A Novel Approach for Document Retrieval System with User Preferences

Sandeep Kaur
Research Scholar (Department of Computer Science and Engineering), RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India

Nidhi Bhatla
Assistant Professor, Department of Computer Science and Engineering, RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India

ABSTRACT
This paper proposes a method for Document Retrieval Systems. The document retrieval system finds information to given criteria by matching text record (*documents*) against user queries. The results generated from information retrieval system must have user preferences. Each user has its own perspectives and cultural context of each word or when the user is searching for highly specific, focussed topic. The probabilistic ranking based on graphic Bayesian statistics is associated with a Kuhn munkres algorithm for it to be really successful to group similar documents. The probabilistic ranking based Kuhn munkres algorithm uses the graphical model such as Bayesian statistics with Bayesian's theorem to find the probability of documents for more relevant results.

General Terms
Document Retrieval System, TextTiling Algorithm, Kuhn Munkres Algorithm,

Keywords
IR, Bayesian Statistics, Bayesian Probability, Graphical models, Relevancy

1. INTRODUCTION
Text Mining is the application of data mining for information extraction, It can be called as discover knowledge from texts available in terms of textual data in different domains such has SMS, chat, wikis, newspaper, eBooks, emails, tweets, blogs. Information retrieval is an approach for accessing the information resources which are most relevant to the user's queries. The searches are done by the users can be approached to metadata or full-text. Information retrieval systems are used to overcome the information overload. The IR applications, web search engines which are use the IR algorithms to calculate the relevancy between similar documents. An information retrieval process begins when a user enters a query into the system. The user's queries are used to describe the information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. User queries are matched against the database information to obtain most relevant results from the entire database according to the user's queries. An optimal matching algorithm is needed to access the most relevant result. The logical and semantically extractions of text from documents done with TextTiling [10] algorithm. The similarity based search on extracted documents done with an optimal matching algorithm such as a Kuhn munkres [11] algorithm. A document retrieval system finds information to given standards by matching text record against user queries. A document retrieval system consists of a database of documents, an optimal matching algorithm to build a full text index, and a user interface to access the database. A document retrieval system has two main tasks: Find relevant documents to user's queries evaluate the matching results and sort them according to relevance, using algorithms such as the Kuhn munkres Algorithm. The retrieval method consists of text extraction and segmentation from different text document formats and source code documents into logical and semantic segments. These segmented documents are used to calculate the similarity against the user queries. The Kuhn munkres [11] algorithm is combined with probabilistic graphical model. The probabilistic retrieval model which ranks the documents based on probability ranking principle. The ranking of a document is based on their probability of relevance to the query. There are many methods to check the relevant or non relevant documents one of them is statistical distribution. The statistical approach provides users with relevant ranking of the retrieved documents. The two measures are used to find out the relevancy of the retrieved documents that are recall and precision.

2. RELATED WORK
Many ranking algorithms have been proposed for document retrieval system in the last few years. Many papers have been ranking algorithms that are used to find more relevant results to the user's queries. Timotej Betina et al. [1] proposed a method to directly facilitate the author's needs during the creation of text documents or source code. The Information retrieval system was integrated with a text editor in order to find similar documents. They analyzed the extraction of logical structure from different text document formats [4], [5] and also from source code documents. The second area was the extraction of semantically coherent blocks of text from documents [6]. They based their solution on the algorithm *TextTiling [10]* originally proposed by Hearst [7]. The third area was the pairwise document similarity based on the extracted document structure. Wan [8] analyzed different approaches and proposed the algorithm for finding the optimal matching solution between two documents using semantically segmented documents is forever. The results confirm it that improved retrieval performance using structured documents [9]. Liangcai Gao et al. [2] combine visual information and semantics of information complementary to solve the problems of article reconstruction. J. M. Bernardo [3] proposed the Bayesian model which has based on logic inference. The interpretation has to be used for finding the probability. The area of interest related to the statistical

inference has to be described when the modification has to be done because of a set of possibilities about the evidence makes by the users and Bayes' theorem defines the concepts how this modification has to be done with making the different views to the problem. To obtain a solution to the problem of extracting semantically coherent blocks of text from documents Timotej Betina et al. [1] used TextTiling [10] algorithm and for finding an optimal matching solution between two documents used Kuhn-Munkres algorithm. Liangcai Gao et al. [2] devised a method in which a bipartite graph, consists the set of vertex of two complementary subsets, and no edge connects two vertexes belonging to the same vertex subset. A matching of a bipartite graph was a subgroup of the graph where each vertex is associated with only one edge. This constraint guarantees that the one to-one relationship in a graph can be found by solving the matching problem. And an optimal matching (OM) of the graph was a match with the maximum weight. The adopted OM algorithm for bipartite graphs is a classic Kuhn-Munkres algorithm. They solve the problem of articles reconstruction from newspapers. Targeting the weaknesses of previous methods, they proposed an optimized solution for reading order detection and article aggregation. The major contributions were formulating article reconstruction as optimal matching of the bipartite graph model, the geometric information and content information are combined to improve the reliability and efficiency, Select the reading order of article blocks as a basic clue to group them.

The first factor which penalized a high number of unmatched segments is the disturbing factor (*df*) [8]. They also need to be aware of the dispersion of assigned segments as in Figure 1. They proposed to use segment disperse factor (*sd*). In (1) a table of content cosine similarity which should represent the important keywords in each document is also used.

$$sim(A, B) = w_{om}om(A, B) + w_{df}df(assignm(A, B)) - w_{sd}sd(assignm(A, B)) + w_{toc}cos(A_{toc}, B_{toc}) \quad (1)$$

In (2) and (3) they were seeing an example of finding a similar document to two documents, B and C. A is found as a similar document. Optimal matching gives us the solution of matching

$$assignm(A, B) = [A2, B1]\ [A3, B2]\ [A8\ B3] \quad (2)$$
$$assignm(A, B) = [A1, C1]\ [A5\ C2]\ [A9\ C3] \quad (3)$$

Their proposed factor favors document B, because the dispersal of its assigned segments to document A is smaller than in the case of document C.

$$sd(x, y) = \frac{\text{Number of unmatched spaces (assignm (x, y))}}{\text{Number of unmatched segments (assignm (x, y))}}$$

A higher value of segment dispersion means a higher penalization. In their case, the values are as follows:

$$sd(A,B) = 1/4 \qquad sd(A,B) = 2/6$$

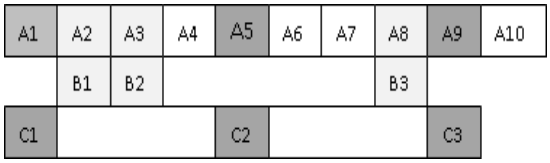

Fig 1: Segment disperse

3. RESEARCH GAP

After conducting a systematic review of the papers solicited from various high impact journals, we found that, typically follow issues remain unsolved and still gaps in this area need some care for example in Timotej Betina et al. [1] Approach was based on the process, in which during the writing of the document, logical and semantic structure previously extracted from existing document were improved to finding more similar and related documents to newly created one. The Author can ask for hints for every paragraph he was creating. Viewing the documents as a set of segments appears to help with that. The first step in the process was the structure extraction from the set of existing documents. The second step was the new document creation process and the third step was comparing the similarity between the new document and the available document set. The text of the new document is processed in real time and segmented into semantic segments. The implemented solution finds similar semantic segments just for one segment, but what is more important is to find whole documents which have a similar semantic structure combined with some logical structure information. They use the *Kuhn-Munkres algorithm* which solves the optimal matching problem based on Wan [8]. It finds the best segment assignments from two compared documents (*assign (A, B)*). It is important to realize that the searching for the most similar document in their new document which is in its creation phase, so it is not a whole document as of yet. So they have to take into account different number of segments between the new document and the searched document. Thomas [12] recovered the reading order using topological sorting of text lines. Pairwise order relationships between certain text line segments were determined through spatial layout-based rules. Chen et al. [13] developed a system for the layout understanding of Chinese newspapers. A set of rules on layout understanding were created based on visual information such as distance, size, color, etc. Opposite to the aforementioned techniques relying merely on the visual information, Aiello et al. [14] introduced a semantic information based method to determine the reading order. A lexical analysis the technique was adopted to rank candidate reading orders based on part-of-speech (POS) probability. Aiello et al. [15] a clustering-based method was proposed to group the text blocks belonging to the same article. In this method, the content similarity of two blocks was defined as their distance on reading order. Their method demonstrated the benefits of the semantic information for newspaper document understanding. Unfortunately, the method tended to fail when several unrelated blocks coincidentally share some textual content.

4. SCOPE OF WORK

The information retrieval system needs human intervention while building a database, query structure and evaluation of the system. The hybrid retrieval system (machine+human based) is useful for searching a most relevant document for user queries. The combination of automatic and manual annotation makes the data more meaningful to understanding of user's queries to the system easier. The results generated from information retrieval system must have user preferences.

The human ranks the results by giving a numerical score or a double star opinion (e.g. "Relevant" or "not relevant") for each item retrieved from database.

- The similarity of class found by the Kuhn munkres [11] algorithm may not suit the preferences of the user. Each user has its own perspectives and cultural context of each word or when the user is searching for highly specific, focussed topic.
- The probabilistic ranking based on graphic Bayesian statistics is associated with a Kuhn munkres [11] algorithm for it to be really successful to group similar documents.
- Probabilistic ranking based Kuhn munkres [11] Algorithm is a hybrid technique in which probabilistic graphical model and Bayesian statistics is combining.
- Bayesian statistics are a subset of the field of statistics in which the evidence about the true state of the world is expressed in terms of degrees of belief or, more specifically, Bayesian probabilities.
- The probabilistic ranking based Kuhn munkres [11] algorithm uses the graphical model such as Bayesian statistics with Bayesian's theorem to find the probability of documents for more relevant results.

The Document Retrieval system used the Kuhn munkres algorithm can be further improved by taking clues from multiuser who may involve in building and using this kind system. For each individual the usage of this kind system, a word/ phrase/ sentence may mean differently in context of culture and language putting them in same semantic class may result in building a corpus, which is correct as the user requirements. The previous work can be enhanced by adding a probabilistic ranking for each text unit assigning it to be a semantic class. Moreover, in the previous work logical structure of tables, figures has also not been associated. This can also be added and improved with probabilistic ranking based Kuhn munkres [11] Algorithm. Build and parse datasets of PDF file repository. Extract the logical structure of a document and build other conceptual, logical blocks to enhance it. Evaluate the system using recall and precision.

5. METHODOLOGY

As mentioned earlier, there is an urgent need to make existing Kuhn Munkres algorithms work more efficiently in terms of their recall and precision values while not ignoring the facts related to resources (bandwidth, time, CPU, utilization etc.) involved in running Kuhn Munkres algorithm. Therefore, we propose the following model shown in figure 2. :

- First step is to make the domain specific PDF document repository.
- Text parsing is done by the process of analyzing a string of symbols, according to the rules of a formal grammar.
- Extract the logical structure of PDF documents using PDF to HTML utility.
- Build retrieval model based on ranking or scoring
- Build probabilistic graphical ranking model.

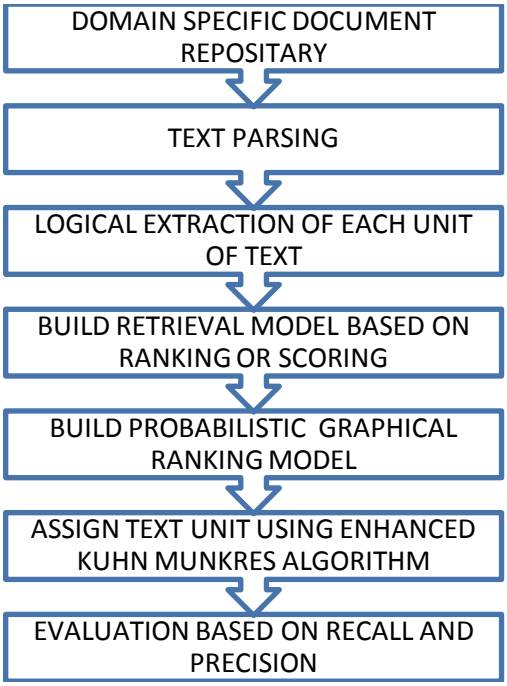

Fig 2: Proposed model for Document Retrieval System

- Probabilistic graphical models are graphs in which nodes represent random variables, and the (lack of) arcs represent conditional independence assumptions. So we will make the graph of the topic (Nodes) in pdf files based on domain.
- Once, this model is made, it is time to build the scoring system or ranking system [mathematical formula] which is based on the similarity of the words/topic nodes
- After the scoring system is built on content (Topic within pdf), now user preferences will be made. The user will also give some score for enhancing the Ranking as per machine + Man.
- Run the final enhanced Kuhn Munkres algorithm.
- Evaluate the system based on recall and precision.

Extraction process has three steps: logical extraction, semantic extraction (does not apply to source code documents) and transforming document representation in order to get the most of the extracted information.

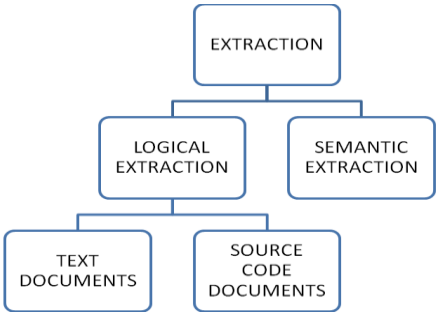

Fig 3: Process for Extraction of documents

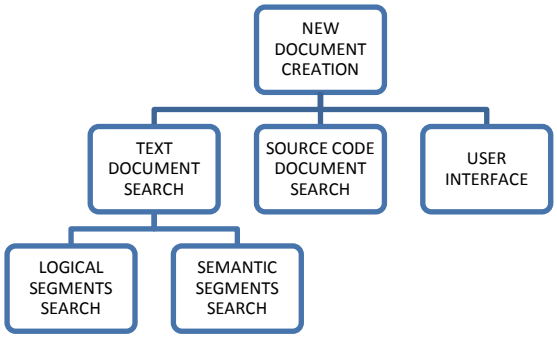

Fig 4: New document creation process

6. CONCLUSION

In this research work, we have come to the conclusion that there is a proper case to improve the Timotej Betina et. Al [1] methods, hence a new methodology has been proposed in section number 5. This proposal overcomes the main problem that would result in accurate results in terms of precision and recall of the overall system. This method helps to take advantage of graph theory. The document data can be represented with the use of SQL as 'No SQL Data' and deep insight relationship can be found within the document context and among the other similar or related documents, which helps to increase the overall experience in research of relevant documents that are more meaning full and useful.

7. FUTURE SCOPE

Based on the current framework and understanding of current perspective, issues in this area for future scope we suggest following

a. Work must be carried out on the domain specific document repository. The domain may be sports, politics etc.

b. Build logical structure of document for extraction of elementary and logical units of contents.

c. Build probabilistic graphical database model which can semantically be used by Kuhn Munkres Algorithm.

d. Use enhanced Bayesian logical functions to obtain high recall and precision for user domain specific queries.

8. REFERENCES

[1] Timotej Betina, Ivan Polasek. Document Creation with Information Retrieval System Support. *14th International Symposium on Computational Intelligence and informatics. 19-21 November, 2013. Budapest, Hungary*.

[2] Liangcai Gao, Zhi Tang, Xiaoyan Lin, Yongtao Wang. A Graph-based Method of Newspaper Article Construction. *21st international conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.*

[3] J. M. Bernardo. *Bayesian Statistics Departamento de Estadística, Facultad de Matemáticas, 46100–Burjassot, Valencia, Spain*.

[4] Anjewierden, A. AIDAS: Incremental Logical Structure Discovery in PDF Documents. In conference *Sixth International Conference on Document Analysis and Recognition*. 10-13 Sep.2001, pp. 374-378. ISBN: 0-7695-1263-1.

[5] Stoffel, A., Spretke, D., Enhancing Document Structure Analysis using Visual Analytics. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. SAC '10, 22-26 March 2010, pp. 8-12. ISBN: 978-1-60558-639-7.

[6] Kaszkiel, M., Zobel, J. Effective ranking with arbitrary passages. In Journal of the American Society for Information Science and *Technology*. Feb. 2001, Vol. 52, Issue 4. Doi:10.1002/1532-2890

[7] Hearst, M. A. TextTiling: Segmenting text into multi-paragraph subtopic passages. In Journal *Computational Linguistics*. March 1997, vol. 23, issue 1. Dostupné na internete: http://dl.acm.org/citation.cfm?id=972687

[8] Wan, X. Beyond topical similarity: a structural similarity measure for retrieving highly similar documents. In *KNOWLEDGE AND INFORMATION SYSTEMS* 2008, vol. 15, NUM. 1, pp. 55-73, DOI: 10.1007/s10115-006-0047-1

[9] Wilkinson, R. Effective retrieval of structured documents. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* SIGIR '94, 1994. ISBN:0-387-19889-X

[10] Marti A. Hearst. TextTiling [10]: Segmenting Text into Multi-paragraph Subtopic passages. Computational linguistics Volume 23 Issue 1, March 1997H.W. Kuhn, On the origin of the Hungarian Method, History of mathematical programming

[11] H.W. Kuhn, On the origin of the Hungarian Method, History of mathematical programming collection of personal reminiscences (J.K. Lenstra, A.H.G. Rinnooy Kan, and A. Schrijv Eds.), North Holland, Amsterdam, 1991, pp. 77–81.

[12] Thomas, M. B. High Performance Document Layout Analysis. In *Proc. Of SDIUT'03*, 2003.

[13] Chen, M., Ding, X. and Liang, J. Analysis, Understanding and Representation of Chinese Newspaperwith Complex Layout. In *Proc. of CIP'00*, 2000.

[14] Aiello, M. and Pegoretti, A. Textual Article Clustering in Newspaper Pages. *Applied Artificial Intelligence*, 2006.

[15] Aiello, M., Monz, C., Todoran. L. and Worring, M. Document Understanding for a Broad Class of Documents. *International Journal on Document Analysis*.