

Extermination of Obsolete Relationship Through KTMIN-JAK-MAXAM Algorithm in Confusion Mining

Kittappa Thiagarajan^{1,*}, Jeyavel Kavitha², Karunakaran Sarukesi³, Avudaiappan Maheshwari⁴

¹Academic Research and Development, Jeppiaar Engineering College, Chennai, India

²Department of Mathematics, R&D Centre, Bharathiar University, Coimbatore, India

³KCG College Engineering and Technology, Chennai, India

⁴Department of Computer Science and Engineering, Jeppiaar Engineering College, Chennai, India

Email address:

vidhyamannan@yahoo.com (K. Thiagarajan), manokavi.j@gmail.com (J. Kavitha), profsaru@gmail.com (K. Sarukesi), 78mahee@gmail.com (A. Maheshwari)

*Corresponding author

To cite this article:

Kittappa Thiagarajan, Jeyavel Kavitha, Karunakaran Sarukesi, Avudaiappan Maheshwari. Extermination of Obsolete Relationship Through KTMIN-JAK-MAXAM Algorithm in Confusion Mining. *American Journal of Data Mining and Knowledge Discovery*.

Vol. 4, No. 1, 2019, pp. 19-23. doi: 10.11648/j.ajdmkd.20190401.14

Received: February 25, 2019; **Accepted:** April 10, 2019; **Published:** May 11, 2019

Abstract: At the present time web contains many indistinguishable documents. Much effort made towards in investigates mechanism with identical detection algorithms, still the retrieved web documents with outmodedlink. In this proposed system, we are successfully identifying and minimize the redundant information and like link in web documents. We introduce the correct graph theory based KTMIN-JAK-MAXAM algorithm filters out the redundant link. From the proposed system, we have relevant information with more accuracy. Using this KTMIN-JAK-MAXAM algorithm accessing of web pages with reduced time and space complication.

Keywords: Data Mining, Degree, Maximum Degree, Minimum Degree, Link, Path

1. Introduction

The data and information available on web is exponentially improving, duplication of web content also increase simultaneously [1, 2]. Retrieving relevant information from web without redundancy is more challenge task nowadays where in web mining communities [3, 4]. Web content mining is the way toward extracting the applicable information, data and learning from World Wide Web. Utilizing customary data recovery [5] 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 [6], Web structure mining, Web usage mining. Traditional web mining algorithms handle with structured document [7-15] than the advanced methodology of mining algorithm can dealthe entire heterogeneous document comprises of images [9], graphs, videos [16], etc.

2. 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.

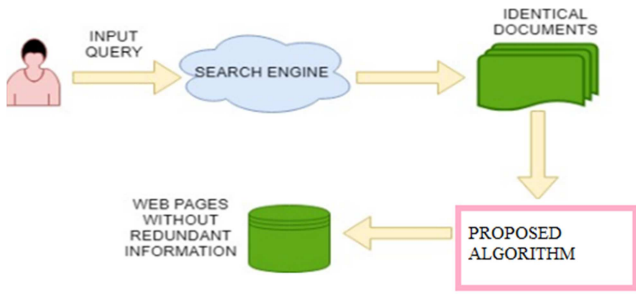


Figure 1. Structural Sketch Chart.

The proposed algorithm shows the procedure to eliminate indistinguishable 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.

Pseudo Code for The Proposed Algorithm KTMIN-JAK-MAXAM

Step1: Compute degree measure for all vertices in the set-up.

Step2: Pick the minimum degree vertex 'v' in the set-up and include in the set U.

Step3: While U doesn't include all vertices

Step3A: Include the entire isolated vertex which is adjacent to the vertex 'v' to U.

Step3B: Find the adjacent vertex 'u' to 'v' which is not in U and has maximum degree. Update 'u' to U.

Step3C: Update the value of degree for all adjacent vertices of 'u'. Iterate through all adjacent vertices if possible.

Repeat step 3 till all nodes are included in the set U.

Step4: Finally network U contains no cyclic information.

Case I: Regular set-up

Connected Regular Setup

Case I: A

Consider the following connected set-up G_1 in figure 2, having 12 nodes having 3 degree in all vertices along with redundant links.

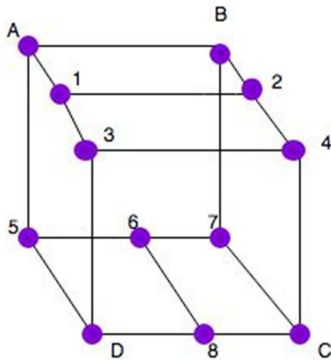


Figure 2. Regular Set-Up G_1 with 3 Degree.

Apply the proposed KTMIN-JAK-MAXAM ALGORITHM to G_1

By step 1: $\text{deg}(\text{All Nodes of } G_1) = 3$
 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 Investigate any unvisited adjacent node from A. We have 3 nodes (B, 1 and 5) and we can pick the minimum 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 Search 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 travel around 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 Discover 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 Survey 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 See the sights 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 Look at 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 survey 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 Look at 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.

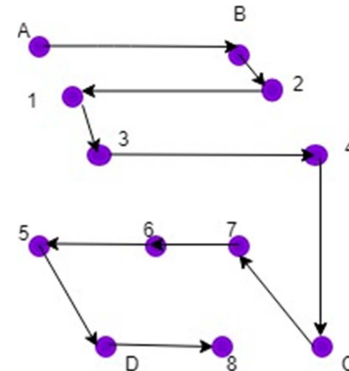


Figure 3. Regular Set-Up G_1 With 3 Degree Without Redundancy Using KTMIN-JAK-MAXAM ALGORITHM.

Case I: B.

Consider the following connected set-up G_2 in figure 4, having 6 nodes having 3 degree in all vertices along with redundant links.

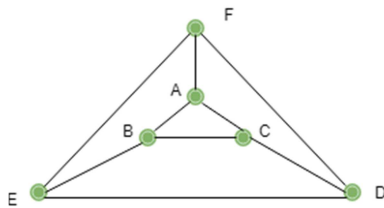


Figure 4. Regular Set-Up G_2 With 3 Degree Without Redundancy Using KTMAX-MINAM ALGORITHM.

Apply the proposed KTMIN-JAK-MAXAM ALGORITHM to G_2

By step 1: $\text{deg}(\text{All Nodes of } G_2) = 3$

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

By step 3: There is no isolated vertex in the given graph G_2

By step 4:

4.1: Investigate any unvisited adjacent node from F. We have 3 nodes (A, E and D) and we can pick the minimum 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 and F.

4.2: Discover the node A, the unvisited adjacent node is from A as B and C. Now the set U consists of the nodes F, A, and C.

4.3: Now travel around the node C, the unvisited adjacent node is from C as B and D. After inclusion of the node B the set U consists of the nodes F, A, C, and B.

4.4: Look at the node B, the unvisited adjacent node is from B as only E. Now the set U consists of the nodes F, A, C, B, and E.

4.5: Search the node E, the unvisited adjacent node is from E as only D. Now the set U consists of the nodes F, A, C, B, E, and D.

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

After Applying The Proposed KTMIN-JAK-MAXAM ALGORITHM To G_2 , we get

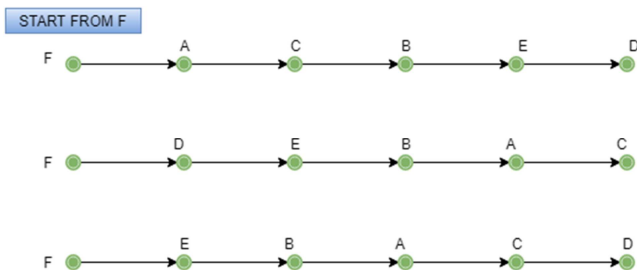


Figure 5. The Solution Is Unique In Symmetrical Structure With Out Any Repetitions.

Note: With reference from the figure 5, it is concluded that the solution is unique.

Case I: C

Consider the following connected set-up G_3 in figure 6,

having $n = 10$ nodes having 3 degree in all vertices along with redundant links.

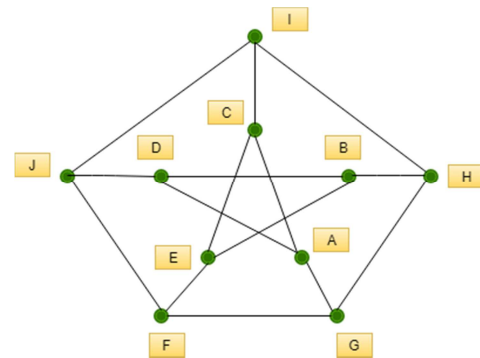


Figure 6. Connected Set-Up G_3 Having 10 Nodes All Having 3 Degree.

After applying the proposed KTMIN-JAK-MAXAM ALGORITHM to the above graph G_3

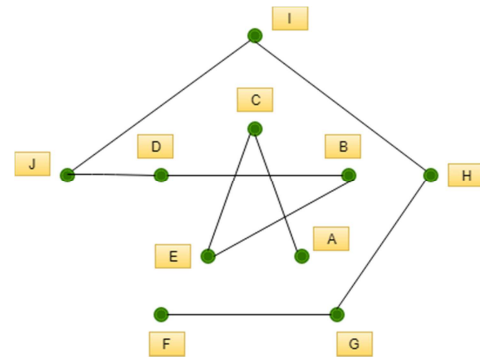


Figure 7. Regular Set-Up G_3 With 3 Degree Without Redundancy Using Ktmin-Jak-Maxam Algorithm.

Here notice that, Regular connected network G_3 , after applying the proposed KTMIN-JAK-MAXAM ALGORITHM to G_3 , we get path of length 9 in figure 7.

Case II:

Connected Complete Network G_4 in figure 8

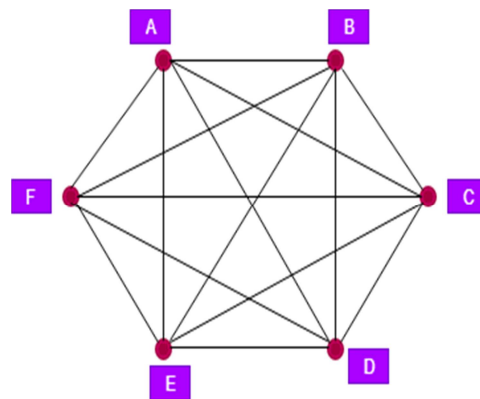


Figure 8. Connected Set-Up G_4 Having 6 Nodes All Having 4 Degree.

After applying the proposed KTMIN-JAK-MAXAM ALGORITHM to the above graph G_4 represented in figure 9.

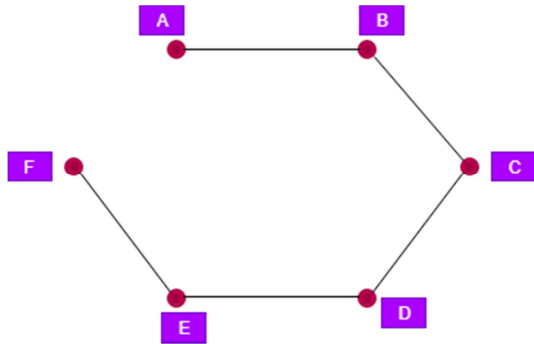


Figure 9. Regular Set-Up G_4 With 5 Degree Without Redundancy Using KTMIN-JAK-MAXAM ALGORITHM.

Here notice that, all complete connected networks G , after applying the proposed KTMIN-JAK-MAXAM ALGORITHM to G , we get linear path of length $(n-1)$.

Case III: Connected Irregular network
Case III: A

Consider the following irregular set-up G_5 in figure 10

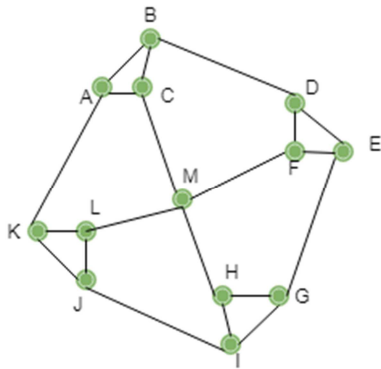


Figure 10. Irregular Set-Up G_5 Containing Vertices With Degree 5, 3 And 2.

After applying proposed KTMIN-JAK-MAXAM ALGORITHM for G_4 , we get the link as follows in figure 11 without repetitions on links.

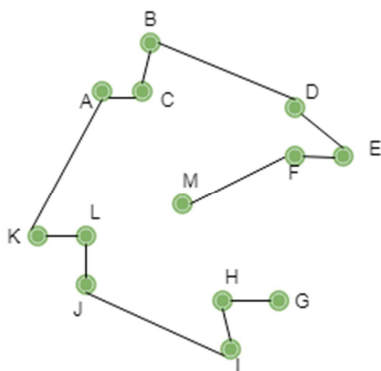


Figure 11. Irregular Set-Up G_5 Without Redundancy Using KTMIN-JAK-MAXAM ALGORITHM.

Case III: B

Consider the following Connected Irregular set-up G_6 in figure 12.

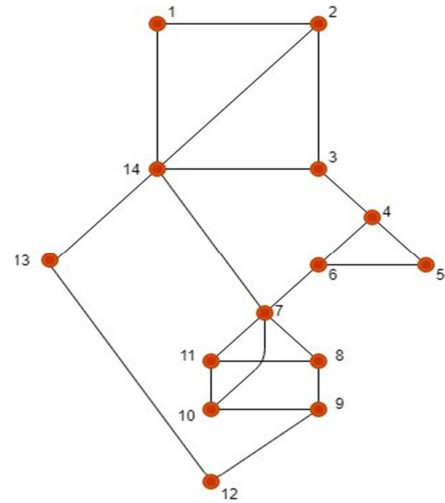


Figure 12. Irregular Set-Up G_6 containing Vertices With Degree 5, 3 And 2.

After applying proposed KTMIN-JAK-MAXAM ALGORITHM for G_6 , we get the link as follows in figure 13 without repetitions on links.

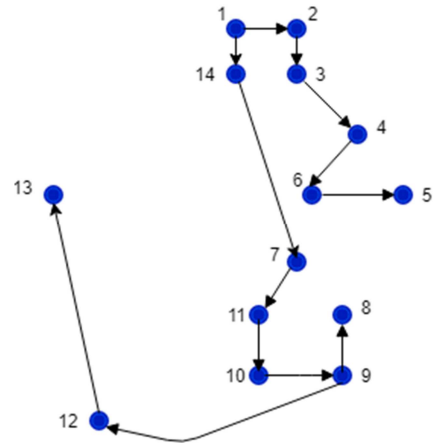


Figure 13. Path With 13 Without Repetitions On Links.

Case III: C

Consider the following Connected Irregular set-up G_7 in figure 14

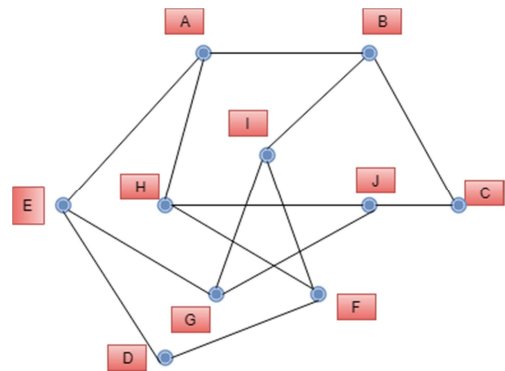


Figure 14. Irregular Set-Up G_7 Containing Vertices With Degree 3 And 2.

After applying the proposed KTMIN-JAK-MAXAM ALGORITHM to the above graph G_7 in figure 15.

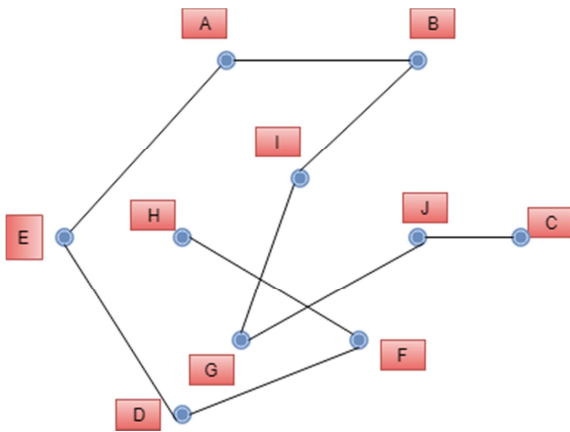


Figure 15. Irregular Set-Up G7 Without Redundancy Using KTMIN-JAK-MAXAM ALGORITHM.

4. 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 graph theoretical based algorithm for detecting and eliminating redundant links. Also we observed derived linked graph need not be unique but this approach will provide the optimized cost analysis report in future in data science field. Future work aims to create a finite automata tool to produce only relevant and without redundant information of web documents in data mining.

Acknowledgements

The Authors would like to thank Prof. Ponnammal Natarajan worked as Director-Research, Anna University-Chennai, 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, GuosunZeng, 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. Alhadjj, 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, ShreyaPatankar, 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.
- [10] D. C. Sancheti and E. K. Kapoor, Statistics (Theory, Methods & Application) By Published by Sultan Chand and Sons, Sixth thoroughly revised Edition, 1990.
- [11] N. P. Gopalan, J. Akilandeswari, Distributed, Fault-tolerant Multi-agent Web Mining System for Scalable Web Search 5th WSEAS Int. Conf. on APPLIED INFORMATICS and COMMUNICATIONS, Malta, September 15-17, 2005 (pp384-390).
- [12] G. Castellano, A. M. Fanelli, M. A. Torsello, Mining usage profiles from access data using fuzzy clustering, 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, September 22-24, 2006.
- [13] AnoopPaharia, YachanaBhawsar, YachanaBhawsar, Developing Web intelligence using data mining 6th WSEAS Int. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Tenerife, Spain, December 14-16, 2007.
- [14] Web Document Classification and its Performance Evaluation: IOAN POP, 9th WSEAS International Conference on EVOLUTIONARY COMPUTING (EC'08), Sofia, Bulgaria, May 2-4, 2008.
- [15] IoanDzitac, IoanaMoisil, Advanced AI Techniques for Web Mining mathematical Methods, Computational Techniques, Non-Linear Systems, Intelligent Systems.
- [16] ZakariaSulimanZubi, Using Some Web Content Mining Techniques for Arabic Text Classification, Recent Advances On Data Networks, Communications, Computers.