find important features and rank them high.

III. PROPOSED SYSTEM

The research includes online forums hotspot detection and forecast using sentiment analysis and text mining approaches. This is developed in two stages: emotional polarity computation and integrated sentiment analysis based on K-means clustering. The proposed unsupervised text-mining approach is used to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span. Data are collected from forums.digitalpoint.com which includes a range of 75 different topic forums. Computation indicates that within the same time window, forecasting achieves highly consistent results with K-means clustering.The large scale forum crawling problem as a URL type recognition problem by recognizing the EIT path through learning the ITF regexes. Index Page, Thread Page and Page-Flipping URLs are identified as well as forums post contents are also extracted. Online data is taken for mining. K-Means clustering approach classifies the forums in to related groups. Not only forums are clustered based on sentiment values, but also posts are clustered to find the number of items belongs to the individual clusters.

The proposed systems mainly focus ITF Regex Learning. To learn ITF regexes, FoCUS adopts a two-step supervised training procedure. The first step is training sets construction. The second step is regexes learning.

1. CONSTRUCTING URL TRAINING SETS

The goal of URL training sets construction is to automatically create sets of highly precise index URL, thread URL, and page-flipping URL strings for ITF regexes learning. Its use a similar procedure to construct index URL and thread URL training sets since they have very similar properties except for the types of their destination pages; to present this part first. Page-flipping URLs have their own specific properties that are different from index URLs and thread URLs.

2. INDEX URL AND THREAD URL TRAINING SETS

A thread URL is a URL that is on an index page; its destination page is a thread page; its anchor text is the thread title of its destination page. It also notes that the only way to distinguish index URLs from thread URLs is the type of their destination pages. Therefore, we need a method to decide the page type of a destination page. The index pages and thread pages each have their own typical layouts. Usually, an index page has many narrow records, relatively long anchor text, and short plain text; while a thread page has a few large records (user posts). Each post has a very long text block and relatively short anchor text. An index page or a thread page always has a timestamp field in each record, but the timestamp order in the two types of pages are reversed: the timestamps are typically in descending order in an index page while they are in ascending order in a thread page. In addition, each record in an index page or a thread page usually has a link pointing to a user profile page.

3. PAGE-FLIPPING URL TRAINING SET

Page-flipping URLs point to index pages or thread pages but they are very different from index URLs or thread URLs. The proposed "connectivity" metric is used to distinguish page-flipping URLs from other loop-back URLs [7]. Page flipping URLs and proposed an algorithm to detect page flipping URLs based on these properties. In particular, the grouped page-flipping URLs have the following properties:

- Their anchor text is either a sequence of digits such as 1, 2, 3, or special text such as "last."
- They appear at the same location on the DOM tree of their source page and the DOM trees of their destination pages.
- Their destination pages have similar layout with their source pages. We use tree similarity to determine whether the layouts of two pages are similar or not. As to single page-flipping URLs, they do not have the property 1, but they have another special property.
- The single page-flipping URLs appearing in their source pages and their destination pages have the same anchor text but different URL strings.

4. K-MEANS CLUSTERING ALGORITHM

K-means [9] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori [11]. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. The k-means approach to clustering performs an iterative alternating fitting process to form the number of specified clusters. The k-means method first selects a set of n points called cluster seeds as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form a set of temporary clusters.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

1. FINDINGS

The proposed system automatically analyzes the emotional polarity of a text, based on which a value for each piece of text is obtained. Social dimensions are extracted to represent the potential affiliations of actors before discriminative learning occurs. The K-means clustering is applied to develop integrated approach for online sports forums cluster analysis. One hundred and forty two forums are extracted from forums.digitialpoint.com Six thousand six hundred and fifty four threads are spread among those forums Posts are found to be valid only if they contain minimum three words. Posts are grouped and average value taken for each thread likewise threads average sentiment values are taken and forum's average value is derived. It is found that more of the forums sentiments values are having more positive percent and less negative percent. In practice parents and students behavior are always hard to be explored and captured and so the proposed methodology efficiently analyze their sentiments [5]. An incomparable advantage of the proposed model is that it easily scales to handle networks with millions of posts while the earlier models fail. This scalable approach offers a viable solution to effective learning of online collective behavior on a large scale. Extending the edge-centric clustering scheme to address the object heterogeneity, not considered in this proposed system, can be a promising future direction. Since the proposed model is sensitive to the number of social dimensions as shown in the experiment, further research is needed to determine a suitable dimensionality automatically [8]. Other behavioral features from social media are not mined, and they can be integrated with social networking information to improve prediction performance in future.

The document is classified as positive and negative statement using K-means algorithm, for the purpose of processing the data for further clarification. The data is classified by the sentiment value, in the positive and negative. Each and every processed in the individual value. Data cleaning is the process is used to refuse the noise data as stem word, stop word and synonym word [13]. In table [1] and table[2] shows the stop and stem word is used to neglect the grammar, verbal and non-verbal words in post and replies. After the data cleaning work, the strength of the data will be reduced.

2. EXPERIMENTAL EVALUATION

A. ASP is a great Technology and with this you have great future...

Table No: 1 Sentimental value for given sentence before data cleaning

WORDS	SENTIMENT VALUE
ASP	2
Is	2
A	2
Great	3
Technology	2
And	2
With	1
This	2
You	4
Have	3
Great	3
Future	2
Total sentimental count	28

ELIMINATE STOP WORD

Table No: 2 Sentimental value for given sentence after data cleaning

WORDS	SENTIMENT VALUE
ASP	2
Great	3
Technology	2
You	4
Have	3
Great	3
Future	2
Total sentimental count	19

B. ASP has not more future, But ASP.Net has a great future

Table No: 3 Sentimental value for given sentence before data cleaning

WORDS	SENTIMENT VALUE
ASP	2
Has	2
Not	-17
More	3
Future	2
But	2
ASP.Net	1
Has	2
A	4
Great	3
Future	2
Total sentimental count	6

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data is shown in table [3].

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.

3. AFTER ELIMINATION OF STOP WORDS

Table No: 4 Sentimental value for given sentence after data

WORDS	SENTIMENT VALUE
ASP	2
Not	-17
More	3
Future	2
But	2
ASP.Net	1
Great	3
Future	2
Total sentimental count	-2

Comparison of Data Cleaning in Post replies with sentiment value

In Fig.1the post replies with sentiment value, before data cleaning and after data cleaning of the forum. The before data cleaning of the forum post is high in ASP is great technology and the after data cleaning of the forum post in ASP is not more future is low.

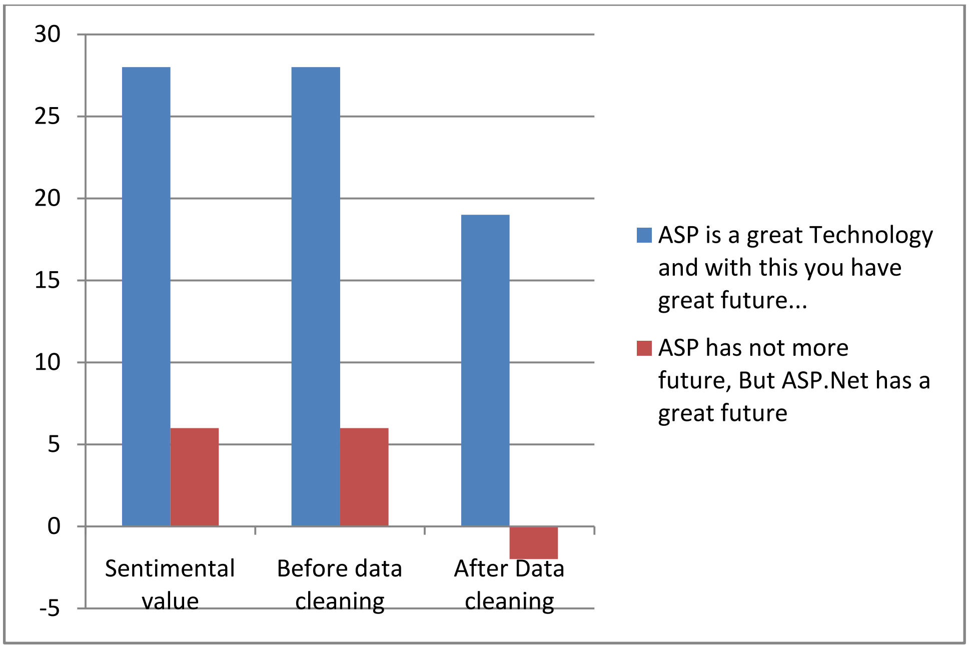

Figure no: 1 Chart representation for post replies with sentiment value, before data cleaning and after data cleaning.

V. CONCLUSION AND FUTURE WORK

In this research, the algorithms are developed to automatically analyze the emotional polarity of a text, based on which a value for each piece of text is obtained. The absolute value of the text represents the influential power and the sign of the text denotes its emotional polarity. This K-means clustering is applied to develop integrated approach for online sports forums cluster analysis. Clustering algorithm is applied to group the forums into various clusters, with the center of each cluster representing a hotspot forum within the current time span. Empirical studies present strong proof of the existence of correlations between post text sentiment and hotspot distribution. In the future, how to utilize the inferred information and extend the framework for efficient and effective network monitoring and application design. The application can be web service oriented so that it can be further developed in any platform. The application if developed as web site can be used from anywhere. At present, number of posts/forum, average sentiment values/forums, positive % of posts/forum and negative % of posts/forums are taken as feature spaces for K-Means clustering. In future, neutral replies, multiple-languages based replies can also be taken as dimensions for clustering purpose. The new system is designed such that those enhancements can be integrated with current modules easily with less integration work. The new system becomes useful if the above enhancements are made in future.

REFERENCES

[1] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-Duping URLs via Rewrite Rules" Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.

[2] M. Henzinger, "Finding Near-Duplicate Web Pages: A Large Scale Evaluation of Algorithms,"Proc. 29th Ann. Int'l ACM SIGIR. Conf. Research and Development in Information Retrieval, pp. 284-291, 2009.

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

[4] L. Zhang, B. Liu, S.H. Lim, and E. O'Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 1462-1470, 2010.

[5] Wilson, Theresa, Janyce Wiebe and Paul Hoffmann,"Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", In Proceedings of HLT/EMNLP, 2005, 2008.

[6] M.L.A. Vidal, A.S. Silva, E.S. Moura, and J.M.B. Cavalcanti, , "Structure-Driven Crawler Generation by Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 292-299, 2006.

[7] Girju, Roxana, Adriana Badulescu and Dan Moldovan.,"Automatic Discovery of Part-Whole Relations" Computational Linguistics, 32(1):83-135, 2006, 2006.

[8] S. Brin, J. Davis, and H. Garcia-Molina, "Copy Detection Mechanisms for Digital Documents,"ACM SIGMOD International Conference on Management of Data, 398-409, 1995.

[9] D. Fetterly, M. Manasse, and M. Najork, "on the evolution of clusters of near-duplicate web pages", In Proc. 1st Conf. on Latin American Web Congress, page 37, 2008.

[10] A. Pereira, R. Baeza-Yates, and N. Ziviani, "Where and how duplicates occur in the web". In Proc. 4th Latin American Web Congress, pages 127–134, 2006.

[11] A. Broder, S. C. Glassman, M. Manasse, and G. Zweig, "Syntactic clustering of the web. Computer Networks", 29(8-13):1157–1166, 1997.

[12] M. Henzinger, "Finding near-duplicate web pages: a large-scale evaluation of algorithms". In Proc. 29th SIGIR, pages 284–291, 2006.

[13] J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W.-Y. Ma,"Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums," Proc. 18th Int'l Conf. World Wide Web, pp. 181-190, 2009.