NAME
Plamen Kozhuharov
(Bulgaria)

ADVISOR
O.Univ.-Prof. Dr.Dr.h.c.mult.
Bruno Buchberger

COMPANY
hagenberg software

fh OBERÖSTERREICH UNIVERSITY OF APPLIED SCIENCES

JOHANNES KEPLER UNIVERSITY LINZ

softwarepark hagenberg

ISI HAGENBERG
International Studies in Informatics

Text Integrated Similarity Retrieval Using Latent Semantic Analysis

Contextual Integrated Similarity
Information Retrieval
Latent Semantic Analysis (LSA)
Semantic Search
Corporate Solution

Nowadays, information is everything. Although an efficient search mechanism in this data is required, the existing search engines mostly provide basic services. For more relevant results, one needs to use semantics for searching. The algorithm called Latent Semantic Analysis (LSA) is a technique for retrieving semantic relationships between a set of documents and the terms they contain.

Introduction

Nowadays information is the key answer to everything. Although a very large amount of data is made accessible to everyone, one of the most challenging questions is how can one find the required knowledge? This immediately leads to the problem of searching.

Motivation

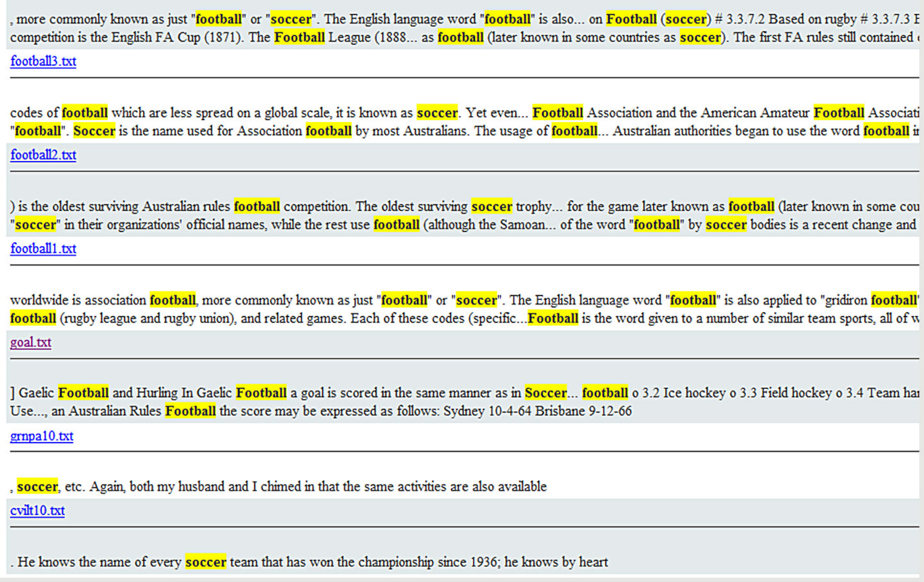
Considering a big server which collects data from hundreds and thousands of people, the question of how to search and retrieve more relevant information appears. When one is searching for a particular word, only the information containing it is sufficient, but it can be the case that one may want to retrieve texts in which this word does not appear but words with similar meaning do appear. The following examples are aimed to describe this aspect in greater details. If one searches for information about football: by typing in the word “football” in the search box, a lot of information will be returned. But suppose that one is interested to find information about the football in the USA. However, the word used for “football” in USA is “soccer” and one might not know this. Consequently, little or even no relevant information will be provided, because the search engine is looking for a sequence of symbols only. So our problem is, when we have a big amount of data, to search in this information and to find the given word and all words with similar meaning. We need a search engine that uses the meaning of words to improve the quality of the retrieved information. We need a semantic search engine.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method for analyzing relationships between documents and the words they contain, by statistical computations applied to a large corpus of files. The simple idea is that the meaning of a text can be presented as a linear combination of the meanings of the words it contains. Also the meaning of a word can be assumed from all the contexts it appears in. Therefore, reversely, the meaning of each word can be represented as LSA produces measures of word-word, word-text and text-text relations, which are very similar to human’s measures for the same examples. These correlations show close similarity between the information extracted by LSA, and the way people’s understanding of the words’ meaning is influenced by their previous knowledge. A practical consequence of this is that LSA allows to closely approximate human judgments regarding the similarity of the meanings between words.

Conclusion

All in all, the aim of this thesis is to give an alternative solution to the semantic search problem, based on the Latent Semantic Analysis method. This method is built upon a large amount of information. This work proves that LSA gives good results. As a conclusion LSA works and it brings the similarity to a different level.



Search

<-- Upload More Files

football Search

Search took 00:00:03.4250000. Found 10 items.

football.txt

, more commonly known as just "**football**" or "**soccer**". The English language word "**football**" is also... on **Football** (**soccer**) # 3.3.7.2 Based on rugby # 3.3.7.3 E
competition is the English FA Cup (1871). The **Football** League (1888... as **football** (later known in some countries as **soccer**). The first FA rules still contained

football3.txt

codes of **football** which are less spread on a global scale, it is known as **soccer**. Yet even... **Football** Association and the American Amateur **Football** Associati
"**football**". **Soccer** is the name used for Association **football** by most Australians. The usage of **football**... Australian authorities began to use the word **football** i

football2.txt

) is the oldest surviving Australian rules **football** competition. The oldest surviving **soccer** trophy... for the game later known as **football** (later known in some cou
"**soccer**" in their organizations' official names, while the rest use **football** (although the Samoan... of the word "**football**" by **soccer** bodies is a recent change and

football1.txt

worldwide is association **football**, more commonly known as just "**football**" or "**soccer**". The English language word "**football**" is also applied to "gridiron **football**"
football (rugby league and rugby union), and related games. Each of these codes (specific...**Football** is the word given to a number of similar team sports, all of w

goal.txt

] Gaelic **Football** and Hurling In Gaelic **Football** a goal is scored in the same manner as in **Soccer**... **football** o 3.2 Ice hockey o 3.3 Field hockey o 3.4 Team har
Use..., an Australian Rules **Football** the score may be expressed as follows: Sydney 10-4-64 Brisbane 9-12-66

grnpa10.txt

, **soccer**, etc. Again, both my husband and I chimed in that the same activities are also available

cvilt10.txt

. He knows the name of every **soccer** team that has won the championship since 1936; he knows by heart