IRSTI 20.23.19

N. Abdinurova[1], G. Tolebi[1], A. Talasbek[1]
[1]Suleyman Demirel University
Kaskelen, Kazakhstan

FEATURES AND ADVANTAGES OF VECTOR SPACE MODEL IN SEARCH ENGINE DESIGN

Abstract. This paper is dedicated to present a design of our project - search engine, which based on one of retrieval models, which have been deeply learned and selected among the several searching models in previous well-known researches. We have explained couple of reasons for choosing certain model, its properties, some advantages, and techniques used to calculate document's relevance to our information need. The principles of ranking algorithm had been interpreted, and the formula for calculating similarity depending on vector values of terms is shown in the part with techniques explication. In conclusion part some benefits of given model are listed and a definite derivation have been drawn.

Key words: Vector Space Model, Lucene, indexing, search engine, relevance, ranking, similarity.

***

Аңдатпа. Бұл мақалада іздестіру модельдерінің біреуіне негізделген іздеу жүйесінің дизайны келтірілген, модель көптеген бастапқы іздеу модельдеріне арналған зерттеулер арасында қарастырылып, іріктелді. Біз белгілі бір модельді таңдаудың бірнеше себептерін, оның қасиеттерін, кейбір артықшылықтарын және біздің ақпараттық қажеттілігімізге құжаттың маңыздылығын есептеу үшін қолданылатын әдістерді түсіндірдік. Ранжирлеу алгоритмінің қағидаттары түсіндірілді және терминдердің векторлық мәніне байланысты ұқсастығын есептеу формуласы техниканы түсіндіру бөлігінде көрсетілген. Қорытындыда берілген модельдің кейбір артықшылықтары келтірілген және белгілі бір қорытынды жасалды.

Түйін сөздер: Векторлық модель, Lucene, индексирование, іздеу жүйесі, өзектілігі, орналасу, ұқсастық.

***

Аннотация. Настоящая статья посвящена описанию процесса разработки авторского проекта – поисковой системы, основанной на одной из моделей поиска, которые были глубоко изучены и отобраны среди нескольких поисковых моделей в предыдущих известных

исследованиях. Мы объяснили пару причин выбора определенной модели, ее свойств, некоторых преимуществ и методов, используемых для расчета значимости документа для нашей информационной потребности. Принципы алгоритма ранжирования были интерпретированы, а формула для вычисления подобия, зависящая от векторных значений термов, показана в части с объяснением методов. В заключение перечислены некоторые преимущества данной модели и определен определенный вывод.

Ключевые слова: Векторная модель, Lucene, индексирование, поисковая система, релевантность, ранжирование, сходство.

Introduction

After development of the World Wide Web people has started to fulfil their information needs with internet. Demand in retrieving spawned the need in searching programs. For this purpose, search engines were created starting from 1992. “Search engines are programs that search documents for specified keywords and return a list of the documents where the keywords were found. A search engine is really a general class of programs; however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web” [1]. From the latest statistics on web developing of China, published in eighteen July, 2007 by CNNIC, it founded that the percentage of using searching systems service is about 74.8 %, which surpasses the percentage of using Email service which is 55.4 %. However, how the information to be retrieved can be rated as relevant and non-relevant? There are several models for information retrieval, which predict and retrieve documents which a user will find relevant respect to a certain query. Depending on that which mathematical model used while ranking, the development of rank algorithm can be considered in 3 steps. The first is the Boolean model, which looks for exact matching. The next stage is the VSM (Vector Support Model). Considering the limitations of the Boolean model, the VSM uses the multi-value method. In other words, the value of keyword ti in document dj is set between 1 and 0. The last stage is the hyperlink analysis model, which is mostly used in the well known SEs currently. In this model, the most famous algorithm is the PageRank algorithm which has been proposed by Lawrence Page and Sergey Brin in 1998 [5].

For this paper we have decided to use Vector Space Model (VSM), since it has high speed for operating with and we will explain techniques and other benefits of using it.

Techniques used for searching and ranking results

The first step in search engine design is to index files to facilitate fast and accurate information retrieval. Lucene Index files source code will be used to index documents. This code uses Standard Analyzer to break up text into

indexed tokens, a.k.a terms, then converts tokens to lowercase and filters out stop words [2].

The user will enter their queries by using Lucene query syntax. Then we will use QueryParser to decode Lucene query syntax to Query object. The QueryParser is constructed with an analyzer used to interpret the query text in the same way the documents are interpreted such as down casing and removing stop words. The Query object contains the results from the QueryParser which will be passed to the searcher [3].

Then the searcher should retrieve documents' list with most relevant one on top. We will use VSM model to evaluate relevance of documents respect to the query.

In VSM [4], set of documents and query are represented as vectors in a common vector space. To find similarity between query and document first we need to know how to represent them as vectors.

Suppose we have set of n terms {t1, t2,. . . , tn}

Document represented as a vector: d =< d1, d2,. . . , dn>

di = weight of term ti in document d (e.g., based on tf x idf)

tf-idf - a combined weight for term t in document d

tf-idf = tf * idf

tf – frequency of occurrence of term in document

idf – inverse document frequency for term t

df – document frequency for term t

N – total number of documents

idf = log (N/df)

Query represented as a vector: q =< q1, q2,... , qn>

qi = weight of term ti in query q (e.g., 1 if tiϵ q; 0 otherwise)

Now we have query and documents as vectors. Next step is to evaluate similarity of each document with query and order them by relevance. Ranking (similarity) based on the angle between the query and document vectors. Distance between vectors is not applicable to use here, because large documents can have more relevant information than smaller ones, but will have more distance and therefore will be rated as less relevant.

Ranking function:

$$\mathrm{R(d,q)} = \mathrm{sim}(\vec{d},\vec{q}) = \cos\alpha = \frac{\vec{d}\times\vec{q}}{\sqrt{\vec{d}^2}\times\sqrt{\vec{q}^2}}$$

Document with maximum R will be considered as most relevant.

Reasons for choosing VSM retrieval model is that:

Simple and fast to compute;

Usually as good as known ranking alternatives;

Term-weighting improves quality of the answer set;

Partial matching allows retrieval of docs that approximate the query conditions;

Cosine ranking formula sorts documents according to degree of similarity to the query;
Document normalization is naturally built-in into the ranking

Conclusion

This paper proposes an algorithm for calculating similarity of documents from set with our information need. Comparing all other models, VSM have been chosen to be investigated for this paper. Couple of advantages, such as high speed, strong and accurate algorithm and especially, built-in document normalization was identified during current research. For future works it can be considered to analyze other models or to improve VSM algorithm.

References:

1 Codetpoint.com [online resource] / Search Engines. - URL: http://codetpoint. com/nielit-ccc/ccc-www-and-web-browsers/6-4-search-engines/;

2 Howstuffworks.com [online resource] /How Internet Search Engines. - URL: https://computer.howstuffworks.com/internet/basics/search-engine.htm;

3Lucene.apache.org [online resource] / Lucene 4.0.0 – URL: http://lucene.apache.org /core/4_0_0/demo/

4 Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008. – P.102

5 Yong Zhang, Long-bin Xiao, Bin Fan. The Research about Web Page Ranking Based on the A-PageRank and the Extended VSM // Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008 – Volume 4 – P. 223-227.