International Journal of Mechanical Engineering & Computer Sciences, Vol.1, Issue 1, Jan-Jun, 2017, pp 12-17

HYBRIDIZED MODEL FOR EFFICIENT MATCHING AND DATA PREDICTION IN INFORMATION RETRIEVAL

BOMA P. W[1] , OGHENEVOR E. E[2], ALABI O. A[3]

[1,2,3] DEPARTMEN T OF COMPUTER SCIENCE

UNIVERSITY OF PORT HARCOURT, NIGERIA

www.arseam.com

ABSTRACT

Information Retrieval (IR) has become a topic of great interest with the advent of text search engines on the Internet. Retrieving information from large databases can sometimes be very difficult and may tend to be very slow. This is especially so when there is need to manage these data or documents. This is because databases contain millions of documents in myriad subject areas. This is why information retrieval (IR) is very important. An IR systems matches user queries to documents stored in a database. Despite all these efforts, retrieving documents and texts is still problematic as none has been fully efficient in terms of speed. In this research work, we combined two known techniques of information retrieval: K-nearest neighbor (KNN) and Support Vector Machine (SVM) to mine and retrieve information more efficiently. We are going to compare our results with other results and see how efficiently our model is over the other models.

Keywords: Information Retrieval, guery.

Introduction:

Retrieval of text-based information also termed Information Retrieval (IR) has become a topic of great interest with the advent of text search engines on the Internet (Baeza-Yates and Ribeiro-Neto, 1999). Traditionally in Information Retrieval (IR), documents and text queries both are represented in a unified manner, as sets of terms, to compute the distances between queries and documents thus providing a framework to directly implement simple text retrieval algorithms (Beitzel et al., 2007).

Materials and methods

Retrieving information from the Internet and from large database is quite difficult and time consuming especially if such information is unstructured. A lot of algorithms and techniques have been developed in the area of data mining and information retrieval yet retrieving data from large databases continue to be problematic. In this research work, we combined two well-known techniques of information retrieval with the hope that our model could capitalize on the advantages of the two techniques and produce a better and more efficient technique or model for information retrieval.

Geng et al. (2008) proposed K-nearest neighbor (KNN) for retrieving information by ranking the pages. The work employs KNN to rank different queries and conducted query-dependent ranking. An online method which created a ranking model for a given feature space and then rank the documents with respect to the query using the model. Then they used two offline approximation of the method to create the ranking models in advance to enhance the efficiency of ranking. They further prove that the approximations are accurate in terms of difference in loss of prediction if the learning algorithm used is stable with respect to minor changes in training examples.

In retrieving information using K-NN technique, the documents are classified based on their similarities. K-NN is then used to retrieve the most similar k documents for each query document. This is done by ranking the candidate categories by vote using the weighted average and then using support vector machine (SVM) to predict the categories of the document as the most voted categories. This is done by finding a set of keywords that have similarity in meanings for the k-nearest neighbor. Queries with similar meanings and contain tags are then searched using closely related keywords. An algorithm is then constructed for the retrieval of documents that are sent to the database based on the queries. As soon as a neighborhood of similar to the words in the search space is established, the algorithm considers on tagged queries or tagged keywords that are used in the query and then assign weight to them by calculating the weight for each tag which we refer to as w_t (i.e., weight for each tag). The average weight of these similar keywords is the calculated and the query tag is applied.

RESULTS

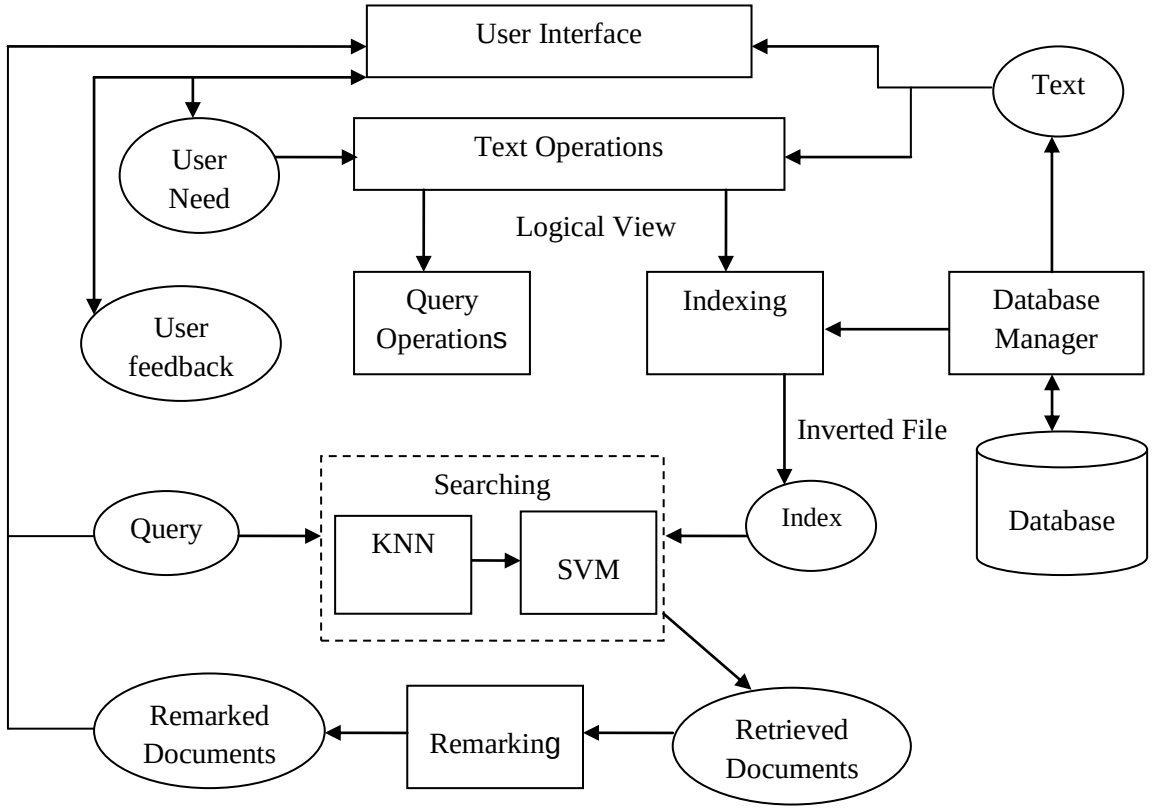


Fig. 1: Architecture of the proposed model

Document from. It has three parts: the database, the data mining application, and the front end which is also the interface where a user can submit his query and get the result of the retrieved information. It also contains the program manager and the metadata layers as shown in figure 3.2. The metadata contains the input data, the staging area, the result of the retrieved information, and the metadata.

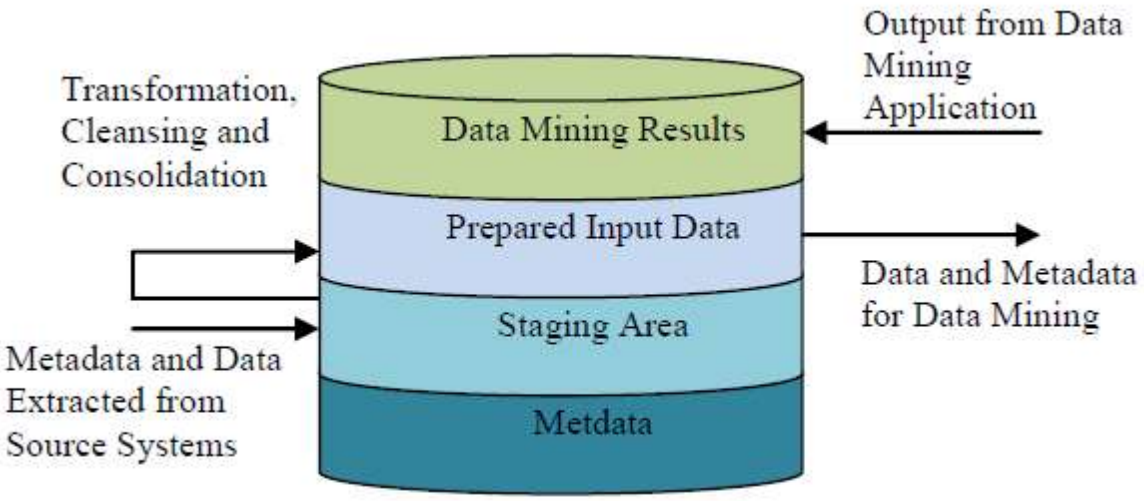
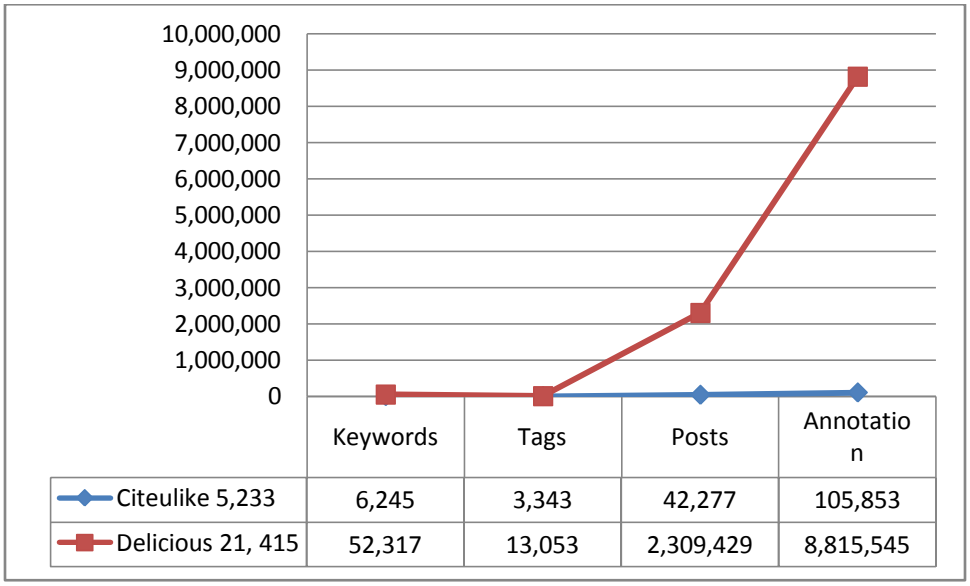


Fig. 2: Metadata layers

The data collected are shown in table 1.

Folksonomy	**Citeulike**	**Delicious**
Users	2,051	18,105
Keywords	6,245	52,317
Tags	3,343	13,053
Posts	42,277	2,309,427
Annotations	105,853	8,815,545



	Keywords	Tags	Posts	Annotation
Citeulike 5,233	6,245	3,343	42,277	105,853
Delicious 21, 415	52,317	13,053	2,309,429	8,815,545

Figure 2:

From the result, we can see that the number of users in Citeulike are 2,051, keywords is 6,245, tags is 3,343, posts is 42,277, annotations is 105,353 while for delicious, the number of users is 18,105, keywords is 52,317, tags is 13,053, posts is 2,309,429, and annotations is 8,815,545. The result shows that there are more users of Delicious and that the keywords and tags are very important in document searching and information retrieval.

Discussion

When we compared this hybridized technique with the techniques used that used KNN and Space vector machine separately, it was found that the performance of this approach is better than when the tools are used separately. That is, our technique is able to retrieve information faster with significant lesser time.

REFERENCES

Anagnostopoulos, A., Broder, A., and Carmel, D. (2005), Sampling Search Engine Results, ACM Journal of Computer Networks, pp. 245 – 256.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. and Raghavan, S. (2001), Searching the Web**,** ACM Transaction on Internet technology, Vol. 1, No. 1, pp. 2 – 43.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999), Modern Information Retrieval, ACM Press.

Barlow, L. (2004). How to Use Web Search Engines, Tips on Using Search Site Like Google, alltheweb and Yahoo.

Brandt, R. (2009). Starting Up, How Google Got Its Groove, Stanford Magazine.

Beitzel, S. M. Jensen, E. C. Chowdhury, A. and Frieder, O. (2007). Varying Approaches to Topical Web Query Classification. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 783–784, New York, NY, USA, 2007.

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Se\arch Engine, WWW7/Computer Networks and ISDN Systems, Vol. 30, pp. 107 – 117.

Broder, A. Z. (2002). A Taxonomy of Web Search , SIGIR Forum, Vol. 36, No. 2, pp. 3 – 10.

Geng, X., Liu, T. Y., Qin, T., Arnold, A., Li, H. and Shum, H.-Y. (2008). Query Dependent Ranking Using K-Nearest Neighbor, SIGIR'08, July 20-24, 2008, Singapore.