Copernic Summarization Technologies White Paper

February 2003

Introduction

Over the past few years, especially with the emergence of the Internet, the exchange of information has increased immensely, affecting all of us. On the one hand, the scientific community makes us aware instantly of its scientific breakthroughs while on the other hand, journalists present reports from around the world in real time.

The growing number of electronic articles, magazines and books that are made available everyday, puts more pressure on professionals from every walk of life as they struggle with information overload. In fact, nowadays, most people have to read daily papers, magazine articles, specialized literature, e-mails and Web pages during the course of their everyday activities. As evidence of information overload, over a million scientific articles are currently available online[1] among the billion pages the Visible Web contains.

With the increasing availability of information and the limited time people have to sort through it all, it has become more and more difficult for them, whether they are business people, journalists, lawyers, researchers or doctors to keep abreast of developments in their respective disciplines. On a larger scale, businesses lose time and money as they struggle to manage the information that must be shared in their organization.

Consequently, now more than ever before, a solution that allows people to process more information in less time and to share it with others easily is needed.

Specialists at Copernic had this objective in mind when they undertook the development of technologies that would serve as the foundation for the firm's summarization solutions. Along the way, they had to face a number of challenges.

The first part of this document describes the challenges that had to be overcome. The second part presents the result of Copernic's development efforts in the field of summarization technologies.

[1] "Online or Visible", Steve Lawrence, Nature, Volume 411, Number 6837, p. 521, May 31st 2001.

Summarization Technologies: Challenges To Take On

The challenges the summarization technologies development team decided to take on were of the following nature:

- Document type.
- Document language.
- Document style.

Document Type

Document Size Factor

No doubt a doctor would find the summary of a 50-page article on a medical treatment more useful than the résumé of a one-page article on the same topic. Unfortunately, the diff iculty of producing a high quality summary is proportional to the length of the document. Just imagine how much more difficult it would be to produce a summary of Tolstoy's 1,500-page novel "War and Peace" compared to summarizing a CNN news article. An automated system faces the same challenge.
Moreover, the automated system must proceed rapidly.

Document Format Factor

The solution must also be able to process documents which can have any one of a variety of formats (Word, text, PDF, Web pages, e-mails, etc.) and which may be available locally, across the Internet or an intranet.

Document Language

Multilingual Factor

Summarizing documents would surely be an easier task if everybody spoke the same language. This is evidently not the case as thousands of languages are spoken worldwide. Challenges posed by the multilingual factor originate from more concrete aspects, namely, grammar and syntax.

Just consider the following examples. In German, new words can be created by combining existing ones. The Japanese, on their part, do not use punctuation. Finally, in English, the use of the Em dash (—) is common while rarely, i f e ver, seen in French.

These are just a few examples. If summarizing a text in German is feasible for a German-speaking person, an automated solution must do better by taking into account the subtleties inherent in a language and still generate summaries of high quality.

Semantic Factor

Ambiguity in the vocabulary constitutes another challenge for the summarization process. Synonyms and related words are sources of vocabulary ambiguity. The word "capital", for example, has very different meanings according to the context: geographical (capital of a country), economic (capital gain), linguistic (capital letters).

This ambiguity also appears on another level in the form of idiomatic expressions. For example, if the author of a document writes "dancing to a different tune", it is not meant to be taken literally of course.

For all of these instances, summarization technologies must be able to automatically make use of the context to ascertain the true meaning and differentiate the words and expressions.

Document Style

Human Factor

Human intervention introduces still more complexity to the summarization process.

Grammar and syntax serve to provide structure to a written language. By following the rules of grammar and syntax, people construct sentences that are easy to read and whose meaning is unambiguous. Of course, writing is essentially a creative process. Furthermore, many people who write texts pay little attention to questions of grammar and syntax. Newsgroups are a case in point: while it is true that some newsgroup messages are clear and concise, many others are characterized by poorly constructed sentences, slang and typing errors.

The text summarization process must take the human factor into consideration and be able to generate superior summaries from poorly written texts as well as from Pulitzer Prize works.

To Sum Up

Optimized summarization technologies must be able to cast the net wide enough to grasp the essential ideas contained in a document, much as person would do, regardless of the document type, language or style. Such technologies must process the ideas and present them to the reader in such a way that the result is a faithful representation of the original text. By reading the resulting résumé, the reader will know if the original document should be read in its entirety.

Specialists from a variety of disciplines at Copernic worked together to develop a solution that tackles all of the issues raised in this paper. Over the next few pages, the summarization technologies that underpin this solution will be discussed.

Our Solution: Copernic Summarization Technologies

To produce document summaries that are balanced and coherent, summarization technologies must incorporate two very distinct components: statistical models and "knowledge intensive" processes.

Statistical Models

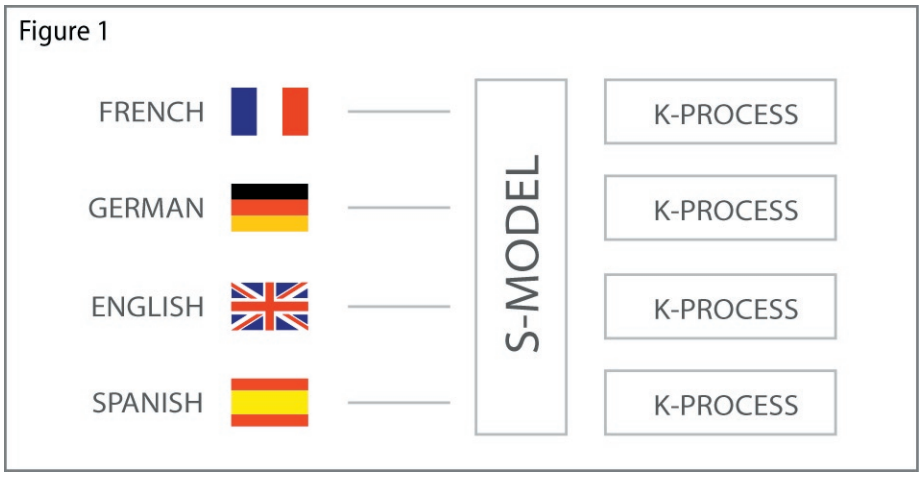
A common statistical model (**S-Model**) (**figure 1**) can be applied to a multitude of languages, to a certain degree, to approximate the topic specific vocabulary. This includes Bayesian estimates and rule systems derived from an analysis of thousands of documents.

Knowledge Intensive Processes

Knowledge intensive processes (**K-Process**) (**figure 1**) take into account the way in which human beings summarize texts. The effectiveness of such processes in generating quality summaries reliably is dependant on how well the summarization technologies integrate and apply the combined knowledge of a variety of specialists—linguistics experts, for example—whose expertise is not necessarily in the field of computer programming. These processes contribute to a better understanding of actions (expressed as verbs), instances of co-reference[2] and recognition of people, places and things.

The following section, "A Step-by-Step Approach", describes how Copernic summarization technologies work and how they integrate both components.

Figure 1



[2] Co-reference refers to cases where two or more words or phrases refer to the same thing. For example, in the sentence "Susan will succeed because she works hard", there is co-reference between "Susan" and "she" because both words refer to the same entity.

A Step-By-Step Approach

When a document is submitted for summarization, an optimization process is launched to collect the information, and then interpret it.

Document Standardization

The first step in the summarization process is the "standardization" of document content: documents, which exist in a variety of formats, must be converted to a common text format before their content can be interpreted. Copernic summarization technologies can convert a wide (and ever increasing) number of document formats. Our latest proprietary algorithms process not only common text formats, but also vectorial representations of texts such as PDF documents.

The firm's patent pending technology, **WebEssence**, identifies images and irrelevant text appearing on Web pages that the software should ignore during summary creation.

Language Detection and Analysis

Documents can be in any one of four languages: English, German, French or Spanish (**figure 1 on page 4**). The language of the document is recognized and identified automatically. Once the language is determined, specific linguistic principles, rule systems and powerful heuristics are applied.

Copernic summarization technologies have little in common with grammar rules and dictionary definitions. Instead, they employ linguistic algorithms that are designed to process written language in "real life" settings. This process involves statistical data about word usage gleaned from studying thousands of documents.

Sentence Boundaries Recognition and Tokenization

Punctuation marks are oftentimes a source of ambiguity, thus causing a real problem for automated systems in determining the beginning and the end of sentences.

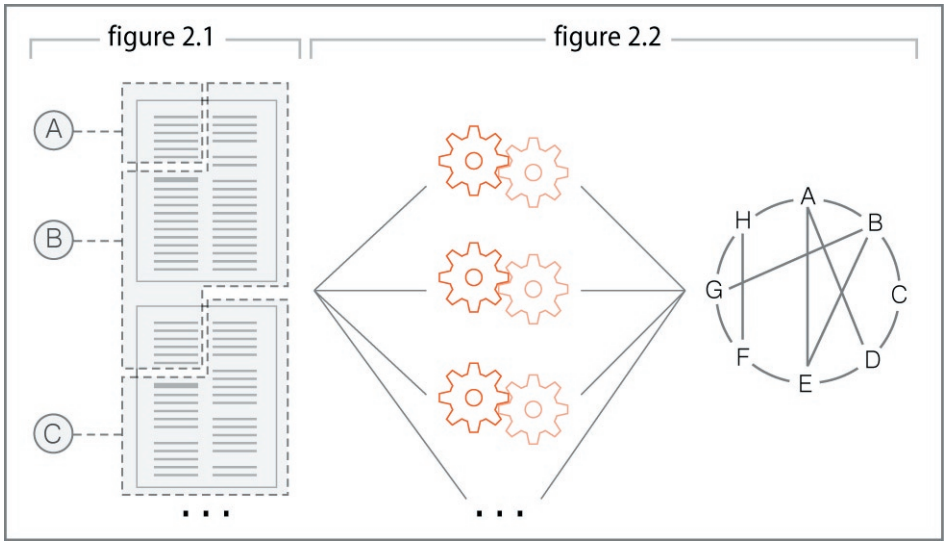
Copernic summarization technologies implement a wide range of heuristics to isolate sentences, bulleted lists, and special strings such as e-mail addresses and scientific formulas. In addition, they tokenize each and every word according to the context in order to identify actions, people, places and things.

Concept Extraction

The set of concepts associated with the document's main topic forms the core information extracted by intelligent summarization. Copernic integrated Extractor, an application developed in collaboration with the National Research Council of Canada (NRC), into its solution. This application extracts, in less than a second, high quality keywords by employing the latest machine learning techniques. Concepts can be looked upon as information at the "atomic" level of the topic description.

Long Document Segmentation

Once the key concepts are determined, Copernic summarization technologies formulate a sort of "picture" of the overall document, and then proceed to divide it into its constituent text segments (**figure 2.1 on page 6**).



These steps are necessary because they make possible the generation of summaries regardless of the length of the original document (while long documents remain a major obstacle for competing systems).

This rather delicate operation subsequently triggers a separate analysis of each segment (**figure 2.2 on page 6**), which is then followed by the integration of all the composite segments into a single, complete representation of the original document.

Sentence Selection

All sentences are weighted according to the relative importance of the information they contain. Sentences that would diminish readability or textual coherence are discarded. The more a sentence exhibits pertinent concepts, the more it is suited to developing important ideas, and consequently, the more likely it will be retained for inclusion in the summary.

Internal Document Representation

The document is thus transformed into an internal representation that can be manipulated again and again.

Because this remarkably flexible structure contains all the processed information, the user can repeatedly adjust the summary length, remove concepts or eliminate sentences in real time.

Conclusion

Copernic summarization technologies create concise text summaries, enabling people to absorb more information in less time. Based on technologies akin to artificial intelligence, Copernic's solution analyses texts, pinpointing key concepts, and produces instant summaries composed of the most important sentences.