IMBIT: Resources and Search Strategy

Claudia Wortmann
IMBIT, Germany

Abstract

The working group "Search Technologies" at the University of Hildesheim has its main research focus in the field of pattern recognition (pattern matching, pattern completion, pattern extraction). The present participation of IMBIT in the CLEF project was based on a particular specification of our engine, namely towards "text-pattern recognition". Here, we mainly differentiate two kinds of patterns aiming at lexicographic and semantic similarities. The first kind is expressed in form of word lists where items are similar in writing (with respect to the input string). This type often is called error- (or fault-) tolerant retrieval. The second kind automatically groups word clusters -without grammar- which define concepts, ideas, events or processes. Such items normally form the content of articles, however following grammatical rules there. The technical basis is a particular artificial neural network, the *SpaCAM* (Sparsely Coded Associative Memory). SpaCAM has proven to work fine in many applications with text-data (like fulltext retrieval, translation memory tasks, terminology extraction etc.), but also in other contexts (like DNA retrieval, signature recognition, machine control tasks etc). Another module is called *DCC-Mindmap* and allows the graphic display of such word groups which mirror ideas or events, described in the normal text in underlying articles or documents. The distance within the (multidimensional) word clusters or between them -expressed via their position in the mindmap- relates to their appearance and importance relative to the search term(s). Based on these tools the search strategy was as follows. We picked a few non-experts (4-6 students), made a little workshop to explain the task, the data, the engines (2 hours) and let them GO! In order to find relevant documents they had to decide intellectually about catch words from the task descriptions. Such manual input for the SpaCAM or the DCC-mindmap lead to automatic results, either as list of hits of documents or as two dimensional grouping of "semantically" related expressions to the input. Normally a few repetitions of such input-output steps brought a state which was defined as final result. The students had to open the best fitting documents (suggested by the engine) and to decide intellectually whether they were relevant or not. If they were, they got their place in the "results-list" (to be sent to CLEF).

Resources

The tools SpaCAM and DCC-Mindmap form full-text retrieval systems, however they differ in their structure. The combined use of the two (extreme) forms offer kind of machine potential in the document retrieval or knowledge management tasks.

The tools can be found and tested under the following adress:
http://147.172.59.61:8790/newmap/servlet/hut.assomap.servlet.AssoMapServlet?basis=frr_w3

SpaCAM is a basic tool for rapid pattern recognition. The term *SpaCAM* stands for *"Sparsely Coded Associative Memory"* and permits fault-tolerant searching. The underlying technology for pattern recognition is based on neuronal network techniques. It is especially useful for the full-text retrieval task. The fundament of the machine´s fault-tolerance and celerity is an associative matrix which is capable of archiving great amounts of data). With the help of a special coding technique the natural data - the contents of the documents - are transformed into an electronically readable format suitable for the matrix, which serves as kind of index. Due to the sparseness of the activated "neurons" the search process it works very fast. The user can choose the number of documents to be shown within the range 1-999 documents. The output list contains extra information, like document number, title, machine ranking (in %). The latter recalculates typing errors, similar wording etc (e.g. each of committee, comittee, comitee, sub-committee contribute towards the ranking, if one of them was given as input, say). Clicking on a document number, the document is shown including heighlights of such words or word groups which are supporters of a high rank.

Search Strategy

Four students were given a detailed instruction in the search engine and then were asked to look for relevant documents for about 13 topics each. They read the topic several times, decided which were the key terms of the topic and started to search. They could alter or augment their search topics with the help of either mindmap or lexicographical mindmap. Starting the query they were given a document list, opened the documents and decided individually whether document was relevant for the topic. If they were relevant, they were added to a result list. This procedure was repeated until the students felt there were not too many unfound documents left. The cutoff was individually different, because the topics were hardly comparable and the students´ motivation was very high.

The documents in the document list had the following form and were marked if they had been opened beforehand: 100% - [fr940530-000939] wolfram schütte

Before finishing the search for one topic students were asked to put all terms used for any search query together in one query, look for the documents found before and write down the overall machine ranking. For technical reasons not all documents were shown following this strategy and thus all the %-values of all anterior queries were written down and ordered - unfortunately independently from the terms used in the query. The value of the machine ranking in decreasing order was deduced from this ranking. RSV-value corresponds to %-values, thus 100% became RSV 1.00, and 84% became RSV 0.84). Since we trusted the machine ranking, the personal ranking was put in increasing order from 0 to 99 in the same order as the RSV-value. In short it can be said that students tried to gather an overview in the mindmap and then looked for documents mostly in SpaCAM. Having different people working on different tasks with two different search modules to choose from, the results should be further verified. The biggest problem with the ranking was the fact that up to five queries within one topics were started and therefore the ranking might not be as adequate as it could if it had been possible doing one query only.

References

Ackermann, Martin (2000): Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. Hildesheimer Informatik-Berichte. 1/2000. (Diss.)

Bentz, H.-J.; Hagström, M., Palm, G. (1989): Information Storage and Effective Data Retrieval in Sparce Matrices, Neural Networks 2, pp. 289-293.

Doyle, L. (1986): Indexing and Abstracting by Association, In: American Documentation, Vol. 13. p.378-390.

Guiliano, V. (1990): The Interpretation of Word Associations. In: Statistical Association Methods for Mechanized Documentation, Miscellanous Publication 269, National Bureau of Standard, p.25-32.

Heitland, Michael (1994): Einsatz der SpaCAM-Technik für ausgewählte Grundaufgaben der Informatik, Hildesheim (Diss.)

Lesk, M.E. (1969): Word-association in document retrieval systems. In: American Documentation, Vol.20, p. 27-38.

Wettler, M.; Rapp, R.; Ferber, R. (1993): Freie Assoziation und Kontiguitäten von Wörtern in Texten. In: Zeitschrift für Psychologie, Bd.201, S.99-108.