into four major groups; namely manufacturing, finance, business and education. The J-Series contained 185 pages and had 10 classifications while the K-Series contained 400 web pages in 20 categories. 800 pages were randomly selected from the original 2100 pages and the number of vectors was set to 40, which doubled the number of categories in the K-Series experiment. The choice of this number of clusters is premised on the fact that it is the most natural number based on the initial tests and observations. The number of maximum nodes per graph was set to be higher to provide improved baseline for the results as shown in Table 2.

Table 2: Performance of Graphs with increasing nodes

Max. Nodes/Graph	A^M (average)
150	0.2218
120	0.2142
90	0.2074
75	0.2045
60	0.1865
45	0.1758
30	0.1617
15	0.1540
5	0.1326

Each row in Table 2 provides results for 10 experiments using the same data sample of 800. The variation in the results is due to randomization in the first stage of the algorithm. Previously obtained data were represented in the graph for better visualization and with a 2.2 GHz processor, it took 7 minutes to represent five nodes per graph. Euclidian distance, δ for point (x,y) was also determined with a view to measuring the vector distance metrics as follows:

$$\delta(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - x) + (y_i - y)^2} \quad (18)$$

x_i and y_i are the i^{th} components of the x and y vectors respectively. The cosine equivalent, β of the distance is obtained as follows:

$$\beta(x,y) = 1 - \frac{x * y}{||x||.||y||} \quad (19)$$

$*$ is the dot operator, and $||$ shows the magnitude of the vector being considered. Comparison of the results obtained from the graphs with those from other techniques is presented in Table 3.

Table 3: Comparison of Graph theoretic approach with other techniques at an instance

Method	A^M (average)
Graphs (current study)	0.222
Extended Jaccard Similarity [34]	0.184
Pearson Correlation [34]	0.178
Cosine Measure [31]	0.178
Random [31]	0.066
Euclidean [31]	0.046

Since larger graphs hold more data, the mutual information is seen to increase as the graph increases in size. The random baseline was used to provide a basis for comparison in the experiment. The Jaccard means was based on the Jaccard similarity and the cosine and Pearson measures were omitted for improved clarity. The graphical representation of experimental values for Graphs, Random, Euclidian and Jaccard methods with the same experimental conditions is shown in Figure 5. It is revealed vividly that the graph theoretic and genetic algorithm-based technique outperforms other techniques especially with increased graph nodes. In other words, as the complexity of the web contents increases, other reviewed techniques could not match up with the proposed technique in the area of mutual information index.

5. Conclusion

With standard tools for web content mining, there is opportunity for extracting only the relevant text from web while unrelated textual noise like advertisements, navigational elements, contact and copyright notes are reliably suppressed. The reported research hybridized graph theoretic and genetic algorithm to formulate a web content mining technique for achieving this purpose. The new technique provides timely search and discovery from large web datasets and experimental results had shown its superiority over other techniques. These suggest the new technique will be very useful in areas where knowledge discovery, web structure and web analytics are required. It is of note that the applicability of the new technique on complex and large number of parameters has not been investigated.

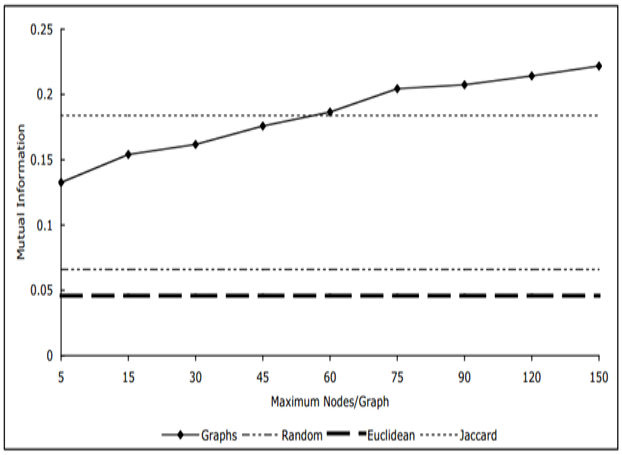

Figure 4.3: Comparison of graph with other techniques

REFERENCES

[1] Liu B., Structured data extraction: Wrapper generation and Web Data Mining, Editorial Issues on Web content Mining. SIGKDD Explorations –Vol. 6, 2005, pp. 363 - 423

[2] Abdelhakim H., Khentout C. and Djoudi M., Overview of Web Content Mining Tools, *The International Journal of Engineering and Science (IJES), Vol. 2, 2013.*

[3] Marghny M. H. and Ali A. F., Web mining based on genetic algorithm, Proceedings of AIML '05 Conference, Cairo, Egypt, 2005, pp 19-21

[4] Ammar S. A., Enhancing recall and precision of web-Search using genetic algorithm, A thesis submitted for the degree of Doctor of Philosophy, School of Information Systems Computing and Mathematics, Brunel University, UK, 2012

[5] Zaiane R. O., Introduction to Data Mining: Principles of knowledge discovery in databases, 1999.

[6] Han J. and Kamber M., Data mining: concepts and techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000

[7] Imielinski T. and Mannila H., A database perspective on knowledge discovery, Communications of ACM, Vol. 39, pp. 58-64.
[8] Rosenfeld L. and Morville P., Information architecture for the World Wide Web, 1st edition, CA, 1998
[9] Callan J., System and method for filtering a document strem", US Press, Patient 6.105.023, 2000
[10] Chen M. S., Han J. and Yu P. S., Data mining: An overview from a database perspective, IEEE Trans. Knowledge and Data Engineering, Vol. 8, 1996, pp 866-883
[11] New York Stock Exchange, 2000. Available at http://www.ecgi.org/codes/documents/nyse_cgreport_23sep 2010 en.pdf. Accessed October 12, 2013.
[12] Piatetsky-Shapiro G., Fayyad U. M. and Smyth P., From data mining to knowledge discovery, An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, pp. 1-35
[13] Sivaramakrishnan J. and Balakrishnan V., Web Mining Functions in an Academic Search Application, *Informatica Economica, Vol.* 13, No. 3, 2009.
[14] Poonkuzhali G., Sarukesi K. and Uma G. V., Web Content Outlier Mining Through Mathematical Approach and Trust Rating, Recent Researches in Applied Computer and Applied Computational Science, 2012.
[15] Silltow J., Data Mining 101: Tools and Techniques", paper presented at The Institute of Internal Auditors (IIA), 247 Maitland Avenue, Altamonte Springs, Florida U.S.A., 2006
[16] Colet E., Clustering and Classification: Data Mining Approaches". Virtual Gold Incorporated, 2002
[17] Galeas P., Web Mining, 2005. Available at: http://www.galeas.de/webmining.html, Accessed January, 2016.
[18] Cooley R., Mobasher B. and Srivastava J., Web mining: information and pattern discovery on the World Wide Web. *Proceedings of 9th IEEE International Conference,* pp. 558 – 567, 1997
[19] Arvind K. S. and Gupta P. C., Exploration of efficient methodologies for the improvement in web mining techniques - A survey, *International Journal of Research in IT & Management,* Vol. 1, No. 3, 2011
[20] Arvind K. S. and Gupta P. C., Study and Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining, *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, Vol. 1 ,No. 8, 2012
[21] Baumgartner R., Gatterbauer W. and Gottlob G., Web data extraction system: Encyclopedia of Database Systems, 2009, pp 3465-3471.
[22] Kosala R. and Blockeel H., Web mining research: A survey," SIGKDD Explorations: Newsletter of the Special Interest Group on Knowledge Discovery and Data (SIGKDD) Mining, ACM, Vol. 2, 2000
[23] Gore M. M. and Mishra A. K., Algorithm for Data Mining. *Proceedings of Winter School on Data Mining*, Allahabad, India, 2001
[24] Ferrara E., De-Meob P., Fiumarac, G. and Baumgartnerd, R., Web Data Extraction, Applications and Techniques: A Survey, 2014. Available at http://www.sciencedirect.com/science/article/pii/S09507051 14002640, Accessed on September, 2015.
[25] Irmak U. and Suel T., Interactive wrapper generation with minimal user effort". *Proceeding of 15th International Conference on World Wide Web*, Edinburgh, Scotland, 2006, pp 553-563
[26] Wang P., Hawk W. and Tenopir C., Users' interaction with World Wide Web resources: an exploratory study using a holistic approach, Information Processing Management, 2000, pp 229 - 251,
[27] Furche T., Gottlob G., Grasso G., Gunes O., Guo X., Kravchenko A., Orsi G., Schallhart C., Sellers A. J. and Wang C., Domain-centric, intelligent, automated data extraction methodology, Companion, Vol. 10, 2012, pp 267-270.
[30] Chen H., Chau M. and Zeng D., Spider: a tool for competitive intelligence on the web, Decision Support System, Vol. 17, No. 34, 2002
[31] Marcov A. Last M. and Kandel A., Model-Based Classification of Web Documents Represented by Graphs, Proceedings of WEBKDD'06, Philadelphia, Pennsylvania, USA, 2006
[32] Cormen T. H., Leiserson C. E., Rivest R. L. and Stein C., The algorithms of Kruskal and Prim: *Introduction to Algorithms"*, 3rd edition. MIT Press, Vol. 23, No. 2, 2009, pp. 631-638.
[33] Artymiuk P. J., Spriggs R. V. and Willett P., Graph theoretic methods for the analysis of structural relationships in biological macromolecules, *Journal of the American Society for Information Science and Technology,* Volume 56, 2005, pp 518 – 528
[34] Schenker A., Last M., Bunke H., and Kandel A., Graph Theoretic Techniques for Web Content Mining, PhD Thesis, College of Engineering, University of South Florida, 2003
[35] Shoreh A. and Mohammad D. J., Deep Web Content Mining, *The Journal of World Academy of Science, Engineering and Technology, 2009,* pp. 49.
[36] Sighn A., Agent Based Framework for Semantic Web Content Mining, *International Journal of Advancements in Technology* (*IJoAT)*, Vol. 3, No. 2, 2012. Available at http://ijict.org/. Accessed September, 2014.
[37] Kaushik M. and Phukon S., A composite Graph Model for Web Document and the MCS Technique, *International Journal of Information Technology and Knowledge Management,* Vol. 4, No.1, 2012, pp. 211-215
[38] Zhenyu W. and Richard L., An Optimal Graph Theoretic Approach to data Clustering, theory and its application to Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 15, No. 11, 1993.
[39] Andrey A. M., A Generalized Graph-Theoretic Mesh Optimization Model, *Proceedings of the 26th International Meshing Roundtable, South Lake Tahoe*, 2005
[40] Wallis W. D., Shoubridge P., Kraetzl M. and Ray D., Graph distances using graph union. *Pattern Recognition Letters*, Vol. 22, No. 6, 2001, pp 701-704.
[41] Bunke H., Recent Development in graph matching. *Proceedings of 15th International Conference on Pattern Recognition,* Vol. 2, 2000, pp 117-124.
[42] Mirtha-Lina F. and Gabriel V., A graph distance metric combining maximum common subg--raph and minimum common super-graph, Pattern Recognition Letters, Vol. 22, 2001, pp 753-758.
[43] Sandhya Chaturvedi M. and Shrotriya A., Graph Theoretic Techniques for Web Content Mining. *The International Journal of Engineering and Science*, Vol. 2, No. 7, 2013, pp. 35-41.
[44] Vikrant S. and Thakur R. S., GA Based Model for Web Content Mining", *International Journal of Computer Science Issues (IJCSI)*, Vol. 10, No. 2, No 3, 2013
[45] Ammar S. A. and Shaker R., Genetic Algorithm Mining for HTML Documents, School of Information Systems Computing and Mathematics (SISCM), Brunel University, UK, 2009