PROJECT PROPOSAL

-SAI AKASH KUTHURU

(sxk3190)

Project Proposal

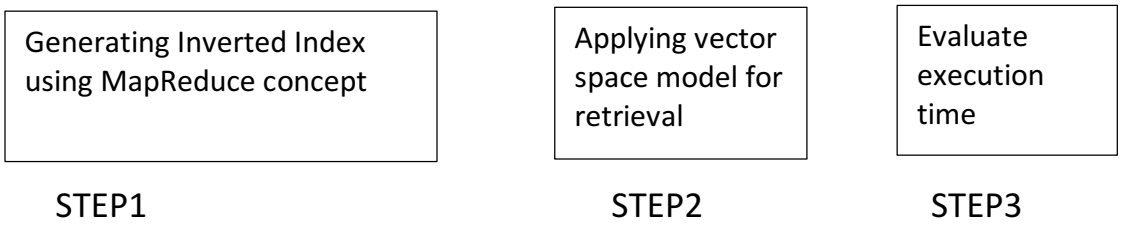
Aim

Creating an Information system for a large dataset and study the execution time of the information system by implementing vector space model.

Problem

Due to large dataset implementation of information system results in serious lack of execution of time. In order to decrease execution time we adopt MapReduce concept of Hadoop.

Sequence of Steps



STEP1:

Inverted Index : Inverted index is data structure which stores index words appearing in the documents which are unique, along with it stores the frequency of the terms appearing in all the documents and the document Id's for that term which is known as posting list's. While creating inverted index we read the data us Map() procedure where we filter and sort data and Reduce() method performs counting or frequencies.

STEP2:

What is vector space model?

Representation of a set of documents and given query as vectors in a common vector space is known as the *vector space model.*

Term Frequency

- Term Frequency for term in a document is defined as how often a term t has occurred in that particular document.
- It is given by the formula $Tf_{t,d}= (1+\log(f_{t,d}))$ for $f_{t,d} >0$

Inverse Document frequency

- Inverse document frequency for the term t is defined as term's scarcity across all the documents.
- It is an inverse measure of the informativeness of t
- It is given by the formula $Idf_{t,d}=(\log_{10}(N/n_t))$.

N:-Total collection of documents.

n_t:-Documents containing term t.

Weighting :- $(1+\log(f_{t,d}))* (\log_{10}(N/n_t))$.

Similarity: It is defined as the cosine angle between query vector and document vectors in which the terms in the query are present.

Is is given by the formula $\text{Cos}(\vec{q},\vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}|*|\vec{d}|} = \frac{\sum_{i=1}^{|v|} q_{i*}d_i}{\sqrt{\sum_{i=1}^{|v|} q_i^2 \times d_i^2}}$

For a given query we find out terms which are present in the documents initially and we find the similarity between the query and all the documents in which the query terms are present in.

Here we convert any dimension vector into unit vector so that similarity would be justified as we are using vectors of unit length.

STEP3

Calculate the starting and ending time which gives the total execution time.

Technical Specifications

Software proposed to use is java SDK and Hadoop on dataset.

Expected Completion

Estimated Project Submission Date: **November 30th 2015**

Tasks to be performed include the following:

Task	**Start Date**	**Estimated Date of Accomplishment**
Indexing and Implementing MapReduce concept on dataset	11/26/2015	11/7/2015
Applying vector space Model	11/8/2015	11/14/2015
Evaluation	11/15/2015	11/21/2015

Results Expected

To show the execution time in series execution and parallel execution (using Hadoop).