Annihilation of Traditional Union through Ktmax-Maxam Algorithm in Disorder Mining

Thiagarajan K, Maheshwari .A, Kavitha J.

Abstract*: Nowadays web contains many identical documents. Many efforts made towards in search engines with identical detection algorithms, still the retrieved web documents with redundant link. The proposed system, we are effectively identifying and minimize the redundant information from web documents. We introduce the degree based** *KTMAX-MAXAM* **algorithm filters out the redundant information. Using proposed** *KTMAX-MAXAM* **algorithm accessing of web pages with reduced time and space complexity.

Index Terms*: Data Mining, Degree, Maximum Degree, Minimum Degree, Link, Path.

I. INTRODUCTION

The data and information available on web is exponentially improving, duplication of web content also increase simultaneously. Retrieving relevant information from web without redundancy is more challenge task nowadays where in web mining communities. Web content mining is the way toward extracting the applicable information, data and learning from World Wide Web. Utilizing customary data recovery and information mining systems it get to the known and obscure data from the Web content. Web mining is categorized into three group Web Content mining, Web structure mining, Web usage mining. Traditional web mining algorithms handle with structured document than the advanced methodology of mining algorithm can deal the entire heterogeneous document comprises of images, graphs, videos, etc.

II. ARCHITECTURE OF PROPOSED SYSTEM

A query is searched in a web search tool to recover some significant and required data for the client, either the search query is known or unintelligible to the client, it generally to reply with relevant data rather than redundant, however we can't guarantee that the reply for the query about the significance and redundancy. Once the input query is requested, the search engine generate the document with multiple web pages along with the links, the user will be unaware about the content of the web pages, the extracted web documents contain multiple web pages either be redundant or not.

The Document retrieved must follow some constraints which have less time & space requirements, based upon the criteria the extracted web document must be preprocessed, for preprocessing & information selection, need to apply some techniques such as stop word removal, Stemming of word, phrasing, normalization of tokens. Once the document is preprocessed, Normalization of tokens is generated to further process the web content document.

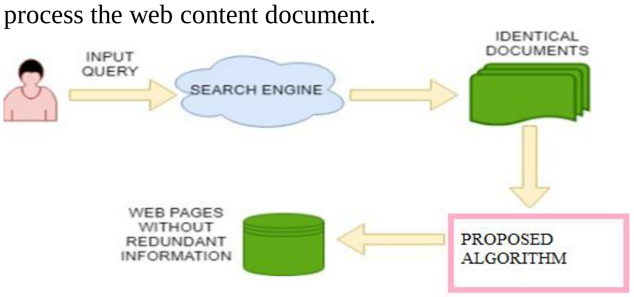

Figure 1(a). Structural Sketch Chart

The following algorithm shows the procedure to eliminate identical pages in the set-up of web pages. Initially calculate measure for all the vertices and maintain the set U which contains a minimum and maximum degree for all vertices and isolated measure vertex. Repeatedly include the minimum measure in the set-up and each measured vertex included only once in the set-up. After applying the above steps the entire vertex without redundant information available in the set U.

III. KTMAX-MAXAM ALGORITHM

```
1. Compute degree measure for all vertices in the set-up.
2. Pick all the maximum degree vertices ‘v’ in the set-up and include in the set U.
3. While U doesn’t include all vertices
    a. Include the entire isolated vertex which is adjacent to the vertices ‘v’ to U.
    b. Find the adjacent vertex ‘u’ to ‘v’ which is not in U and has next maximum degree.
    c. Update ‘u’ to U.
    d. Update the value of degree for all adjacent vertices of ‘u’. Iterate through all adjacent vertices if possible.
    e. Repeat step 3 till all nodes are included in the set U.
```

Finally network U contains no cyclic information.

Revised Manuscript Received on July 05, 2019.

K. Thiagarajan, Academic Research Professor, Anna University, Jeppiaar Engineering College, Chennai, India, vidhyamannan@yahoo.com..
A. Maheshwari, Department of Computer Science and Engineering, Jeppiaar Engineering College, Chennai. 78mahee@gmail.com
J. Kavitha, Department of Mathematics, Ph.D-CB-DEC2013-0334, Research & Development Centre, Bharathiar University, Coimbatore. India. manokavi.j@gmail.com.

Retrieval Number: B10610782S419/2019©BEIESP
DOI: 10.35940/ijrte.B1061.0782S419

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Figure 1(c). Pseudo code for the Proposed Algorithm (KTMAX-MAXAM)

Case I:
Regular set-up

Consider the following connected set-up G_1, having 12 nodes having 3 degree in all vertices along with redundant links.

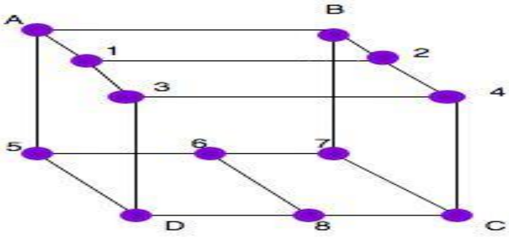

Figure 2a. **Regular set-up G_1 with 3 Degree**

Apply the proposed **KTMAX- MAXAM ALGORITHM** to G_1

By step 1: deg (All Nodes of G_1) = 3

By step 2: Mark the node A as visited and put it onto the set U.

By step 3: There is no isolated vertex in the given graph G_1

By step 4:

4.1 Explore any unvisited adjacent node from A. We have 3 nodes (B, 1 and 5) and we can pick the maximum degree node. Here all nodes have an equal degree (3). So pick any one of the node among the 3 adjacent nodes. Now the set U consists of the nodes A, B.

4.2 Explore the node B, the unvisited adjacent node is from B as 2 and 7. Now the set U consists of the nodes A, B, 2.

4.3 Now explore the node 2, the unvisited adjacent node is from 2 as 1 and 4. After inclusion of the node 1 the set U consists of the nodes A, B, 2, 1.

4.4 Explore the node 1, the unvisited adjacent node is from 1 as only 3. Now the set U consists of the nodes A, B, 2, 1, 3.

4.5 Explore the node 3, the unvisited adjacent node is from 3 as 4 and D. Now the set U consists of the nodes A, B, 2, 1, 3, 4.

4.6 Explore the node 4, the unvisited adjacent node is from 4 as only C. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C.

4.7 Explore the node C, the unvisited adjacent node is from C as 7 and 8. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C, 7.

4.8 Explore the node 7, the unvisited adjacent node is from 7 as only 6. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C, 7, 6.

4.9 Explore the node 6, the unvisited adjacent node is from 6 as 5 and 8. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C, 7, 6, 5.

4.10 Explore the node 5, the unvisited adjacent node is from 5 as only D. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C, 7, 6, 5, D.

4.11 Finally explore the node D, the unvisited adjacent node of D is only 8. Now the set U consists of the nodes A, B, 2, 1, 3, 4, C, 7, 6, 5, D, 8.

By step 5:
Finally network U contains without cyclic information.

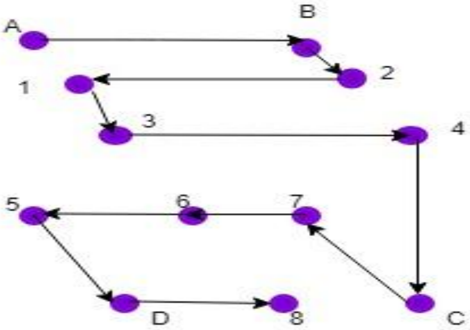

Figure 2b. Regular set-up G_1 with 3 Degree without Redundancy using KTMAX-MAXAM ALGORITHM.

Case II:
Irregular set-up
Consider the following irregular set-up G_2

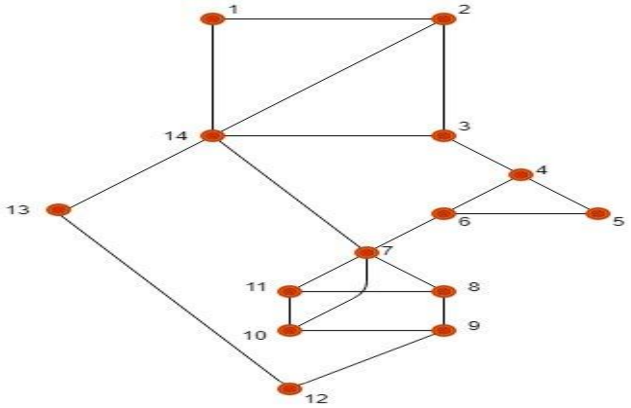

Figure 3a. **Irregular set-up G_2 containing vertices with degree 5, 3 and 2**

After applying proposed **KTMAX- MAXAM ALGORITHM** for G_2, we get the link as follows in Figure: 3b without repetitions on links.

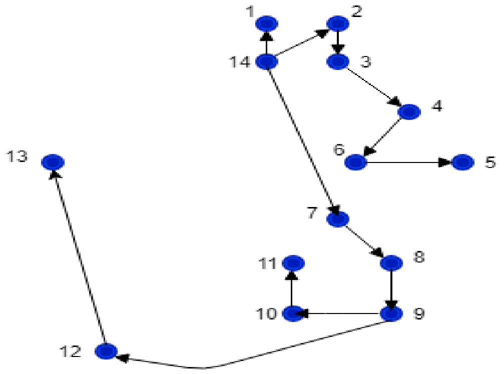

Figure 3c. Irregular set-up G_2 without redundancy using KTMAX-MAXAM ALGORITHM.

IV. CONCLUSION

In search engine generate relevant information but most of the time the information is not redundant. While getting more redundant web pages for a single search query, it’s more difficult to recognize the redundant links. We propose a mathematical based KTMAX-MAXAM algorithm for detecting redundant links and eliminating redundant links. Future work aims to create a finite automata tool to produce only relevant and without redundant information of web documents.

ACKNOWLEDGMENT

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, GuorongXu , A Web Page Segmentation Algorithm for Extracting Product Information , Information Acquisition, 2006 IEEE International Conference on Publication Date: Aug. 2006.
3. Bing Liu, Kevin Chen- Chuan Chang , Editorial: Special issue on Web Content Mining , SIGKDD Explorations, Volume 6, Issue 2.
4. JaroslavPokorny, JozefSmizansky, Page Content Rank: An approach to the Web Content Mining.
5. 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
6. 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.
7. Jiang Yiyong, Zhang Jifu,CaiJainghui, Zhang Sulan, Hu Lihua , The Outliers Mining Algorithm Based On Constrained Concept Lattice, Internal Symposium on Data Privacy and E.commerce , IEEE 2007.
8. Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, A Survey on Web Content Mining and Extraction of Structured and Semistructureddata,First International Conference on Emerging trends in Engineering and Technology, 2008.
9. J.P. Tremblay and R. Manohar, “Discrete Mathematical Structures with Applications to Computer Science”, TMH, 1997.