

Creating Document Surrogates with Lexical Cohesion

Drs. Bas van Gils Dr. Hans Paijmans
University of Nijmegen Tilburg University
Bas.vanGils@cs.kun.nl J.J.Paijmans@uvt.nl

November 18, 2002

1 Introduction

The Field of Information Retrieval (IR) traditionally deals with representation, storage, organization and finding information items in (large) collections, and has been extensively studied in e.g. (Rijsbergen, 1979; Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999). Closely related to Information Retrieval are Text Classification and Text Categorization: all three largely depend on the identification of features that can be used as keywords to be included in a query or classifier. For the purpose of our paper, we will focus on the *indexing process*, or rather on the creation of so-called Document Surrogates¹.

Deciding which parts of the text are good candidates for the indexing process is a complex task, for items in the index must not only reflect the ‘aboutness’ of the represented documents as well as possible, but they also can be part of classifiers that divide the group of documents in two classes: relevant and not-relevant with respect to a query (which is not necessarily the same). A large part of this ‘aboutness’ is captured by the words used in the text (the *lexicon*). In (Morris and Hirst, 1991; Kozima, 1993) it is explained that a text is more than a random sequence of words, it has a coherent structure. Our hypothesis is that this structure can be used in the creation of a document representation. More specifically, we try to use *lexical cohesiveness* of (expository) text for this purpose (Morris and Hirst, 1991).

In text, lexical cohesion (Lcoh) is the result of chains of related words (or phrases) that contribute to the continuity of lexical meaning. A ‘count’ of the number of active chains is taken to be a metric for the (amount of) cohesiveness of text. A variation of this concept, in combination with the *tf.idf* weights of the occurring words, has been exploited by e.g. Hearst (Hearst and Schütze, 1993) to detect topical shifts in text.

In other publications (e.g. (Paijmans, 1994) (Paijmans, 1997)) we have already proposed that in documents exist so-called *Gravity Wells of meaning*; passages that contain more information about the topic of the document than other passages, and that such passages are marked by certain properties of the text, such as position, or the occurrence of *cue words* or *- phrases* or other surface properties (see also (Paice, 1990)).

The goal of our recent experiments is to find out whether passages with a higher (or lower) degree of Lcoh are better document surrogates than randomly chosen passages.

¹For the sake of clarity: we define the actual input to the indexing process to be the document surrogate; the result of this process is the document representation.

Preliminary experiments, in which we compared extracts created by Microsoft Word with extract based on lexical cohesion, showed a performance that was as good or better than the Word extracts, when compared with human performance².

2 Setup of the experiments

At this moment we are still working with a dataset of 200 \LaTeX documents, collected from the arXiv-website³. These documents are evenly distributed over the subjects *Computer Science* (CS) and *Astrophysics* (APH). The main reason for using \LaTeX documents is the fact that they are easily parsed. Furthermore, these documents are longer than the newspaper articles that are customarily used in experiments of this kind (e.g. the REUTERS corpus), and well structured (in the sense that they are divided in sections, subsections etcetera). Hence, they are good representatives for the scientific and scholarly documents in the retrieval of which we are naturally interested. Nevertheless we ran into serious problems with this dataset, which we will describe later.

A first naive approach is to get rid of the \LaTeX codes in the file and work on the remaining text. However, this leaves the problem of (lists of) mathematical equations, graphics and so on. Furthermore, since the authors of the documents differ, the internal layout of the files differ as well. For example, some authors tend to wrap their lines at 80 characters, whereas others put the entire paragraph on a single line. As the grammatical sentence is one of the units that we wanted to compute the Lcoh over, we had to write a program to normalize for these differences.

In the current version of our program, we ignore big mathematical environments and graphics, but keep inline maths. We try to normalize the text so that every sentence is on a single line, with an efficiency of approximately 80%. The errors that we still have in our normalization occur mainly with references such as "...In Fig. 3 ...", and bibliographical references; we will fix such problems as we go along. Maths are a problem in itself. Documents that did not display a coherent text after 'detexing' were skipped, so that the final database consisted of 200 documents, 100 from each class. As measure for performance we take the number of documents that were correctly classified.

2.1 The classifier

The primary setup of the experiments is that of a classification task within the vector Space Model. Word-document weights are computed with the *atc*-variant of Smart. The classifier then is created automatically by computing the centroid of the positive examples in the database. Classification finally is obtained by comparing this centroid or a derivative of it with all document vectors. In our experience the centroid of the *atc* values of a class is not a particular good classifier, but in this database it performed very well, perhaps too well: between 90 and 98%. This happened regardless of the similarity function that was used to compute the difference between each document and this classifier; only the dice coefficient scored lower. Superficial checks of the documents gave no obvious reasons other than the fact that the documents covered very different topics,

²Unpublished "onderzoeksproject" by L. Flinkenflögel, N. Konings and C. Koolen, computational linguistic students at Tilburg

³<http://lanl.arXiv.org>

but we will have to go into this problem again, and perhaps compile a totally different database, before we draw any final conclusions.

In any case we are not interested in absolute performance, but in the differences in performance between text segments with low or high Lcoh and random segments.

2.2 Chaining

Another program that we wrote *chains*, computes for every sentence or text window the number of active word chains, i.e. words that reoccur within a certain number of units. Such units are either sentences or windows of an equal number of tokens. The actual words that are taken into consideration can be controlled in several ways, including stopwords and lists of synonyms. Because of the many possibilities, finding the optimal algorithm for chaining is a problem in itself. For the first series of experiments we did not use the possibility of synonym lists, but concentrated on *identity of reference* or literal strings, using as units either complete sentences or word windows. As chaining with only *identity of reference* is easy to implement (but gives mediocre results), it offers a good opportunity to test our apparatus and general concepts.

So, which lexical cohesion properties we should expect in our *gravity wells* compared to the rest of the document?

- A difference with respect to non-function words. Intuitively one would assume that an author is more 'focussed' when he is describing concepts that are central to his discourse. This would lead to less but longer chains of content-bearing words and to a lower Lcoh score for such passages when the count is done with short chains.
- Also, one would expect these words to contain more information, which we can measure with e.g. the *tf.idf*.

3 Evaluation

The goal of our experiments is to find out whether document surrogates with an atypical number of active lexical chains are better document surrogates than randomly chosen passages (of the same length). Hence, we start experiments by creating three classes of document surrogates: the full text of the documents, parts selected according to Lcoh cohesion algorithms (the 'selection') and finally document surrogates that consist of random passages ('random'), but of a length equal to the selection files. The complete database is only used for reference; the real work is done with the 'selection' and the 'random' databases.

The first check of differences between the two databases gave promising results. We created a selection with low Lcoh and replaced every word in this set and in the random set by its *atc* value (as computed over the complete database). The average *tf.idf* of the words in the selection was consistently a few percents lower than that in the random database. This seems to point to a situation where the author uses less different words in 'interesting' passages that are central to his discourse.

The second measure of success would be that the document fragments selected by looking for passages with a different degree of lexical cohesion, would be easier to classify into the original classes than those of the the random database. To our chagrin, correct

classification was constantly very high and showed no actual difference between the two databases.

Former work (Apté et al., 1994), (Paijmans, 1998) had shown us that classification could be improved by using so-called *local dictionaries*. In that case only a very small number of keywords (typically between 50 and 100) is selected from the centroid and the rest is zeroed. This added one more technique to our arsenal. With this 'local dictionaries' variant of the centroid, the results changed somewhat, in that differences began to show between the selection and the random database. However, these differences were not consistent.

4 Conclusions

The first and most urgent conclusion of our work so far is that we may have selected the wrong database to experiment with. As we already mentioned, the experimental text databases that have been used by the IR community are almost exclusively short newsfeed documents, such as the Reuter collection or the documents that come with TREC. We were not able to find a publicized database of longer documents that were in a usable format and pre-classified. Creating a database of our own, and cleaning it, brings many unexpected problems and questions, not the least being the question how to handle maths and other "fremdkörper" in IR.

The experiments so far seem to suggest that lexical cohesion may be one of the surface properties of text that indicate emphasis on the topic or 'aboutness' of a document. The evidence is as yet very weak, and only expressed in differing *tf.idf* values. We hope to get a better grip on the classification experiments after we have included synonym lists and lists of related terms, so that the reiteration by semantic relations can also be measured.

References

- Apté, C., Damerau, F., and Weiss, S. M. (1994). Toward language independent automated learning of text categorization models. In *SIGIR94*. To appear.
- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley. ISBN: 0-201-39829-X.
- Hearst, M. A. and Schütze, H. (1993). Customizing a lexicon to better suit a computational task. In *Proc. ACL SIGLEX Work. Acquisition of Lexical Knowledge from Text*, Columbus, Ohio.
- Kozima, H. (1993). Text segmentation based on similarity between words. In *Meeting of the Association for Computational Linguistics*, pages 286–288.
- Morris, J. and Hirst, G. (1991). Lexical cohesions computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):1991, 21–48.
- Paice, C. D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information processing and management*, 26(1):171–186.

- Pajmans, J. J. (1994). Relative weights of words in documents. In Noordman, L. G. M. and de Vroomen, W. A. M., editors, *Conference proceedings of STINFON*, pages 195–208. StinfoN.
- Pajmans, J. J. (1997). Gravity wells of meaning: detecting information-rich passages in scientific texts. *Journal of Documentation*, 53(5):520–536.
- Pajmans, J. J. (1998). Text categorization as an information retrieval task. *South African Computer Journal*, (21):4–15.
- Rijsbergen, C. v. (1979). *Information Retrieval*. Butterworths, London, United Kingdom, 2nd edition.
- Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, New York.