A New Concept in Search

By John Challis, CEO, CTO, Concept Searching

With the exponential increase in unstructured information, enterprises are seeking new ways to improve not only the search and retrieval process but to identify tools to manage, capitalize on and leverage their information assets to improve organizational performance. Moving beyond keyword identification and traditional taxonomy approaches, the use of compound term processing or identifying "concepts in context" effectively addresses the issue of managing unstructured content and enables organizations to more effectively find, organize and manage their information capital.

Compound term processing automatically identifies the word patterns in unstructured text that convey the most meaning and uses these higher order terms to improve precision with no loss of recall. The algorithms adapt to each customer's content and they work in any language regardless of vocabulary or linguistic style. The technology was originally developed by Concept Searching in 2002 and is similar in many ways to the "phrase-based indexing" techniques detailed in various U.S. patents filed in 2004 and to which Google subsequently acquired the rights.

Keyword Search versus Concept Search

Knowledge workers need to identify content in the context of what they are seeking. The fundamental problem with most enterprise search solutions, and *all* statistical search solutions, is that they are based on an index of single words. Yet most queries are expressed in short patterns of words and not single words in isolation which are highly ambiguous.

A concept search engine can isolate the key meaning that is normally expressed as proper nouns, nouns phrases and verb phrases. Although linguistic products can do this, their performance is highly variable depending upon the vocabulary and language in use. A statistical-based, language-independent concept search can accept queries in natural language with the user typing words, phrases or whole sentences. The system then analyzes the natural language query to extract the keywords and phrases to identify the main concepts and retrieve content that is highly relevant.

Precision and recall are the two key performance measurements for information retrieval. *Precision* is the retrieval of only those items that are relevant to the query. *Recall* is the retrieval of all items that are relevant to the query. Yet most information retrieval technologies are less than 22% accurate for both precision and recall. The ideal goal is to have them balanced. Compound term processing has the ability to increase precision with no loss of recall.

Managing Content

Taxonomy development and maintenance has traditionally been a laborious and on-going challenge, not to mention costly. The most effective approach is to use rules-based categorization, providing enterprises complete control of rules-based descriptors unique to their organization. Since all rules can be defined and managed, error-prone results utilizing "training" algorithms typically found in other approaches are eliminated.

"Precision and recall are the two key performance measurements for information retrieval."

A concept-based automatic classification process identifies, during indexing, the categories each document belongs to. Each category is identified by a unique descriptor and is associated with key descriptive words and/or phrases held in the database. This approach enables a rapid implementation of a corporate taxonomy with all documents classified to multiple nodes at index time. Ideally, the taxonomy can be used to browse the document collection or as a filter when running ad hoc searches.

An easy-to-use taxonomy and automatic classification tool creates the framework to classify content based on concepts to one or more nodes in the taxonomy. Features that enable subject-matter experts to interact with the taxonomy can simplify on-going maintenance. For example: automatically generating compound term clues from the document corpus; dynamically showing the effect of changes on the taxonomy; and class weighting influenced by parent, child and sibling can reduce taxonomy development and on-going maintenance by 66%-80%.

Semantic Metadata Generation

The metadata generation issue is increasingly a growing concern in large enterprises. A comprehensive approach that requires more than syntactic metadata and that requires end users to add rich metadata is haphazard and subjective at best. Since the suggested approach is no longer restricted to keyword identification, compound-term metadata can be automatically generated either when the content is created or ingested. The generation of metadata based on *concepts* extracts compound terms and keywords from a document or corpus of documents that are highly correlated to a particular concept. By identifying the most significant patterns in any text, these compound terms can then be used to generate non-subjective metadata based on an understanding of conceptual meaning.

Compound-term processing is a new approach to an old problem. Instead of identifying single keywords, compound-term processing identifies multi-word terms that form a complex entity and identifies them as a concept. By forming these compound terms and placing them in the search engine's index, the search can be performed with a higher degree of accuracy because the ambiguity inherent in single words is no longer a problem. As a result, a search for "*survival rates following a triple heart bypass*" will locate documents about this topic even if this precise phrase is not contained in any document. A concept search using compound-term processing can extract the key concepts, in this case "*survival rates*" and "*triple heart bypass*" and use these concepts to select the most relevant documents.

Compound-term processing can address many challenges facing large enterprises and provide many benefits. Identification of concepts within a large corpus of information removes the ambiguity in search, eliminates inconsistent meta-tagging and automatic classification and taxonomy management based on concept identification, simplifies development and on-going maintenance. ▮

John Challis has had success with several ventures involving the management of unstructured data. In 1990, he founded Imagesolve International which became the UK's leading supplier of document image processing and workflow products. In 1995 he launched ImageFirst Office for BancTec in the US. He was also CTO at Smartlogik, the company behind the world's first probabilistic search engine.

Providing advanced search, auto-classification, taxonomy and semantic metadata-tagging solutions, Concept Searching is the first and only statistical search and classification company that uses compound-term processing to identify concepts within unstructured content.