conceptSearching

Compound Term Processing White Paper

Prepared by:
Concept Searching
8300 Greensboro Drive
Suite 800
McLean
VA 22102
USA
+1 703 531 8567

9 Shephall Lane
Stevenage
Hertfordshire
SG2 8DH
UK
+44 (0)1438 213545

info-usa@conceptsearching.com
http://www.conceptsearching.com
Twitter:@conceptsearch
Concept Searching Blog

John Challis
Chief Executive Officer and Chief Technology Officer
john@conceptsearching.com

March 2015

Abstract

This White Paper provides an introduction to Concept Searching's compound term processing technology. The major features are described and a range of applications discussed.

Author Information

John Challis, Chief Executive Officer and Chief Technology Officer at Concept Searching, 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.

John then launched ImageFirst Office for BancTec in the USA in 1995, closing over $5m new business in the first 12 months. Prior to Concept Searching he was Chief Technology Officer at Smartlogik, the company behind the world's first probabilistic search engine.

He is the originator of the company's compound term processing technology and is the driving force behind the product strategy.

Table of Contents

Overview

This White Paper provides an introduction to Concept Searching's compound term processing technology. The major features are described and a range of applications discussed.-

A Brief History of Enterprise Search Technologies

Keywords

The first search engines were based on keyword matching – if a document contains the search terms then it is returned in the results, with ranking of results based on the 'within document frequency' (wdf).

Phrase searching could be used, although users must specifically identify phrases using quotation marks. Phrase searching will improve precision but will reduce recall because any document that does not match the phrase exactly is ignored completely.

Boolean operators added a degree of sophistication, but also tend to improve precision at the expense of recall because any document that does not match the Boolean expression is ignored. The vast majority of search users are unable to formulate even basic Boolean expressions.

Probabilistic Model

Products based on the probabilistic model started to appear in the 1980s and these introduced the idea that the importance of each search term is related to its 'within collection frequency' (wcf). The theory is reputedly based on Bayesian inference, although in practice the initial search weighting are only trivially different from a more simple inverse term frequency analysis. The Bayesian inference logic only makes a difference when relevance feedback is implemented – but I am not aware of a single example of a successful commercial search engine that implements relevance feedback today, because users do not like it. As a result, we find that the probabilistic model did improve search precision without loss of recall, but was really little more than a keyword search engine that implemented weightings for single words based on inverse term frequency. The probabilistic model did nothing to address the problems caused by the ambiguity of single words processed in isolation because it continues the assumption that terms are mutually independent.

The most significant improvement in search results at this time was arguably due to the research by Robertson and Spärck Jones that led to the BM25 weighting function. Although based on the probabilistic framework, its power lies in the combination of wdf, wcf and 'normalised document length' (ndl) parameters, rather than a tenuous link to the work of Thomas Bayes.

Latent Semantics

In the late 1990s we started to see the appearance of a search engine based on latent semantic indexing and then probabilistic latent semantic analysis. These two techniques are quite different, but both are based on complex matrix algebra. Both approaches claim to improve search precision by resolving the basic problem that keywords can be highly ambiguous. Whilst both approaches recognise that terms do not occur independently, it remains true that the underlying algorithms remain too complex for most practitioners and, as a result, very few people have the required knowledge to

tune the algorithms for different applications. For example, if the results are disappointing should the matrix dimensions be increased or reduced, Landauer and Dumais (two of the original latent semantic indexing patent holders) state that any number between 50 and 1,000 dimensions is suitable depending on the size and nature of the document collection. As a result of its recondite nature, latent semantics has largely failed to deliver products that work for the mass market.

The leading search technology emerging at this time was probably Autonomy. Whilst the algorithms utilized by Autonomy are not in the public domain it seems likely that much of their matching technology is based on matrix algebra.

Faceted Navigation

From the year 2000, we have seen rapid growth of search interfaces that incorporate faceted navigation. This type of search filtering is becoming prevalent on the Web, as well as many popular enterprise search products. There is no doubt that many users find such search filters useful, although generating the required metadata can be a challenge, especially when the facets are based on taxonomy structures.

"In 2003 Concept Searching introduced compound term processing, which was a breakthrough technology that could identify and weight multi-word concepts based on a purely statistical analysis. At last we have a technology that understands the relationships between words but which is independent of vocabulary, grammatical style and language."

Linguistic Processing

Linguistic processing seeks to improve search results by analysing sentence structure and using the results to move beyond the assumption of term independence. Convera was arguably the leading vendor in this area with its semantic network technology. Convera originally focused on neural networks technology (as Excalibur Technologies) but switched to a linguistic approach in 2002, after it acquired Semantix. Whilst Convera received many accolades for its technology, it was attempting to solve an intractable problem. Its manually constructed semantic network attempted to define a synonym ring for every concept used in any vertical sector and translated into every language. Over $1Bn was invested by shareholders but this investment was lost when Convera finally stopped trying to 'boil the ocean' in August 2007, when it was sold to FAST for $23m. The product was subsequently withdrawn from the market.

Compound Term Processing

In 2003 Concept Searching introduced compound term processing, which was a breakthrough technology that could identify and weight multi-word concepts based on a purely statistical analysis. At last we have a technology that understands the relationships between words but which is independent of vocabulary, grammatical style, and language.

Prior to IDOL 7, Autonomy's products for search, classification and query expansion were based entirely on single word processing. IDOL 7, released in 2008, introduced a multi-word concept identification technology that appears comparable to the Concept Searching technology released more than five years earlier.

To the best of our knowledge, Concept Searching and Autonomy remain the only vendors with enterprise search and auto-classification products available today that incorporate statistical technology for the identification and weighting of multi-word concepts in unstructured text.

What is Compound Term Processing?

Simple Multi-word Queries

Consider the following query text: "John Major the UK Prime Minister".

A document containing this text is clearly relevant:

"Sir John Major served as Prime Minister of the UK from 1990 to 1997"

Unfortunately, many search engines available today would rank the following text equally:

"John Prescott, a Labour minister, stated that the major spending on infrastructure in the UK was a prime example of Labour's policies"

This is because most search engines treat each search word independently.

Note that a phrase search would find neither document. Even this phrase search would find neither document: "UK Prime Minister".

Some search engines, such as FAST, do implement linguistic techniques that can identify noun phrases and proper nouns and these are added to their lexicon as multi-word concepts. These can then be used to improve the weighting of the first document relative to the second, but only because the query text contains proper nouns.

"Crucially, compound term processing greatly improves search precision but with no reduction in recall – a unique characteristic."

Linguistic techniques used in this way will only ever detect a tiny proportion of all multi-word concepts found in unstructured text. Why limit search improvement to a tiny proportion of all searches based on 'cherry picking' concepts, when you can improve all searches by extracting all concepts?

With compound term processing the following search terms are extracted from the query shown above:

John; Major; UK; Prime; Minister; John Major; UK Prime; Prime Minister; UK Prime Minister.

These search terms will locate both documents but will rank the first document significantly higher, due to the fact that it matches more of the query terms.

Note that neither document contains all of the query terms, but this does not prevent both from being found. Crucially, compound term processing greatly improves search precision but with no reduction in recall – a unique characteristic.

Complex Multi-word Queries

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.

"Only compound term processing offers a solution that will improve precision, by adding higher-order query terms that boost the ranking of relevant documents, with no loss of recall because the less relevant documents are still available towards the bottom of the hit list."

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

All traditional search techniques will struggle with this type of matching. If techniques that favour precision are used (e.g. Boolean expressions or phrase searching), then recall will suffer. If techniques that favour recall are used (i.e. single word processing), then precision will suffer.

Only compound term processing offers a solution that will improve precision, by adding higher-order query terms that boost the ranking of relevant documents with no loss of recall because the less relevant documents are still available towards the bottom of the hit list.

Compound term processing achieves this remarkable feat using a combination of techniques.

Concept Identification

The above query will generate the following query terms: "insider; dealing; shares; directors; access; unpublished; price; sensitive; information; insider dealing; unpublished price; price sensitive; sensitive information; unpublished price sensitive; price sensitive information".

The same concept identification happens when documents are indexed and so the Concept Searching index contains a lexicon of all single words and all multi-word concepts.

In addition, the technology is smart enough to identify various forms of multi-word concepts such as:

- Euro-zone economies
- King of the mountain
- Carta de crédito (in an Italian document)
- Juguetes de los niños (in a Spanish document)

The concept identification works in any language based on the Roman alphabet. And because the technology is statistical, it works regardless of the vocabulary (i.e. vertical sector) or the grammatical style employed.

Weighting of Compound Terms

Concept identification alone is not sufficient to deliver superior matching, term weighting is also vitally important. And traditional term weighting techniques, such as the probabilistic model or matrix algebra, cannot be used because they will skew search results incorrectly.

In order to utilise compound terms in the ranking algorithms, it is necessary to understand the incremental value of higher order terms with regard to their lower order component parts.

For example, how much additional information is conveyed by the term "price sensitive" compared to its component parts "price" and "sensitive". And how much additional information is conveyed by the term "price sensitive information" compared to its component parts "price sensitive" and "sensitive information".

It is not always the case that higher order compound terms score more highly than lower order or single word terms. In fact we often find that the incremental value of the highest order compound terms is negligible. For example, the term "Apple computers" may appear commonly in an IT environment, but less often in a retail environment where the term "apple juice" may be more common. So, what is the incremental value of the term "Apple computers"? The answer is that really all depends on the corpus.

Compound term processing answers these questions using proprietary weighting algorithms that were original developed in 2002 and subsequently tuned for over a decade.

Compound Term Processing – Not Just for Search

As we can see, compound term processing can dramatically improve search results, but it has many other applications. Whilst some of our customers do use compound term processing for enterprise search applications, the majority of our customers today use it for one of the following:

Auto-classification

The benefits of metadata applied to otherwise unstructured content are widely understood. The metadata can be used to improve findability, automate corporate policies, identify records, migrate documents from one platform to another, identify sensitive content, etc.

"Only Concept Searching offers a simple to use rule building environment that can automate the generation of rules using its integrated search engine. This approach is unique because Concept Searching is the only rules-based auto-classification vendor today that implements Compound Term Processing."

But metadata generation, especially that based on corporate taxonomies, remains an intractable problem – humans just cannot be relied upon to do it accurately and consistently. The only viable solution is to automate the process.

Unfortunately, automated metadata generation is difficult to achieve consistently with high precision and recall. Many auto-classification products on the market today require complex rules to be generated often involving search syntax. Some even require a document training set for every term to be processed. These techniques create a very high initial cost both in terms of the time taken and also the level of qualified staff required. Most of these products employ linguistic techniques that will not perform consistently across different vertical markets. And the grammatical style of a legal contract or patent application is very different to that of a news article or a typical web page. These differences contribute to the difficulties found when using products based primarily on linguistic processing

Only Concept Searching offers a simple-to-use rule building environment that can automate the generation of rules using its integrated search engine. This approach is unique because Concept Searching is the only rules-based auto-classification vendor today that implements compound term processing.

Search Suggestion

Search suggestion (aka incremental search) has become popular due to its convenience and accuracy found when searching the Web with, for example, either Google or Bing. Various attempts have been made to implement similar functionality in an enterprise search environment. Unfortunately, the vocabulary used within the enterprise is very different to that used by popular Web searches. As a result different techniques are required to identify the vocabulary to be used by any search suggestion product.

Existing products tend to be based on either single words or require a history of searches to be available. Single word suggestions are of very limited value and a reliance on search history means that it will take time for a suitable lexicon to be built and the lexicon will never contain relevant concepts.

The requirements for search suggestion do include some additional challenges compared to most matching applications. For example, this query text "Hong Kong airport" will generate query terms "Hong", "Kong", "airport", "Hong Kong", "Kong airport" and "Hong Kong airport" when using compound term processing. This is ideal for ranking search results, but the search suggestion should not offer the term "Kong airport" since it is not meaningful to a human.

Only Concept Searching offers a search suggestion mechanism that can include suggestions based upon every word and also every multi-word concept used by an enterprise. This approach is unique, because Concept Searching is the only search suggestion vendor today that implements compound term processing.

Predictive Coding

Predictive coding takes input from human beings, normally in the form of document training sets (aka pre-classified documents), and uses this information to cluster and rank unclassified documents. Equivio and Recommind are two of the leading vendors in this area, especially for eDiscovery and legal applications.

"Only Concept Searching offers a search suggestion mechanism that can include suggestions based upon every word and also every multi-word concept used by an enterprise. This approach is unique, because Concept Searching is the only search suggestion vendor today that implements compound term processing."

Compound term processing can be used for predictive coding applications, based on its ability to extract the key concepts found in one or more documents, and then uses this list of concepts to deliver a ranking of all available documents based on their relevance to the training set. Concept searching is not unique in this area, but compound term processing probably offers an alternative approach that may offer different advantages compared to the other leading competitive products.

Automated Query Expansion

Query expansion remains one of the great challenges faced in information retrieval. The basic idea is to retrieve relevant documents even if they do not contain the actual words used in the query text. For example, a search for "laptops" should perhaps return a document about "mobile computing devices" even if that document does not mention "laptops".

It is fairly simple to create synonym rings to do this, but as with Convera discussed above, it is impossible to construct synonym rings for every concept in every language. And any synonym list will require constant manual effort as new terminology emerges.

Query expansion can be implemented explicitly (with the user selecting terms from a list suggested by the search engine) or implicitly (with the search engine automatically appending the most relevant suggested terms).

Autonomy is famous for offering automated query expansion in its IDOL product line. Prior to IDOL 7 (released in 2008) its query expansion was always based on single words added to the query in isolation. As a result, the accuracy of Autonomy's query expansion was highly variable prior to the release of IDOL 7, when its performance improved dramatically.

Concept Searching has offered automated query expansion since it launched its products in 2003. Today, only Concept Searching and Autonomy offer this facility with any degree of accuracy and consistency.

Summary

Compound term processing has the potential to be widely adopted as the next major advance in information retrieval.

Following its acquisition by HP in 2011, the Autonomy product line seems to be in decline and may never recover its market leading position.

Concept Searching's compound term processing has the potential to step into the void and become the underlying technology across a range of information retrieval applications.

About Concept Searching

Concept Searching is the industry leader in advanced semantic metadata generation, auto-classification, and taxonomy management. Its award winning products are the only statistical metadata generation and classification technologies that use compound term processing to generate intelligent metadata from unstructured and semi-structured data. Compound term processing, or identifying 'concepts in context', solves a variety of business challenges. Using the concept identification capabilities, organizations can transform content into business assets to improve performance.

Concept Searching's Smart Content Framework™ for information governance is a combination of best practices and underlying products that encompass the entire portfolio of unstructured information assets, resulting in increased organizational performance and agility. The output from the Smart Content Framework™ delivers intelligent metadata enabled solutions that are being used to enable concept based searching, automatic declaration of documents of record, identification and protection of privacy and confidential data, intelligent migration, content management, granular identification of content for text analytics, and improved delivery of social content. The solutions are deployed in diverse industries, Fortune 1000 companies, and smaller companies that need to meet strict compliance, data privacy, and information governance regulations.

Concept Searching has a Microsoft Gold Application Development competency and is a participant in the global Business-Critical SharePoint program. Although platform independent, the Concept Searching Microsoft suite of products uses a single code base, supporting all versions of SharePoint, Office 365, and OneDrive for Business, providing clients with the choice of on-premise, cloud based, or hybrid environments to best meet their needs.

Headquartered in the US, with offices in the UK, Canada, and South Africa, Concept Searching solves the problem of finding, organizing, and managing information capital. For more information about Concept Searching's solutions and technologies please visit www.conceptsearching.com and follow on Twitter and LinkedIn.

Business-Critical SharePoint Partner

Microsoft Partner
Gold Application Development

Americas
+1 703 531 8567
info-usa@conceptsearching.com

Europe
+44 (0)1438 213545
info-uk@conceptsearching.com

Canada
+1 703 531 8567
info-canada@conceptsearching.com

Australia
+61 (0)2 8006 2611
info-australia@conceptsearching.com

New Zealand
+64 (0)4 889 2867
info-nz@conceptsearching.com

Africa
+27 (0)21 712 5179
info-sa@conceptsearching.com

Marketing and PR
International: +1 703 531 8564
Europe: +44 (0)1438 213545
marketing@conceptsearching.com

Follow us on Twitter
@conceptsearch

www.conceptsearching.com