Applied Concept Modeling Techniques for Semantic Data Retrieval

Đorđević, Nemanja; Meštrović, Nenad; Mijović, Đorđe; Dimovski, Boris; and Stančin, Sara

Abstract – The *main topic of this paper is a general overview of currently applicable concept modeling approaches for extracting useful data from large quantities of text that can be accessed nowadays. The paper will begin by discussing the main ideas of concept modeling and the usability it can provide in order to retrieve the actual context from raw texts. Paper also includes current implementation overview as well as a few examples that present the usage of the system, as well as mentioning of issues that yet need to be resolved in the future.*

Index Terms – *concept modeling, semantic, text mining*

1. Introduction

The Internet and the increasing development of powerful computer systems both for personal and professional usage led to unimagined increase in the amount of available data in various forms. While database systems evolved the capability to provide flexible, fast, and scalable data management, the problems with large texts as data storage units did not diminish. On contrary, with emerging of fast Internet search systems such as Google®, that provide mechanisms to search a vast amount of data syntactically, the need to semantically structure the text has even more decreased. However, the need for semantic text mining did not decrease and users lose much more time doing semantic searches via text-based search tools, which can sometimes be very inefficient. One solution is to the problem is semantic text mining. Semantic text mining is a technique that allows us to create meaningful relationships between abstractions within the text. The created relationships are in fact representations of brand new abstractions. In this paper, we will often refer to these abstractions as concepts, although concepts do not have a limitation on a number of abstractions that they bind. A concept is virtually any meaningful abstraction, which can exist without dependencies towards other abstractions. The idea of the project is to create and extract as many concepts from texts as possible, as long as they deserve to be called a concept (i.e. to have a meaning). In the remainder of the paper we will introduce some observations that make the concept creation process possible and usable.

2. Text characteristics overview and its influence on concept modeling

Generally, texts come in various forms. Besides conforming to language standards like grammar and vocabulary there is nothing else that can be used to structure a text in some generic manner. However there is one characteristic that is common to almost all known world languages, and that is the following: abstractions that form a concept are related not only by meaning, but by distance of words within the text. Concepts are therefore made of abstractions, or better say words, that are adjacent. The idea of concept modeling is to establish connections not only between abstractions to form concepts, but to form connections between concepts as well. Therefore the idea is not only to search for words that are adjacent to the observed

Manuscript received May, 27, 2009.

Đorđević, Meštrović, Mijović and Dimovski are with the School of Electrical Engineering, University of Belgrade, Serbia.

Sara Stančin is with the University of Ljubljana, Slovenia.

word, but also words that appear in its wider neighborhood. Multiple appearances of a candidate concept increase its chance of being promoted to actual concepts. Although the idea is language-independent, there are some issues that make the problem actually bound to a specific language, or better say, to the structure of a that language. There is always a certain probability that we can extract abstractions that do not have a meaning and therefore do not form a concept. In order to increase our chances of retrieving useful content, the text must undergo a filtration stage. The filtration stage is necessary in order to remove language elements we certainly know that will not help us form true concepts. Most of these elements are binders, propositions, pronouns, separators, numbers etc. Besides increasing the chance of retrieving concepts, the side effect of this stage is also noticed as a performance boost to the complete search process, because fewer words need to undergo further processing. The next chapter will introduce some techniques that are used in the implementation. Note that in this stage of research only double-worded concepts are established.

3. Exploiting the idea

a. Implementation details

In order to achieve the desired effects with the filtering stage mentioned in the text above, we need to provide the engine a list of words that need to be filtered and excluded from the search from that point onwards. This is the part of the solution that is language-dependent, so if there is need to make the engine run for languages other than English, one has to assemble a list of the unwanted words (commonly called stop-words), as mentioned above. This particular implementation is based on English texts only, in order to simplify the research process by using only the standard ASCII character palette. The engine was given a list of around 500 words that need to be excluded. Be aware that the filtering stage as the first stage of the process affects all subsequent stages, and thus have a considerable effect on final results. Failing to properly deduce the stop-words can severely downgrade the quality of the output.

The output of the filtering stage is considered as the actual input to the processing engine. In order to help fulfilling the processing task, a specially designed database is used to store intermediate data. The main elements of this data consist of words that are related by distance, which is their main relationship parameter, and number of appearances of the concept within the text. Also there is a record of the frequency of the concept along various documents. The frequency counting is essential for concept relevance evaluation process. Relevance evaluation has critical impact on the results, and by using appearance counting in two levels (counting against the text alone, and counting against the number of appearances in different documents) the overall quality of the output is dramatically increased. By setting the low-limit to the number of appearances a potential concept needs to breach we can fine-tune the results quality vs. quantity characteristic as we shall see in the example outputs later.

b. Tuning the test

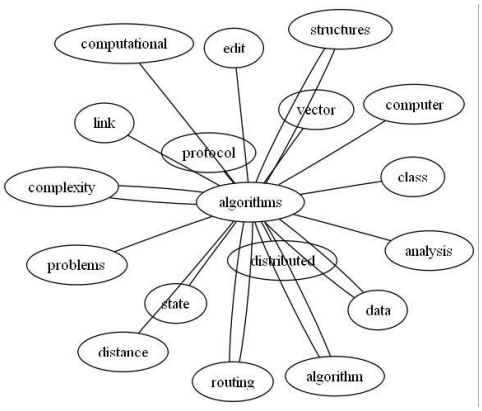
During the white box testing phase, we tested vast amount of documents with different concept distance and appearance limits. It was attempted to establish concepts between words that are up to 10 words away. As expect, this turned out to be much more than needed, as the median distance for all real concepts was about 5.4, though it is important to mention that many concepts were found on great distances, even on 10. The reason for this is the fact that by removing the stop-words, we also delimited sentences, so words from different sentences could have been connected. However, these concepts always appeared much more frequent on lower distances, so the exclusion of these candidates did not degrade the results. Naturally, the same frequency breaching low-limit could not be used for documents that varied greatly in total number of words in text. Therefore it was attempted to establish a connection between a number of words within the text and the frequency low-limit that give optimal results. The targeted optimal result limit was 80%, meaning that only 1 out of 5 concepts

candidates that are promoted to real concepts are allowed to be false. The discovery of function between the number of words and the desired breaching limit turned out to be much more difficult task then expected, because it varied greatly along different types of texts. The testing itself used Wikipedia® as primary source. There are two reasons for this decision. One reason is that these articles were compiled by a number of people, making it much less standardized because no single form is applied to all of the available documents. The second reason is that these texts are suitable for concept modeling as they generally include many concepts and are thus suitable for testing as more useful data can be collected in less time. The low-limit value we managed to agree on was set to 0.5% of total word count as no other function produced better results to justify its use for all the tested documents. This means that for example, the text that consists of 3000 words will have a breaching limit of 15. Someone might say that the limit is too high, but the number of false concepts increases in similar manner to the number of real concepts, so any lower value would threaten the output quality. The linear function turned out to be a simple yet effective solution to the problem.

4. Analyzing the results

The output of the processing is parameterized by minimum number of documents a candidate concept must be found in. This particular test was undergone on Wikipedia® articles that are related to computer science in general. Also, articles related via hyperlinks were processed too, to give a total of 45 documents. Due to the time complexity of the entire processing we were unable to process a larger amount of documents with currently implemented methods and available hardware, but nevertheless, the concept table contained 3118 records which proved to be relevant sample after all. First 3 outputs are produced using "algorithms" keyword while the 4th was an output to keyword "digital". Results are shown in a form of a graph, connecting related words into concepts.

Test 1

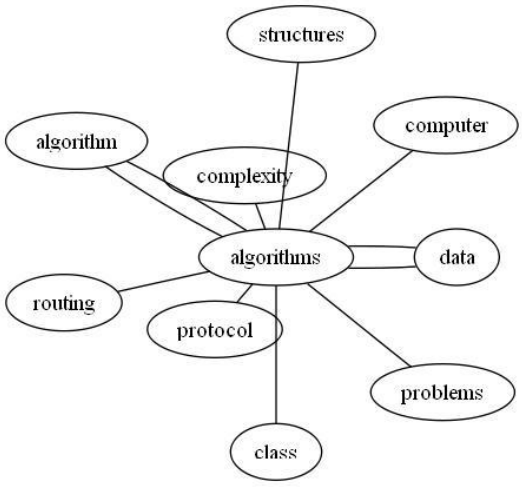


- keyword - "algorithms"
- document low-limit - 1

As you can see, there are some false concepts that were outputted, meaning that the breaching limit is not high enough.

Test 2

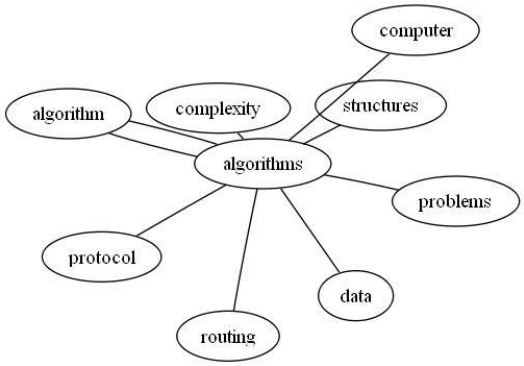
.



- keyword – "algorithms"
- document low-limit – 2

Now we can see an improvement in favor of quality rather than quantity.

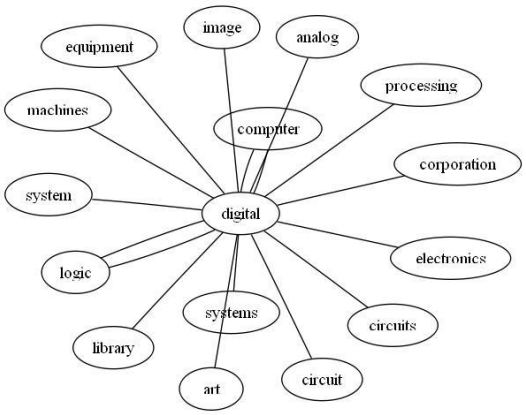
Test 3



- keyword – “algorithms”
- document low-limit – 3

In this case only a small improvement is noticed with the disappearance of the “class” abstraction. As you can see, there is no linear improvement of results with increasing of document low-limit boundary.

Test 4



- keyword – “digital”
- document low-limit – 1

In the last example, we received outstanding results beyond our own expectation, even with minimum document low-limit of 1. Almost all of the connections make real concepts. So, the general conclusion is that the document low-limit parameter has yet more to be examined, and the way it effects the results.

5. Future improvements

The current implementation suffers from many limitations. Many of them are caused by limited hardware resources due to much extended processing times. Although there is still some room for optimizations, it is undoubtedly that the technique itself needs powerful hardware in order to provide accurate results and to forms large concept networks, which would increase the relevance of searches in general. The next generation concept modeling engine will benefit from standards like HTML and XML that provide hyperlinks and custom structuring that can ease relevance determination, providing even better results. Nevertheless, what is essential to the technique is that large quantities of texts need to be processed, and that the quality of the results is maximized by creating a proper list of stop-words.

Although still far away from real-world exploitation, the concept modeling text mining technique will surely be a tool the fast text search engines will find very useful.

Acknowledgement

Authors are thankful to Prof. Tomazic of U. of Ljubljana, for his help.

References

[1] OMEROVIC, Sanida, TOMAZIC, Saso, MILUTINOVIC, Veljko. Concept modelling. V: AURER, Boris (ur.), BACA, Miroslav (ur.), RABUZIN, Kornelije (ur.). 19th Central European Conference on Information and Intelligent Systems, September 24-26, 2008, Varazdin, Croatia. Conference proceedings. Varazdin: University of Zagreb, Faculty of Organisation and Informatics, 2008, str. 33.

[2] OMEROVIC, Sanida, JAKUS, Grega, FILIMONOVA, Tatjana, TOMAZIC, Saso. Zapis vecjezicnih besedil v e-sperantu. Elektroteh. vestn., 2007, letn. 74, st. 4, str. 151-157, ilustr. [COBISS.SI-ID 6001492]

[3] TOMAZIC, Saso. Multilingual web with E-speranto. IPSI BGD Trans. Internet Res.. [Print ed.], Jul. 2007, vol. 3, no. 2, str. 13-15.