IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4770

An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management

Rocio Abascal, Béatrice Rumpler and Jean-Marie Pinon
INSA of Lyon - LISI
7 Avenue J. Capelle Bât 502 – Blaise Pascal
F-69621 Villeurbanne cedex – France
rabascal@lisi.insa-lyon.fr, beatrice.rumpler@insa-lyon.fr

ABSTRACT

In our project about the digital library (DL) of scientific theses, we need to allow the user an access to the most pertinent information. Therefore, it is important to extract the main concepts to improve the information retrieval in this area.

This article represents an empirical evaluation of four tools for automatically extracting concepts from documents. We have compared these tools by using different document collections. For each document, we have extracted manually a list of concepts tied to the main topics. The tools are evaluated according to the degree of similitude between the concepts defined manually and the concepts automatically extracted by these tools. The four evaluated tools are: (1) TerminologyExtractor of Chamblon Systems Inc., (2) Xerox Terminology Suite of Xerox, (3) Nomino of Nomino Technologies, (4) Copernic Summarizer of NRC. This article presents the criteria of evaluation, a comparative study of the tools, an evaluation of the results and a proposed tool to annotate documents based on the concepts extracted.

INTRODUCTION

The rapid increase of scientific information in the web makes difficult the access to relevant sources to the user. The automatic extraction of concepts can improve scientific communication by quickly supplying understood labels or metadata in a user interface where is a need to display pertinent information. In addition, the extraction of concepts can help an author or an editor who wants to supply a concept list to a document.

This article evaluates four different tools able to extract concepts from documents. Each tool takes an electronic document as input and produces a list of phrases as output. We evaluate the tools by comparing their results with the list of concepts generated manually called the "reference list". By the term *concept*, we mean a generic idea generalized describing the document. For our evaluation, the concepts are composed of phrases of two or more words.

The task we consider here consists in taking a document as input and automatically generating a concept list (in no particular order) as output. This task could be called "*concept generation*", however the four evaluated tools perform "*concept extraction*". This means that the concepts they propose already exist in the body of the document. The tools used to make the extraction of concept cannot achieve 100% agreement with the concepts done by the human because at least 10% of these concepts do not appear in the input text. Therefore, the concept extraction may be viewed as a problem of classification. A document is visualized as a phrase bag where each phrase belongs to one of two possible cases: either it is a *concept* or it is a *non-concept*.

The goal of our work is to provide an objective evaluation of four tools based on the concept extraction. This requires a precise formal measure of a concept list. Following the presentation, Section 1 discusses our measure of performance in detail. The four tools are described in Section 2. Section 3 presents the tests, the results and a proposed tool to annotate the document with the concepts extracted. Finally, in the Section 4, we discuss the related works to finish with the conclusion and the future work.

1. MEASURE OF PERFORMANCE OF THE TOOLS

In our approach, we measure the performance of the four tools by comparing their output list of concepts to the handmade list. The measure of performance is based on the number of corresponding terms between the concepts extracted by the tool and the concepts selected manually. In the following subsections, we define what we mean by *"matching concepts"* and we describe how the measure of performance is calculated from the number of matches.

1.1 Criteria for Matching Concepts

The selection of a concept is done only when there is a correspondence between the same sequences of stems. A stem is obtained by removing the suffixes from a word. By sequences of stems, we mean sequences of words appearing in the same order. For example, if the concept extracted manually is "digital libraries" and the concept extracted by the tool is "digital library" we are going to count this like a "matching concept".

Stemming can be used to ensure that the greatest number of pertinent matching concepts is included in the results (Carlberger et al., 01).

For matching concepts we have one "*reference list*" which contains concepts that where suggested by the author to represent the document. We had enriched this list by adding concepts that we thought were important. Some examples of the concepts that we are going to have in the reference list are: "*document conception*", "*information system*", "*method of indexation*", etc.

1.2 Measure of Words Frequency

The measure of *word frequency* is based on the number of words common to each document (Lawrence et al., 99). The word frequency occurrence in a document, furnishes a useful measurement of word significance (Luhn, 58). In some tools like TerminologyExtractor, we have removed the high frequency words named "stop" words or "fluff" words. The input text is then compared with the "stop list", which can be updated according to the characteristics of the working area. In our evaluation, we have selected the words or phrases with a high frequency that does not appear in the "stop list".

1.3. Evaluation Method

The performance of a concept extraction tool is evaluated by comparing the output of the tool by the relevance judgments of a human expert. This means that we have to classify the results of the tools in *relevant* or *irrelevant* concepts. This idea leads to use the evaluation method based on information retrieval:

$$\text{Precision} = \frac{a}{a+b} \quad (1)$$

$$\text{Recall} = \frac{a}{a+c} \quad (2)$$

where *"a"* are concepts extracted by the human, *"b"* are wrong concepts extracted by the tools but selected as important and *"c"* are significant concepts extracted by the human, but which are not selected by the tool. This way, the *recall* is the percentage of *ALL* the relevant concepts found by the tool, even if it includes some irrelevant concepts. *Precision* is the percentage of *ONLY* the relevant concepts, even if it skips irrelevant concepts. In general, higher *precision* indicates that most of the relevant concepts are retrieved. Higher *recall* indicates that most of the retrieved concepts are relevant and it possibly exists a large amount of irrelevant records.

In the following experiments, we have evaluated the performance of each tool by making a comparison between the precision and the recall values obtained.

2. THE CONCEPT EXTRACTION TOOLS

In this section, we present the four tools tested for extraction of concepts by describing their functioning.

2.1 Xerox Terminology Suite

Xerox Terminology Suite (XTS) of Xerox (XTS, 01) is a terminology management system that allows the automatic creation of a multilingual or monolingual dictionary, the creation of a web-based knowledge and it also helps to build and update terminology.

Two components of XTS have been used: Xerox TermFinder and Xerox TermOrganizer. TermFinder builds automatically a database of terms and enables the user to create semi-automatically multilingual terminology. It is based on linguistic components and especially it uses the *noun* phrase extraction tool. By the term *noun*, we mean a word or sentence constituting the distinctive designation of a person or a thing. The *noun* phrase extraction tool consists of several modules: tokenizer (to instance linguistic expressions), a part-of-speech disambiguator and a non-phrase mark-up (NP). The NP applies finite-state automata describing *noun* phrase patterns. For example, a very simple *noun* phrase description for a given language may consist in a sequence of adjectives followed by a *noun* and another sequence of adjectives. The *noun* phrase extraction leads only to the selection of candidate terms. The terminology extraction takes place by monolingual NP extraction with alignment techniques based on statistical methods.

The Xerox TermOrganizer manages the terminology database created with TermFinder. It is possible to modify it, add and remove terms, and add specific information.

2.2 TerminologyExtractor

TerminologyExtractor of Chamblon Systems Inc. (TerminologyExtractor, 02) is a tool that extracts word and collocation lists with their frequency percentage. It uses several algorithms to provide the best output. For example, it transforms plurals and conjugated verbs into singulars and infinitives. To avoid collocations such as "you are", "of the", etc., it uses a stop list with pronouns, prepositions, articles, etc.

One of the main features of TerminologyExtractor is the capacity to differentiate the relevant words from the irrelevant words. This is possible by the use of a dictionary. The string is marked as relevant when the word is found in the dictionary. Otherwise, it is marked as irrelevant. The irrelevant list named *stop list* contains abbreviations, proper names, misspelled words and words which are very specific to the text area. This capacity to differentiate is immediately spotted without having to go manually through a long list of words.

The collocation lists produced by TerminologyExtractor contain all words and non-words sequences that appear more than once in the text. A special algorithm allows seeing collocations that appear within longer collocations. For example, in a text about "multimedia" we may find terms like "multimedia data" and "multimedia data repositories". These terms will both appear in the collocation list with their respective frequency.

2.3 Nomino

The search engine called Nomino of Nomino Technologies (Nomino, 01) uses a highly sophisticated system for linguistic analyses.

Nomino takes a document or a group of documents as input. For each document, it displays a list of terms and the document sentences that best summarize the main idea of the document.

Nomino allows to build a knowledge base by using the natural language as information support. Nomino extracts a *general index* in the form of concepts list and links them towards the sentences containing them. It also produces an *outstanding index* or *remarkable index* with the most interesting concepts for Nomino. The outstanding calculation is based on two principles: the *gain to reach* and the *gain to express*. The principle of the *gain to reach* stipulates that a concept is more important if it is very rare. For example, the word "ontology" brings more information about the document content than the word "system", not matter if it appears less frequently. The *gain to express*, will classify the concepts according on the specific character of the located information. For example, if a paragraph is about only one concept, this concept is very representative for the document.

2.4 Copernic Summarizer

Copernic Summarizer (CopernicSummarizer, 01) uses the extraction algorithm of the NRC (National Reseach Council). It takes a document as input and generates a list of concepts as output. It is tri-lingual (English, French and German) and it can summarize different formats of text based on *key concepts*. These key concepts are integrated by more than one word (Ribeiro et al., 01) and are determinated by a statistical analysis. In this case, we do not evaluate the summarizer but only the concept extraction capabilities.

In the next part of this paper, we will use the names XTS, TerminologyExtractor, Nomino and CopernicSummarizer to mean automatic concept extraction (generally called "extractor of terms"). By this, we do not mean the whole software package but only the part making the extraction of terms.

3. EXPERIMENT AND TOOLS EVALUATION

This section presents the experimental design and the results of the evaluation. We begin with the description of the corpus used and then we discuss the experimental process and the results.

3.1 The Evaluated Corpus

To evaluate the four tools we have chosen two different collections of document. The first one correspond to five articles in RTF format with about eight pages each one. The second collection corresponds to two PhD dissertations containing about 150 pages each one. The articles present the advantage to be read and treat more quickly than a PhD dissertation. The articles, as well as the dissertations, are from the data processing field. This allows us to easily evaluate the results.

The Scientific Articles

We have selected five articles from the data processing field. The full text of each article is available on the web. The authors have supplied keywords and sometimes *keyphrases* used as an initial list of concepts extracted manually.

The first step was to compare the concepts extracted by the tool with the first list of concepts extracted by the human. One interesting point was to compare long phrases by using the tools. For example, for one of the articles one concept extracted by the human was " Système d'information en ligne". For this case, only Nomino was able to extract this term. This is because Nomino is the only tool able to extract long *syntagms* (sentences composed of more than 3 words). As described above, Nomino produces the outstanding index composed of very significant concepts. For example, for the article number five, one of the concepts extracted by the human was "ontologies". In this case, the Nomino outstanding index has extracted *syntagms* like: "connaissance ontologique", "representation d'ontologie" and "ontologie formelle".

The second step was to make five tables for each article containing the results of the concepts extracted by the tool and to compare them.

The third step was to analyze the values obtained in the evaluation. These values were:

- the total number of concepts extracted by the tool,
- the total number of concepts extracted by the tool present in the reference list (the list with the concepts extracted by the human),
- the total number of concepts extracted by the tool that did not appear in the reference list,
- the total number of concepts extracted by the human and not extracted by the tool.

The precision and the recall calculated are discussed in Section 3.2.

The Scientific Dissertations

We have also selected two PhD dissertations to evaluate the different tools. We have followed the same way as for articles to evaluate the tools. One difference between articles is that scientific dissertations are longer documents (more than 150 pages).

3.2 Comparison of the Tools

In this section, we present the different performance of the tools and we give our interpretation about the experimental results. In the Table 1, we present the average of precision and the recall values (articles and PhD dissertations) obtained by each tool. The highest precision percentage is about 83,4% and it was obtained by the Nomino outstanding index. This index extracts less irrelevant concepts.

The lowest precision percentage was obtained by XTS which extracts a high number of irrelevant concepts. Only about 2,8% of the concepts extracted by XTS were present in the reference list.

The highest recall percentage was produced by the Nomino general index. However, only 9% of all the extracted concepts are relevant.

Copernic Summarizer has a rate of recall of 51% for correct concepts extracted by the tool versus 33,9% of precision (the concepts that appear in the reference list). In this case, Copernic Summarizer and the Nomino outstanding index extract a little list relative to the other tools. For example, for one of the dissertations, Copernic Summarizer has extracted 99 concepts versus 6932 concepts extracted by XTS, 3137 concepts extracted by TerminologyExtractor and 15618 concepts extracted by Nomino general index. The list with fewer concepts is generated by the Nomino outstanding index with only 71 extracted concepts.

From the results shown in Table 1, we can say that the tool with better recall is obtained by the general index of Nomino, it always extract the best concepts. The main difficulty is that it generates many concepts and it is very difficult to analyze them.

Graph 1, shows the difference between the two numbers obtained by each tool. In this case, we can see that the difference between the values of recall and precision obtained by XTS and the Nomino general index are more distant (i.e. XTS has 2,8% of precision and 90,5% of recall) than the values of the Nomino outstading index (83,4% if precision and 65,1% of recall). However, Copernic Summarizer gives also a very good result in precision and recall. The precision is higher in the Nomino outstanding index so we decided to use this tool for extracting pertinent concepts.

Table 1. Recall and precision percentage

	XTS (TermFinder)	Copernic Summarizer	Terminology Extractor	Nomino (outstanding index)	Nomino (general index)
Precision	2,8%	33,9%	6,8%	83,4%	9%
Recall	90,5%	51%	64,8%	65,1%	100%

3.3 The Tool Proposed for the Annotation of Document

From the results presented above (Section 3.2) we decided to select Nomino to extract relevant concepts. Therefore, we have used the index generated by Nomino to build a tool for annotation of scientific documents.

Our annotation tool shows to the user, in alphabetic order, the concepts selected by Nomino in the current document. This tool allows the management of the concepts proposed by Nomino, the indexation and the extraction of the pertinent paragraphs of the document according to some research criteria. We briefly present our prototype.

In the first window (Image 1), the user can make modifications in the proposed list of concepts. The concepts can also be erased and new concepts can be added to the list. This helps the user to modify the concepts or to erase the irrelevant one. When one modification has been done, a confirmation is required. Therefore, the modifications can be saved and new concepts are going to be indexed to the initial list.

Graph 1. Recall and precision values.

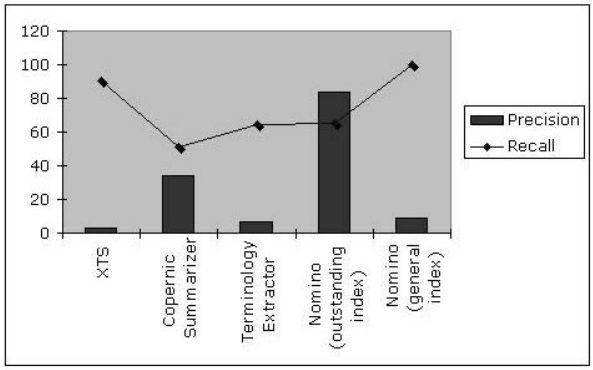

Image 1. List of concepts of the annotation tool

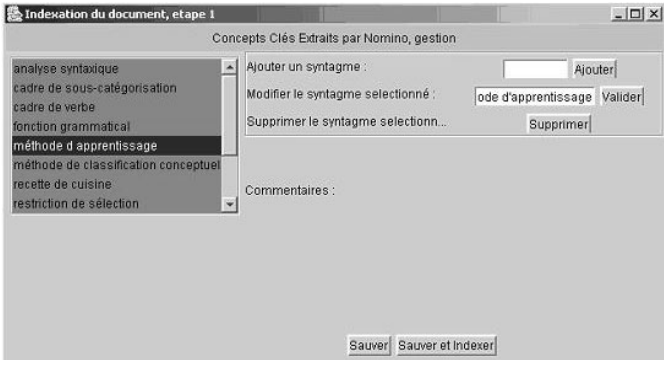

Image 2. Search using concepts proposed by Nomino

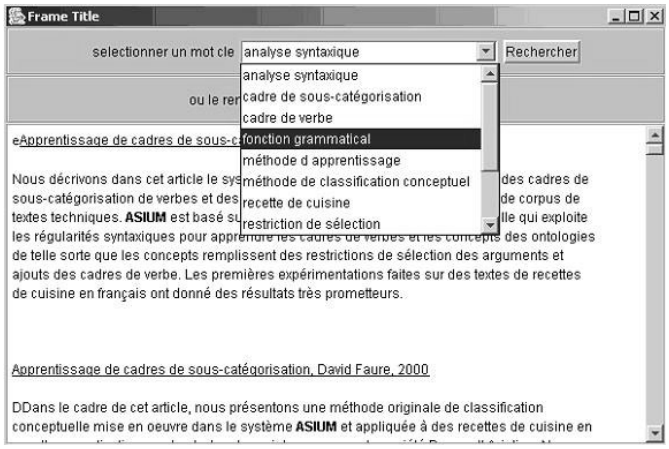

Image 2 shows the indexation phase. The concepts are included like tags to the initial document. The annotation will allow to find the concepts and to retrieve the paragraph containing the pertinent information. We can make the search session by using the concepts of the list.

At present, we continue our work to improve the annotation tool by including a thesaurus and a dictionary.

In Section 4, we present the related works, especially others concepts extraction tools we had not evaluate.

4. RELATED WORK

Ribeiro and Fresno (Ribeiro et al., 01) says that a textual characterization should be made through the extraction of an appropriate set of relevant concepts. Several implementations and evaluations of concepts extraction tools have been carried out like in (Ribeiro et al., 01) where Copernic Summarizer was evaluated versus a tool named IAI-extractor in terms of size and heterogeneity.

Several work have been done to evaluate and to implement new algorithm of extraction of concept and new methods like in (Witten, 99) and (Frank et al., 99).

Others works have been developed in the area of document summarization by analyzing the different tools and principally the techniques of extraction used. The extraction of concept is one of the bases of summarization allowing the production of sentences for a summary. One evaluation of tools for summarization of documents is carried out in (Jones et al., 02), tools are evaluated in terms of precision and recall.

CONCLUSION

The evaluation of tools had been done by comparing the concepts defined manually with the concepts extracted by the tools. We have compared each list generated by the tool with our reference list. We have decided to simplify the analysis when there was more than 1000 terms in the list.

From our evaluation, we have noticed that:

1. Every reference list can be enriched with the concepts extracted by the tools, for example by using the proposed terms extracted by the outstanding index of Nomino.
2. Nomino and XTS are efficient tools to extract the best concepts. However, for an analysis of long documents, the generated list by these tools is very long and it does not allows a simple analysis.
3. Thanks to Nomino terminological extraction qualities combined with the possibility of indexation, Nomino is the best suited tool to our work. Nevertheless, we recommend the use of the outstanding index rather than the general index whose precision is very low.

The results obtained cannot be generalized in other working situations without making an additional analysis. We only tested the feature of concepts extraction in the tools listed above. Some tools, such as Nomino and Copernic Summarizer propose other treatments for corpus like summarization but that was not the object of our test.

REFERENCES

1. (Carlberger et al., 01) Carlberger J., Dalianis H., Hassel M., & Knutsson Ola. (2001) Improving Precision in Information Retrieval for Swedish using Stemming, in the Proceedings of NODALIDA'01 – 13th Nordic Conference on Computational Linguistics. May 21-22, 2001, Uppsala, Sweden.

2. (Copernic Summarizer, 01) Copernic Summarizer, Copernic Technologies Inc. (2001) Version 2.0 updated in December 2001. Available in http://www.nrc.ca/corporate/english/.

3. (Frank et al., 99) Frank E., Paynter G., Witten I.H., Gutwin C., & Nevill-Manning C. (1999) Domain-Specific Keyphrase Extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, Morgan Kaufmann, 668-673.

4. (Jones et al., 02) Jones S., Lundy S., & Paynter G. W. (2002) Interactive Document Summarisation Using Automatically Extracted Keyphrases. Proceedings of the 35th Hawaii International Conference on System Sciences.

5. (Lawrence et al., 99) Lawrence S., Lee Giles C., & Bollacker K. (1999) Digital Libraries and Autonomous Citation Indexing, IEEE Computer, Volume 32, Number 6, pp 67-71.

6. (Luhn, 58) Luhn, H.P. (1958) The automatic creation of literature abstracts, IBM Journal of Research and Development, 2, 159-165.

7. (Nomino, 01) Nomino version 4.2.22 updated the 25 July 2001. Available in http://www.nominotechnologies.com.

8. (Ribeiro et al., 01) Ribeiro A. & Fresno V. (2001) A Multi Criteria Function to Concept Extraction in HTML Environment. In proceedings of the IC'01, Las Vegas, Nevada, USA, Volume 1, pp 1-6.

9. (TerminologyExtractor, 02) TerminologyExtractor version 3.0. (2002) Available in http://www.chamblon.com.

10. (Witten et al., 99) Witten I., Gordon P., Eibe F. et al. Ian H. Witten I. H., Gordon P., Eibe F., Gutwin, & Nevill-Manning C. G. (1999) KEA: Practical Automatic Keyphrase Extraction. In ACM DL.

11. (XTS, 01) Xerox Terminology Extractor version 2.0 updated in February 2001. Version called *XTS the Terminology Suite*. Available in http://www.mkms.xerox.com.