# Advanced Document Description, a Sequential Approach

Antoine Doucet
University of Helsinki
Department of Computer Science
*Antoine.Doucet@cs.Helsinki.FI*
*http://www.cs.helsinki.fi/Antoine.Doucet/*

**Keywords:** Lexical Cohesion, Text Data Mining, Information Retrieval, Term Dependence, Phrases, Collocations, N-grams, Multi-Word Units, Sequential Patterns, Maximal Frequent Sequences, Document Retrieval, Automatic Indexing, Information Systems.

This dissertation addresses the problems of the extraction, selection and exploitation of word sequences, with a particular focus on the applicability to document collections of any type and written in any language.

Its main contribution is the definition of a formula and an efficient algorithm to address the problem of computing the probability of occurrence of a discontinued sequence of items. An application of this result is the possibility of a direct evaluation of word sequences through the comparison of their expected and observed frequency. This evaluation is unsupervised and does not depend on the intended use of the phrases.

To be able to perform efficient document processing, information systems need to use simple models of documents that can be treated in a small number of operations. This problem of document representation is not trivial. For decades, researchers have tried to combine relevant document representations with efficient processing. Documents are commonly represented by vectors in which each dimension corresponds to a word of the document. This approach is termed "bag of words", as it entirely ignores the relative positions of words. One natural improvement over this representation is the extraction and use of cohesive word sequences. Evidently, two words are more likely to be related if they occur next to each other than if they are separated by three book chapters of 20 pages each. Unfortunately, most document models, based on the bag of words principle, do not take this fact into account.

Previous studies attempted to get around this weakness by adding new dimensions to the document vector. To supplement the values representing single words, these extra dimensions contain values to represent the importance of multiple words occurring together in a document. A difficulty is that the number of ways to combine words can be enormous and the representation of each of those associations by a dimension of the vector space causes efficiency problems. Even if we decided to restrict these associations to adjacent word pairs, their number would often be too high. At the same time, one may observe that using only adjacent word pairs already means leaving a considerable amount of information out. For instance, if the word "and" occurs between two other words, they are certainly related but they do not form an adjacent pair.

The example of the use of adjacent word pairs is very representative of the problem of finding a good phrasal description. We easily end up with too many descriptors that are paradoxically insufficient. This motivates research in the area of *multi-word unit* extraction, where the goal is to extract from text cohesive units of several words, and ignore the majority of joint word occurrences that do not form cohesive units.

Following a look at the state of the art of advanced document representations, the dissertation focuses on these very problems of the extraction and selection of multi-word units. Its last part is an exploratory work on the exploitation of multi-words units in information retrieval. An important particularity of this research is the development of techniques that are entirely language-independent.

**Main Contributions**

The thesis presents three main results that respectively contribute to 1) the extraction, 2) the selection, and 3) the exploitation of the sequential nature of text. These results are described below.

1. Maximal frequent sequences (MFSs) are word sequences that are more frequent than a frequency threshold. They are maximal in the sense that no longer sequence that contains an MFS is frequent. The interest in MFS is due to the fact that they permit a compact document representation. Their extraction is, however, difficult, as it requires counting and comparing numerous word sequences with each other and even $MineMFS$, the current best-performing technique to extract the MFS set of a text collection sometimes fails to produce results in a reasonable amount of time, especially when the collection is large.

   We introduce $MFS\_MineSweep$, a partition-rejoin technique that uses $MineMFS$ as a black-box, and permits obtaining an approximation of the set of MFSs of a document collection. This method permits extracting descriptors even from collections with which $MineMFS$ fails. It effectively increases the scope of use of MFSs as document descriptors to document collections of virtually any size. Even for smaller collections, our experiments indicate that $MFS\_MineSweep$ can extract a more exhaustive phrasal description and that it does it faster.

2. The main contribution of our work is the definition of a formula and an efficient algorithm to address the problem of computing the probability of occurrence of a discontinued sequence of items. We formalized the problem to a simple Markov process, and exploited the specificities of the corresponding transition matrix through techniques of linear algebra. This technique goes well beyond the scope of this thesis as it can be applied to any type of sequential data. In text, it is common to estimate the probability of occurrence of word sequences, but the sequences are often defined with fixed relative positions of their word constituents, or sometimes by a maximal distance between the first and last word. To propose probabilities without constraints on the distance between words is new.

   A neat application of this work to textual data is the following. We have extended our technique of computation of the probability of occurrence of a discontinued sequence towards an efficient algorithm for the calculation of the expected document frequency of such a sequence in a given document collection. The expected document frequency of a word sequence can then be compared to its actual frequency using statistical significance techniques. This provides a general-purpose technique to directly evaluate and rank a set of word phrases. The evaluation of word sequences has always been indirect, heavily relying on the intended application and on the subjective judgment of human assessors. Our technique provides an alternative to evaluate the quality of word phrases from a general point of view, regardless of their intended use.

3. Our third contribution permits us to exploit a phrasal document description in information retrieval. As a result of an exploratory attempt to use MFS-based document descriptors in a document retrieval framework, we developed a novel technique to measure the phrasal similarity of documents. The descriptors can be matched more loosely, and a set of parameters is proposed to loosen or tighten constraints, such as the distance between words, their possible use in inverse order, and so on. A number of retrieval experiments were attempted, using MFS-based descriptors with radically different document collections: news-feed articles written in four languages (English, Japanese, Chinese, and Korean), and computer science journal articles in English.

   This exploratory research could not demonstrate the intrinsic quality of MFSs as descriptors that would be particularly suited for document retrieval applications, but the phrasal similarity measure we developed showed a significant improvement on all three Asian language collections.