Managing the Information Overload

New automatic summarisation tools are good news for the hard-pressed executive

Introduction

In this information age, an essential ingredient in the success of any business is access to many sources of material, including technical journals, newspapers, TV and radio broadcasts. Computers, plus powerful software products, have been the primary enabler for the massive growth in available information, freeing the creative talents of people by providing them tools to generate, store and process information. What is more, thanks to modern telecommunications systems, this information can be delivered, usually within minutes or hours, anywhere in the world. In the not-too-distant future we will see the provision of 'super information highways' capable of delivering orders of magnitude more information than today's simple systems. However, each technological advance brings its own problems, and in this case, we have reached a state where people simply cannot cope with the rapidly accumulating mountains of information: we will suffer from 'information overload'.

Natural Language Processing is one technology that will help people find and use relevant information, without being overloaded. Recent developments at BT Laboratories include an Information Agent, which "learns" the interests of its human user and selects appropriate articles from a database, and a Text Summariser, which interactively abridges articles by extracting the "most important" sentences.

Information Agent

An Information Agent is a program that acts like a personal assistant, selecting information that it thinks you will find useful. The prototype system currently operates on a list of journal articles that is prepared weekly, each containing the titles and author details of over 400 new entries. The output is in the form of a spreadsheet, which can be scrolled through, and articles of interest selected. The Agent then automatically generates requests for copies of those articles. What is more important, your Agent "learns" what interests you, building up a profile, which is updated whenever you select further articles, and will automatically put subsequent lists in what it considers is your order of preference.

The main benefit over standard document retrieval techniques is in ease of use. It is no longer necessary to think up keywords or search strings. By choosing an article of interest one is essentially saying to the Agent: "find me more like that".

Text Summariser

When a selection of articles has been made, the next task is to read through them to extract the information of interest. Many people carry out an initial scan with a highlighter pen, marking the parts of text and sentences that are of particular interest.

To help in the task of quickly extracting information from documents, we have automated the highlighting process. Called the Summariser, the program accepts any document in machine-readable form and will automatically highlight the sentences that it considers make up the "most

important" part of the text (see Figure 1). Alternatively, it can extract these sentences from the text to produce an abridgement of the article.

The process is interactive, allowing the user to choose longer or shorter extracts at will, from a single sentence to the full text. Typically, one begins with a very short abridgement, to see if the article is relevant. If so, the length of the abridgement can be increased to see more details.

The basic technique that lies behind the Summariser is very robust, giving it the ability to work on any text, independent of subject. For example, equally good results have been achieved with articles on such disparate subjects as semiconductor lasers and red squirrels!

How well does the summariser work ?

The usefulness of a summariser depends on how well it extracts the key information from an article. We have evaluated the Summariser on a number of technical articles with author-written abstracts. On the material we have tested so far, we find that an abridgement of typically only 5% of the original article will contain roughly 70% of the information in an author-written abstract, while a 25% abridgement contains essentially all of the information in the abstract (Figure 2).

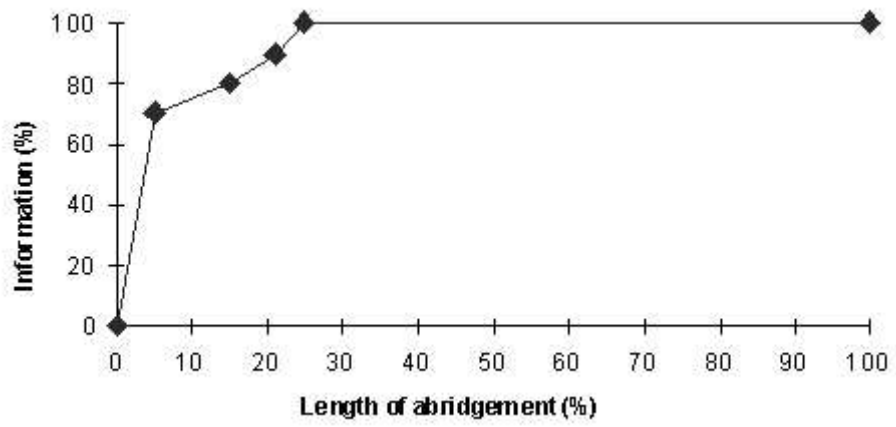

Figure 2: Evaluation results

The Future

Although our work to date has concentrated on text, we are investigating ways of extending these information management techniques to cope with multimedia information. We are also carrying out research into approaches involving deeper analysis and "understanding" of articles.

The capabilities described so far are just the first steps towards building the intelligent information networks of the future. These networks will not simply transport information from place to place, but will increasingly understand, filter and process information to meet the needs of modern businesses.