



Content in context: a data-driven approach

Judi Vernau

AND Classification Data Ltd, UK.

Abstract

The dramatic increase in document availability via the Internet means that traditional methods of cataloguing, classifying and retrieving information have been unable to cope with the demand. In order to solve this problem, the emphasis has shifted from the previous labour-intensive methods to technical solutions which rely on statistical and algorithmic means of identifying topics and proper names in documents. While these various solutions have provided a great improvement in retrieval of relevant information, it is still not possible to identify the exact nature of particular terms or relationships between terms and topics. However, using a specially-devised ontology covering topic and type, and a vocabulary set classified using this ontology, it is possible to disambiguate between different senses of terms, and therefore to identify the main topic of a document and the nature of the lexical terms and proper names within it, and to create genuine relationships.

Introduction

From the third century BC to the mid-twentieth century, retrieval of 'documents' (scrolls, papyri, books, pamphlets, journals, etc) has been founded primarily on what was once cataloguing information and is now often referred to as metadata. Metadata for documents may include author, title, date, publisher, subject, etc, and until the advent of shared cataloguing systems at the turn of the last century, these elements had to be assigned manually by each library. The development of computerised cataloguing systems in the early 1960s, and the advent of MARC in 1966, began to take a great part of the strain off institutions trying valiantly to keep up with user needs to retrieve relevant documents, but the arrival and fast take-up of the Internet has turned document retrieval on its head again. Since it is impossible to provide reliable metadata for even a fraction



of the documents on the Internet (although organisations such as Yahoo! and the Open Directory Project do their best to provide subject access to as many sites as possible), retrieval across the greatest number of documents currently relies on full text indexing.

Most approaches to full text indexing use statistics-based technology. This paper explores how both full text indexing results and subject access can be improved by the use of data in the form of classified vocabulary sets.

Disambiguation

There are several different ways of carrying out and ranking the results of a full text search, based on such mechanisms as statistical analysis of word frequencies, or heuristically-based algorithms for clustering documents on similar subjects [1], for example. However, none of these approaches take into account the actual semantic meaning of the search term, because that is not definitively identifiable by purely algorithmic methods, although interesting work has been carried out in this area, such as the work on 'ambiguation' at Microsoft Research [2]. If the semantic meaning cannot be identified with certainty, there is the danger that irrelevant documents will be retrieved, because, regardless of the number of times the term appears in the document, the meaning of that term is not the meaning intended by the searcher.

Identifying different meanings

In 1997 David Crystal, a linguistics expert and encyclopedist, proposed the concept of classifying vocabulary in order to allow a searcher to select for his or her self the meaning, or context, of the term being sought. So when an ambiguous search term is entered, the searcher is presented with a number of contexts in which the term might occur. For example, Oasis has at least the following meanings:

- a pop band
- an organisation supporting standards for structured information
- a make of car
- a fertile spot in the desert
- a product made of foam to facilitate flower arranging.

Each one of these contexts is clearly different from the other and will be associated with a discrete classification category. Each category is itself associated with a list of keywords which characterise the relevant subject area. So in the above example, Oasis the pop band is classified as 'mind - arts - music - performance - popular music', which is associated with keywords such as music, rhythm, tune, song, band, singer, etc, and Oasis the make of car is classified as 'mind - technology - transport - road vehicles - cars', which is associated with keywords such as travel, drive, engine, speed, model, etc. By automatically adding the appropriate keywords to the search term, the free text search is made significantly more accurate.



In order to achieve this facility, three data sets are required:

- a classification scheme
- keyword sets for each category in the classification scheme
- classified vocabulary

so in 1998 an AND project was set up called ALFIE (AND Lexical Filter for Internet Enquiry) to build each of these data sets in order to test the concept.

The classification scheme

The classification scheme was designed specifically to support disambiguation, and is a hierarchical scheme deriving from ten main topics (Earth, Environment, Natural history, Human geography, Human history, Humanity, Mind, Recreation, Society, Universe). The scheme currently comprises about 1100 subject categories, covering all areas of knowledge.

An additional schema for classifying terms is also used. This assigns a type, rather than a subject, the top level types being organisation, place, person, product or general lexical item. So in the previous example, Oasis the band and the standards group are organisations, the make of car and the flower arranging foam are products, and the fertile area is a general lexical term.

The keyword sets

The concept of creating sets of terms which characterise a subject area (variously called signatures, collocations, lexica, etc) has been explored by several institutions [eg 3], primarily for the purpose of automatically classifying documents. In these cases, the sets are usually constructed based on the frequency of occurrence of certain terms in document sets known to be on a particular subject.

The purpose of the keyword sets in the AND project is to extend the search term in order to focus the search more accurately. Keyword sets generally consist of about forty terms, and it is important that as many of the terms as possible are added to the search, as the greater the number present, the more reliable the search results will be. The terms are selected mainly intuitively using encyclopedic and internet resources, with some input from statistical sources, but because the terms are collocates, they are not necessarily unique to each category. The word 'house' may be a keyword for politics, architecture or building: it is its appearance with other collocating keywords which will determine the subject accurately.

The keyword sets are also used to test the distinction between different subject categories, in that the process of identifying the most useful keywords highlights whether or not categories in the same branch of the hierarchy are truly conceptually discrete.



The classified vocabulary

The vocabulary is a database of English terms (single or multi-word), with their variants and inflections, which have been classified by subject and type. The initial stages of the project focussed on lexical terms, but these are now being supplemented by proper names. For each term, it is important to consider not just the basic semantic concept, but also the contexts in which the concept may appear. For example, the herb rosemary is inherently a plant, but the context of the search might be gardening, cooking or cosmetics. Each of these contexts must be recorded, because each will be linked to a different category.

There are currently about 200,000 unique term/context units in the database, but this is expected to increase dramatically in 2000 as many more proper names are added. Lists of current web search terms (such as those published by Voyeur.com) are also consulted regularly to make sure that popular terms are included.

Automatic classification

Knowing the context of terms also provides a means for identifying the subject matter of running text. As mentioned above, some commercial and research systems use keyword sets (lexica, etc) to automatically classify documents. The AND project looked at using our own keyword sets for this purpose, but results of testing showed that comparing the words in the documents to the classified vocabulary produced more accurate results. One important reason for this is that nouns and noun phrases are generally the most accurate predictors of subject [4], and the inclusion in the data set of so many proper names reinforces this.

In order to arrive at a subject classification, the AND classified terms are compared to all the words in a given document, and terms which are found in both places are assigned a weighting. The most successful method for weighting has so far been to assign one point to each term found. This means that monosemous terms score one full point, and polysemous terms have the point divided up between the number of relevant categories, which gives a weighting in favour of the monosemous terms. By using an equation for calculating standard deviation $\sqrt{(\Sigma(x - m_a)^2/n)}$ where

x = an individual result

n = the number of test results

m_a = the mean of the theoretical normal distribution curve for all results

m_n = the mean of n results

a final score is then allocated to each category.

This method of automatically classifying documents is currently only tested at a fairly broad level, since the classification scheme is itself still broadly based. For example, the section of the scheme dealing with animals will distinguish between mammals and reptiles, but does not yet have the level of granularity to be able to identify cats and dogs. However, it has sufficient breadth to work across a range of documents to provide at least a distinction between history and ancient history, country and continent, classical and popular music.



Since the classification scheme is hierarchical, it is possible to make the connection between broader and narrower topics, and this is supplemented by an index which links unpreferred terms to the preferred category names. Work is also underway to test linking the categories to those of other schemes.

Information extraction

As well as automatically classifying documents, by comparing terms in the document with terms in the classified vocabulary it also possible to compile indexes of proper names mentioned in the text, and to use the classification of the proper names to create links to related material. For example, an article is classified as being about baseball. The text includes the string 'Michael Jordan'. According to the classified vocabulary, there is a Michael Jordan who is a baseball player. We can then presume with sufficient certainty that the person mentioned in this text is Michael Jordan the baseball player, and he can therefore be indexed and linked to other baseball players. Similarly in an article on computing we can assume that Apple refers to the computer company, and not to a record label, and we can create a link from this Apple to any other mention of Apple that occurs within articles on computing. All of these relationships are being created by the data alone, with no recourse to syntactic analysis or other parsing methods.

Afterword

These methods of disambiguation, automatic classification and information extraction, relying as they do entirely on the data, can only work really effectively when the relevant data is present. As stated above, the current classified vocabulary consists of 200,000 term/context units, and will increase enormously in the coming years. Our challenge is to identify and add the most useful terms in the areas of most interest to potential users.

References

- [1] Zamir, O, and Etzioni, O., Web Document Clustering: A Feasibility Demonstration, *Proceedings of ACM SIGIR '98*, 1998.
- [2] Dolan, W.B., Word Sense Ambiguation: Clustering Related Senses, *Proceedings of COLING-94*, pp. 712-716.
- [3] Jenkins, C., Jackson, M., Burden, P., Wallis, J., Automatic Classification of Web Resources using Java and Dewey Decimal Classification, *Computer Networks and ISDN Systems*, Volume 30 646-648, 1998.
- [4] Godby, Jean, *Two Techniques for the Identification of Phrases in Full Text*, <http://www.oclc.org/oclc/research/publications/review94/part1/twotech.htm>