Document clustering and Ranking

1. Project Description

An application to collect the documents, form groups out of the given documents based upon all the users query previously, later retrieve the exact and the most relevant document(s) among the clusters, depending upon the search query from the user. We are expecting to provide the user with the most relevant document for the given query through this application.

2. Existing System

The existing system has a facility that is when given a query; it has to search each and every line of the document for the particular word given by the user. This process takes up long duration for searching each and every line of the document and we know that every user expects to retrieve the search result in less time but the existing system is taking a long time for the activity as mentioned above.

3. Proposed System

We propose this work of document clustering and ranking, which is an automatic grouping of documents (.pdf) into clusters. Retrieving the information and documents as per the user requested query which are relevant to what the user is expecting and also help tracking the frequently searched keywords and maintaining the cache.

4. Analysis

Our proposed system includes the following activities:

a) Login –

This module helps the valid person to enter this application in order to provide with the genuine documents and to retrieve the required information based on the query posed.

b) Inserting the document -

In this we remove the stop words and sort the remaining words in the chronological order. To these words we will maintain separate page count of every word. The process of maintaining the page count is the index of the words of the document which will allow us to retrieve the document as quick as possible.

c) Searching the document –

When the user gives the query, then indices are searched and the book containing the highest frequency of the page count is placed at the top of the result, to which the user can access. This process is predicted to take less time as the indices for each word is already maintained.

5. Project Objectives and Expected Business Benefits

The main objective is to reduce the time taken to retrieve the documents as per the user expected query and also provide user with the most relevant information.

Business Need	***Solution***	***Expected Benefits***
Time reducing and maintain frequent updates of documents	Maintain the updated set of the documents and use the relational database as the cache memory to store the frequency of the given word in the document.	Time for processing the documents would be reduced by using the cache memory and the accurate results will be obtained as we maintain the updates documents into this application.

6. Technologies Used

- Java
- Hadoop Distributed File System
- JSP
- HIVE

7. Key Requirements

6.1 Software requirements

- Linux
- JAVA 1.8

6.2 Hardware requirements

- 8 GB RAM
- 1.7 - 2.4 GHz processor speed
- 500GB HDD

8. Project Lifecycle

Developing the project involves software development life cycle which has the following phases like planning, analysis, design, integration, implementation and acceptance stages in iterative manner. Ability to use key performance metrics to track resource efficiency from the project initiation to completion and enhance the project in the best way possible.