Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

Research Paper

A Comparative Study on Ranking the Documents

Mathematics

KEYWORDS : Latent Semantic Indexing, Random walk, Novel ranking algorithm

Dr. P. Sunitha | Department Of Mathematics, Anna University of Technology Tiruchirappalli

ABSTRACT

During the past two decades researchers have drawn much attention in ranking the documents using various techniques. Xiaojin Zhu et al (2007) have introduced a novel ranking algorithm called Grasshopper to improve diversity in ranking. Benyu Zhang et al (2005) used affinity graph for improving web search results. Most of the results derived here is an integral part of stochastic process involving m-state Markov chains. The fundamental matrix plays an important role in determining the expected number of visits to evaluate the rank. The aim of this paper is to rank the document by Latent Semantic Indexing.

INTRODUCTION:

The underlying idea in many natural language processing is that the aggregate of all the words contexts in which a given word does and does not appears provides a set of mutual constraints that largely determines the similarity of meaning of words and passages by analysis of large text corpora.

Latent Semantic Indexing is a theory and method for extracting and representing the contextual- usage meaning of words by statistical computations applied to a large corpus of text (Landauer and Dumais, 1997). It is an indexing and retrieval method that uses a mathematical technique called Singular value decomposition which identifies the patterns in the relationships between the term and documents contained in an unstructured collection of text.

In this a document is conceptually represented by a matrix consisting of keywords extracted from the document, with associated weights representing the importance of the keywords in the document and within the whole document collection ; likewise, a query is modeled as a list of keywords with associated weights representing the importance of keywords in the query.

THE MODEL:

It begins by constructing a term document matrix to identify the occurrences of the m-unique terms within a collection of n documents. In a term document matrix, each term is represented by a row, and each column contains the document, with each matrix cell, aij, representing the no of times the associated term appears in the indicated document. This matrix is generally very large.

The mathematical technique Singular value decomposition is used to decompose the given term document matrix into three new matrices as with the following identities to be satisfied: $U^TU=I_{mxm}$ and $V^TV=I_{nxn}$. Here S is a rectangular matrix with the same dimensions as A consisting of singular values along the diagonal arranged in descending order.

The Latent Semantic Indexing modification to a standard Singular value decomposition is to reduce the rank to size k, effectively reducing the term and document matrix sizes. The Singular value decomposition operation along with this reduction has the effect of preserving the most important semantic information in the text while reducing other undesirable terms of the original space of A. This reduced matrices is denoted by $A_{K=UKSK}$.

$$A_K = U_K S_K V_K^T .$$

The computed UK and VK matrices define the term and document vector spaces, which with the computed singular values SK, represents the conceptual information derived from the document collection. The similarity of terms or documents within these spaces is a factor of how close they are to each other, typically computed as a function of the angle between the corresponding vectors denoted by

$$\text{Sim}(q,d) = \frac{q.d}{|q|.|d|}$$

$$\text{where } q = q^T U_k S_k^{-1} \text{ and } d = d^T U_k S_k^{-1}$$

Depending on the cosine similarity value, ranking is given to the items i.e., top rank is given to higher similarity value and so on.

Here we review the GRASSHOPPER algorithm which is most similar to Latent Semantic Indexing approach. Zhu et al(2007) introduced a novel ranking algorithm called GRASSHOPPER which ranks the items in the documents.

A state Si of a Markov chain is called absorbing if it is impossible to leave it ie, pii=1 and a Markov chain is absorbing if it has at least one absorbing state and if from every state it is possible to go to an absorbing state. In an absorbing Markov chain a state which is not absorbing is called transient.

Initially a nxn transition matrix was formed by normalizing the rows using where are the weights of the given transition matrix.

The walk was made into a teleporting random walk P by $P=\lambda P^{-}+(1-\lambda)1r^T$ where $1r^T$ is the outer product of and an all-1 vector.

$$\left(\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n}\right)^T$$

Chapman Kolmogrov equation can be used to determine the stationary distribution. Consider the state with largest probability value from the stationary distribution as the first rank g_1.

For an absorbing Markov chain, rearrange the states so that the transient state come first. If there are t transient states and r absorbing states the transition matrix will have the following canonical form

$$P = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix}$$

where the first t states are transient and last r states are absorbing. Here I is an rxr identity matrix, 0 is an rxt zero matrix, R is a non zero txr matrix and Q is an txt matrix.

For an absorbing Markov chain P, the matrix $N= (I-Q)^{-1} = I+Q+Q^2$ +----- is called the fundamental matrix for P. The entry nij of N gives the expected no of times that the process is in the transient state si. Average over starting states to obtain

$$V = \frac{N^T 1}{n - |G|}$$

where V_j is the expected number of visits to j. The state with the largest expected number of visits is selected as the next item

$$g_{|G|+1} = \arg max_i v_i$$

This process is repeated until all items are ranked by turning the ranked items into absorbing state.

Volume : 2 | Issue : 12 | December 2013 • ISSN No 2277 - 8179

DISCUSSION:

From the review, it is clear that we must select and rank sentences originating from a set of documents about a particular topic. Our ultimate aim is to produce a summary that includes all conceptual facts avoiding repetition. Even though GRASSHOPPER algorithm is similar to Latent Semantic Indexing technique, we can conclude that Latent Semantic Indexing is more advantageous than GRASSHOPPER. The fact is that Latent Semantic Indexing is applicable for a rectangular matrix while GRASSHOPPER is applicable only for a square matrix. The other fact is that in GRASSHOPPER algorithm transition from one state to other is involved where such restrictions are not involved in Latent Semantic Indexing technique. So Latent Semantic Indexing provides a unified better approach for ranking the multi document.

REFERENCE

1. Xiaojin Zhu et al (2007), Improving Diversity In Ranking Using Absorbing Random Walks. | 2. Benyu Zhang et al (2005), Improving Web Search Results Using Affinity Graph. | 3. Landauer and Dumais(1997), Latent Semantic Analysis Theory of the Acquisition, Induction | and representation of Knowledge | 4. I.J Good, (1969). Some applications of the singular decomposition of matrix, Technometrics, Vol 11,no 4,pp 823-831. |