ISSN 2395-1621

Document Relevance Ranking Using Machine Learning.

[#1]Aifaz Khan, [#2]Pranav Hondrao, [#3]Pranjal Dimbale

[1]ayanaifazkhan@gmail.com
[2]pranjaldimbale9890@gmail.com
[3]pranavhondrao@gmail.com

[#123]Department of Computer Engineering

Dhole Patil college of Engineering, Pune.

ABSTRACT

The paper reviews and analyzes existing solutions for determining the meaning of text documents. Typical questions and problems of determining the meaning of text documents used in modern systems are discussed. The most used models and methods of semantic text processing are considered, as well as the classical process of text processing in semantic analysis. In the course of the analysis, it was found that when using the models considered, in practice, a partial loss of meaningful meaning of the text occurs, which in turn is not always justified, although it allows performing certain procedures of semantic analysis, although it misses the features. The most frequently used libraries of semantic analysis are analyzed in different programming languages such as Java, Rybi, C , PHP. It should be noted that the considered libraries of semantic analysis operate in the same way, the results of their work are quite similar. Therefore, when developing, it is necessary to take into account specific goals and tasks that the developer must solve, as well as in what functional-style types of texts are analyzed. It is also expedient to perform a statistical analysis of the work of these libraries, on the same tasks and texts. It was found that it is advisable to develop a single system for solving the problem of analyzing and evaluating texts according to different criteria, for example, analyzing texts for text borrowings and borrowing ideas, as well as for determining the authorship of the text, taking into account its paraphrasing.

Keyword-
Definition of the meaning of the text Semantic analysis, Latent-semantic Analysis., Text Analysis, Text Fragment.

ARTICLE INFO

Article History

Received: 9th December 2019

Received in revised form :

9th December 2019

Accepted: 11th December 2019

Published online :

12th December 2019

I. INTRODUCTION

Due to the Huge domains in the IT industry, it's very difficult to find Ranking relevance of classify the domains from a huge number of PDFs. We are going to generate a short Description of that paper so that we can save the time to find out their relevance. Ranking of queries is one of the fundamental problems in information retrieval. Ranking relevance is that the work of uncertain documents into classes supported their content. There area unit several classification and ranking strategies for documents. The purpose of the ranking algorithm is to retrieve from the collection of documents the most relevant ones Ranking directly affects retrieval quality. For each document, a score is computed which reflects his relevance concerning the given query. Then, documents are ranked according to this score and returned as a result.

This project uses machine learning techniques for classification and ranking the relevance of documents. Machine Learning enables systems to recognize patterns based on existing algorithms and data sets and to develop adequate solution

concepts. Machine Learning undoubtedly helps people to work more creatively and efficiently. Therefore, in Machine Learning, artificial knowledge is generated based on experience. In this project, by classifying a document, one or more categories are assigned to a document, making it easier to manage and sort then find ranking. This is especially useful for publishers, news sites, blogs or anyone who deals with a lot of content**.**

Machine Learning uses different algorithms to train systems like Support Vector Machine (SVM), Naïve Bayes, K-nearest neighbor (KNN), TF-IDF Decision tree, K-means, etc. In this system, the Naïve Bayes algorithm is being used by comparing different algorithms, Naïve Bayes shows the highest efficiency and it is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. TF_IDF these algorithms suggest the frequency of a particular document.

II. LITERATURE SURVEY

Paper Name:- Classification of text documents based on naïve Bayes using n-gram features
Author Name:-. Mehmet Begin
Year:- 2018
Paper Description:- Text and document classification processes are often used in areas such as sentiment analysis, text summarization, etc. The author Mehmet Baygin has performed document classification using the Naive Bayes approach.
Paper Limitations:- Need to 2-gram, 3-gram and 4-gram features of all documents were extracted and the test procedures were performed separately for each feature.

Paper Name:- An outcome-based comparative study of different text classification algorithms
Author Name:- 1.A.Helen Victoria, 2.M.Vijayalakshmi
Year:-2018
Paper Description:- Text classification has become more important due to the growth of big data with which we could obtain huge data daily. It has many applications like information retrieval, spam detection, language identification, sentiment analysis and plays a major role in natural language processing as well
Paper Limitations:- The Decision trees require the features that have to be checked in a specific order, which limits their ability to exploit features that are relatively independent of each other.

Paper Name:- Ranking of Text Documents using TF-IDF Weighting and Association Rules mining
Author Name:- Siham JABRI, Azzeddine DAHBI
Year:-2018
Paper Description:- In this paper, we have tried to explore a novel ranking model which provides the result-set according to the user query by keeping the vector space model and using an association rules technique based on dominance relations presented in

Paper Name:- Document clustering: tf-idf approach
Author Name:- 1.Prafulla Bafna, 2.Dhanya Pramod, 3.Anagha Vaidya
Year:-2016
Paper Description:- Clustering techniques can be applied only to structured data. So unstructured data need to be converted to structured data. But while converting unstructured data into structured data the algorithm efficiency decreases. So to increase the efficiency For this, we are going to use the TF-IDF approach.
Paper Limitations:- Need to use better semantic relativity concepts that might be domain-specific but provide better results.

Paper Name:- Text classification and classifiers: a survey
Author Name:- 1.Vandana Korde, 2.C Namrata Mahendar
Year:-2012
Paper Description:- In this paper, Namrata Mahender and Vandana Korde have tried to give the introduction of text classification its process and also the overview of classifiers. They also tried to compare some existing classifiers.The existing classification methods are compared and contrasted based on various parameters. They also found that the performance of the classification

algorithm is greatly affected by the quality of data source

Paper Limitations:- From the above discussion, it is understood that no single representation scheme and classifier can be mentioned as a general model for any application.

III. PROBLEM STATEMENT

To do text Summarization and Processing of Large Documents providing the facility of the text summary.

IV. BLOCK DIAGRAM

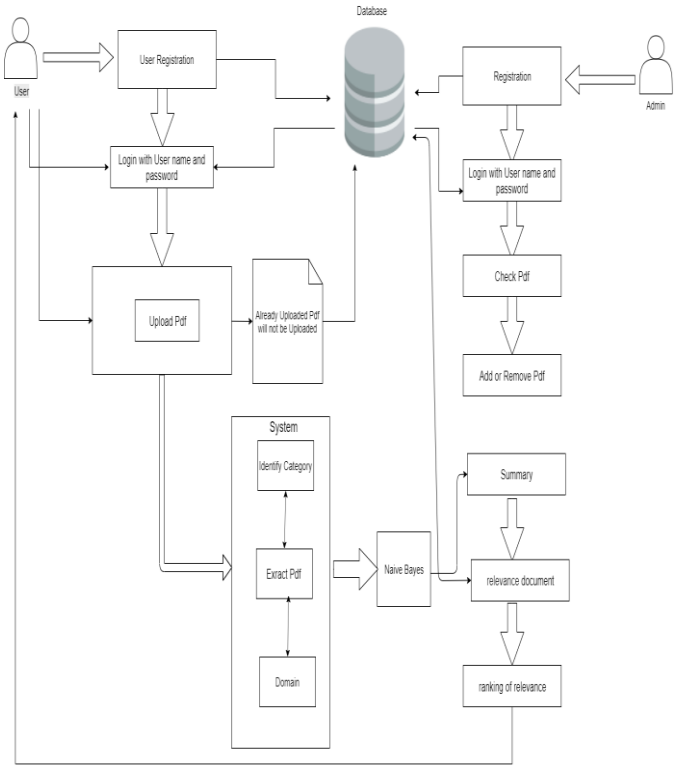

Fig 1. System architecture

V. CONCLUSION

In this system, we have developed an application by which we can identify the domain of our document. And also relevance. Due to the use of our system, we can easily be finding the domain of our document which is useful business or education purposes.

This paper expressed the extraction of fields document related to IT research paper. It applied the naive Bayes algorithms to classify documents automatically. This classifier gives a correct and accurate result.

REFERENCE

[1]. Mehmet Baygin 2018 "classification of text document based on naive bayes using N-grams Features"

[2].Prafulla Bafna,Dhanya Pramod,Anagha Vaidya(2016)"Document clustering:TF-IDF Approach“

[3]A.Helen Victoria, M.Vijayalakshmi.An Outcome based Comparative study of different Text Classification Algorithm. Volume 118 No. 22 2018, 1871-187

[4]Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

[5]Badgujar, M. G. V., & Sawant, K. (2016). Improved C4. 5 Decision Tree Classifier Algorithms for Analysis of Data Mining Application. International Journal, 1(8).

