conceptSearching

Technology Overview
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

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

October 10th 2011

conceptSearching

Abstract

This White Paper discusses the technology framework utilized for the development of the Concept Searching suite of products that provides significant advantages to organizations wanting to capitalize on their information assets.

Author Information

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

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.

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 CTO 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.

Overview

Connecting knowledge workers with relevant and timely information in an efficient manner has been proven to provide considerable time and cost savings. In order to gain this advantage in today's business environment, management must implement processes and toolsets that effectively maximize the return on intellectual and information assets through the re-use of existing knowledge within their organization.

The Concept Searching toolset helps clients develop and integrate resources for intellectual capital management. A web-based delivery model developed to virtually share, collaborate, distribute and exchange knowledge among members of an organization, the products support an intelligent portal for content management, workflow, e-learning, and collaboration.

Concept Searching technologies provide:

- Automatic Query Expansion
- Relevance ranking based on the Probabilistic Model (Bayesian Inference)
- Dynamic summarization
- Concept identification based on Shannon's Information Theory
- Cross platform compatibility via web services
- All Application Programming Interfaces (APIs) based on XML
- Transparent access to system internals including the statistical profile of terms
- True relevance ranking of compound (i.e. multi-word) items
- Simple GUI for building and maintaining taxonomies

What does this mean to the knowledge worker?

- High recall and high precision
- Contextualization - finding hidden relationships between documents within the enterprise
- Exceptionally high sorting rates
- Language independence
- Scalability

Most knowledge management platforms are robust document management systems. They are heavily dependent on subject matter experts (SMEs) who upload and organize information within their respective business domain. This is both time consuming and subjective, resulting in a stove-piped, static presentation of information. Retrieval of information is dependent on active searches conducted by end-users. This can be a hit or miss approach and is compounded by a lack of integration between existing knowledge management platforms and document management systems. Often little or no connectivity exists to other valuable data sources.

In almost every organization, unstructured information is categorized and indexed by various individuals using 'subjective' factors in a manner that is very time consuming and ineffective. The Concept Searching approach uses 'objective' criteria (established and accepted taxonomies) to automate the classification function. Unstructured information is automatically indexed and relationally organized according to organizational, functional taxonomies. Individuals can be automatically notified any time relevant information is added.

Knowledge Discovery Technology

Bayesian Inference

Thomas Bayes was an eighteenth century mathematician who devised a theory for conditional probability:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Conditional probability is the probability of some event given that some other event has already occurred. In the above equation the left hand term P(A/B) is known as the posterior probability or the probability of some event A occurring given that event B has occurred is equal to the probability of event B occurring given that event A has occurred, multiplied by the probability of event A occurring and divided by the probability of even B occurring.

The Probabilistic Model interprets Bayes' Theorem in an Information Retrieval (IR) context where the probability that certain query terms are better differentiators between relevant and non-relevant documents than other query terms evaluated given implicit or explicit relevance feedback.

Probabilistic Model

The Probabilistic Model was pioneered at Cambridge University during the 1970's and 1980's. The model is an application of Baye's Theorem and defines a system of weighting individual query terms and documents based on:

- The frequency of terms across the document collection (wcf)
- The frequency of terms within a given document (wdf)
- Normalized document length (ndl)
- Explicit or implicit feedback on document relevance

In 1976 Professor Stephen Robertson and Karen Sparck Jones devised a formula for computing term weights and document weights and subsequently performed extensive evaluations on relevance feedback techniques using standard document collections. In 1994 Robertson introduced an extended model that was no longer based on a binary independence model and this work has strongly influenced the design of the Concept Searching products.

Why is the Probabilistic Model superior to traditional free text systems?

Traditional free text systems are based on simple keywords and Boolean logic (primarily the AND, OR and NOT operators). While this technique is very precise it does not perform well when the number of documents is too large to examine thoroughly. In this case the ability to rank documents, with the most important ones at the top of the list, is of paramount importance. Over time, traditional systems have introduced various ways to rank results, but this is not based on a sophisticated model of term profiles across the collection of indexed documents and tends to rely too heavily on a 'within document frequency' (wdf) analysis. The statistical model of term frequency across the document collection is unique to the Probabilistic Model. This model not only allows initial relevance ranking to be more accurate but it also provides a mechanism for iterative searching based on relevance feedback.

Automatic Query Expansion

Automatic Query Expansion is the ability to locate documents that are relevant to the user's query even if they do not contain any of the words in the user's query text. It is also has the ability to ignore documents that do contain words from the user's query but are not relevant.

Automatic Query Expansion is achieved by:

- Relevance ranking the documents matched by the initial query

- Extracting the distinguishing concepts from the most relevant documents

- Expanding the query to include selected related concepts

The inclusion of related concepts can be done explicitly (user decision) or implicitly where related concepts are included automatically based on an understanding of the application area and/or user personalization.

Imagine searching for 'portable computer' and finding documents that were about 'laptops', 'the Toshiba Tecra' and 'notebooks' where some of the retrieved documents do not contain any words from the original query - that's Automatic Query Expansion.

Relevance Feedback

Traditional information retrieval systems provide a static mechanism to index documents and service retrieval requests. Relevance feedback is used to describe dynamic mechanisms that allow the retrievals to be tuned over time based on explicit or implicit feedback from the user(s). An example of implicit feedback would be where a user identifies individual documents that are relevant to their query. Another example of implicit feedback would be where the system monitors the user's activity to see what documents they examine; how long they spend looking at individual documents; what documents they author or perhaps a common pattern to their retrieval activity. The Probabilistic Model allows this type of explicit or implicit feedback to be injected into the retrieval process so that the weightings applied are modified, or tuned, automatically to suit a particular user's requirements.

Concept Searching versus Simple Keyword Searching

An Automatic Query Expansion implementation that works on the basis of words appearing in documents independently from other words will provide a reasonable level of accuracy. However, if the implementation understands that the co-location of words is relevant and should form part of the weighting process then a significant improvement in the relevance ranking can be achieved.

For example, consider the following query:

'dangerous dog attacks baby'

A human would interpret this phrase as being about a wild animal attacking an infant. However, a simple information retrieval system that assumes that words appear independently from each other would assume that any document containing the phrase:

'dangerous virus attacks baby dog'

would be 100% relevant to the above query on the basis that it contains all of the words. Most humans would disagree.

Shannon's Information Theory

Concept Searching uses Shannon's Information Theory to compute the incremental value of compound terms based on an analysis of the probability of the joint occurrence.

Claude Shannon, a scientist working at Bell Labs, published his Information Theory in 1948 and this had an immediate and lasting impact on data communication technology. Shannon demonstrated that the value of a piece of information is proportional to its probability and the entropy of a joint event is given by:

$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j)$$

Concept Searching interprets this in an IR context to compute the incremental value of a two-word term over its single components. Higher order compound terms are evaluated using their lower order compound components.

It is no coincidence that the majority of compound terms are in fact proper nouns, noun phrases and verb phrases and it is these sentence fragments that convey the key concepts in most text. However, the concepts are identified without any linguistic analysis and so the products work with any vocabulary and are language independent. The mathematical approach works because Shannon's theory can be applied to any human language communication.

The ability of an IR system to identify clusters of words that identify specific concepts represents a major advancement over systems that fail to do this.

Language Stemming

Often a user will type in a query with one form of a word but would like to match other forms of what is essentially that same word. In 1980 Dr. Martin Porter, a member of the team working on a Probabilistic Model at Cambridge University developed a suffix-stripping algorithm that has been very widely adopted for normalizing words in information retrieval systems.

Using Porter's algorithm the following words can be matched:

> 'dangerous' with 'danger'; 'dangers' and dangerous'
>
> 'attacks' with 'attack'; 'attacks', 'attacker', 'attackers' with 'attacking'
>
> 'baby' with 'baby' and 'babies'

In addition, Concept Searching uses a fuzzy stemmer the following words can also be matched:

> 'misspelt' with 'mispelt'
>
> 'commission' with 'commision', 'comission', 'commissioning' and 'comisioned'
>
> 'accommodate' with 'accomodate' and 'acomodation'

Concept Searching uses language stemming as part of its concept matching process, although individual words and phrases may be left un-stemmed by enclosing them with double quotes. This means that by default stemming broadens the matching process but where a particular word should be interpreted verbatim it can be easily excluded from the stemming process.

Dynamic Summarization

When a document is retrieved we normally need to display an extract from the document as an aid to the user when reviewing the returned document set. Most systems will display a static summary that is the same regardless of the user's query. Concept Searching can display static summaries. However, it can also apply a modified weighting system to identify short extracts that are most relevant to the user's query. The number, length and relevance threshold for these extracts are all configurable. The extracts will normally comprise whole sentences or short paragraphs.

Supported Document Formats

Concept Searching has the following collectors:

- HTTP collector - for spidering web pages
- File collector - for documents located on file systems
- SharePoint collector
- SQL collector - for documents held in a SQL database (e.g. SQL Server or Oracle)
- XML collector - for custom document types
- Exchange collector

Concept Searching has native file conversion facilities for the following document types:

- All HTML and XML formats
- Adobe Portable Document Format (PDF)
- Microsoft Word and Rich Text Formats
- Microsoft Excel
- Microsoft PowerPoint
- Any other files in text format (e.g. TXT, CSV, etc.)
- Corel WordPerfect

In addition, third party iFilters can be used to convert virtually all other popular document formats (e.g. Microsoft Visio, email file formats, StarOffice documents, etc).

Database Support

Concept Searching stores its probabilistic index in a proprietary database. However, the indexer uses a SQL database to manage the queue of documents to be indexed. The SQL database contains all information necessary to perform indexing, such as the individual filenames and URLs, access criteria, re-indexing frequency, inclusions and exclusions, etc. The SQL database may also be used to store any application specific meta-data.

The SQL database can be either Microsoft SQLServer (2000 or later) or Oracle (8i or later).

Supported Languages

Concept Searching can index any text in the Roman alphabet including full support for diacritics. The use of diacritics within documents or queries is entirely optional so that fitchée will match fitchee and vice versa. All information is exchanged and managed internally; using UTF-8 and support for non-roman alphabets (e.g. Kanji or Arabic) can be accomplished based on client requirements. The following languages are automatically detected and processed:

- Afrikaans
- Danish
- Dutch
- English
- Finnish
- French
- German
- Hungarian
- Italian
- Norwegian
- Portuguese
- Spanish

Scalability

The developers of Concept Searching products have many years' experience in implementing proprietary file systems and custom databases. In particular the database format has been designed to allow concurrent indexing at full speed while allowing simultaneous access for retrievals. This concurrency has been achieved in part by reducing the amount of file restructuring typically found in competitive systems, which are often based on B-tree structures. The selected design tends to produce an index database a little larger than some alternatives but with faster retrieval. The proprietary database format has been designed to provide optimum performance and concurrency.

For testing and development the entire system can be installed on a single computer. For live implementations the Query Server, Index Server and the Web Application would normally be distributed. A multi-server configuration will be capable of indexing about a million pages per day while simultaneously providing retrieval to thousands of concurrent users. For very large implementations multiple Query Servers could be configured with shared access from a pool of application servers.

There is also a Distributed Query Server so that very large indexes can be partitioned over a number of servers to improve indexing performance.

High Recall AND High Precision

Recall is a measure of how many of the documents that are relevant get found, with high recall indicating that most of the relevant documents are found. Precision is a measure of how many documents in the returned set are relevant, with high precision indicating that most of the documents returned are relevant. Concept Searching offers high recall and high precision where other products offer only high recall or high precision.

Classification and Taxonomy Support

Concept Searching's classifier modules can be used by subject matter experts and analysts to easily build taxonomies and classify documents into predefined categories based on a small number of descriptors or clues. Once classified the documents can then be applied to a corporate taxonomy and used for browsing the database or as a filter when running ad hoc queries.

Summary

Concept Searching's products were developed to overcome the weaknesses inherent in currently available search and classification products. Utilizing a combination of sophisticated and proven algorithms Concept Searching has been able to successfully overcome the limitations of competitive offerings. Delivering robust solutions, organizations can now more fully realize the value of their information assets resulting in cost efficiencies, better decision making, and competitive advantages.
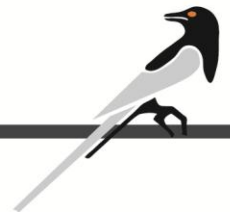
About Concept Searching

Founded in 2002, Concept Searching provides software products that deliver conceptual metadata generation, auto-classification, and powerful taxonomy management from the desktop to the enterprise.

Concept Searching is the only platform independent statistical metadata generation and classification software company in the world that uses concept extraction and compound term processing to significantly improve access to unstructured information. The Concept Searching Microsoft suite of technologies runs natively in SharePoint 2007 and 2010, FAST, Windows Server 2008 R2 FCI, and in Microsoft Office applications.

Headquartered in the US, with offices in the UK, South Africa and Canada, Concept Searching solves the problem of finding, organizing, and managing information capital far beyond search and retrieval. The technologies are being used to drive intelligent migration, enable effective records management, identify and lock down sensitive information and enhance governance. For more information about Concept Searching's solutions and technologies please visit http://www.conceptsearching.com.

Microsoft Partner

Gold Independent Software Vendor (ISV)
Silver Content Management
Silver Search
Silver Portals and Collaboration

2011 Concept Searching

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