Survey on Information Retrieval Algorithm

Sandeep Kaur[#1], Nidhi Bhatla[*2]

[#] *Research Scholar (Department of Computer Science and Engineering), RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India*

[1]chohansandeep78@yahoo.in

[*] *Assistant Professor, Department of Computer Science and Engineering, RIMT College of Engineering & Technology, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India*

[2]engineernidhi@yahoo.com

***Abstract*— Document retrieval systems find information to given criteria by matching text record (*documents*) 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.**

***Keywords*— IR, TextTiling, Kuhn Munkres, Bayseian Statistics, Bayseian Probability, Relevancy**

I. Introduction

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 or an ordinal score or a binary judgment (e.g. "Relevant" or "not relevant") for each item retrieved from database. The machines rank the results using a ranking model with the association of ranking algorithm such as *TextTiling [10]*, *Kuhn munkres [11]* algorithm. Fig.1 has shown Ranking Model. The retrieval method consists of two main tasks: text extraction, segmentation of different text document formats and source code documents into logical and semantic segments. These existing segments are used to calculate the similarity with the new document using an optimal matching algorithm [1]**.** The articles are the main concepts in a newspaper which has to be reconstructed. The reconstruction process has two steps article aggregation and reading order recovery. The reconstruction of articles from newspapers is a difficult task to handle the multi-article page layout because of complexity [2]. Mathematical statistics have two major models, conventional and Bayesian. Bayesian methods provide a complet*e* model for both Statistical inference, decision making to compete with uncertainty [3].

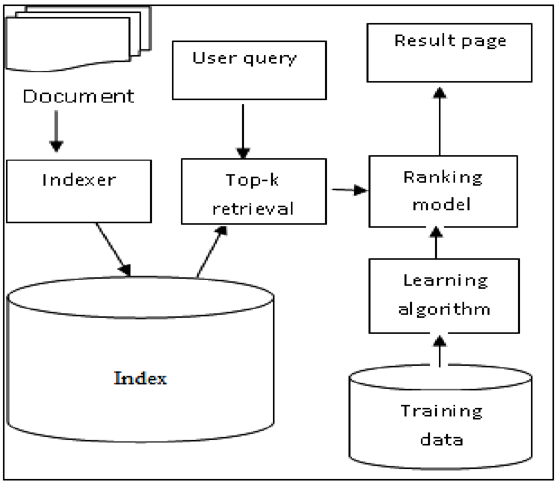

Fig 1 Ranking Model

II. Related work

Many ranking algorithms have been proposed for document retrieval system in the last few years. Like in many papers these ranking algorithms 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 in the created document during its writing phase. They analyzed the extraction of logical structure from different text document formats [4], [5] and also from source code documents . The second area is 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 is 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. His results confirm improved retrieval performance using structured documents [9]. Liangcai Gao et al. [2] Combine visual information and semantics of information complementary to better 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. Timotej Betina et al. [1] Approach was based on the process, which is shown in Figure 2. 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. 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 are 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.

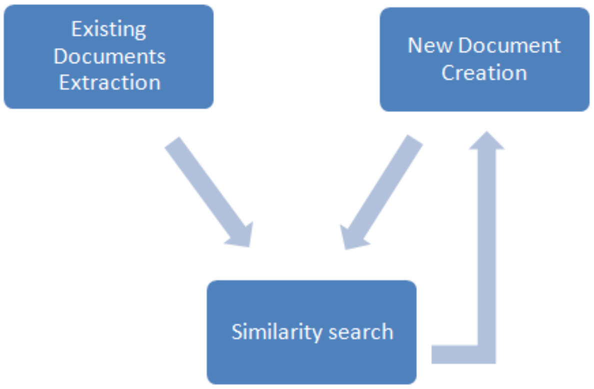

Fig. 2 Document creation process

III. Information Retrieval Algorithms

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 extraction 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. TextTiling [10] Algorithm

TextTiling [10] is a technique which subdivides the document's text into paragraphs that represent passages, or subtopics. The user's cues have to be used for identifying the subtopics according to patterns of lexical co-occurrence and distribution. The algorithm works as the group discussion in which set of words is in use when subtopic discussed, but when that subtopic changes, a significant proportion of the vocabulary changes according to subtopic as well.

B. Kuhn munkres [11] Algorithm

A Kuhn munkres [11] algorithm is a combinatorial optimization algorithm that solves the assignment problem in polynomial time. For example, assuming that numerical scores are available for the performance of each of N persons on each of N jobs, the "assignment problem" is the quest for an assignment of persons to jobs so that the sum of the N squares so obtained is as large as possible.

IV. Conclusion

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.

V. Future work

The Docuemt 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.

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.