Mining Research Abstracts for Exploration of Research Communities

Mahalakshmi G.S.
Dept. of Computer Science and Engineering, Anna University, Chennai,
India.
mahalakshmi@cs.annauniv.edu

Dilip Sam S.
Dept. of Computer Science and Engineering, Anna University, Chennai,
India.
sdilipsam@gmail.com

Sendhilkumar S.
Dept. of Information Science and Technology, Anna University, Chennai,
India.
ssk_pdy@yahoo.co.in

ABSTRACT

Research abstracts are the ‘information scents’ which attract a novice researcher. To read through the entire research paper and to decide the suitability of the paper to one’s research problem is a tough and abstract task. Many times, researchers do not know whether they are citing the relevant (but original?) research articles. It has been only a trial and error approach so far. To enable researchers to correctly target at the relevant and yet quality research literature, mechanisms to organise collections of research papers are essential. Though a considerable effort has been attempted earlier in this context, establishing research communities concentrated on citation based recommendations only. However, the quality and originality of research articles have not been taken into account until now. In this paper, we propose the evolution of research communities by analysing the research abstracts. We utilise Fuzzy Concept Map based approach in detecting the originality of scientific abstracts. By K-means clustering, we establish a research article hyper graph from the qualified abstracts. Later, we evolve the author clusters for every topic cluster and analyse them for redundancy. Further study on relevant bibliometrics helps us to identify a ‘nucleus author’ for every topic cluster.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retreival—*Information Filtering*; H.3.7 [**Information Storage & Retrieval**]: Digital Libraries—*User Issues.*

General Terms

Experimentation, Human Factors.

Keywords

Fuzzy Cognitive Maps, Originality of Research Publications.

1. INTRODUCTION

The diffusion of knowledge from an individual to others in a social network environment demands higher significance in the research field. This distributed knowledge need to be acquired as a collective object for a researcher in order to gain the higher impact of a concept. Here comes the most important characteristic of knowledge network i.e. the linkage between the knowledge that the individuals possess. Research communities are special applications of social network environment where the purpose of communication is to share knowledge. A knowledge network is an exceptional one within social network that the actors in the knowledge networks are related to academic research locale. Knowledge network varies from social network where actors sounding better in their research oriented aspect are inter-related to form a specific group. The categorization of ‘who knows who knows what’ is likely to be achieved via knowledge networks [51]. The ability to gain knowledge from others is not simple as this knowledge does not exist readily in a single place. Social networking elements like blogs and discussion forums are not valued much due to their improper standards. The academic related information/knowledge lags in these elements. Though semantic social network exhibit relationship among people, it does not provide certain features like author bonding, knowledge transition of authors, domain specific author quality, levels of author contribution, author centrality, etc. Hence researchers obviously trust standard bibliographic repositories like DBLP, Cite seer for better knowledge acquisition. This implicitly demands for quality of researchers present in those repositories. The feasibility for the transition of social network to knowledge network is not simple because knowledge network simply possess academic oriented aspects whereas social network elements are not restricted to any particular aspect.

Hence evolution of knowledge network becomes the need of the hour for researchers to pursue their work in an intelligent way. The actors join or leave a knowledge network on the basis of tasks to be accomplished, and their levels of interests, resources, and commitments. The links within the knowledge network are also likely to change on the basis of evolving tasks, the distribution of contributions that the author made, or changes in the actors' cognitive knowledge network. The various levels of contribution of actors can be easily identified via knowledge networks confined to a particular domain. The knowledge of a researcher shall be weighted according to their previous research contributions. In this context, the research abstracts of a researcher shall be analysed for determining one’s capacity as a researcher.

In this paper, we propose the evolution of knowledge network from digital bibliographic repositories such as DBLP. In order to achieve this technological infrastructure, each individual (author)

Compute2012, January 23--24, 2012, Pune, Maharashtra, India.
Copyright 2012 ACM, ISBN 978-1-4503-1440-4

is categorized towards a specific domain and their research abstracts are analyzed to grade their level of quality. The fluidity of knowledge among the actors is analyzed based on their contributions in every domain. The bonding between the researchers in the knowledge network is also identified based on the 'betweenness' among them as authors in their respective publications. Finally, a research hyper graph is arrived thereby bringing in the concept of knowledge network.

2. RELATED WORK

Extracting information from bibtex data for potential research use has been the focus of data mining and information retrieval research. Evolving research communities which are made up of authors, representing different research groups, that are linked with different type of relations has the concept of social network in it. Hence, viewing and understanding the research relationship between the nodes of the network is an essential part of social network analysis.

Various Social Network Analysis (SNA) methods [58] exist to analyse citation based research networks. Community Mining [40, 46] has received considerable attention over recent years. Identifying the connections existing between the nodes of the communities with nodes sharing similar properties with each other is very interesting yet challenging task [53]. Our idea is to find potential collaborating researchers by discovering communities in an author-centric research network. However, we tend to differ from the formation of research communities of Osmar et.al. [53].

In community mining, the closeness of related concepts is usually measured by 'relevance score'. For this measure, relationships between the entities need to be identified. With the possibility of multiple, multi-level and multi-variant relationships between the nodes of research communities, quantifying a relevance score would be more approximate or would be done under varying assumptions. Euclidean distance or Pearson correlation [58] could be used for such purpose. Since social networks could be modeled as graphs, usage of traditional graph algorithms such as spectral bisection method [2] which is based on eigen vectors, or Kernighan-Lin algorithm [8] which greedily optimizes the number of internal and interface level community edges suffer from graph bisection problems. The decision on when to stop the graph bisection is of prime importance. Hierarchical clustering [28] could be a better bet, however, if nodes of the communities are not close to one another, then forming the clusters would be a major problem.

Random walk approach [37, 43] is widely used to determine the relevance score between the entities of a community network. Another variation of Random walk approach, called Random walk with restart (RWR) [53] is used by considering the traditional random walk with a restart probability. Using this iterative random walk algorithm, the relevance score is computed for recommendation of potential research collaborators. In addition, analysing the co-author relation might reveal interesting results [1, 42]. However, a community discovery to recommend potential collaborating researchers should not end up with directing only a colleague or fellow researcher as a collaborator. Co-authorship information is something which is directly available with the bibliographic data. Rather deriving other implicit information about the researchers would be more difficult because of the volume of bibliographic data.

DBLife [19], DBConnect [18] and Libra [38] are some projects experimented over DBLP for evolving heterogeneous information networks. They provide related researchers and related topics to a given researcher. Rapid understanding of scientific publications has been made possible with the advancement of text mining and NLP research. The Action Science Explorer [21] is a similar tool designed to support exploration of a collection of papers so as to rapidly provide a summary, while identifying key papers, topics, and research groups. Existing systems provide some of these features in various combinations, though none allow users to leverage all of them in a single analysis. For their initial exploration, users frequently use academic search tools like Google Scholar [25] and Microsoft Academic Search [47]. Subscriber-only general databases are used frequently at universities and research labs, such as ISI Web of Knowledge [64] and SciVerse Scopus [22]. Additionally, many field specific databases exist such as PubMed [48] for Life and Biological Sciences. Computer and Information Sciences have databases like the web harvesting CiteSeer [10, 24], arXiv [17] for preprints, and the publisher-run ACM Digital Library [7] and IEEE Xplore [30]. These search tools and databases generally provide a sortable, filterable list of papers matching a user-specified query, sometimes augmented by faceted browsing capabilities and general overview statistics. An emerging category of products called reference managers enhances the paper management capabilities by supporting additional search, grouping, and annotation features, as well as basic collection statistics or overview visualizations. Some examples are JabRef [31], Zotero [13], EndNote [63], and Mendeley [45].

Academic research tools apply bibliometrics to help users understand collections through network visualizations of paper citations, author collaborations, author or paper co-citations, and user access patterns [49]. Many standard bibliometric analysis and visualization approaches are integrated in Network Workbench [52]. Another tool designed for analyzing evolving fields is CiteSpace [14, 15, 16], which is targeted at identifying clusters and intellectual turning points. Similarly, semantic substrates can be used for citation network visualization [6], showing scatter plot layouts of nodes to see influence between research fronts. Unfortunately these visualizations are weakly integrated into the rest of the exploration process and are yet to be widely used. However, the quality of research abstracts and the potential relevance of research abstracts to the corresponding titles and /or the rest of the paper is of a major concern. Therefore, the accuracy of relevance score is diminishing without the 'quality' perspective.

With the only available short-text, the title of the paper, it is hard to extract the correct research topic of the paper. In addition, the practice of researchers naming their paper with metaphoric / unrelated words / acronyms / question phrases will worsen the level of accuracy. The possible solution recommended by Osmar et al. [53] is to implement a hierarchy of topical words. In this paper, we have the objective of performing 'quality based' community discovery to recommend 'authentic' and not popular potential collaborating researchers. The idea is to bring the knowledge of a researcher as a prime component in the information network. Therefore, we tend to name it as 'knowledge networks'. This knowledge is hidden in massive links of the research network [32] and for the same reason link mining of research network only would lead to identifying knowledge out of the research community.

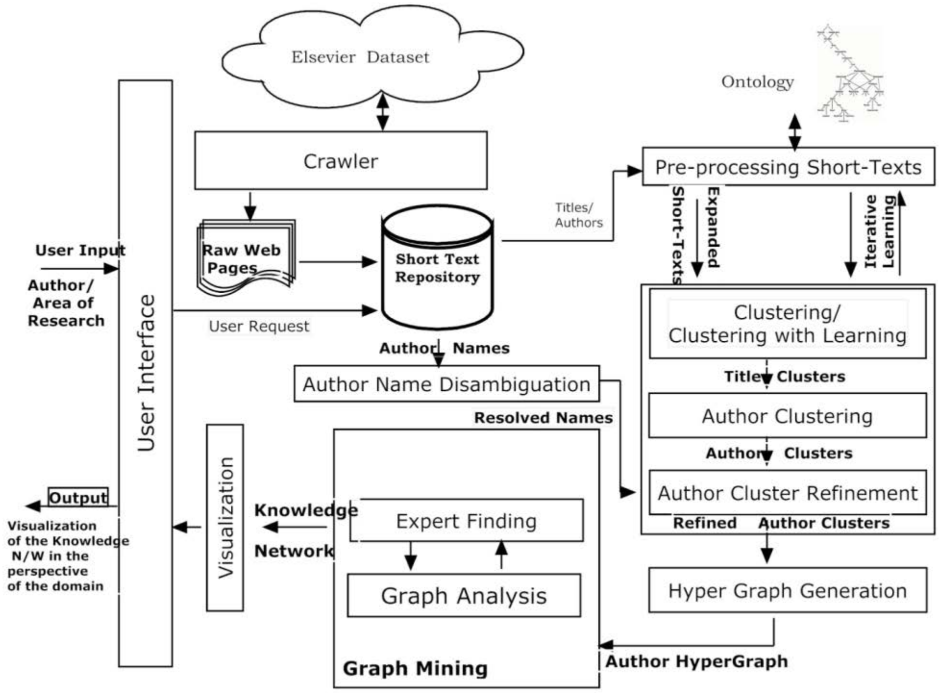

Figure 1 Evolution of Knowledge Network.

Since the research community has to evolve from time to time, the position of researcher should also be looked upon with respect to the research period, and the domain of research. To assess the knowledge of the researcher, content analysis of one’s research articles is essential. To lead a novice researcher into the nucleus of knowledge networks, assessing the research impact and productivity on quality (and not of popularity) perspectives becomes mandatory. Therefore, we attempt at evolving a research network where every researcher finds some place with defined values assigned for their research contributions. We believe that instead of following the substitutive measure like ‘Journal Impact Factor’ [62], an integrative system of research evaluation would be more appealing to the research community.

CiteSense [9] is such a system which helps researchers to quickly locate and analyse the relevance of citations within the research publication. The system includes a crawler, research paper database, an NLP module, a knowledge network, a sentiment module, and an author social network. However, the author social network is constructed based on the relationship of co-authors and the quality of research contribution is not a major concern. In this paper, we suggest an integrated approach for evolution of research communities which is discussed in the following section.

3. EVOLUTION OF RESEARCH COMMUNITIES

This system aims in providing the user with a comprehensible network, conceived by the knowledge extracted from the content available in the digital bibliographic repositories. This knowledge network will supply the user with inherent relationships among authors, research based upon the area of interest. In this system, a subject-specific search in the repository would return the relationship between authors and between papers based on the relevance of the release to the subject of interest. This will include the extraction of short-texts available in the digital bibliographic sites to find key- phrases and arrive upon the degree of relation between the paper and the subject of interest to the user, thereby building a hyper graph. This system can be enhanced by content retrieval of papers from other related websites, analysis of abstracts. Upon applying reasoning to the hyper graph, the relationship between authors, research limited to the subject of interest is provided to the user.

3.1 Crawler

The crawler uses a modified KPS algorithm [26]. Pages from digital bibliographic sites in the Internet are crawled for publications. From the extracted pages, information available about the journal, authors and the title / abstract of the article are extracted. The crawler is implemented as an independent agent to perform incremental crawling [57]. Incremental crawling would ensure the freshness of the information.

3.2 Author Name Disambiguation

Authors’ name disambiguation grows complex with the amount of information available to the system. The problem compounds many factors including the same author using different names, typographic errors, authors sharing the same name etc. This system provides an adapted K-way spectral clustering [27] method by indexing authors and deals with ambiguity by heuristics which include the order in which the first, middle and last names appear in the journal. The ambiguities will be resolved by co-authorship and the chronological information which would be available in the bibliographic sites.

3.3 Clustering by Conceptual Similarity

In text mining, ontological approach is used as a specific characteristic to detect document similarity. Ontology can serve as repository of all concepts in a domain. This knowledge pertaining to a domain can be harnessed in detecting similarity between documents belonging to the same domain. In the context of applying ontology to detect similarity, we chose to apply Fuzzy cognitive maps. Fuzzy Cognitive Map (FCM) is a mental map or mind map which is often used in representing the relationships between concepts. Fuzzy Cognitve Map (FCM) is a directed graph with concepts as nodes and causality as edges [54]. Fuzzy Cognitive Maps are evolved from offline ontology. The development of FCM typically includes two steps: The identification of concepts, which is followed by the identification of causal relationships among these concepts [41]. Therefore, as an initial step, we have constructed an offline ontology based on valid concepts from ACM classification system [3]. More than 1500 concepts and the respective concept details are present in the ontology specifically in computer networks and related domain. The offline ontology is built using Protégé editor [55]. The edge weights are assigned among concepts based on combinations of concepts in matrix format [36] in case of FCM.

The underlying text is pre-processed for removal of stop words. Later, quality terms are selected for further processing. The term quality is measured from the traditional metric introduced by Salton and McGill [60]. The terms with high quality are clustered together by using the k-means clustering [4, 35]. The k value is chosen at random to conclude the number of clusters. The individual terms obtained from the clusters are given as input to the offline generated ontology and the neighborhood concepts are extracted from the constructed ontology. The extracted concepts thus form the Ontology Set (Ontoset) [59]. The Onto sets of text documents thus obtained are merged to form a matrix with concepts of Onto set1 in rows and those in Onto set2 in columns. The relationship between the concepts are measured as the edge value where concepts as nodes. The metric used by Makoto et al [59] is adapted to compute Ontosets similarity. The similarity is computed for all the concepts of Ontoset1 and 2, thereby contributing towards determination of the document's similarity. The abstracts are grouped into clusters based on the conceptual similarity index and a hyper-graph is generated with details of authors for every cluster. The number of clusters is determined dynamically so that the documents within the clusters have maximum silhouette.

3.4 Graph Mining

The hyper-Graph generated will have clusters of authors based on the domain. Mining the graph would provide the user with a comprehensible network. Partitioning the hyper-graph is done using the spectral hyper-graph partitioning algorithm [65]. Further analysis of hyper-graph will lead to finding the nucleus of the domain of research. This would enable the system to recommend the highest contributor of the underlying research domain.

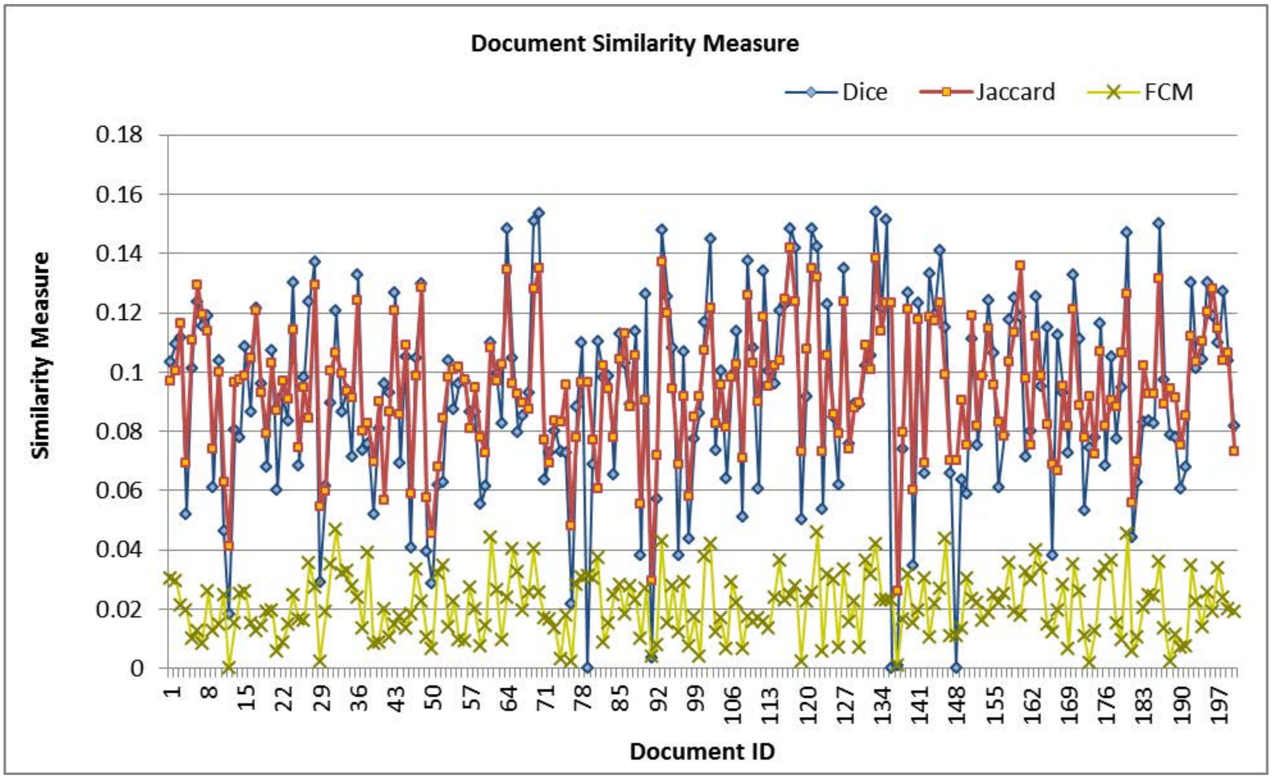

Figure 2a Comparison of statistical and ontological approach in detecting abstract based document similarity for 200 scientific abstracts of Volume 53 Elsevier 'Computer Networks' across abstracts corpus of size = 5000. Fuzzy Cognitive Maps outperforms by reducing the extreme fluctuations in the conducted experiments.

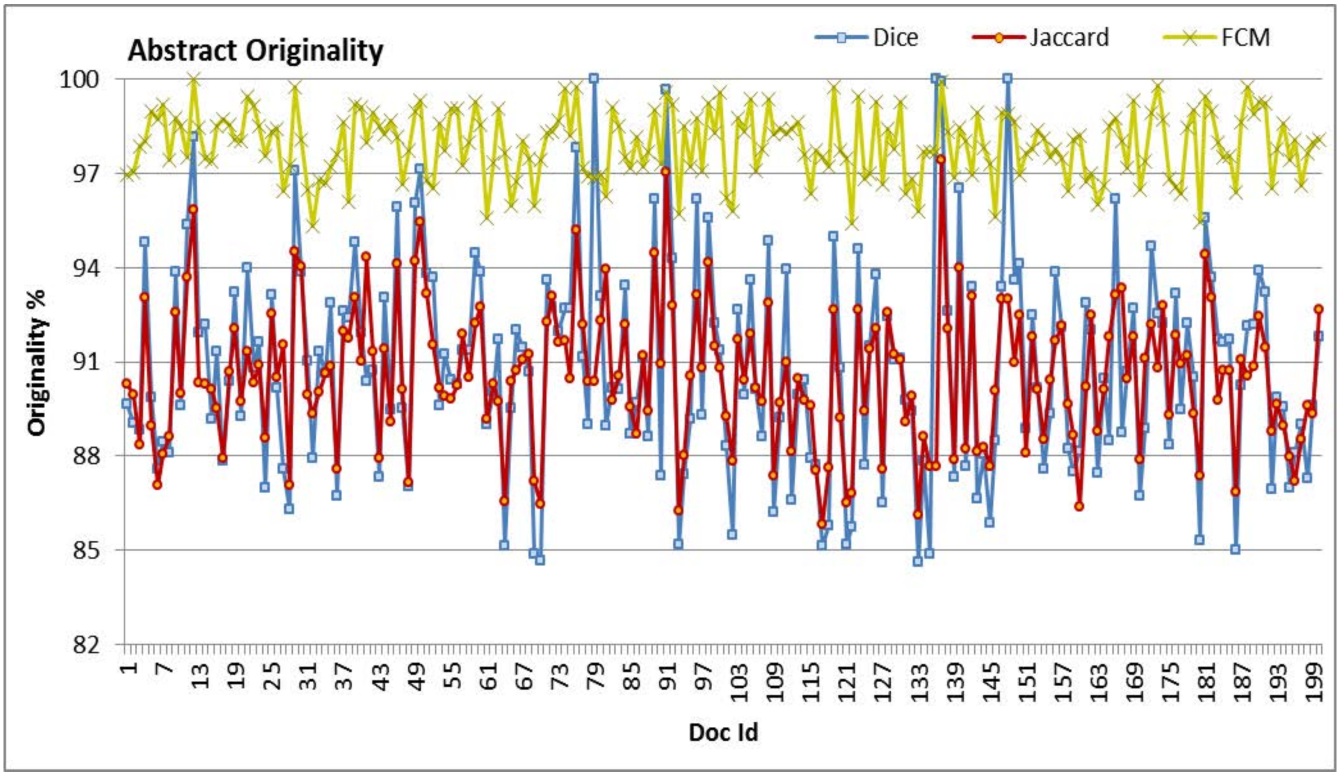

Figure 2b Comparison of statistical and ontological approach in detecting abstract based document originality for 200 scientific abstracts of Volume 53 Elsevier 'Computer Networks' across abstracts corpus of size = 5000. FCM outperforms in almost every experiment

4. RESULTS AND DISCUSSION

Research publications of Volume 53, Elsevier 'Computer Networks' journal was used as a data set for our work. The offline ontology constructed has 2000 concepts with relations and other relevant details. The concepts of the upper layers of the ontology are based on the standard ACM classification system [3]. 5000+ research papers in 'Computer Networks' domain were collected automatically via 'Google Search' (by framing search queries with concepts in the ontology) to form the 'Publication corpus'. From this, the abstracts were automatically extracted to form the 'Abstracts corpus'.

The quality assessment of research abstracts is twofold. These two dimensions of quality of research abstracts are selected to ensure that the abstract of the research article conveys the research (relevance) and the research in itself is novel (originality).

(1) The originality of research abstract

- To identify 'abstract originality' we have compared the abstracts of the data set with that of the 'abstracts corpus'.

(2) The relevance of research abstract across the respective research publication

- To identify 'abstract relevance' we have compared the abstracts of the data set with that of the respective research publications.

4.1 Empirical Analysis of Research Abstract Originality

Abstracts are said to contain the essence of any research publication. Therefore, finding the originality of the research publication via abstracts is equally intelligent [23]. Therefore, we have taken the abstracts of Volume 53 of Elsevier 'Computer Networks' Journal as input (sampling with replacement) and tabulated the results in Figure 2. We have compared the FCM based abstract similarity measures (figure 2a) with Dice's co-efficient and Jaccard co-efficient [56]. FCM based analysis yet outperforms every other approach in detecting the originality (figure 2b) of scientific publication with respect to the abstract, except for a few surprises. The reason where abstract based document originality goes for a failure would be for those abstracts which do not carry a detailed representation of the underlying idea, and, many times during our experiments, we found inconsistency of ideas as expressed in the research publication across the abstracts, i.e. the abstracts were found to contain inspiring thoughts as well but the respective research description in the paper was not so convincing. (Probably, this was the reason to proceed to analyse 'abstracts' relevance' which

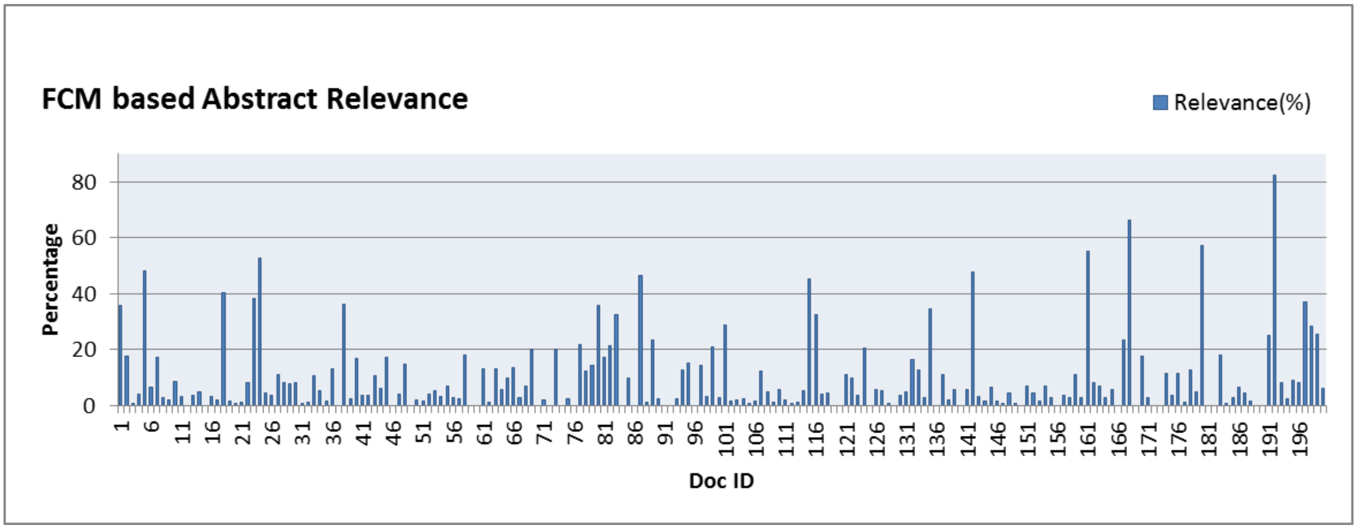

Figure 3 Analysing Abstract Relevance via Fuzzy Cognitive Map (FCM)

is our next step). Another reason may be that the abstracts are actually the short texts and therefore, to analyse the originality with the short text would be misleading.

4.2 Analysis of Relevance of Research Abstract

Table 1 Precison, Recall for FCM based 'Abstract relevance'

Method	Precision	Recall
Dice	0.556604	0.62766
Jaccard	0.572816	0.608247
FCM	0.608247	0.572816

Abstracts reflect the idea of any research article. Therefore, the completeness and originality of idea should reflect as well in research abstracts. In this context, the relevance of research abstract to the rest of the respective research paper was measured using FCM (refer figure 3). Research abstracts from the Elsevier dataset were considered as input. From the precision and recall values (refer table 1), it is understood that the FCM serves as a unique yet precise approach to finding abstract relevance.

4.3 Discussion on 'full-text vs. abstracts' Approach for Establishing Knowledge Networks

Problems of research abstracts are evident (as discussed in section 5.2) and therefore, analysing the entire research article to assess the research contribution of the author would be more appealing. Generally, conceptual similarity has the following behaviours: 1. supports the decision of syntactical similarity 2. Outperforms any other syntactical similarity by normalising the variations in similarity. The reason may be that word level variations are transformed into conceptual variations and therefore the normalised results.

Normalizing the obtained values, the FCM based originality calculations run on the corpus of abstracts alone suggested 51.5% (i.e. 103/200) of the publications to be original whereas the calculations based on full-text corpus suggested only 41% of the publications to be original. This reduction in the number of original documents can be attributed to the presence of the related work section in the full-texts. The elaboration of the related work in the full-text documents brings down the originality measure of the documents owing to the increase in the similarity that exists through the related work section. This could be overcome by not considering the related-work section of the full-texts, thereby providing a level ground for research articles with elaborate related work. Another method of overcoming this issue would be to consider the titles alone. The articles which do not surpass the originality threshold in the corpus were not being considered for the knowledge network

4.4 Generation of Abstract Clusters via K-means Clustering

With the quality and originality of the research abstracts determined, the abstracts are then clustered to form topic clusters. Clustering is done by using k-means algorithm with the Euclidean distance as a similarity metric. The number of clusters is determined by the Silhouette values of the data in the clusters. For the 103 documents which had better originality levels in the analysis of research abstracts, the process produced seven clusters, the information of which is presented below (figure 4). The clusters are labelled with the most frequent bi-gram appearing in the cluster. However, this approach created ambiguity in the cluster topics as there was more than one cluster which had the same label. This problem could be overcome by considering tri-grams or using external knowledge labels to arrive on descriptive labels for the clusters.

Table 2 Cluster Info

Cluster No.	Cluster Label	Number of Documents
1	Select share	9
2	Sensor nodes	6
3	Communication	8
4	TCP	15
5	Configuration	42
6	Energy Efficiency	14
7	Real Time	9

4.5 Obtaining Author Clusters from Topic Clusters

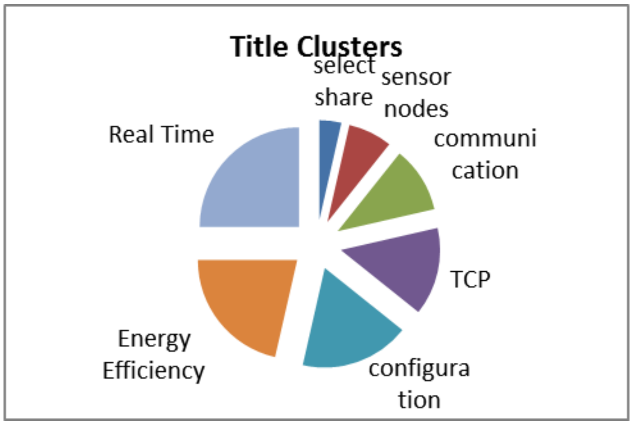

Figure 4 Topic clusters based on Research Abstracts

The authors of every research article in a cluster are analysed and a weight is assigned based on various parameters. These include position of author and citations earned by the author. Position of author is obtained automatically by analysing the subsequent lines of research abstract which follows the title. However, citations earned by every author are fed manually by analysing the values recorded in the SCOPUS database [22], since the ambiguity in author names would induce anomaly at this stage.

In addition, the important assumption that we make is about the percentage of research publications declared as original contributions. This implies that an author having equal percentage of selection and rejection with respect to 'abstract originality' is considered lower than an author having higher percentage of selection with respect to rejection. In other words, we indirectly penalise the author for not having written a quality article, whatever the number of research articles published may be. i.e. we do not take into account the research productivity [11, 44] which is yet another serious issue in determining the research impact [12, 20, 29, 34, 61] and thereby the impact factor. However, an author with no abstracts rejected will be considered as the highest and qualified contributor according to our assumption. This need not be true, since we only consider a small portion of the published articles. Once the clusters are formed, the clusters can then be analysed to identify the author with the highest contribution to the topic.

Two observations are obtained from the results.

(1) An author being present more than once in the same topic cluster

- This could be due to high research productivity (and thereby the publishing sentiment) of author towards the respective journal. The nodes indicating same author could be merged and the values assigned may be aggregated which may lead to finding 'author nucleus'.

(2) An author being present in more than one topic clusters

- This is a normal scenario. If an author is present in more than one topic cluster, the author's importance in every cluster is calculated separately.

With the author importance calculated, every cluster would contain unique authors with weights assigned. These weights are referred to as 'author research index'. Later, the author clusters are merged to form a single author hyper graph. The nucleus of author hypergraph is determined by ranking the author research index. By default, the author with highest 'research index' becomes the nucleus. Upon conflict, the ranking of authors is determined with respect to originality score and 'knowledge network' acceptance / rejection ratio of every author. With author research index and nucleus determined, the distance and polarity of authors with respect to the nucleus is determined. This (distance, polarity) pair is referred to as 'author quality index'. There is very less probability of more than one author possessing the same 'author quality index'. This is the reason behind our success of depiction of collaborative researchers via visualisation. Originality and acceptance/rejection ratio into the research community are assumed to be the 'x' and 'y' axes with the nucleus author at the convergence point. The author nodes are plotted in the four quadrants and a link with the nucleus indicating the link strength ('distance') is drawn. By updating the co-author information which has to be present obviously within the community network, the visualisation of research community is fulfilled (figure 5). Work is in progress to automate the visualisation of community network. By applying graph based techniques to the community network, we could further analyse the network to reach at surprising results.

5. CONCLUSION

The technological advancements and open source policies have equally had a negative impact on promoting high quality research. Hence there is a need to identify quality researchers for the benefit of the research community. Quality arises from, not the quantity of research contributions but from the research originality. Therefore, we believe that detecting the originality of scientific research abstracts will contribute more towards preserving the quality of scientific research. FCM-based method provides a cognitive perspective to the calculation of document originality and facilitates the researcher to quantify the grey influence of relationships between research concepts. The research network thus evolved will be a reliable recommendation system to find the authors who have had an impact on a specific research area. In future, we tend to incorporate more aspects of visualisation which enables a novice researcher to easily navigate through the community. In addition, the author citation count has to be analysed for citation senses, i.e. polarity of citations which could be attempted through opinion mining of research citations. Through citation sense making, the guest citations would be eliminated from the author citation count, thereby impacting the author quality. More ideas in the direction of Fuzzy grey cognitive maps [33] for document originality need to be analysed such that with originality as the first filter, relevance rating could serve as the next level of filtration to evolve more eminent research community network.

Table 3 Research Articles in Cluster 2

S.No	Document ID	Title
1	science_006	A modeling framework of content pollution in Peer-to-Peer video streaming systems
2	science_056	Using Shared Risk Link Groups to enhance backup path computation
3	science_110	Support vector regression for link load prediction
4	science_167	Support Vector Machines for TCP traffic classification
5	science_179	Zero config residential gateway experiences for next generation smart homes
6	science_188	Two and three-dimensional intrusion object detection under randomized scheduling algorithms in sensor networks

Table 4 Authors in Cluster 2

Doc Id	Author 1	Author 2	Author 3	Author 4	Author 5	Author 6	Author 7
science_006	Ye Xia	Shigang Chen	Chunglae Cho	Vivekanand Korgaonka			
science_056	Yan Hu	Dah-Ming Chiu	John C.S. Lu				
science_110	Falko Dressler	Isabel Dietrich	Reinhard German	Bettina Krüge			
science_167	Yang Xiao	Yanping Zhang	Miao Peng	Hui Chen	Xiaojiang Du	Bo Sun	Kui W
science_179	Lluís Parcerisa Giné	Ian F. Akyildi					
science_188	Tridib Mukherjee	Ayan Banerjee	Georgios Varsamopoulos	Sandeep K.S. Gupta	Sanjay Rungt		

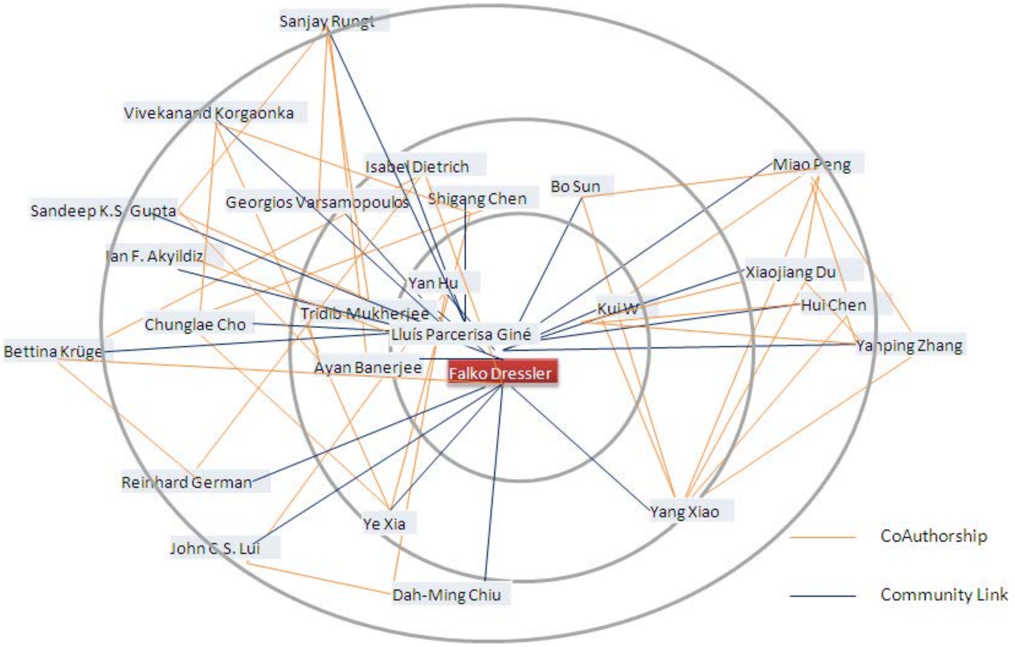

Figure 5. Visualisation of Research Community Network

6. REFERENCES

[1] F. Smeaton, G. Keogh, C. Gurrin, K. Mcdonald, and T. Sodring. 2002. Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century. *SIGIR Forum, 36(2),* 39–43.

[2] Pothen, H. Simon, and K. P. Liou. 1990. Partitioning sparse matrices with eigenvectorsof graphs. *SIAM J. Matrix Anal. Appl.*, 11,430–452.

[3] ACM Classification System, http://www.acm.org/about/class/1998

[4] Andrew W. Moore. 2001. K-means and Hierarchical clustering.

[5] http://www.autonlab.org/tutorials/kmeans09.pdf.

[6] Aris, A., Shneiderman, B., Qazvinian, V., & Radev, D. 2009. Visual overviews for discovering key papers and inuences across research fronts. *JASIST: Journal of the American Society for Information Science and Technology*, 60(11). 2219-2228.

[7] Association for Computing Machinery. 2011. ACM Digital Library. Retrieved from http://portal.acm.org

[8] W. Kernighan and S. Lin.1970. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, *49,* 291–307.

[9] Bi Chen, Baojun Qiu, Yan Qu, Xiaolong Zhang, John Yen. 2009. Intelligent Sensemaking of Scientific Literature through Citation Scent. *Sensemaking Workshop associated with CHI*.

[10] Bollacker, K. D., Lawrence, S., & Giles, C. L. 1998. CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications. *In AGENTS '98:proc. second international conference on autonomous agents New York*.116-123.

[11] Carayol, N. and Matt, M.2006. Individual and collective determinants of academic scientists' productivity, *Information Economics and Policy 18* (1).

[12] Castelnuovo G. 2008. Ditching impact factors: Time for the single researcher impact factor. *Bmj*. 336-789.

[13] Center for History and New Media, G. 2011. Zotero [Software]. Retrieved from http://www.zotero.org

[14] Chen, C. 2004. Searching for intellectual turning points: Progressive knowledge domain visualization. *PNAS: Proc. National Academy of Sciences of the United States of America*, 101(90001), 5303-5310.

[15] Chen, C. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST: Journal of the American Society for Information Science and Technology, 57(3)*. 359-377

[16] Chen, C., Ibekwe-SanJuan, F., & Hou, J. 2010. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *JASIST: Journal of the American Society for Information Science and Technology, 61*. 1386-1409.

[17] Cornell University Library. 2011. ArXiv. Retrieved from http://arxiv.org

[18] DBconnect: mining research community on DBLP data. 2007. *in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, COPYRIGHT ACM.*

[19] DBLife, http://dblife.cs.wisc.edu/

[20] Deepika J. and Mahalakhsmi G.S. 2011, Journal Impact Factor – A measure of quality or popularity? – *in proc. of int. conf. IICAI spl session on Advances in Web Intelligence and Data Mining – December 2011*.

[21] Dunne, C., Shneiderman, B., Gove, R., Klavans, J. & Dorr, B. 2011, Rapid understanding of scientific paper collections: integrating statistics, textual analysis, and visualization. *University of Maryland*.

[22] Elsevier. (2011). SciVerse Scopus. Retrieved from http://scopus.com/

[23] G. S. Mahalakshmi, S. Sendhilkumar, AlaguIrulappan, PreethamMirinda. 2009. Ontology Based Relevance Analysis for Automatic Reference Tracking. *International Journal of Computer Applications in Technology (IJCAT): Special Issue on Computer Applications in Knowledge-Based Systems, ISSN 0952-8091, Vol. 35, Nos. 2/3/4*. 165-173.

[24] Giles, C. L., Bollacker, K. D., & Lawrence, S. 1998. CiteSeer: an automatic citation indexing system. *In DL '98: proc. 3rd ACM conference on digital libraries New York, NY, USA: ACM*. 89-98.

[25] Google. Google Scholar. 2011. Retrieved from http://scholar.google.com

[26] Guan, T. & Wong, K.F.1999. KPS a Web information mining algorithm. *WWW8*.

[27] Han, H., Zha, H., & Giles, C. 2005. Name disambiguation in author citations using a k-way spectral clustering method. *Proceedings of the 5th ACM/IEEECS Joint Conference on Digital Libraries*. 334-343.

[28] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. 2009. 14.3.12 Hierarchical clustering (PDF). *The Elements of Statistical Learning (2nd ed.). New York: Springer*. 520–528.

[29] Hirsch, J. E. 2009.An index to quantify an individual's scientific research output, *Proc Natl Acad Sci USA, Vol. 102, No. 46, Springerlink, Scientometrics*. 16569–16573.

[30] Institute of Electrical and Electronics Engineers. 2011. IEEE Xplore. Retrieved from http://ieeexplore.ieee.org

[31] JabRef Development Team. 2011. JabRef [Software]. Retrieved from http://jabref.sourceforge.net

[32] Jiawei Han. 2009. Mining Heterogeneous Information Networks by Exploring the Power of Links, *J. Gama et al. (Eds.): DS 2009, Springer-Verlag Berlin Heidelberg, LNAI 5808.* 13–30.

[33] Jose Salmeron.2010. Modeling grey uncertainty with Fuzzy Grey Cognitive Maps, *Expert Systems with Applications 37 (2010)* 7581–7588.

[34] K. Satyanarayana.2010. Impact Factor and other indices to assess science, scientists and scientific journals. *Indian Journal of Physiology and Pharmacology, Vol.54, No.3.* 197-212.

[35] Kanungo. T, Mount, D. M., Netanyahu, N. S,Piatko, C. D.; Silverman, R.; Wu, A. Y.2002. An efficient k-means

clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence 24*. 881–892.

[36] Kosko B. 1986. Fuzzy Cognitive Maps. *Int. Journal of Man-Machine Studies*, Vol. 24. 65-75.

[37] Lars Backstrom, Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. *WSDM 2011*. 635-644.

[38] Libra , http://libra.msra.cn/

[39] Makoto Takeya, Hitoshi Sasaki , KeizoNagaoka and Nobuyoshi Vonezawa 2004. A Performance scoring method based on quantitative comparison ofConcept maps by a teacher and students. Concept maps: theory, methodology, technology , *Proc. Of the first int. Conference on concept mapping, A. J. Cañas, J. D. Novak, F. M. González, eds, Pamplona, Spain.*

[40] M. E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review, 45(2)*.167–256.

[41] Manjula Dissanayake and SimaanM.AbouRizk. 2007. Qualitative simulation of construction performance using fuzzy cognitive maps,IEEE, *Proceedings of the 2007 Winter Simulation Conference S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds*.2134 -2140.

[42] Mario A. Nascimento, J¨org Sander, and Jeffrey Pound. 2003. Analysis of sigmod's co-authorship graph. *SIGMOD Record, 32(2)*.57–58.

[43] Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *PNAS 2008 105 (4)*. 1118-1123

[44] Meho, L.I., & Spurgin, K.M. 2005. Ranking the research productivity of LIS faculty and schools: *An evaluation of data sources and research methods*. 1314-1331.

[45] Mendeley Ltd. 2011. Mendeley [Software]. Retrieved from http://www.mendeley.com

[46] Michelle Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *In Proceedings of the National Academy of Science USA, 99*. 8271-8276.

[47] Microsoft Research. 2011. Microsoft Academic Search. Retrieved from http://academic.research.microsoft.com

[48] National Center for Biotechnology Information. 2011. PubMed. Retrieved from http://ncbi.nlm.nih.gov/pubmed

[49] Newman, M. E. J. 2001. The structure of scientific collaboration networks. *PNAS: Proc.National Academy of Sciences of the United States of America, 98(2)*. 404-409.

[50] Nevin Heintze. 1996. Scalable Document Fingerprinting. *Proceedings of the Second USENIX Workshop on Electronic Commerce, Oakland, California, November 18-21*.

[51] Noshir S. Contractor, Peter R. Monge. 2002. Theories of Communication Networks. Oxford University Press.

[52] NWB Team. 2006. Network Workbench [Software]. Retrieved from http://nwb.slis.indiana.edu

[53] Osmar R. Za¨iane, Jiyang Chen, and Randy Goebel. 2007. Mining Research Communities in Bibliographical Data. *In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD '07). ACM, New York, NY, USA*. 74-81.

[54] Pena A. Sossa H. and Gutiérrez, F.2005. Knowledge and Reasoning Supported by Cognitive Maps, *Proceedings of Mexican International Conference on Artificial Intelligence 2005 (MICAI'05), Lecture Notes in Artificial Intelligence, Vol.3789. Springer, Monterrey, Mexico*.

[55] PROTÉGÉ, 2011, http://protege.stanford.edu

[56] R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. *In Proceedings of AAAI-06*.

[57] Rosy Madaan, Ashutosh Dixit, A.K. Sharma, Komal Kumar Bhatia. 2010. A Framework for Incremental Hidden Web Crawler, (IJCSE) *International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010*. 753-758.

[58] S. Wasserman and K. Faust. 1994. Social network analysis: Methods and applications. Cambridge University Press.

[59] SalhaAlzahrani and NaomieSalim. 2010. Fuzzy semantic-Based String Similarity for Extrinsic Plagiarism Detection. *Lab Report for PAN at CLEF 2010*.

[60] Salton, G. and McGill, M. J. 1983. Introduction to Modern Information Retrieval. *McGraw-Hill, New York, NY*.

[61] Seglen PO. 1997. Why the impact factor of journals should not be used for evaluating research. *BMJ, Vol.314, No.7079*. 498-502.

[62] Thomson Reuters Impact Factor. 1994.

[63] Thomson Reuters. 2011a. EndNote [Software]. Retrieved from http://www.endnote.com

[64] Thomson Reuters. 2011b. ISI Web of Knowledge. Retrieved from http: / /isiwebofknowledge.com

[65] Zhen, L., Jiang, Z. 2010. Hy-SN: Hyper-graph based semantic network. *Knowledge-Based Systems 23(8)*. 809–816.