

White Paper:

Using Linguistics in Search Engines

Version 1.0 (30th July 2003)

Graham Whitehouse

In2itive Business Group

An FSDC plc Group Company

Registered Office: 15 The Metro Centre, Welbeck Way, Peterborough PE2 7UH

Registered in England No: 4132210

Using Linguistics in Search Engines

Background

There are many 'search solutions' on the market today that offer users the ability to locate and retrieve documents, files or 'Knowledge', and many people consider that the 'speed' of a search or the number of results as the differentiators. **In reality however, the technology that is at work 'locating' the results of a search can be the difference between finding what you are looking for and having to re-create something.**

A majority of the more common search tools (often 'add-ons' to content or document management systems) use very 'traditional' techniques, like Boolean searches, as the method for finding what you need.

As the quantity of the data increases however, so the accuracy of the search can decrease, with slower performance and the system offering more and more results. The effect is that the users have to 'sub-search' the results, or build another query to find the information. In many cases, the preferred solution to improving the search is to throw more hardware at the problem, but this purely gets the vast list of results back more quickly.

Quality of Results

However, this is really a poor measurement of this kind of technology; surely it is more about the quality of the results rather than how fast they are returned? Research from the Penn State School of Information Sciences and Technology (IST) found that more than 80% of people visit only the first three results from a query, with more than half checking out only one result. They also found that 50% never even looked at the second page of results.

Getting a list back within half a second is great, but if that just provides 3,000 hits, where has it got you – other than being good at building Boolean searches?

Conversely, the use of traditional search tools can also limit the results – by excluding results that are not an exact match for the words that you entered – imagine searching for 'car' and missing all the documents containing 'automobile' that the US company provided!

This paper deals with the use of **Natural Language Processing (NLP) and advanced linguistics** to help deliver fewer, but higher quality and more accurate results.

Natural Language Processing

The use of Natural Language Processing techniques allows incoming documents to be analysed, and word meanings to be extracted for use at a later stage. This

means the system can look beyond the 'words' of a document, and start to understanding the meanings as well.

The techniques used in the use of natural language processing can be applied to questions that are entered or to documents being indexed. However, due to the massive use of traditional search tools on the web, people find entering 'natural language' in a search tool quite a stretch.

Synonymy and Morphology

One interesting effect of utilising linguistics is that it can add a level of 'fault tolerance' to entry that can ensure a system offers the best possible chance of locating the information that you require. Understanding the linguistic use of words means that a system can use morphology, synonymy and semantic information to both select and increase the number of results – **so helping find information that would have been overlooked in a traditional word search.**

The implementation of a combination of both synonymy and morphology is vital to a search tool, as each offers benefits and flexibility to the task of searching.

Morphology allows the system to take other forms of the same word into account. Some companies call this process 'stemming', where a word is reduced to its base form before a search is run. This means that instead of just searching for an exact match on 'insurance', the system will be able to find all forms of the word, such as 'insure' or 'insured'.

Synonymy brings another expansion of the search terms, as the system takes on a 'natural' response from a question to look for similar terms and word uses. This means that the above search for insurance could grow to include words such as 'policy', 'premium' or 'indemnity'.

These expansions of the search criteria mean that a system can offer the broadest possible base to search upon, however, this can also expand the list of results, meaning that more understanding of the language should be used.

Part of Speech

Many words in the English language have multiple syntactic uses, meaning, for example, that they can be used as nouns, verbs or adjectives. The Natural Language Processing that takes place can isolate how, and in what sense, words are being used, and part-of-speech recognition means that these different categories can be stored (or tagged). This information is stored along with the words as documents are indexed, meaning that even more accurate searches can be performed.

The most advanced part-of-speech recognition systems can also spot the use of proper nouns within the text, and index these accordingly. An example could be given when looking for information about the Microsoft product 'Access'. Traditional searches bring back every occurrence of the word 'access' in the

library, whereas if you could specify a search for proper nouns only, this number would not only be greatly reduced, but would also be more specific to the product.

Going one stage further, some systems have the ability to recognise proper noun phrases, such as 'Old Trafford' or 'Microsoft Access'. This again will allow the search tool to only return results related to that specific 'unit' of text, rather than every occurrence of the words throughout the whole library.

Changing Results

In practice, these functions can have dramatic effects on the result set delivered for a query. The following examples were taken from a fairly small set of test documents, and are the results of searching for the word 'Break'.

	No Synonymy	With Synonymy
All	42	553
Noun only	9	42
Verb only	33	532

The table shows three vital aspects of linguistic use:

- 1) There can be as much as a ten fold increase in the number of results when synonymy looks for other 'like' words. When you are unsure for exactly what you are searching, this will greatly increase your chance of locating the desired document.
- 2) By searching for a specific syntactic use of a word, accuracy can be improved by a factor of between four and ten.
- 3) Combining the technologies can provide increased search breadth whilst also delivering decreased noise in the results.

Summary

When considering a large range of documents and information, and the need to find particular data, any tools or technology that can assist are vital business tools, they reduce wasted time and enhance end user satisfaction.

Allowing the user to find information when they have entered the 'wrong' word brings a level of flexibility and tolerance that could be the difference between finding information and starting again.

However, the ability to filter all of the noise that is stored down to the bear essentials that are needed to find what you are looking for could save hours of wasted 'browsing'.

Understanding a document's use of linguistics and can provide the foundations to enable better, faster and more revealing searches.

Where do In2itive fit in?

In2itive started a project just over two years ago to build a knowledge management solution that could both be used to extend their portal offerings, sold as a stand-alone solution as well as via OEM partners .

After some extensive research, they decided to work in close co-operation with the University of Manchester Institute of Science and Technology (UMIST), as they believed that great value could be derived from use of advanced linguistics in a Knowledge Management engine.

Many of the items described in this paper were actually developed during this period, and indeed, patents have now been granted to In2itive on the way that the linguistics are used in this environment.

Utilising theorists and experts as well as quality programmers meant that foundations of the linguistic algorithms are of excellent breeding, and they are within the control of the group for building future upgrades and developments.