Signed Approach for Mining Web Content Outliers

G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma

Abstract**—**The emergence of the Internet has brewed the revolution of information storage and retrieval. As most of the data in the web is unstructured, and contains a mix of text, video, audio etc, there is a need to mine information to cater to the specific needs of the users without loss of important hidden information. Thus developing user friendly and automated tools for providing relevant information quickly becomes a major challenge in web mining research. Most of the existing web mining algorithms have concentrated on finding frequent patterns while neglecting the less frequent ones that are likely to contain outlying data such as noise, irrelevant and redundant data. This paper mainly focuses on Signed approach and full word matching on the organized domain dictionary for mining web content outliers. This Signed approach gives the relevant web documents as well as outlying web documents. As the dictionary is organized based on the number of characters in a word, searching and retrieval of documents takes less time and less space.

Keywords**—**Outliers, Relevant document,, Signed Approach, Web content mining, Web documents..

I. Introduction

WITH the exponential growth of information available on the internet, updating incoming data and retrieving relevant information from the web quickly and efficiently is a growing concern. Most of the web search engines typically employ conventional information retrieval and data mining techniques to discover automatically useful and previously unknown information from web content. With the enormous growth on the web, users get easily lost in the rich hyper structure. In addition, as most of the data in the web is unstructured, and contains a mix of text, video, audio etc, there is a need to mine information to cater to the specific needs of the users[9]. Efforts are being made to make such data available, usually in some structured form as in matrix form for further manipulation. Web mining is an emerging research area focused on resolving these problems. The proposed work in web mining aims to develop new methodology to effectively mine useful knowledge or information from the web documents quickly. In general, web mining tasks can be classified into three major categories, web structure mining, web usage mining and web content mining. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web usage mining refers to the discovery of user access patterns from web usage logs. Web content mining aims to extract/mine useful information from the web pages based on their contents [1],[4],[10],[11]. Two groups of web content mining are those that directly mine the content of documents and those that improve on the content search of other tools like search engine. For Web content mining data can be image, audio, text and video [15]-[16].

Existing web mining algorithms do not consider documents having varying contents within the same category called web content outliers. Generally, Outliers are the data that obviously deviate from others, disobey the general mode or behavior of data and disaccord with other existing data. Outliers may also reflect the true properties of data, such as the rare disastrous weather recorded in meteorological database, which often contains one or more properties whose values seriously deviate from the normal values. However, these data may contain more valuable information than normal data.

Researches on outlier detection broadly fall into following categories:

A. Distribution based methods are conducted by the statistics community. These methods deploy some known distribution model and detect as outliers points that deviate from the model.

B. Depth based algorithms organize objects in convex hull layers in data space according to peeling depth and outliers expected to be with shallow depth values[13].

C. Deviation based techniques detect outliers by checking the characteristics of objects and identify an object as that deviates these features as outlier.

D. Distance based algorithms give a rank to all points, using distance of point from *k*-th nearest neighbor, and orders points by this rank. The top *n* points in ranked list identified as outliers. Alternative approaches compute the outlier factor as sum of distances from *k* nearest neighbors.

E. Density based methods rely on local outlier factor (LOF) of each point, which depends on local density of neighborhood. Points with high factor are indicated as outliers.[12]

G.Poonkuzhali is Assistant professor in the Department of Computer Science and Engineering with the Rajalakshmi Engineering College, Affiliated to Anna University Chennai, Tamil Nadu, India, phone: 9444836861, email : Kuzhal_s@yahoo.co.in

K.Thiagarajan is Senior Lecturer in the Department of Mathematics with the Rajalakshmi Engineering College, Affiliated to Anna University Chennai, Tamil Nadu, India, email : vidhyamannan@yahoo.com

K.Sarukesi is Vice Chancellor with the Hindusthan University – Chennai, email: profsaru@yahoo.com

G.V.Uma is Professor in the Department of Computer Science and Engineering with the Anna University-Chennai, email: gvuma@annauniv.edu

Unlike traditional outlier mining algorithm designed only for numeric data sets, web outliers mining algorithm should be applicable to various types of data including text, hypertext, image, video etc. Web pages that have different contents from the category in which they were taken constitute web content outliers.[7]-[8] Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from the web documents[10]-[11] Also, web content outliers mining can be used to determine pages with entirely different contents from their parent web sites.

In the proposed system, web documents are extracted from the search engines by giving query by the user to the web. Then the obtained web documents D is preprocessed, i.e., stop words, stem words and except text other data such as hyperlinks, sound, images etc are removed. The output is a set of documents with white-spaced separated words and it is indexed in two dimensional format (i,j), where 'i' represent web pages and 'j' represent words. Therefore, first word from first web page is indexed as (1,1), second word from the first page is indexed as (1,2) etc,. The domain dictionary is arranged in such a way that, all 1-letter word will be indexed first, followed by 2-letter words, then 3-letter words similarly up to 15-letters word which is a very reasonable upper bounds for number of characters in a word.

Each page is mined individually to detect relevant and irrelevant documents using signed approach. Finally, a relevant web document is obtained which contains required information catering to the user needs.

Outline of Paper

Section 2 presents the overview of the signed approach for relevancy computation. Section 3 presents architectural design and flow diagram of the proposed system. Section 4 presents the algorithm for retrieving relevant and irrelevant web documents. Section 5 presents observations. Finally, Section 6 presents conclusions and future work.

II.SIGNED APPROACH FOR RELEVANCY COMPUTATION

The proposed algorithm explores the advantages of full word matching and signed approach using organized domain dictionary where the indexing is done based on the length of the word. First, the input web document is preprocessed and separated into white spaced words. The full word profile for the document is generated in matrix form (i.e., $W_{1,4}$ - represents 4th word in 1st page). Then the jth word from ith page is taken and its length is calculated (i.e., | W_{ji} |) and depending on the number of characters, the respective index on the domain dictionary is searched. If the word (W_{ji}) is found in the dictionary, then positive count is incremented by one else negative count is incremented by one. This process is carried out for all words in that web page. Finally, positive count is compared with the negative count to check the relevancy of that web page. If the positive count is less than the negative count, then that page is irrelevant, otherwise it is considered as more relevant.

III.ARCHITECTURE OF THE PROPOSED SYSTEM

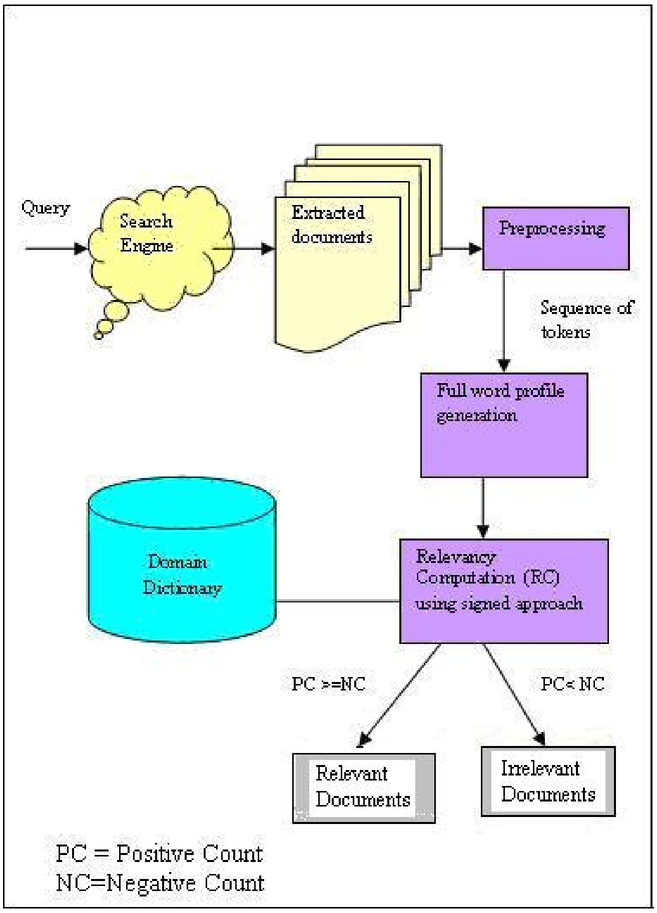
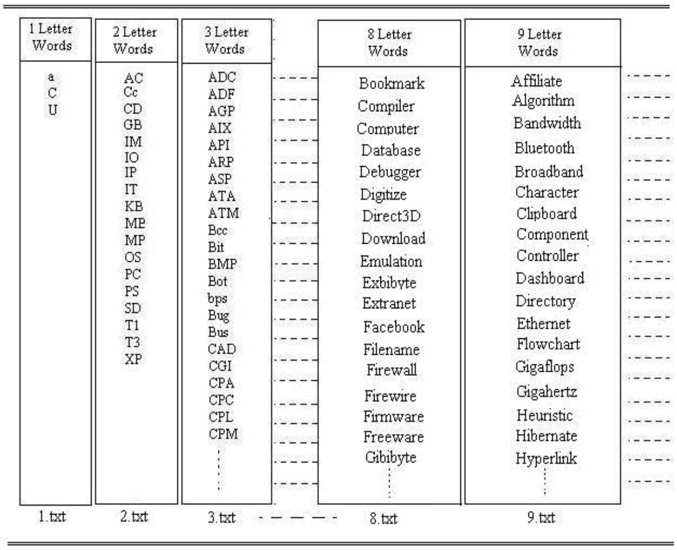


Fig. 1 Architectural Design of the proposed System



1 Letter Words	2 Letter Words	3 Letter Words		8 Letter Words	9 Letter Words
a	AC	ADC		Bookmark	Affiliate
C	Cc	ADF		Compiler	Algorithm
U	CD	AGP		Computer	Bandwidth
	GB	AIX		Database	Bluetooth
	IM	API		Debugger	Broadband
	IO	ARP		Digitize	Character
	IP	ASP		Direct3D	Clipboard
	IT	ATA		Download	Component
	KB	ATM		Emulation	Controller
	MB	Bcc		Exbibyte	Dashboard
	MP	Bit		Extranet	Directory
	OS	BMP		Facebook	Ethernet
	PC	Bot		Filename	Flowchart
	PS	bps		Firewall	Gigaflops
	SD	Bug		Firewire	Gigahertz
	T1	Bus		Firmware	Heuristic
	T3	CAD		Freeware	Hibernate
	XP	CGI		Gibibyte	Hyperlink
		CPA		⋮	⋮
		CPC			
		CPL			
		CPM			
		⋮			
1.txt	2.txt	3.txt	- — — — -	8.txt	9.txt

Fig. 2 Domain dictionary for computer terms

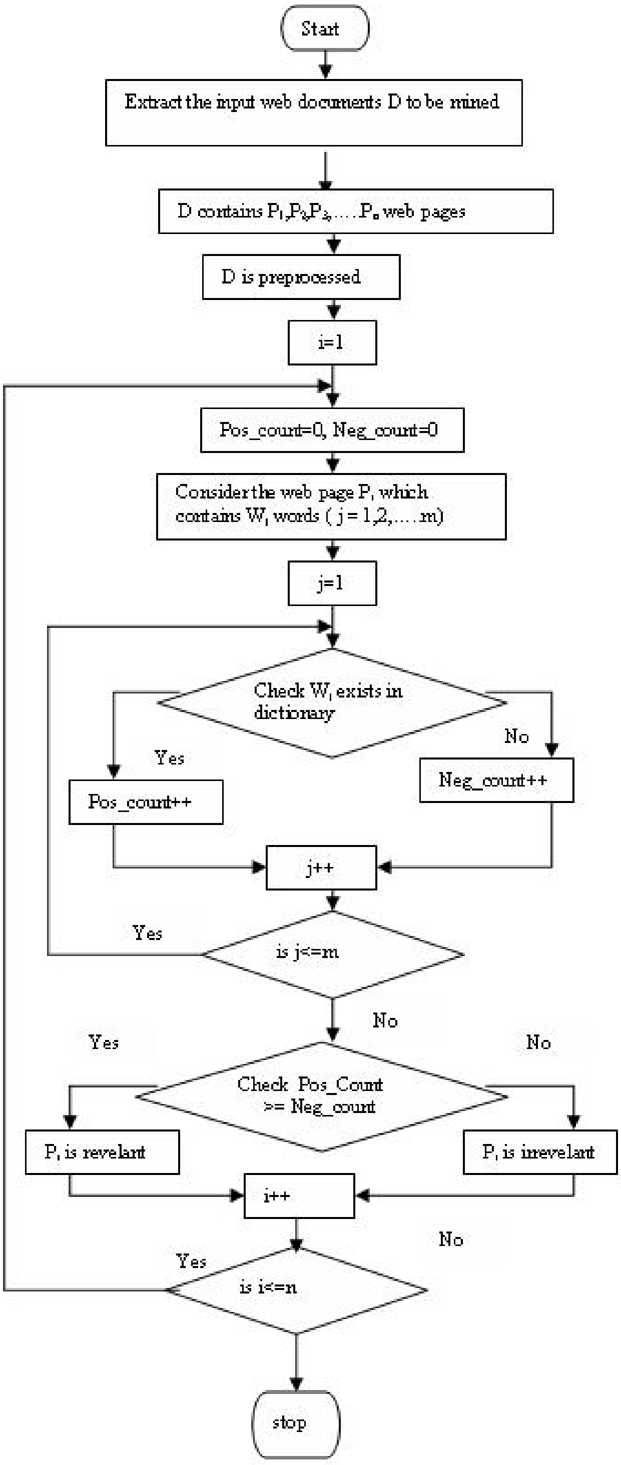

Start
Extract the input web documents D to be mined
D contains P1,P2,P3,....Pn web pages
D is preprocessed
i=1
Pos_count=0, Neg_count=0
Consider the web page Pi which contains Wj words (j = 1,2,....m)
j=1
Check Wj exists in dictionary
Yes
No
Pos_count++
Neg_count++
j++
Yes
is j<=m
No
Yes
No
Check Pos_Count >= Neg_count
Pi is revelant
Pi is irrevelant
i++
No
Yes
is i<=n
stop

Fig. 3 Flow Diagram of the Proposed System

IV. ALGORITHM

Input: Domain Dictionary, Web Document D_i

Output: Relevant Pages and Irrelevant Pages

Other Variable: Pos_count, Neg_count

```
Extract the input web document D after preprocessing.
Read the contents of web page $P_i$
Generate full word profile.
for ( i=1;i<=n;i++)
{
    Pos_count=0; Neg_count=0;
    for(j=1;j<=m;j++)
    {
        if ( jth word exists in dictionary)
        {
            Pos_count++;
        else
            Neg_count++;
        }
    }
    if (Pos_count >= Neg_count)
    {
        Print $P_i$ as relevant web page ;
    else
        Print $P_i$ as irrelevant web page ;
    }
}
```

Nomenclature

D – Web document to be mined.

P_i – Web page

$W_{j,i}$ - j^{th} word in i^{th} web page

V.OBSERVATIONS

Experimental results ensure that the memory space, search time and run time gets reduced by using organized domain dictionary than normal indexed dictionary for checking the relevancy of the web documents. As the efficiency of web content is increased, the quality of the search engines also gets increased. This method is very simple to implement. The proposed algorithm is used by business personals to keep track of all the positive and negatives aspects related to their business.

TABLE I
DOCUMENTS RETRIEVED ORIGINALLY WITH RESPECT TO COMPUTER DOMAIN

File name	Positive count	Negative count	Status
File1.html	52	39	relevant
File2.html	100	145	not relevant
File3.html	127	107	relevant
File4.html	13	49	not relevant
File5.html	68	140	not relevant
File6.html	184	113	relevant
File7.html	169	120	relevant
File8.html	36	75	not relevant
File9.html	87	200	not relevant

TABLE II
DOCUMENTS RETRIEVED AFTER RELEVANCY COMPUTATION

File name	Positive count	Negative count	Status
file1.html	52	39	relevant
file3.html	127	107	relevant
File6.html	184	113	relevant
File7.html	169	120	relevant

A. *Precision*

It is the ratio between the number of relevant documents returned originally and the total number of retrieved documents returned after eliminating irrelevant documents. Here the relevant documents indicate the required documents which satisfy the user needs.

$$\text{Precision} = \frac{\text{Relevant} \cap \text{Retrieved originally}}{\text{Retrieved after refinement}}$$

B. *Recall*

It is the ratio between the number of relevant documents returned originally and the total number of relevant documents returned after eliminating irrelevant documents

$$\text{Recall} = \frac{\text{Relevant} \cap \text{Retrieved originally}}{\text{Relevant after refinement}}$$

C. *Time Taken*

The time taken by the entire process is the sum of the initial time taken by the general purpose search engines plus the time taken by the refinement algorithm to process the results. As the searching time to match the word with the dictionary is reduced, the overall time taken to refine the documents also gets reduced drastically.

VI. CONCLUSION

Web mining is a growing research area in the mining community because of the great patronage the web continues to enjoy. Retrieving relevant content from the web is a very common task. However, the results produced by most of the search engine do not necessarily produce result that is best possible catering to the user needs. This paper proposes a new algorithm using signed approach for improving the results of web content mining by detecting both relevant and irrelevant web documents. Future work aims at experimental evaluation of web content mining in terms of reliability and to explore other mathematical tools for mining the web content. Also, a comparative study of this algorithm with existing algorithms is to be done.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ponnammal Natarajan, former Director of research, Anna University, Chennai, India. Currently Advisor, (Research and Development), Rajalakshmi Engineering College, for her intuitive ideas and fruitful discussions with respect to the paper's contribution..

REFERENCES

[1] Bing Liu, Kevin Chen- Chuan Chang , Editorial: Special issue on Web Content Mining , SIGKDD Explorations, Volume 6, Issue 2.

[2] Changjun Wu, Guosun Zeng, Guorong Xu , A Web Page Segmentation Algorithm for Extracting Product Information , Information Acquisition, 2006 IEEE International Conference on Publication Date: Aug. 2006.

[3] Cheng Wang, Ying Liu, Liheng Jian, Peng Zhang, A Utility based Web Content Sensitivity Mining Approach, International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT), IEEE/WIC/ACM 2008.

[4] Hongqi li, Zhuang Wu, Xiaogang Ji, Research on the techniques for Effectively Searching and Retrieving Information from Internet, International Symposium on Electronic Commerce and Security, IEEE 2008

[5] Jaroslav Pokorny, Jozef Smizansky, Page Content Rank: An approach to the Web Content Mining

[6] Jiang Yiyong, Zhang Jifu,Cai Jainghui, Zhang Sulan, Hu Lihua , The Outliers Mining Algorithm Based On Constrained Concept Lattice, Internal Symposium on Data Privacy and E.commerce , IEEE 2007.

[7] kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, A Survey on Web Content Mining and Extraction of Structured and Semistructured data,First International Conference on Emerging trends in Engineering and Technology, 2008

[8] Malik Agyemang, Ken Barker, Rada S. Alhajj, Framework for Mining Web Content Outliers , 2004 ACM Symposiumon Applied Computing.

[9] Malik Agyemang Ken Barker Rada S. Alhajj , Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams , 2005 ACM Symposium on Applied Computing

[10] G.Poonkuzhali, K.Thiagarajan, K.Sarukesi,Set theoretical Approach for mining web content through outliers detection, International journal on research and industrial applications, Volume 2, Jan 2009.

[11] G.Poonkuzhali, K.Thiagarajan, K.Sarukesi, Elimination of redundant Links in web pages- Mathematical Approach, Proc. Of World Academy of Science, Engineering and Technology, Volume 40, April 2009, pp 555-562

[12] Peng Yang, Biao Huang, A modified Density Based Outliers Mining Algorithm for large Dataset, 2008 IEEE, International Seminar on Future Information technology and Management Engineering.

[13] Peng Yang, Biao Huang, Density Based Outliers Mining Algorithm with Application to Intrusion Detection, 2008 IEEE, Pacific asia workshop on computational Intelligence and Industrial Application.

[14] Ramaswamy S, Rastogi R, Shim k, Efficient Algorithm for mining outliers from large data sets, proc. Of ACM SIGMOD 2000, pp 127 – 138.

[15] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD, July 2000

[16] Ricardo Campos , Gael Dias, Celia Nunes, WISE : Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques, International conference on Web Intelligence, IEEE/WIC/ACM 2006.

[17] R.P. Grimaldi, "Discrete and Combinatorial Mathematics", Pearson Edition, New Delhi 2002.
[18] Kenneth H. Rosen, "Discrete Mathematics and its Applications", Fifth Edition, TMH, 2003.
[19] J.P. Tremblay and R. Manohar, "Discrete Mathematical Structures with Applications to Computer Science", TMH, 1997.
[20] M.K. Venkataraman, N. Sridharan and N.Chandrasekaran, "Discrete Mathematics", The National Publishing Company, 2003.
[21] J.W.Han, M.Kamber, Data Mining: Concepts and Techniques Newyork kaufmann publishers 2001.

G.Poonkuzhali received B.E degree in Computer Science and Engineering from University of Madras, Chennai, India, in 1998, and the M.E degree in Computer Science and Engineering from Sathyabama University, Chennai, India, in 2005. Currently she is pursuing Ph.D programme in the Department of Information and Communication Engineering at Anna University – Chennai, India. She has presented and published 3 international journals and authored 5 books. She is a life member of ISTE (Indian Society for Technical Education) and CSI (Computer Society of India).

K.Thiagarajan working as Senior Lecturer in the Department of Mathematics in Rajalakshmi Engineering College- Chennai-India. He has totally 14 years of experience in teaching. He has attended and presented research articles in 33 National and International Conferences and published one national journal and 26 international journals. Currently he is working on web mining through automata and set theory. His area of specialization is coloring of graphs and DNA Computing.

Dr. K. Sarukesi has a very distinguished career spanning over 38 years. He has a vast teaching experience in various universities in India and abroad. He was awarded a commonwealth scholarship by the association of common wealth universities, London for doing Ph.D in UK. He completed his Ph.D from the University of Warwick – U.K in the year 1982. His area of specialization is Technological Information System. He worked as expert in various foreign universities. He has executed number of consultancy projects. He has been honored and awarded commendations for his work in the field of information technology by the government of Tamilnadu. He has published over 30 research papers in international conferences/journals and 40 National Conferences/journals.

G.V.Uma, received her M.E. from Bharathidasan University, India in year 1995 and Ph.D. from Anna University, Chennai, India in 2002. She has rich experience in teaching and research; currently working as a Professor in the Department of Computer Science & Engineering in Anna University. Her research interests include Software Engineering, Genetic Privacy, Ontology, Knowledge Engineering & Management, and Natural Language Processing. She has organized many Workshops, Seminars and Conferences in national and International level.