RECOMMENDER SYSTEM ON STRUCTURED DATA

Muhammad Tahir[1], Shahzad[2], Nazia Azim[3], Izaz Ahmad Khan[4], Syed Roohullah Jan[5], Fazlullah Khan[6]

[1,2,3,5,6]*Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan*
[4]*Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan*

ABSTRACT

*The publications of research papers are increasing exponentially and to find the relevant paper to a research area or according a user query is crucial task. It is due to the large corpus of text data in different formats. If a user query on search engine to read a research paper, it will provide him a bundle of papers, now to find the relevancy of the paper; user will read one by one each paper. There are some techniques already exist as TF*IDF with many variations and content match search cited by and co-citation. Document ranking and the vector space model is nearest to our system. In this paper, we used a new approach for recommender system , it examine identify paper section then assign weights to each section , applying Paper section weights to determine what documents are more relevant to a user query. When papers retrieved then rank it on the basis of section weights. each paper ha four main sections are Paper Title, Abstract, Keywords, Introduction and Conclusion. In online survey we get best results against paper content match.*

Keywords *:— section weight, term frequency, rank document.*

1. INTRODUCTION

Finding papers for a related research work is time consuming task. When scholars start their research they need some related work to match his work with the previous work, or to add more details in his work, it can only be done by finding relevant literature on the net. As all we know each day new fields emerging and literature of all fields grow rapidly, so finding research papers related to a user is very complicated issue, also It is due to the large datasets, structure or format of papers. Many techniques in literature we found for user query relevant document. Some existing systems although consider the overall paper text i.e. Tf*Idf [1,2], but issue is the complexity of TF*idf and time consuming, also mostly it make parse vectors which has no benefit just slowing processing, also a chance may that words may match in the paper references or acknowledgement.

Another approach is Vector-based methods for performing query retrieval also show good results[4]. suggest performing query retrieval using a popular matrix algorithm called Latent Semantic Indexing (LSI). In essence, the algorithm creates a reduced-dimensional vector space that captures an n-dimensional representation of a set of documents. When a query is entered, its numerical representation is compared the cosine-distance of other documents in the document space, and the algorithm returns documents where this distance is small. Here issue dimensional reduction which can degrade or missed some data by reducing the size of documents or terms

In this paper we present recommender system, which uses content-based techniques with little modifications. In our recommender system, first we will limit ourselves to searching a collection of English documents (research papers). Then each and every paper will be converted to XML format. Simply we can describe this problem more formally. We have a set of documents D, by performing data preparation steps on D, We store it in data base, now user entering a query q = w1, w2, ., wn for a sequence of words wi. Then we wish to return a subset D* of D by applying section weights.

2. RELATED WORK

Retrieval and ranking document is also done by different Methods of term weighting for documents, using measures of term importance within an entire document collection, term importance within a given document, and document length[1]. On the basis of these methods a value of cosine is calculated for each document with a user query. Values in range of 0 and 1, highest value is consider as a relevant document to a user query.
Examining the weighting of document terms and the weighting of query terms raises two questions:

1. What sections, terns or factors in a document are important in measuring document-query similarity
2. How should these sections terms or factors be measured and combined.

For similarity of query and documents four major factors are:
1. The number of term matches between the query and a document,
2. The importance of a given term within a document collection,
3. The importance of a given term within a given document,
4. The length of a document.

As TF-IDF calculates values for each word in a document the word appears in that particular. Words with high TF-IDF numbers imply a strong relationship with the document they appear in paper. The overall approach works as follows. Given a document collection D, a word w, and an individual document d є D, we calculate wd = (fw, d) * log (|D|/fw, D)
Where (fw, d) equals the number of times w appears in d, |D| is the size of the corpus, and (fw, D) equals the number of documents in which w appears in D [1].
In terms of synonyms, notice that TF-IDF does not make the jump to the relationship between words. When the user want to find information about word “priest”, so the document containing “reverend” will not be considered as relevant while both words are similar. TF-IDF could not equate the word “drug” with its plural drugs.
Text document clustering groups similar documents that to form a cluster, while documents that are different have separated apart into different clusters[5].
Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful clusters, therefore same pattern or closest documents or kept in one Cluster; cluster may be one or more than one. It provides simple access to same document if a user query matched to a cluster. Simple K-mean algorithm is used here.
For clusters row objects must satisfy the following four conditions.
Let x and y be any two objects in a set and d(x, y) be the distance between x and y. The distance between any two points must be nonnegative, that is, d(x, y) >= 0. The distance between two objects must be zero if and only if the two objects are identical, that is, d(x, y) = 0 if and only if x = y. Distance must be symmetric, that is, distance from x to y, is the same as the distance from y to x, i.e. d(x, y) = d(y, x). The measure must satisfy the triangle inequality, which is d(x, z) = d(x, y) + d(y, z). Here is partial cluster may also be built; one document may also belong to more than one cluster. This paper not discussing this issue [7-15].

In vector-space model, a document is conceptually represented by a vector of keywords extracted from the document, then finding weights of the terms in documents. These weights describe the importance of the keywords in the document and within the whole document collection. The user query is likewise is modeled as a list of keywords with associated weights representing the importance of the keywords in the query [6].

In this document’s vector representation is only conceptual because the full vector is rarely stored internally as is because it is long and sparse. Document vectors are stored in an inverted file that can return the list of documents containing a given keyword of the user query. Problem is direct comparison between the vectors is slow because it would incur N vector comparisons. Cost increases exponentially in large corpuses. This method assigns high weights to terms that appear frequently in a small number of documents in the document set

3. PROPOSED IDEA

Our systemuses a two layer process to find a set of documents to relevant to user query. In the first layer, the system preprocess data, the second layer searches a collection of over a million papers, and returns the top 10 most relevant papers to the user query. These layers are

i. Data preparation Layer

ii. Application Layer

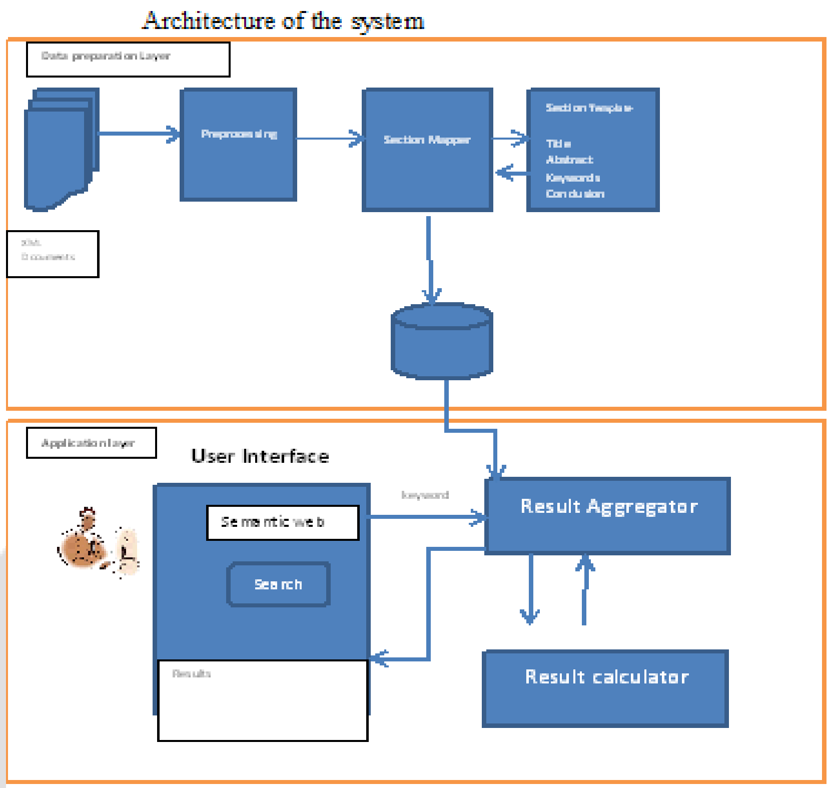

Fig-1: Architecture of the system

This recommender system consist of two layers

i. Data preparation layer

ii. Application layer

First layer name describes that this layer prepare data, data comes from JUCS data set, each papers is converted to XML formats. Then pre-processing step remove some noise word i.e. "a, the, then e.t.c.". This speed up our process and also less space required for storage. Section mapper and section template find section in each paper, section template consist of section names i.e. Abstract, title, keywords and so on. Section mapper extracts that section and then stores it in database. For each section a table is maintained in databases with the paper Id. Data preparation step is primary step or this recommender system [16-28].

Second layer is the application layer it is related to user interaction with system. A user will enter a keyword to search. Then will submit search button, the query will go the result aggregator, result aggregator will fetch results from database of the paper, it will just count the keywords in each section of the paper in database, the total words matched in each section pass to the result calculator. Here the section weights are assigned and aggregate weight is calculated, result return to the aggregator it will give papers to user in a ranking order [29-36].

4. EXPERIMENTS AND RESULTS

For this recommender system we used the corpus of JUCS papers nearly 500 papers, each paper is divided in five sections are Title, abstract, keywords, Introduction, Conclusion Two methods results are compared total content match and the new developed system When I entered the query as "Knowledge management" so I retrieved ten papers for both systems. The content match give the top most relevant paper is "Process Oriented Knowledge Management: A Service Based Approach". And our recommender system gives us the top most relevant paper "Reconciling Knowledge Management and Workflow Management Systems [37-40]: The Activity-Based

Knowledge Management Approach". The content match considers the paper header and footer so it give importance to the paper du header footer matching. Our new system just considers the section weights and gives most relevant papers. Online survey gives response that the recommender system result is better than the content match

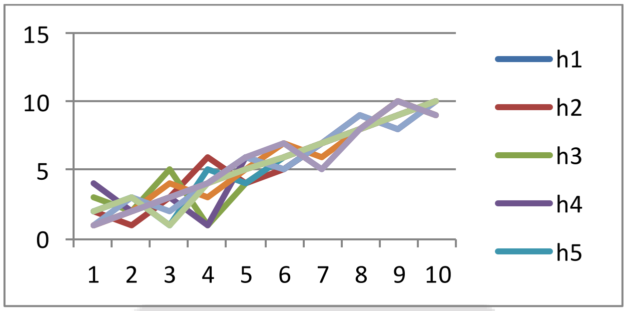

Chart-1: Knowledge Management

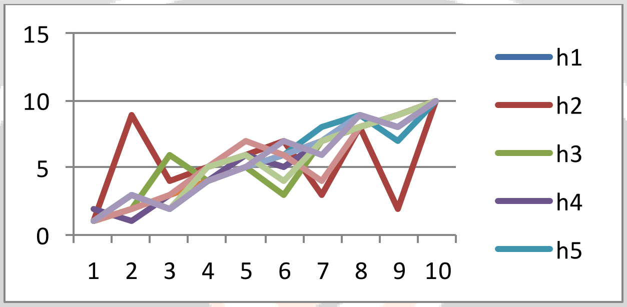

Chart-2: Informational Retrieval.

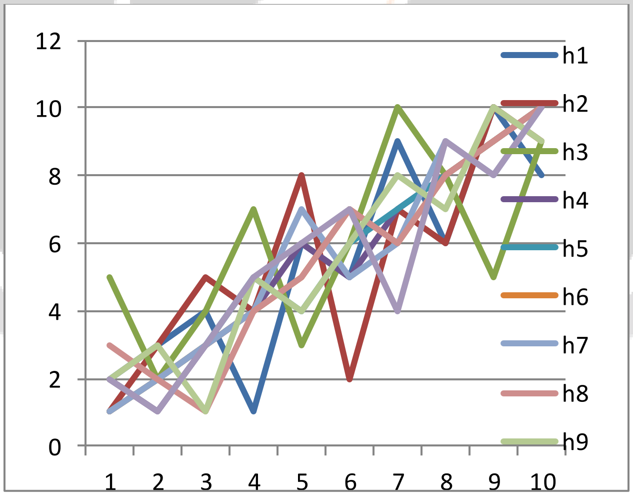

Chart-3-: Information Visualization

5. CONCLUSION AND FUTURE WORK

This system is a recommender system for research papers, finding papers relevant to a user interest or query. This tool can save researchers a great deal of time and effort in the process of a literature search. The system is in first stages. However, there are directions in which we intend to further develop and enhance this system by adding new techniques as tf*idf with it, or more enhance it by partial match of the keywords and some synonyms can improve result more better. We can also make it intelligent by learning from the user interest, once learned the user interest system will automatically recommend the papers. One direction that give an portion on user interface where he can assign weights to each section dynamically so if user interest in related work of the papers then he will assign more weights than other sections. Adding semantic to this system can enhance more results.

6. REFERENCES

[1]. Khan. F., Bashir, F. (2012). Dual Head Clustering Scheme in Wireless Sensor Networks. in the IEEE International Conference on Emerging Technologies (pp. 1-8). Islamabad: IEEE Islamabad.

[2]. Syed Roohullah Jan, Farman Ullah, Hashim Ali, Fazlullah Khan, " Enhanced and Effective Learning through Mobile Learning an Insight into Students Perception of Mobile Learning at University Level", International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.674-681, March-April 2016. URL : http://ijsrset.com/IJSRSET1622209.php

[3]. Khan. F., Nakagawa. K. (2012). Cooperative Spectrum Sensing Techniques in Cognitive Radio Networks. in the Institute of Electronics, Information and Communication Engineers (IEICE), Japan , Vol -1, 2.

[4]. Puthal, D., Nepal, S., Ranjan, R., & Chen, J. (2015). A Dynamic Key Length Based Approach for Real-Time Security Verification of Big Sensing Data Stream. In Web Information Systems Engineering–WISE 2015 (pp. 93-108). Springer International Publishing.

[5]. M. A. Jan, P. Nanda, X. He, Z. Tan and R. P. Liu, “A robust authentication scheme for observing resources in the internet of things environment” in 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 205-211, 2014, IEEE.

[6]. Khan. F., Nakagawa, K. (2012). Performance Improvement in Cognitive Radio Sensor Networks. in the Institute of Electronics, Information and Communication Engineers (IEICE) , 8.

[7]. Puthal, D., Nepal, S., Ranjan, R., & Chen, J. (2015, August). DPBSV--An Efficient and Secure Scheme for Big Sensing Data Stream. InTrustcom/BigDataSE/ISPA, 2015 IEEE (Vol. 1, pp. 246-253). IEEE.

[8]. M. A. Jan, P. Nanda and X. He, “Energy Evaluation Model for an Improved Centralized Clustering Hierarchical Algorithm in WSN,” in Wired/Wireless Internet Communication, Lecture Notes in Computer Science, pp. 154–167, Springer, Berlin, Germany, 2013.

[9]. Khan. F., Kamal, S. A. (2013). Fairness Improvement in long-chain Multi-hop Wireless Adhoc Networks. International Conference on Connected Vehicles & Expo (pp. 1-8). Las Vegas: IEEE Las Vegas, USA.

[10]. Puthal, D., Nepal, S., Ranjan, R., & Chen, J. (2016). A dynamic prime number based efficient security mechanism for big sensing data streams.Journal of Computer and System Sciences.

[11]. M. A. Jan, P. Nanda, X. He and R. P. Liu, “Enhancing lifetime and quality of data in cluster-based hierarchical routing protocol for wireless sensor network”, 2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC & EUC), pp. 1400-1407, 2013.

[12]. Jabeen. Q., Khan. F., Khan, Shahzad, Jan. M. A., Khan. S.A (2016). Performance Improvement in Multihop Wireless Mobile Adhoc Networks. in the Journal Applied, Environmental, and Biological Sciences (JAEBS), Print ISSN: 2090-4274 Online ISSN: 2090-4215

[13]. Khan. F., Nakagawa, K. (2013). Comparative Study of Spectrum Sensing Techniques in Cognitive Radio Networks. in IEEE World Congress on Communication and Information Technologies (p. 8). Tunisia: IEEE Tunisia.

[14]. Puthal, D., Sahoo, B., & Sahoo, B. P. S. (2012). Effective Machine to Machine Communications in Smart Grid Networks. ARPN J. Syst. Softw.© 2009-2011 AJSS Journal, 2(1), 18-22.

[15]. Khan. F. (2014). Secure Communication and Routing Architecture in Wireless Sensor Networks. the 3rd Global Conference on Consumer Electronics (GCCE) (p. 4). Tokyo, Japan: IEEE Tokyo.

[16]. M. A. Jan, P. Nanda, X. He and R. P. Liu, “PASCCC: Priority-based application-specific congestion control clustering protocol” Computer Networks, Vol. 74, PP-92-102, 2014.

[17]. Khan. F. (2014). Throughput & Fairness Improvement in Mobile Ad hoc Networks. the 27th Annual Canadian

Conference on Electrical and Computer Engineering (p. 6). Toronto, Canada: IEEE Toronto.

[18]. Mian Ahmad Jan and Muhammad Khan, “A Survey of Cluster-based Hierarchical Routing Protocols”, in IRACST–International Journal of Computer Networks and Wireless Communications (IJCNWC), Vol.3, April. 2013, pp.138-143.

[19]. Khan. S., Khan. F., (2015). Delay and Throughput Improvement in Wireless Sensor and Actor Networks. 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW) (pp. 1-8). Riyadh: IEEE Riyad Chapter.

[20]. Khan. Shahzad, Khan. F., Jabeen. Q., Arif F., Jan. M. A., Khan. S.A (2016). Performance Improvement in Wireless Sensor and Actor Networks. in the Journal Applied, Environmental, and Biological Sciences Print ISSN: 2090-4274 Online ISSN: 2090-4215

[21]. Puthal, D., & Sahoo, B. (2012). Secure Data Collection & Critical Data Transmission in Mobile Sink WSN: Secure and Energy efficient data collection technique.

[22]. Mian Ahmad Jan and Muhammad Khan, “Denial of Service Attacks and Their Countermeasures in WSN”, in IRACST–International Journal of Computer Networks and Wireless Communications (IJCNWC), Vol.3, April. 2013.

[23]. Qamar Jabeen, Fazlullah Khan, Muhammad Nouman Hayat, Haroon Khan, Syed Roohullah Jan, Farman Ullah, " A Survey : Embedded Systems Supporting By Different Operating Systems", International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.664-673, March-April 2016. URL : http://ijsrset.com/IJSRSET1622208.php

[24]. M. A. Jan, P. Nanda, X. He and R. P. Liu, “A Sybil Attack Detection Scheme for a Centralized Clustering-based Hierarchical Network” in Trustcom/BigDataSE/ISPA, Vol.1, PP-318-325, 2015, IEEE.

[25]. Jabeen. Q., Khan. F., Hayat, M.N., Khan, H., Jan., S.R., Ullah, F., (2016) A Survey : Embedded Systems Supporting By Different Operating Systems in the International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.664-673.

[26]. Syed Roohullah Jan, Syed Tauhid Ullah Shah, Zia Ullah Johar, Yasin Shah, Fazlullah Khan, " An Innovative Approach to Investigate Various Software Testing Techniques and Strategies", International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.682-689, March-April 2016. URL : http://ijsrset.com/IJSRSET1622210.php

[27]. Khan. F., Khan. F., Jabeen. Q., Jan. S. R., Khan. S., (2016) Applications, Limitations, and Improvements in Visible Light Communication Systems in the VAWKUM Transaction on Computer Science Vol. 9, Iss.2, DOI: http://dx.doi.org/10.21015/vtcs.v9i2.398

[28]. Syed Roohullah Jan, Fazlullah Khan, Muhammad Tahir, Shahzad Khan., (2016) “Survey: Dealing Non-Functional Requirements At Architecture Level”, VFAST Transactions on Software Engineering, (Accepted 2016)

[29]. M. A. Jan, “Energy-efficient routing and secure communication in wireless sensor networks,” Ph.D. dissertation, 2016.

[30]. Syed Roohullah Jan, Faheem Dad, Nouman Amin, Abdul Hameed, Syed Saad Ali Shah, " Issues In Global Software Development (Communication, Coordination and Trust) - A Critical Review", International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.660-663, March-April 2016. URL : http://ijsrset.com/IJSRSET1622207.php

[31]. An Experimental Study of Factors Important in Document Ranking by Donna Harmsn Lister Hill National Center for Biomedical Communications National Library of Medicine Bethesda, Maryland, 20209

[32]. Using TF-IDF to Determine Word Relevance in Document Queries Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 0885

[33]. Scienstein: A Research Paper Recommender System Bela Gipp1, Jöran Beel1, Christian Hentschel2 1 Otto-von-Guericke University, Dept. of Computer Science, Magdeburg, Germany Fraunhofer Institute for Telecommunications, Berlin, Germany

[34]. M. A. Jan, P. Nanda, X. He and R. P. Liu. 2016. A Lightweight Mutual Authentication Scheme for IoT Objects, IEEE Transactions on Dependable and Secure Computing (TDSC), “Submitted”.

[35]. M. A. Jan, P. Nanda, X. He and R. P. Liu. 2016. A Sybil Attack Detection Scheme for a Forest Wildfire Monitoring Application, Elsevier Future Generation Computer Systems (FGCS), “Submitted”.

[36]. M. A. Jan, M. Usman, P. Nanda and X. He. 2016. PAWN: A Payload-based mutual Authentication scheme for Wireless Sensor Networks, in 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-16), “accepted”.

[37]. M. Usman, M. A. Jan and X. He. 2016. Cryptography-based Secure Data Storage and Sharing Using HEVC and Public Clouds, Elsevier Information sciences, “accepted”.

[38]. Berry, Michael W. et al. (1995). Using Linear Algebra for Intelligent Information Retrieval. SIAM Review, 37(4):177-196

[39]. Similarity Measures for Text Document Clustering, Anna Huang Department of Computer Science The University of Waikato, Hamilton, New Zealand

[40]. Document Ranking and the Vector-Space Model, DIK L. LEE,HUEI CHUANG,KENT SEAMONS Hong Kong University of Science and Technology