



LATERAL THINKING IN INFORMATION RETRIEVAL

JOHN CHALLIS, CEO/CTO
CONCEPT SEARCHING

This white paper looks at the problem of searching unstructured information with particular emphasis on knowledge discovery.

About the Author

John Challis is an experienced entrepreneur having had success with several previous ventures involving the management of unstructured data. In 1990 he founded Imagesolve International which quickly became the UK's leading supplier of document image Processing and workflow products.

He then launched ImageFirst Office for BanTec in the United States in 1995 and in the first twelve months achieved over five million dollars in new business. Prior to Concept Searching he was CTO at Smartlogik, the company behind the first probabilistic search engine.



Introduction

Unstructured information (i.e. office documents, emails, news feeds, web pages, etc) accounts for about 90% of all digital information within most organisations. Products that can search and classify this type of information have been available for many years but most tend to focus on static structures and require the user to execute a very precise and relatively simple search.

It is argued that traditional search technology does not adequately support the more complex search scenarios where the topic does not align with prior structures and is difficult to articulate. A new statistical approach that supports lateral thinking techniques in information retrieval applications is introduced.

Lateral Thinking

Edward de Bono is widely recognised as the inventor of lateral thinking techniques and in his 1970 classic book on the subject he states:

“Because of the way the mind works to create fixed patterns we cannot make the best use of new information unless we have some means for restructuring the old patterns and bringing them up to date. **Vertical thinking** is concerned with proving or developing concept patterns. **Lateral thinking** is concerned with restructuring such patterns (insight) and provoking new ones (creativity).”

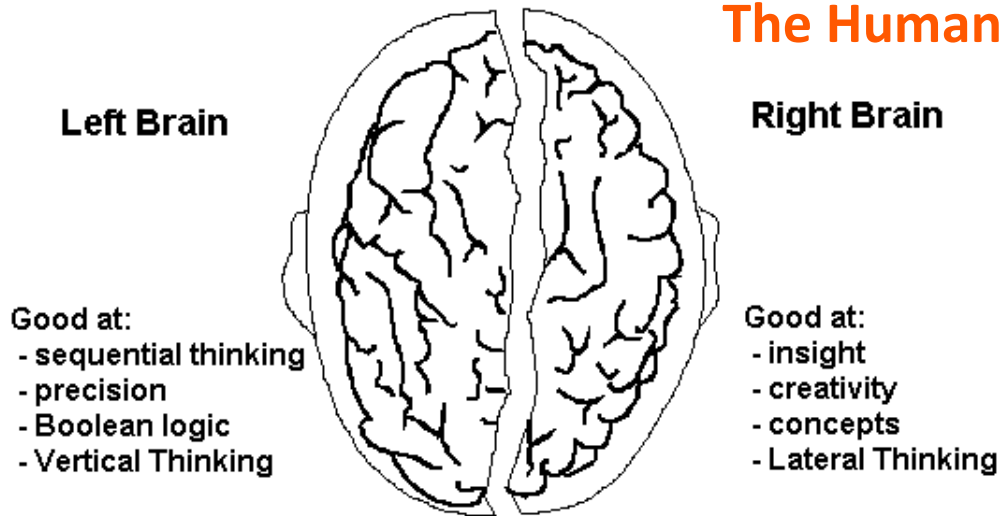
“Vertical thinking is selective, lateral thinking is generative”

“Vertical thinking is a finite process, lateral thinking is a Probabilistic one”.

In an information retrieval context, vertical thinking is used when we know precisely what we are looking for and selecting the finite set of relevant documents is relatively straightforward. In contrast, lateral thinking would be required where the requirements are less well defined and the process of locating relevant information requires a degree of trial and error. Knowledge discovery applications often require some lateral thinking in order to refine queries and to follow different lines of enquiry in order to construct as complete a picture as possible from the available information.



The Human Brain



The human brain is divided into two halves as shown in the diagram.

The left-brain excels at sequential thinking where the desired outcome is achieved by following a logical sequence of actions. In contrast, the right brain is optimised for creativity where the desired outcome may require a degree of non-linear processing.

Most people are familiar with searching highly structured data, typically in a relational database, where the query is very specific (e.g. find all invoices from Acme plc received during March 2002). This is classic left brain activity. Unfortunately, many people expect to search unstructured information in the same way and are often disappointed when the documents they expect to find are not returned. The problem is that unstructured data is highly variable in layout, terminology and style whilst the queries tend to be more difficult to define.

Most information retrieval activity, and virtually all supporting information technology, is focussed on the requirements of the left brain, which is most comfortable when searching with precision.

Searching with Precision

When searching for specific information traditional techniques can be used to find documents that contain the required keywords, perhaps combined with Boolean logic (the AND, OR and NOT operators) and phrase or word proximity searching. This is analogous to vertical thinking – keywords are either present in a given document or they are not.

For example, if we were interested in the effects of taking aspirin on blood pressure we might enter a search like:

aspirin AND "blood pressure"

The left brain easily understands this type of search and the results tend to be satisfactory regardless of the specific product in use. If taxonomy were available (i.e. a pre-existing classification system) then we might also be able to further restrict the output to documents pre-classified as "medical research" or "cardiovascular" and the results would be even more precise.

Typical vendors supporting this type of search include: Verity, Hummingbird, Microsoft, Oracle, FAST, Alta Vista and Open Text. These systems offer virtually no assistance to the user's right brain.



Concept Searching

Sometimes we are looking for information about a particular topic but the concept is nebulous and difficult to articulate precisely. For example, consider the following topic:

Insider dealing of shares by directors with access to unpublished price sensitive information.

With this type of query it is going to be difficult to specify our search so that all of the best documents are found without too many irrelevant ones. Issues to consider with this topic include:

- ◆ “insider dealing” may be referred to as “insider trading”
- ◆ “shares” may be referred to as “company securities”
- ◆ not all “insiders” are “directors”
- ◆ etc

The difficulties are compounded if there is uncertainty about the presence of documents and the exercise is designed to gather evidence, or to prove the absence of, information about the selected topic.

A successful outcome is likely to involve some right brain activity as we iterate the process with carefully modified search criteria. Unfortunately, traditional techniques, employed when searching with precision, do not provide much assistance with this type of problem and the user is left to try query after query until they have exhausted all permutations.

Linguistic, or semantic, approaches to information retrieval analyse the sentence structure and can potentially resolve ambiguities introduced by the same word being used as noun in one context and as a verb in another. In addition, semantic networks can be used to automatically expand queries and to create linkages between documents. Convera is a good example of a linguistic system and does offer some assistance to the user’s right brain with options for query expansion and document linking. Like all linguistic products it is highly language specific and its performance will vary depending on the application, vocabulary and language.

The probabilistic (Bayesian inference) approach to information retrieval was first implemented in the 1980’s and provides a statistical model for word frequencies based on the documents in the index collection. The model can then be used to weight multi-word queries so that the results can be ranked in importance and explicit or implicit relevance feedback can be exploited. Vendors of this type of product include: Autonomy, and APR/Smartlogik.

Autonomy provides automated query expansion and can also identify related documents automatically. APR/Smartlogik can provide a list of related words, although these are displayed in truncated form based on stemmed versions of the words.

It is important to understand that all of these products perform their statistical analysis based on an analysis of discrete words with the assumption that words occur in documents on an independent basis. Therefore, Autonomy’s query expansion always adds isolated words and APR/Smartlogik’s list of related topics are always single words.



Consider the following examples:

Isolated Words	Concepts
Price	Price sensitive information
Dealing	Insider dealing
Company	Company securities

The isolated words have no real meaning since they are ambiguous out of context whilst, in contrast, the two and three word phrases have clear meaning. Professor Stephen Robertson, widely regarded as the father of the Probabilistic Model, has stated repeatedly that the assumption of word independence is “patently not justified” and exists simply as a “matter of mathematical convenience”.

Concept Searching Systems

We all know that it is the combination of words that convey the main concepts in any language and that words used in isolation, or taken out of context, can be confusing and ambiguous. And yet all of the established probabilistic vendors perform their analysis on the basis of word independence.

In order to qualify as a true concept search engine a product must be able to isolate the key meaning that is normally expressed as proper nouns, noun phrases and verb phrases. Linguistic products can do this but their performance is highly variable depending upon the vocabulary and language in use.

One way to do this mathematically is to apply Shannon Information Theory to compute the incremental value of compound terms (i.e. multi-word terms) over their lower order component parts. In this way a phrase like “New York” would receive a significant weighting given that the two-word concept has significant value (statistically) over and above the individual words. In contrast, “New Zealand” would receive very little, or zero, weighting since the value of this two-word phrase is entirely contained in the word “Zealand” (assuming the word “Zealand” never appears except as part of the two word phrase).

Lateral Thinking in Information Retrieval

In order to provide some assistance to the right brain when searching unstructured information the ability to automatically identify multi-word concepts is absolutely fundamental. Without this ability the system is simply analysing individual word frequencies that are unlikely to make much sense to a human brain when taken out of context.

A true concept search engine can accept queries in natural language with the user simply typing words, phrases or even whole sentences in order to state what is on their mind. The system then analyses the natural language query to extract the key words and phrases that identify the main concepts. No complex syntax is required and the user can expand and refine their search simply by editing the free text description that they have entered.

Getting users to type in more than a word or two when searching can be difficult and one way to counter this is to offer facilities to drag and drop text from an existing document or from within the hitlist and use this as the basis of a new search. Alternatively, a ‘more like this’ facility that locates related document could use an entire document as the basis for a new query. Traditional systems can also support these features to a limited extent but only a true concept search engine will extract the multi-word phrases that capture the essence of these potentially very large queries.

Often a user will have a concept in their mind but are unsure what terminology is being used by the document collection they are searching. This is especially a problem if the documents are from external sources or if the documents are not subject to the requirements of a controlled vocabulary. In these cases it would be ideal if the user could conduct a search and then have the system suggest highly correlated terms found in the matching documents.

As an example, suppose that we are searching the UK government web site at <http://www.hmsso.gov.uk> which publishes UK Acts of Parliament and related documents dating from 1988. We might begin a search with a simple two word query like:

insider dealing

Of course the system returns a relevance ranked list of matching documents each of which may have a dynamic summary showing the query concepts in context since this is far more useful than static summaries when deciding which documents are worth opening and reading in detail. Each document may also be used as the basis for a new query effectively finding related documents at the touch of a button.

Now, imagine if the system were to offer a list of matching, multi-word, concepts that were found in the matching documents and were highly correlated to the user's original query. It might look like this:

Related Topics:	
To expand the search select entries and then search again or click a link to start a new search	
<input type="checkbox"/>	Company Securities
<input type="checkbox"/>	Regulated Markets
<input type="checkbox"/>	Insider Dealing Order
<input type="checkbox"/>	DEALING ORDER
<input type="checkbox"/>	Insider Dealing Act
<input type="checkbox"/>	unpublished price
<input type="checkbox"/>	unpublished price sensitive

In fact, this list of related topics was generated automatically from an index of the HMSO data, which can be seen at: <http://www.conceptsearching.com/conceptHMSO/>.

The list of related topics can be based on a pure statistical analysis or it can be generated from a personalisation profile that identifies topics that the user is interested in. This approach to user personalisation can automatically identify concepts in any search that would otherwise be overlooked.

The related topics tend to be compound terms and consist largely of proper nouns, verb phrases and noun phrases – even though there is no semantic analysis. So, this technology works with any vocabulary and in any of the 12 supported languages.

Now the user can switch the search to a related topic by clicking on any one of the hyperlinks. Or, they can expand the query by selecting any number of the checkboxes alongside the related topics and re-issuing the query. In this way a complex query can quickly be constructed just by typing a couple of words and then clicking the mouse a few times.

At last, some food for the right brain...