International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

Elimination of Redundant Links in Web Pages – Mathematical Approach

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

Abstract—With the enormous growth on the web, users get easily lost in the rich hyper structure. Thus developing user friendly and automated tools for providing relevant information without any redundant links to the users to cater to their needs is the primary task for the website owners. Most of the existing web mining algorithms have concentrated on finding frequent patterns while neglecting the less frequent one that are likely to contain the outlying data such as noise, irrelevant and redundant data. This paper proposes new algorithm for mining the web content by detecting the redundant links from the web documents using set theoretical(classical mathematics) such as subset, union, intersection etc,. Then the redundant links is removed from the original web content to get the required information by the user..

Keywords—Web documents, Web content mining, redundant link, outliers, set theory.

I. Introduction

WITH the exponential growth of information available on the web, 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. 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[2]. Efforts are being made to make such data available, usually in some structured form such as table, for querying and further manipulation. Web mining is an emerging research area focused on resolving these problems. . 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[4]-[5].

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

Web content mining is the process of mining, extraction and integration of useful data, information and knowledge from Web page contents. Some of the areas of doing research in web content mining is listed below:

- *Structured Data Extraction*
- *Unstructured Text Extraction*
- *Web Information Integration and Schema matching*
- *Building Concept Hierarchies*
- *Segmentation and Noise Detection*
- *Opinion extraction*

This paper focuses on segmentation and detection of noise issue, which implies outliers mining. Generally, Outliers are observations that deviate so much from other observations to arouse suspicion that they might have been generated using a different mechanism or data objects that are inconsistent with the rest of the data objects. Outliers identified in web data are referred to as *web outlier.*

Existing web mining algorithms do not consider documents having varying contents within the same category called web content outliers. 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. Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from the web documents[6]-[7]. Also, web content outliers mining can be used to determine pages with entirely different contents from their parent web sites. 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.

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]

Outline of paper

Section 2 presents the flow diagram of the proposed system. Section 3 presents the algorithm for detecting and eliminating redundant links on the web documents. Section 4 presents observations. Finally, Section 5 presents conclusions and future work.

II. Architecture of the Proposed System

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 divided into 'n' web pages based on the links. Then all the pages are preprocessed , by stemming process, stop words removal and except text, images, audio are also removed. Each page is mined individually to detect redundant links using set theory concepts. Initially, the contents of first page is taken and compared with the content of the second page. This process is repeated till n^{th} page. n general, In general, page P_i is compared with P_{i+1} to P_n. If any redundant links is noted, then that particular web page itself is removed from that web document. Finally, a modified web document is obtained which contains required information catering to the user needs.

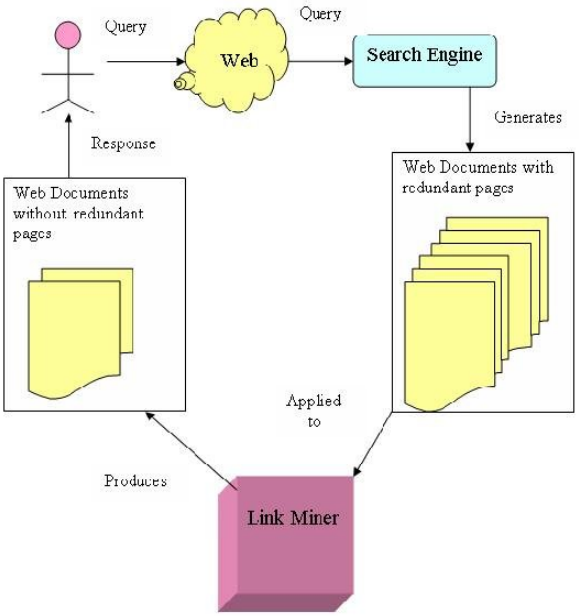

Fig. 1 Architecture of the proposed system

III. Flow of the Proposed System

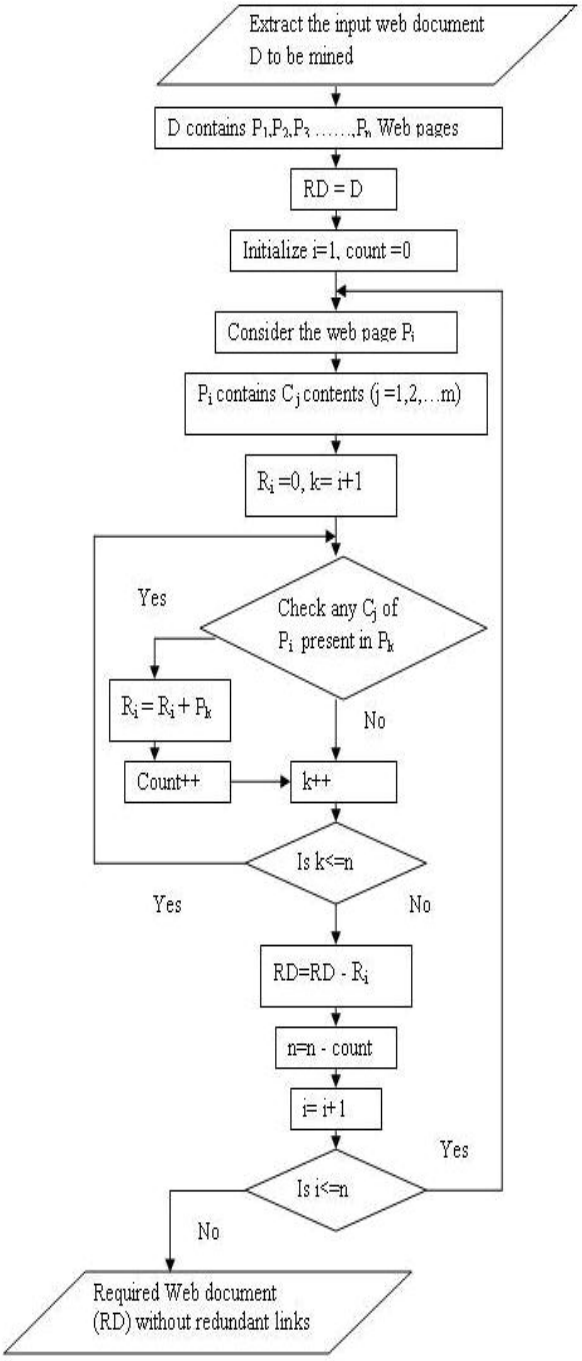

Fig. 2 Flow of the proposed system

IV. Algorithm of the Proposed System

Step 1: Enter the query on the web.

Step 2: Get the input web document D to be mined.

Step 3: Document D contains

$$D = \bigcup_{i=1}^{n} P_i$$

where i = 1,2,…..,n web pages.

Step 4: Assign required document(RD) as D

RD=D

Step 5 : Initialize the count as 0

Step 6: Our aim is to detect and eliminate the redundant links. Assume that P_i does not contain redundant links.

$$\text{Let } P_i = \bigcup_{j=1}^{m} C_j$$

where C_j = 1,2,…..,m web contents.(May be text, image of hypertext etc)

$$\text{such that } \bigcap_{j=1}^{m} C_j = \Phi$$

Step 7: Initialize Redundant set as 0 , increment k by i.

$R_i = 0$, k = i+1

Step 8: Check whether any C_j of P_i is present in P_k

$$\text{i.e., } C_j \bigcap [\bigcup_{k=1}^{n} P_k]$$

Step 9:→If present, add that P_k to the redundant document set R and increment count by one. Then goto step 11.

i.e., $R_i = R_i + P_k$

count = count+1

Step 10 → If not present, go to step 11

Step 11: increment k and repeat step 8 until k<=n.

Step 12: update the required document as

RD=RD - R_i

Step 13: compute n= n – count and increment i by one.

Step 14: Repeat from step 6 for all web pages in D. (i.e., i<=n)

Step 15: Required web document (RD) is obtained after removing Redundant set R from the original web document D.

V. Nomenclature

D – Web document to be mined.
P_i – Web page
C_i – Web content of any type (text / image / hypertext).
R_i - Contains the union of all redundant links. (Redundant Set)
RD - Contains Mined web content required by the end user.

VI. Observations

Experimental results ensure that the memory space, search time and run time gets reduced after eliminating the redundant links from the retrieved documents drastically. Also removal of redundant links makes the zipping of large data easier. As the efficiency of web content is increased, the quality of the search engines also gets increased. Precision of the refined document increases considerably.

Precision is one of the statistical measures for finding the success (quality) of the refined pages retrieved. It is the ratio between the number of relevant documents returned originally and the total number of relevant documents returned after eliminating redundant links. Here the relevant documents indicate the required documents which satisfy the user needs.

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

VII. Conclusion and Future Work

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 and mathematical set formulae for improving the results of web content mining by detecting and eliminating redundant links. 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.

VIII. Acknowledgements

The authors would like to thank Prof. Ponnammal Natarajan worked as Director – Research , Anna University- Chennai. Currently Advisor, (Research and Development), Rajalakshmi Engineering College, for her intuitive ideas and fruitful discussions with respect to the paper's contribution.

REFERENCES

[1] S.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
[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] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD, July 2000
[4] Bing Liu, Kevin Chen- Chuan Chang , Editorial: Special issue on Web Content Mining , SIGKDD Explorations, Volume 6, Issue 2.
[5] Jaroslav Pokorny, Jozef Smizansky, Page Content Rank: An approach to the Web Content Mining.
[6] 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
[7] 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.
[8] 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.
[9] 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
[10] J.P. Tremblay and R. Manohar, "Discrete Mathematical Structures with Applications to Computer Science", TMH, 1997.
[11] Kenneth H. Rosen, "Discrete Mathematics and its Applications", Fifth Edition, TMH, 2003.
[12] R.P. Grimaldi, "Discrete and Combinatorial Mathematics", Pearson Edition, New Delhi 2002.
[13]] M.K. Venkataraman, N. Sridharan and N.Chandrasekaran, "Discrete Mathematics", The National Publishing Company, 2003.
[14] 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

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 research papers in international conferences and 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 of nearly 38 years. He has a vast teaching experience in various university 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 specializations are 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.