INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

Personalized Document Retrieval Using Text Mining

Meda Sai Kheerthana
Department of Computer Science Engineering
R.V. College of Engineering
Bangalore-560059, Karnataka, India.
medakheerthanaap12@gmail.com

Sushmitha K. S
Department of Computer Science Engineering
R.V. College of Engineering
Bangalore-560059, Karnataka, India.
sushmithakirani@gmail.com

Geethika .R
Department of Computer Science Engineering
R.V. College of Engineering
Bangalore-560059, Karnataka, India.
geethika012@gmail.com

Abstract: The data produces in the last two years has outweighed all the data existing up until then. Therefore, there is a need to organize and classify this information so that its retrieval is ideally relevant and smooth. The project in hand employs text mining and machine learning techniques to offer a solution to the problem. The project enables a user to upload a document and search for the document. A graphical user Interface is developed to enable a user to upload and type his search query. The documents are stored in a database. The naïve Bayesian classification algorithm is used to classify the uploaded documents into respective categories. A novel algorithm is developed based on tf - idf and cosine similarity and used for searching the database and retrieving documents relevant to the user's query by considering user's personal interests.

Keywords: Document Retrieval, Text Mining, Personalization, TF-IDF, Cosine Similarity, Personalized Search.

I. INTRODUCTION

Document retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual. User queries can range from multi-sentence full descriptions of an information need to a few words. Document retrieval is sometimes referred to as, or as a branch of, text retrieval. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. Machine learning algorithms allow for classification of documents into respective domains.

Inspired by the implicit personalization technology that has been successfully employed to improve user experience in searching the World Wide Web, this paper explores the implementation of this technology to improve the student experience in searching for documents in student teacher portal.

This paper addresses the question of whether incorporating student enrollment information, that is, the information about the units they are currently enrolled in can effectively be used as a basis for identifying their learning needs and customizing their document search results accordingly.

Motivated by the large and ever-growing volume of resources in digital libraries, coupled with students' limited experience in searching for these resources, particularly in translating their information needs into queries, this research investigates the potential of incorporating student enrollment information, that is, published information on the units/subjects they are enrolled in, to identify student's learning needs and produce personalized search results.[4]

The large investments libraries make in their collections would be best utilized if supported by effective search engines through which students could find the resources to satisfy their information needs. Unless the portal can offer efficient, customized and easy-to-use services, students may turn away from them. This would result in a significant waste of effort toward developing such portal, reducing the discovery of resources in which the university has made large financial investments. More importantly, it may result in students not using high-quality resources, as they may turn to commercial products instead of the university-related documents, the latter of which filters resources such that only relevant documents are provided.

II. RESEARCH METHOD

The objective of machine learning is to make the computer aware of the data and the environment it's working with that would further improve the expected output. The objective of the project can be divided into four components.

The first component involves developing a framework that helps in efficiently managing data. The second objective is to classify the documents when they are uploaded. Next, an algorithm has to be developed to mine data from the documents and rank them according to their relevance considering users personal information. Lastly, GUI has to efficiently manage the functioning of the server.

Initially, it was assumed that users were well informed and could effectively convert their need to a search query but these days university's digital libraries are more heterogeneous in terms of the collections they hold and their expected users. Students may be from various domains, backgrounds, skills, and interests. It is possible that the user may encounter results which are not very apt to his background and interests but relevant to his query. This occurs when a single word in the query can be linked to various domains. In order to avoid displaying results which are not apt to his interests and background, personalization is considered necessary. Personalization involves taking into consideration all the interests specified by the student in his profile and displaying only the results which are apt to his interests.

The system is broken down into smaller modules and these modules work collectively to make the system as a fully functional unit. The design for the system takes all the constraints into consideration. Only with accurate and correct design of the system can the model be realized. This section describes the general constraints in the development of the project. The constraints considered include integration of components, time taken to classify documents and to sort results according to the relevance.

The time taken to classify the documents into a relevant category must not be considerably high as it could be a hindrance to the uploader. In order to reduce this training set was reduced to half for all subsequent runs. The results should be produced with a minimum delay even if the document size is large.[6]

First, the system was looked at as a whole to see what is inputted to the system and what is outputted.The system was broken down into constituent modules. Each module performs an independent task. However, all modules have to the working sequence for the whole system to work. Each module is carefully analyzed to define its methodology.

Searching and uploading are the two major functions of the system. The phrases entered during searching and uploading are preprocessed invariably to remove stop words. In the case of uploading the database is updated.[2]

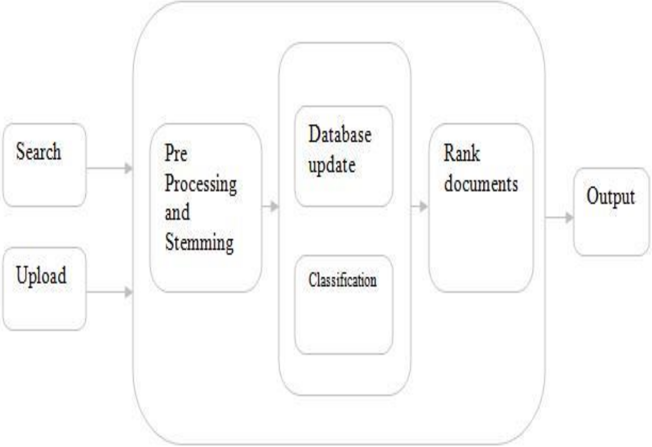


Fig 1 System Architecture

Fig 1 shows the system architecture. The documents are classified when it is uploaded using the machine learning algorithm. The query is processed in case of searching and the documents are ranked to present the user with most relevant results. The user can select from the list of relevant documents presented to him. The system is broken down into It's constituent modules. Each module performs an independent task that contributes to the overall functioning of the system. The internal working of each component is explained as follows

Graphical User Interface

Purpose: The aim of this module is to enable the user to interact with the system by proving them the required interface to search and upload documents.

Functionality: The user types the query or requests to upload via an HTML form which is sent to the server using POST request. If the uploading fails it is intimated to the user

The Searching System

Purpose: Enter the search phrase to obtain relevant results.

Functionality: This system enables the user to search for documents that have been uploaded by the teachers. Tf-idf and cosine similarity is used to search for relevant documents and the top ten documents are shown to the user.

The Uploading System

Purpose: The purpose of this system is to enable to user to upload a file.

Functionality: The uploading page enables the user to upload a file. The file can be of different formats.Once the user uploads the classification [7] system classifies the document into the respective category for providing the users with accurate results when a search query is typed.

The Database Management System

Purpose: The aim of this system is to store the uploaded documents in a respective folder.

Functionality: when the user uploads documents, it is stemmed and classified into a respective folder which is stored in the database. The user's query is preprocessed and searched with respect to the documents in the database to obtain relevant results.

Result Ranking System

Purpose: The aim of this document is to retrieve and display the list of relevant documents.

Functionality: The user provides a search phrase for retrieving relevant documents. The search query should not contain the stop words, so a novel program has been written in python to remove the stop words. A Scoring algorithm like tf-idf and cosine similarity is used to rank the documents considering various parameters including the search query [8]. The most relevant or high scored documents are sent to the user to choose from them.

Personalization

Purpose: The aim of this module is to provide the user with documents in accordance with his personal interests.[5]

Functionality: The user's interests are converted into an array and cosine similarity is found out between the retrieved documents from the ranking module and his interests.

III. RESULTS AND ANALYSIS

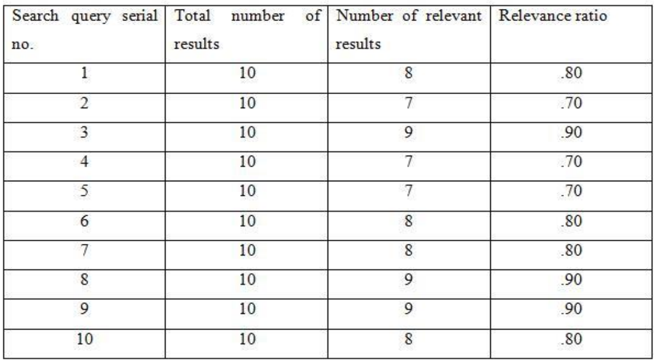
The analysis of the system is performed to determine which inputs provided the best results and how these inputs affect the overall efficiency of the system. In the testing phase, various modules of the system are tested individually by providing sample inputs and observing the outputs they produce. Evaluation metrics are the criteria used for testing the efficiency of the developed system. They are the set of predefined rules that analyze the results obtained from the various test cases to get the quantitative measure of the accuracy of the system.

The metric used in the project is the accuracy score. The accuracy score is calculated by taking the ratio of the number of relevant results to the total number of results when a search query is entered. The results are decided to be relevant based on the TF-IDF and cosine similarity score. The content of the document is also taken into consideration while deciding if it is relevant or not.

After conducting numerous tests and analyzing the results, the system was found to have accuracy 80%. This implies that whenever a search query is entered belonging to a particular category, about 80% of the results belong to that category. The better statistics would be to say that this 80 % of the results which are the relevant ones rank higher than the 20% irrelevant ones.

Thus, the system is adept at ranking the results and provides almost spot on accuracy when it comes to showing the important results before the probably irrelevant ones. The irrelevant results are displayed at the bottom due to exact keyword matches and much cannot be done about this. This is a problem associated with any search engine.

Table I Relevance ratio table



Search query serial no.	Total number of results	Number of relevant results	Relevance ratio
1	10	8	.80
2	10	7	.70
3	10	9	.90
4	10	7	.70
5	10	7	.70
6	10	8	.80
7	10	8	.80
8	10	9	.90
9	10	9	.90
10	10	8	.80

Table I shows a comparison of how many of the total results were relevant where 10 random search queries were made. The table shows a variation of the accuracy score from 0.70 to 0.90. This depends on the Tf-Idf and cosine similarity.

CONCLUSION

The Documents Retrieval System has been proposed and implemented based on machine learning and data mining techniques. This system provides relevant results based on the context of the documents. The system developed enables the user to upload documents of different formats and search for documents by providing a search phrase. In an institution, this system would benefit the teachers and students to a great extent. The faculty can upload important documents to be shared with the students. The documents can be in picture format as well. The students can input a search query to find relevant documents. The list of relevant documents is displayed to the student. This system can serve as a platform exclusively for this purpose. Thus, a student need not have to surf the net and go through tons of useless data to find the relevant document. Furthermore, the faculty needs to upload the notes only once, it is stored permanently. Hence, the system provides great use.

ACKNOWLEDGEMENTS

We thank R.V. College of Engineering for giving the opportunity and providing the required resources for working on this project.

REFERENCE

[1]. Palvi Arora and TarunBhalla, "*Recommendation based on Deduced Social Networks in an educational digital library.*" Paper presented at the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL).

[2]. AMINI, B., IBRAHIM, R., OTHMAN, M. S., & RASTEGARI, H. (2011). *Incorporating scholar's background knowledge into recommender system for digital libraries.* Paper presented at the 2011 5th Malaysian Conference in Software Engineering (MySEC), Johor Bahru, Malaysia.

[3]. ASHRAF, T., & GULATI, P. A. (2010). Digital Libraries: A Sustainable Approach. In T. Ashraf, J. Sharma & P. Gulati (Eds.), *Developing Sustainable Digital Libraries: Socio-Technical Perspectives* (pp. 1-18). Hershey, PA: IGI Global.

[4]. BRUSILOVSKY, P., FARZAN, R., & JAE-WOOK, A. (2005). *Comprehensive personalized information access in an educational digital library.* Paper presented at the Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. JCDL '05.

[5]. CALLAN, J., SMEATON, A., BEAULIEU, M., BRUSILOVSKY, P., CHALMERS, M., RIEDL, J.TOMS (2003). *Personalization and Recommender Systems in Digital Libraries*: Joint NSF-EU DELOS Working Group Report.

[6]. JÄRVELIN, K., & KEKÄLÄINEN, J. (2002). *Cumulated gain-based evaluation of IR techniques.ACM Trans. Inf. Syst., 20*(4), 422-446. doi: 10.1145/582415.582418

[7]. JOHN, G. H., & LANGLEY, P. (1995). *Estimating continuous distributions in Bayesian classifiers.* Paper presented at the Proceedings of the Eleventh conference on Uncertainty in artificial intelligence.

[8]. JOMSRI, P., SANGUANSINTUKUL, S., & CHOOCHAIWATTANA, W. (2012). A Personalized Reranking Technique for Academic Paper Searching Based on User Profiles. *International Journal of Digital Content Technology and its Applications (JDCTA), 6*(16). doi: 10.4156/jdcta.vol6.issue16.62