Information Retrieval Techniques
In Commercial Systems

Universidade
do Porto
Faculdade de
Engenharia
FEUP

Mestrado em Gestão de Informação

Disciplina: Armazenamento e Recuperação de Informação

Docente: Mark Sanderson

Autor: Filipe da Silva

Information Retrieval Techniques in Commercial Systems

Abstract

Looking for the commercial papers of commercial IR systems we find a group of terms like "innovative", "fast", "exactly what you are looking for".

As "fast" is only relative, what really do they mean of "innovative"? Do they really find "exactly what you are looking for"?

In this paper we study some commercial papers of commercial IR systems to find if they really can manage that.

Introduction

In this report we will outline the IR techniques being used by some commercial IR systems. We based our study in the publicity material and White Papers available in the product company Web Site.

This products aim normally the company documentation systems and so the universe of the information is narrowed that a web search engine, but needs to have a better precision as several similar documents can be found.

We should notice that we based our study in commercial papers and not in technological descriptions of the system, and so we will that some assumptions about the real techniques that the system are applying, finding if the systems are really innovative as they claim.

We do not test the systems and even if you could test them we could not compare them unless we used the same test collection [Salto et al, 1997].

The systems we analyzed were Verify, Autonomy, WebTop and Dolphin Search.

Verify Search analysis

Verify Search uses what they call fuzzy search, which is in reality the use of stemming, synonym expansion (using thesauri), wildcard search and terms addition like spelling and typographic al errors and sound like and phonetics like; and using a concept extraction which could be done using Latent Semantic Indexing, finding the terms that commonly co-occur in the collection. According to a paper from a person connected to another company in this concept extraction Verify uses also an artificial neural network [Roitblat].

Verify Search also allows using a part of a given document as the entry to find related documents, which is really a way to increase the size of the question term and by that narrowing the search.

Verify Search says it constructs an index with all the words found in the documents, mapping the words to documents and some short of location within the document. To do this it can use an index with word plus location type. To use word proximity the collection index must also store additional information.

The use of an all word index means that the stemming will be done later in query mode, and this technique is used because verify provides an operator that allows the search by the exact word. This should slow the system in query mode if no exact match is needed.

The multi-language support means the use of dictionaries and thesaurus (it doesn't say about the use of a master language but that should be "natural" and should be English) and maybe some way (using words relationships) to improve the weak result that gives, or even the use of multi-dictionaries (across languages).

In this multi-language support, we think that not all the search algorithms are available. Language depended operations like stemming, grammatical analysis and phonetics are not easily implemented for one language, and we do not believe that is applied for all the 24 supported languages. We should again notice that it is a commercial paper and not a technological description of the system.

Verify also uses a collection administration to fine the index relations. The manual indexing and relation concepts are then also an important factor of the value of the system. One way to do this pseudo-automatically would be to present to the administrator the result of finding some high-related term (found via some method like statistical or neural net) to be validated, or fine tuning around a business area.

Besides the operators that can provide explicit weight to question terms, that could also be implicit in long questions, it does not seem that Verify search give to people new features/technology, besides the neural nets that are supposedly used, in terms of retrieving. Verify has the ability of using filters to apply the system to a big source of documents types.

Autonomy analysis

Autonomy uses a technology they called Dynamic Reasoning Engine (DRE) based in the Bayesian Inference and Claude Shannon's principles of information theory to get a "concept" of the document. This is really a value of mathematical relationship of the terms in the query and the high-related terms in the collection and the inverse-frequency of the term in the collection.

Autonomy affirms it does not use any natural language analysis, which confirms they claim of language independency it's only for the total collection (no multi-language support).

Autonomy appears to use only simple statistical means used already in some web engines.

WebTop

WebTop uses a Linguistic Inference algorithm in their commercial product that provides the system of ways to extract and identify concepts, identify user interests and integrative refinement.

The explanation of what the system is doing is presented in the appendix of its white paper [WebTop]. It doesn't index stop words (words that appears to many times in the collections and in the documents), and uses a mathematical formula with the inverse frequency and intra-document frequency of the question terms. Also a proximity algorithm is used which suggests the broken of pieces of the documents. This is a rather, nowadays, common algorithm of Information Retrieval.

What WebTop provided in its commercial product is the storing of a user profile that will enlarge the query terms used, and allows the user to tell the system that a document is useful but that other doesn't have any interest.

This will provided the system with a growing term query that will narrow the search.

This idea is somewhat similar to the Verify Search notion of using a portion of a document to enhance the query.

Because WebTop system doesn't index stop words search like "to be or not to be" will return no Shakespeare related documents (if they are presented in the collection), but this kind of question will not arise many times in an organization (if it's not a library!).

Dolphin Search analysis

Dolphin Search claims it uses a ratter different technology.

Their idea is the use of artificial neural networks to provide the system with a concept "intuition".

As words have many meanings dependent of the context, and that people store in their memories a "concept image" and not the real words, the idea of using the artificial neural networks seems rather interesting.

Dolphin Search uses two neural networks to accomplish that. Each piece of text is translated to a vector in which an element codes a word in the systems vocabulary (0 if the word isn't present in the text/paragraph being represented or a positive value if otherwise). One neural network learns the pattern of the words relations in the documents and learns the meaning of those words in terms of vectors. Another neural network learns the relationship between the words of the vocabulary and the documents.

Dolphin Search claims that this method implements the Bayes Rule.

Each document is then represented by a "semantic profile" and Dolphin Search uses then neural networks to find the documents with similar patterns and similar to the pattern of a question. Dolphin Search does not say how the real artificial neural network training is being made. It appears that some kind of interesting vocabulary index should be provided and unless the neural network is ready to learn itself it will need a test collection to figure out the weight connections of the neurons.

We found this method the most interesting or innovative but because we did not have access to it, we could not make some tests with it.

Conclusion

Many commercial papers, like the Verify Search and Autonomy, talk about Boolean search as the worst thing that could be made. That because it can only return documents with no weights, forgetting that nowadays everybody is expecting and getting Boolean search that uses term weights (inside and between the documents) and relationships between words, because already most internet used search engines use this kind of simple technology. The simple Boolean search is still available in many other systems, but those ones uses normally thesaurus.

As we expected some of the innovations, most of the company's claims are really simple application of ideas and techniques already available on information retrieval for some time.

Retrieving precision is improved in some of the systems by the use personal or group type profiles, by interacting with the user, or by the use of manual categorization.

It appears that there is a large market for IR system in organizations, but the most commercial products only provided means to search a variety of documents formats and are not really trying to improve the retrieval precision with any new IR technique) besides the Dolphin Search).

References

[Autonomy] Autonomy's technology White Paper, http://www.autonomy.com/echo/userfile/Autonomy_Technology_WP(0401).pdf, visited on 2/06/2001

[Dolphin, 2000] Dolfin Search, Detailed technology white paper, 2000, http://www.dolphinsearch.com/downloads/dsWhitePaper2001.pdf, visited on 2/06/2001

[Salto et al, 1997] Gerard Salton, Christopher Buckley "Term-Weighting Approaches in Automatic Text Retrieval", in Information Processing and Management, 24, 513-523. Reprint in "Readings in Information Retrieval, Karen Sparck-Jones and Peter Willet, Morgan-Kaufmann 1997, pp. 323-328.

[Roitblat] Roitblat, Herbert L., "Biomimetic Systems for Information Retrieval", http://www.dolphinsearch.com/downloads/InfoTodayTalk.doc, visited on 5/06/2001

[WebTop] WebTop, "Introducing Linguistic Inference", http://www.webtop.com/docs/linginf.doc, visited on 2/06/2001

Web Pages of the companies

Autonomy http://www.autonomy.com/

Verify Search http://www.verity.com/

WebTop http://www.webtop.com/

Dolphin Search http://www.dolphinsearch.com/