

TopQuadrant Technology Research

Dictionary of Search Terminology

TQTR-Search02_color.doc	Date 4/10/2003	Page 1 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant		

Table of Contents

TopQuadrant Technology Research 1
Dictionary of Search Terminology 1
Search Technology Overview..... 5
 How Search Works 5
 Categorization and Search 6
 The Reasons for publishing and using a Dictionary of Search Terminology 6
Dictionary 6
 Adaptive probabilistic concept modeling (APCM) 6
 Boolean Search 6
 Bayesian Inference or Bayesian Statistics 6
 Capitalization 6
 Case Based Reasoning 6
 Categorization 6
 Controlled Vocabulary 6
 Corpus 6
 Dublin Core 6
 Fuzzy Search..... 6
 Genre Detection..... 6
 Grammatical analysis 6
 Guided Search 6
 Inbound Link..... 6
 Index File 6
 Information Gain 6
 Information Visualization 6
 Inverse Document Frequency (IDF)..... 6
 Inverted File 6
 Keyword Search 6
 Keyword targeting 6
 Knowledge Extraction 6
 Knowledge Model..... 6
 Knowledge Representation Language..... 6
 Language Identification..... 6
 Lexical analysis or Tokenizing..... 6
 Link Tracking..... 6
 Log File Analysis..... 6
 Metadata 6
 Meta Search Engine..... 6
 Meta Tag 6
 Natural Language Query..... 6
 Natural Language Processing 6

[Navigational Search](#) 6
[Ontology](#) 6
[Ontology Model](#) 6
[Parametric Search](#) 6
[Pattern Matching](#) 6
[Phonetic Analysis](#) 6
[Phrase Extraction](#) 6
[Precision](#) 6
[Pragmatic Analysis](#) 6
[Proximity Search](#) 6
[Query by Example](#) 6
[Ranking](#) 6
[Recall](#) 6
[Relevance](#) 6
[Relevance Modeling Technology](#) 6
[Results Management](#) 6
[Semantic Analysis](#) 6
[Semantic Web](#) 6
[Similarity Measures](#) 6
[Spiders or Crawlers](#) 6
[Stemming](#) 6
[Soundex Search](#) 6
[Summarization](#) 6
[Syntactic Analysis](#) 6
[Taxonomy](#) 6
[Term Frequency \(TF\)](#) 6
[Term Vectors](#) 6
[Thesaurus](#) 6
[Word Exclusion and Meaningless Terms](#) 6
[Word Location](#) 6
[Word Proximity](#) 6
Emerging Standards 6
[Knowledge Representation](#) 6
DAML 6
OIL 6
OWL 6
RDF 6
RDF Schema 6
TopicMaps 6
Metadata 6
Dublin Core 6
ISO/IEC 11179 6
About TopQuadrant 6
Additional TopQuadrant Technology Briefings are Available 6

TQTR-Search02_color.doc	Date 4/10/2003	Page 4 of 23
<p>Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant</p>		

Search Technology Overview

The problem of Information Retrieval continues to attract increasing attention as the oceans of unstructured data organizations have already captured, and are continuing to capture, keep on growing. At the same time accurate and speedy access to the information is becoming ever more difficult.

The benefits of implementing the right search technology can be significant: recent Delphi survey of 450 executives, IT or managers at large enterprise organizations with at least 50,000 documents shows that most spend more than 2 hours per day searching for information and 73% say that finding information is difficult. Effective search has been identified as essential to e-commerce: Market researcher Jupiter Media Metrix Inc. found that 80% of online users will abandon a site if the search function doesn't work well.

Information Retrieval, a technology behind text search tools, relies on a variety of mathematical models and algorithms to insure retrieval quality. The main purpose of these models is to establish the relevance of a document to a query. For a computer to do this it needs a model within which relevance decisions can be quantified. A number of technologies in this category are using clustering algorithms to determine similar documents and connection between documents. Another trend is to take in to account popularity of documents and user feedback. When the information is semi-structured fuzzy match strategies can be effectively supplemented with a more precise approach where inferencing can take in to account defined concepts and their relationships.

In selecting the right search tool many factors need to be taken in to account, including:

- ❖ Nature of typical search queries (studies have found important differences between e-commerce product search, customer service / tech support search and other types of searches)
- ❖ Form, format and location of the information and knowledge sources
- ❖ Organization's information publishing processes
- ❖ Availability of the metadata, custom dictionaries and taxonomies
- ❖ Readiness of an organization to engage in a continuous process necessary to achieve search precision

How Search Works

Information Retrieval is different from Data Retrieval, a deterministic process, where we are normally looking for an exact match. In Information Retrieval the exact matches may sometimes be of interest but more generally we want to do a fuzzy match: find the items which partially match the request and then select from those a few of the best matching ones. This makes the match process more probabilistic and explains why many search tools today rely on Bayesian logic to carry out inferences. Tools based on Bayes' Theorem are language independent since they do not attempt to "understand" the text. Autonomy is one of the best known technologies representing this approach. Search engines based on Bayesian logic and other probabilistic algorithms require careful training using small sets of representative documents in order to be effective.

Other approaches to inferencing over unstructured data include:

TQTR-Search02_color.doc	Date	4/10/2003	Page	5 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

❖ **Syntactic and semantic analysis**

These are most effective when they can take advantage of custom dictionaries of common terms and synonyms specific to the domain they search over.

❖ **Clustering algorithms**

Clustering algorithms are intended to determine similar documents and connection between documents

❖ **Ranking based on popularity of documents and user feedback**

A well known example of this approach in action is the Google Search Engine. Inside enterprise walls Google algorithm for popularity ranking can be taken advantage of only if the content consists of mostly HTML pages that are highly cross-referenced. If a sizeable portion of the content is in documents (MS Word, Adobe PDF and other formats), a different strategy for determining popularity and incorporating user feedback will be needed.

When the information is semi-structured i.e., XML documents and any files that have metadata or standard templates, fuzzy match strategies can be effectively supplemented with a more precise approach where inferencing can take in to account defined concepts and their relationships. This approach is very effective where a high quality metadata exists or can be derived.

The emerging trend in the search technology field is to improve the precision of search results by combining multiple approaches to Information Retrieval and including in the search query user profiling information - contextual data about end user activities and goals.

The figure below shows basic components of a search solution including index builder, user interface and a matching engine. As explained above different search engines employ different algorithms to determine what constitutes a match.

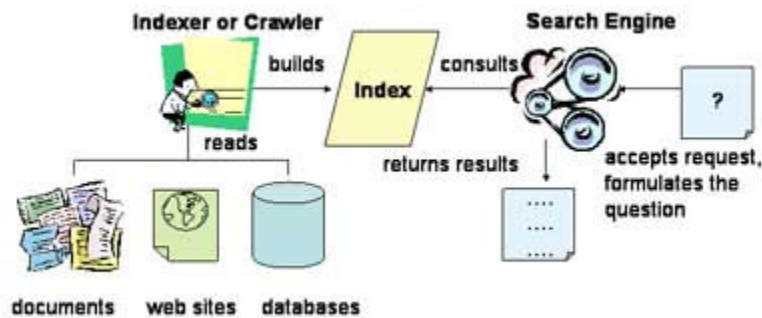


Figure 1: Basic components of the search process

Categorization and Search

A growing army of vendors is developing software to automatically index and categorize digital content. They use a taxonomy, or formal classification scheme, to determine where

TQTR-Search02_color.doc	Date	4/10/2003	Page	6 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

and under what headings content is filed, just as a library stamps a number on the spine of a book and files it in the stacks accordingly.

To end users, taxonomies offer a way to browse items in categories of interest-just as you might browse a section of books in a library. Full-text key word searches are good at pinpointing specific documents that match very specific queries, but they won't necessarily return related documents that don't contain the specific search term. In this way, a taxonomy not only brings order to digital collections for those maintaining the content, but also helps keep the collection accessible as it grows.

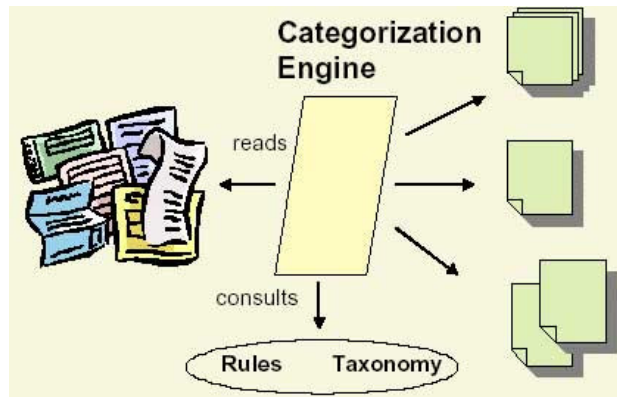


Figure 2: Categorization Engine

Auto-categorization technology is still in an early stage of maturity. Most vendors have proprietary products. Few adhere to standards such as the ability to import and export classification schemas from other repositories using XML and RDF or the ability to leverage standard classifications for specific industries.

Tools alone are not sufficient to make a categorization process effective. Most companies underestimate the amount of organizational change necessary in order for a taxonomy to make a meaningful difference in their organization. One of the key secrets to realizing business value of a taxonomy is to consider how it will be maintained over time, and to put that system in place early. The figure below shows the lifecycle stages of a taxonomy-based solution.

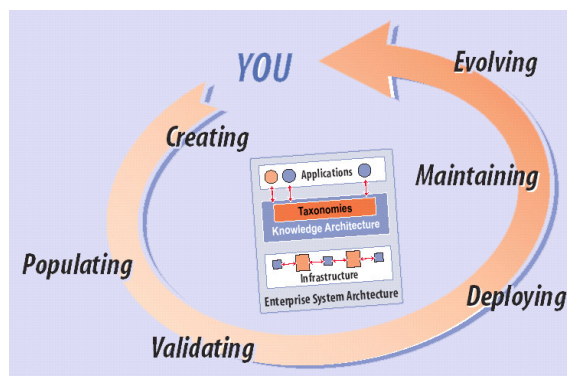


Figure 3: Taxonomy-based Solution Lifecycle

TQTR-Search02_color.doc	Date	4/10/2003	Page	7 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Most experts recommend piloting alpha and beta versions of the taxonomy before implementing the first production system. Taking a lifecycle view early enough in the taxonomy planning process ensures that the evolution of the taxonomy can be handled in a systematic way, and the personnel that will maintain the taxonomy get on-the-job experience right away.

To take in to account end user goals and points of view as they access corporate information, classification schemas employed by the auto-categorization tools have evolved from the traditional taxonomies (rigid hierarchies of concepts) in to light weight ontologies (flexible classifications enriched with the understanding of semantic and structural relationships between concepts).

The Reasons for publishing and using a Dictionary of Search Terminology

Language typically used by search and categorization consultants and vendors can be highly technical. This dictionary seeks to explain and provide some standard definitions for most commonly used terms.

Dictionary

Adaptive probabilistic concept modeling (APCM)

Analyzes correlations between features found in documents relevant to an agent profile, finding new concepts and documents. This is an extension of the Bayesian theory. Using this technique, concepts important to sets of documents can be determined, allowing new documents to be accurately classified.

Boolean Search

An advanced form of keyword search that enable the end user to author queries in a specific language using logical constructs (also known as Boolean form) i.e. "Dog AND (NOT (Sleeping OR Rabid OR Running*)) AND Sheep."

Bayesian Inference or Bayesian Statistics

Invented by Thomas Bayes, an 8th century English cleric known for his works on mathematical probability. Bayes' work centered on calculating the probabilistic relationship between multiple variables and determining the extent to which one variable impacts on another. A typical problem is to judge how relevant a document is to a given query or agent profile. Bayesian theory aids in this calculation by relating this judgment to details that we already know, such as the model of an agent. Pattern recognition based on Bayesian algorithms is entirely language independent. This approach is very different from the semantic analysis approach.

Capitalization

Identifies case variants (capitalization) of a word as separate words. This, for example, enables us to distinguish between documents containing the company name "NeXt" and documents containing the word "next".

TQTR-Search02_color.doc	Date	4/10/2003	Page	8 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Case Based Reasoning

Case-Based Reasoning, or CBR, is a technology that is often thought of in connection with help-desk applications. In fact, CBR has been employed in many domains where new problems can be resolved by consulting knowledge of past experiences held in "cases" CBR systems are essentially "learning" systems - each time a new experience is captured it is in the case-base to help future problems - a model, highly analogous to how people solve real life problems, enabling incremental and sustained learning.

Two forms of CBR can be distinguished: "conversational" and "parametric". In both, a problem is matched with a known solution through a guided inquiry or through "information gain" with the user. Conversational CBR systems are typically based on unstructured or informal models that emphasize the association of question and answer pairs with cases to promote (or demote) their relevance. On the other hand, parametric CBR systems have a knowledge model with a categorization schema using name-value pair attributes and strategies for constructing similarities and computing relevance. A query is expressed using the same knowledge model and relevant cases are determined by applying the knowledge model. The query need only be constructed over a partial cover of the knowledge model. Many conventional search engines use categorization schemes, but none have the benefits that stem from using the CBR strategies for computing similarity measures and determining information gain. CBR systems can also employ rules for completing cases and rules for adapting cases. These benefits have resulted in a number of impressive CBR-based knowledge systems that consistently demonstrate retrieval of relevant results from queries.

Categorization

Categorization places structure around similar document objects. The categories should be contextually appropriate for a specific audience or community of interest. For example, taxonomy is a form of categorization.

Automated categorization uses technology to organize content into groups. The result of automatic categorization is either a content collection clustered into groups (possibly a candidate taxonomy), or content categorized according to a pre-existing taxonomy. The best results are obtained by defining a business process that combines manual and automated processing so that technology is leveraged and human editorial input is optimized.

Controlled Vocabulary

This term refers to any organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search. Taxonomies, ontologies and thesauri are controlled vocabularies.

Corpus

A body or collection of document objects, usually textual.

TQTR-Search02_color.doc	Date	4/10/2003	Page	9 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Dublin Core

A set of 15 metadata elements (the [Dublin Core Metadata Element Set](#)) used to describe and catalog content so it can be discovered and retrieved. The Dublin Core is the de facto standard for cataloging web content.

Fuzzy Search

There are multiple definitions of the term. The most accepted definition is:

A search algorithm that allows some form of latitude such as a matching threshold in a pattern recognition engine like Excalibur's, or in the vector similarity matching that takes place in products like Aptex or Autonomy.

However, Verity's definition of the fuzzy search is that it expands queries to include terms that, for example, sound like or are typographically similar to the term requested. This capability is supported by [stemming](#), [phonetic analysis](#) and synonym expansion.

Genre Detection

Identifies the genre of a document. Knowing that a document is a newspaper article, a scholarly publication, a letter to the editor, or email message can make a difference. Users derive important information about the context, structure, contents, and value of a document from its genre.

Grammatical analysis

Identifies the part of speech for each word. Many more classes of words exist than would be commonly thought and these distinctions can make a difference in how a text-processing system might treat the word. For example, people can meet (verb) each other at the office, or compete against each other at a track and field meet (noun). Grammatical analysis determines whether the word "meet" is being used as a verb or a noun.

Autonomy calls grammatical analysis "tagging".

Guided Search

Guided search or dialog-based search is an interactive approach for refining a search by having the system ask questions that provide a way to narrow down on the search result. Guided search is very often built using information gain mechanisms and is also a feature of CBR systems. Also see [Parametric Search](#).

Inbound Link

When site A links to site B, site A has an outbound link and site B has an inbound link. Inbound links are counted to determine link popularity. Also see [Link Tracking](#).

Index File

A file created by a search indexer program, designed to store information in a format that makes fast retrieval possible. Index file can be as large as one fourth of the size of the document collection it indexes.

TQTR-Search02_color.doc	Date	4/10/2003	Page	10 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Information Gain

Information gain is a search refinement strategy that uses knowledge of how concepts occur in a search result set in order to guide the further narrowing down of the result set. For example, in some systems, parametric CBR, a knowledge model constructed over a set of attributes can allow the system to know that asking the user to specify a specific attribute can partition the space of results into the most optimum subsets for further refinement.

Information Visualization

Is about using visual, or graphical techniques to present information and support interactions. Visual techniques take advantage of the human brain's exceptional capacities for detecting subtle relationships and changes. Visualization is particularly useful for representing multidimensional data and large or complex data sets. By leveraging the strengths as well as respecting the constraints of our cognitive, perceptual, and motor systems, it is possible to support interaction with much larger amounts of information than possible with non-graphical interfaces.

Inverse Document Frequency (IDF)

A measure of how rare a term is in a collection of documents.

Inverted File

A file that represents a collection of documents or database. The inverted file lists all words that appear in all documents in the database, as well as a reference to the document where the word appears.

Keyword Search

Using this method, a user enters a keyword or term into a text field, for example "dog". The keyword search engine then searches through its index for documents that contain that word. In order to improve the search precision, keyword search engines often utilize lists of keyword associations or "topics" such as dog =hound =canine or dog is 90%canine, 10%furry. Typically, such lists require manual maintenance.

Keyword targeting

The practice of optimizing certain pages of a web site to rank well in a search for specific keywords. Keyword targeting is part of search optimization.

Knowledge Extraction

Is about extracting attributes or structure from information objects (documents or datasets) or collections (e.g. corpora or databases) of the kind that we naturally use in sense-making and in decisions about how we are going to use our time. This extracted information, often called metadata or even knowledge, is valuable in guiding access and use of the information and often for automating portions of the work.

TQTR-Search02_color.doc	Date	4/10/2003	Page	11 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Knowledge Model

A Knowledge Model is a model of the concepts, attributes and relationships within a domain of interest. At its least formal, it may simply be a list of keywords and synonyms. More formal are taxonomies and thesauruses. At its most formal, it is a full [ontology model](#).

Knowledge Representation Language

A formal language for expressing knowledge structures. Examples are CycL, LOOM and KIF. Currently there is much work happening in KRLs for the Semantic WEB, notably markup languages based on XML, such as [RDF](#) and [RDF Schema](#), and the [DAML+OIL](#) and W3C [OWL](#) initiatives.

Language Identification

Identifies the language of the document. This is an important capability for systems that will be used in global markets by multi-national companies and users of the Internet.

Lexical analysis or Tokenizing

Breaks a stream of characters and punctuation into discrete words, sentences and paragraphs. In English this would be apparently easy, but special characters and punctuation present in other languages can present a challenge. In Japanese, for example, there is no space between words.

During lexical analysis each word is associated with a word class or part of speech. When a word is of more than one class, for example is both a noun and a verb, it is associated with multiple classes. Lexical analysis is typically followed by syntactic analysis that identifies a unique word class for each word by taking in to account the context of the sentence.

Link Tracking

A type of indexing designed to track inbound links to a document. Many search engines offer ways to easily track inbound links. At Google, for example, simply type "link:www.your-domain-here.com" (without the quotation marks) for a list of sites linking to www.your-domain-here.com.

Log File Analysis

Referring to the analysis of records stored in the web site log file. In its raw format, the data in the log files can be hard to read and overwhelming. There are numerous log file analyzers that convert log file data into user-friendly charts and graphs. A good analyzer is generally considered an essential search optimization tool because it can show search engine statistics such as the number of visitors received from each search engine, the keywords each visitors used to find the site, visits by search engine spiders etc.

Metadata

Is data about data. Knowing about a thing can help a person decide whether it is worth spending time on it. Consider how you decide whether to go to a movie. Most likely you would want to know what the movie is about, who the actors and the director are, what kind of reviews it is getting. Similarly in dealing with the massive amounts and diverse

TQTR-Search02_color.doc	Date	4/10/2003	Page	12 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

sources of information available to a knowledge worker, knowing about enables more effectively selecting and filtering documents and sources. For example, users can search for only those documents created after January 1, 1999 by someone partially named "Clancy" to access the latest information written by that author. Often this kind of data is called meta-data or meta-information, which can be a catchall category covering many different types of features or descriptions of documents or collections of documents. Metadata can be readily present in the document, for example its date and its author or it can be derived, for example its language, genre and usage statistics.

Meta Search Engine

A type of search engine. Meta search engines usually do not maintain their own indexes. Instead, they query other search engines and return results from all of them - often with a mention of the search engine next to the each result. In combining the results meta search engines use their own algorithms to re-rank and re-order results that have already been ranked by the individual search engines.

Meta Tag

An HTML tag placed in the head section of a web page. The tag provides additional information that is not displayed on the page itself. The initial idea was that webmasters should use these tags to help search engines index the page correctly by providing an accurate description of the page content and a list of keywords associated with the page.

Natural Language Query

A way to enter a query to a search engine using conversational language structures, such as a sentence

Natural Language Processing

Natural Language Processing (NLP) uses the rules of native languages to examine the content and meaning of text. NLP commonly uses artificial intelligence and a trained rule base of meaning of words. This approach has been in existence for a long time, but has yet to prove its effectiveness as a search technology. There currently are efforts to incorporate this technology as an addition to other approaches such as neural network search engines to improve overall performance.

Navigational Search

This approach to search is based on navigating topical directories, or taxonomies, present on most Internet and Intranet portal. Well organized navigational structures help users narrow in on the general neighborhood of the information they seek.

Ontology

The study of the categories of things that exist or may exist in some domain. There is an emphasis upon "knowledge representations".

Ontology Model

An ontology model is a specification of the concepts, properties and their relationships within a domain of interest. The model provides a vocabulary and a shared meaning to the

TQTR-Search02_color.doc	Date	4/10/2003	Page	13 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

domain. Building on this shared understanding, an ontology is a framework that provides a common way to interpret and communicate what things mean with a domain of discourse.

Parametric Search

Parametric Search is a search using attributes that are defined over one or more knowledge sources. For structured knowledge, parametric search is relatively straight forward, Parametric Search is also possible with unstructured knowledge, where intelligent miners can discern the concepts represented with the knowledge artifacts.

Pattern Matching

Identifies the patterns that naturally occur in text, based on the usage and frequency of words, terms, or even letter patterns that correspond to specific ideas or concepts. Usually utilizes probabilistic algorithms such as [Bayesian Inference](#) or neural networks.

Phonetic Analysis

Phonetic representations of a word can be stored in collections. This allows users who don't know exact spellings to look for words that "sound like" another word (see [soundex search](#)).

Phrase Extraction

Identifies unique and important linguistic concepts that often appear as multi-word phrases in documents. These multiple word combinations often have a specific linguistic value within a particular document or an application that would be lost by treating the component words separately.

Precision

The number of correct documents retrieved as the result of a query or search divided by the number of documents retrieved (i.e., [$\#$ correct retrieved] / [$\#$ retrieved]). Search engines strive to have the right balance between precision and [Ranking](#)

Referring to the position of a web page or a document on the search results for a particular query. For example, a page that is listed third for the term "bubblegum" is said to have a ranking of 3 for that term.

Recall.

Pragmatic Analysis

Organizes semantic information in a way so it becomes useful for a particular purpose. For example, the extracted object, subject and action can be organized in a problem-solution format.

Proximity Search

A search method using the proximity of one term to another within a document

Query by Example

Query by example (QBE) uses a specific result from one search to find other results that are similar. The process of finding similar documents varies widely depending upon the kind of search engine being used.

TQTR-Search02_color.doc	Date	4/10/2003	Page	14 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Ranking

Referring to the position of a web page or a document on the search results for a particular query. For example, a page that is listed third for the term "bubblegum" is said to have a ranking of 3 for that term.

Recall

The number of documents retrieved containing correct content as the result of a query against a corpus divided by the possible number of documents in the corpus containing correct content (i.e., [# correct retrieved] / [# of possible correct]).

Relevance

A subjective measurement ranking which is supposed to represent the significance of information to an individual user

Relevance Modeling Technology

Identifies the degree of relevance, or ranking, of individual item in the search result set to the search query. Sophisticated relevance technologies take in to account the text of the document, its source, its associated links and other unique document characteristics when determining its relevance.

Results Management

How search results can be further processed. Includes saving common queries, refining queries, searching within the returned result set.

Semantic Analysis

Applies rules of grammar and lexicons to explicitly understand textual information, particularly in regard to the structure of the text and the relationship of words within a sentence or paragraph. This approach represents a combination of [grammatical](#), [lexical](#), [word proximity analysis](#) and [phrase extraction](#) techniques. The sentence is analyzed to identify its major structural elements – Object, Subject, and Action. Some search vendors call this process parsing.

Semantic Web

The next generation of the WEB with ontology-based markup languages allowing more precise search and inter-operability (URL: <http://www.SemanticWeb.org>).

Similarity Measures

Measures for determining how the target set of attribute values is close in meaning to the query. Similarity measures can be based on tables, functions, or taxonomical/enumeration/simple scalar value closeness.

Spiders or Crawlers

Crawling systems spider websites and file systems to build indexes and identify changed specific documents. Many feature ability to adjust revisit frequencies, link depth, and directory depth, and dynamically created links. Usually they are able to cross firewalls and index password-protected sites. Special connectors are needed to crawl certain database formats, such as Lotus Notes.

TQTR-Search02_color.doc	Date	4/10/2003	Page	15 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Stemming

Identifies the base form of each word in the source text. This converts a surface form of a word like survive, surviving, or survived into the underlying root, to survive.

Soundex Search

A search technique, which allows a user to find words in documents, which sound like, or are commonly pronounced like another word.

Summarization

Reduces document volume and complexity while still retaining the essential information of the document. Research has shown that summaries containing only 20% of the original content can be as informative as the full text. This research also indicates that even shorter summaries can retain the essence of the document meaning. Summarization allows users to determine whether a document is relevant before retrieving the entire file.

Syntactic Analysis

Completes the process of associating each word in the sentences with a single word class (part of speech) started during lexical analysis by taking in to account the context of the sentence. Accomplished by applying grammatical rules and statistical information to determine most likely sequence of word classes in the sentence.

Taxonomy

An original definition of a taxonomy is a hierarchical system of classification representing structural differences. A classic example of taxonomy is the "Taxonomy of Life" which separates mammals from birds and spiders from insects based on structural differences. With its use on the Web, the original meaning of 'Taxonomy' as a strict type hierarchy organized as a generalization/specialization relationship among concepts – a so-called 'is-a' hierarchy, has evolved to a more generic meaning of a scheme for categorization that facilitates browsing of a rich space of content. For example, on Yahoo (probably the better known example of a modern taxonomy) the 'Government' category includes 'National Songs and Symbols' subcategory which clearly doesn't have an 'is-a' relationship with its parent. Web taxonomies often contain cross-links and place a given object in more than one category. Web taxonomy is an ontology that does not explicitly define the nature of relationship between its concepts.

Term Frequency (TF)

A measure of how often a term is found in a collection of documents. TF is combined with [Inverse Document Frequency](#) (IDF) as a means of determining which documents are most relevant to a query. TF is sometimes also used to measure how often a word appears in a specific document.

Term Vectors

In the classic vector-space retrieval model, documents and queries are converted to term vectors to allow documents to be matched to queries and ranked based on the number of times the search terms occur in each document.

TQTR-Search02_color.doc	Date	4/10/2003	Page	16 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Thesaurus

A list of subject headings or descriptors, usually with a cross-reference system for use in the organization of a collection of documents for reference and retrieval. For each word in a thesaurus there is a list of broader, narrower and related terms.

Word Exclusion and Meaningless Terms

Sometimes it is useful to exclude certain words from document processing and indexing. For example, if all documents in a collection include the word "company," excluding it from indexes will improve search speed, but will not affect search precision. The list of such terms is usually called a "Stop word list".

Word Location

Storing the location of words within a document allows search terms to be highlighted when returned documents are viewed.

Word Proximity

Word proximity analysis determines how close searching terms are to each other.

TQTR-Search02_color.doc	Date	4/10/2003	Page	17 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

Emerging Standards

Knowledge Representation

DAML

The DARPA Agent Markup Language (DAML) Program, officially began in August 2000. The goal of the DAML effort is to develop a language and tools to facilitate the concept of the semantic web.

The World Wide Web (WWW) contains a large amount of information which is expanding at a rapid rate. Most of that information is currently being represented using the Hypertext Markup Language (HTML), which is designed to allow web developers to display information in a way that is accessible to humans for viewing via web browsers. While HTML allows us to visualize the information on the web, it doesn't provide much capability to describe the information in ways that facilitate the use of software programs to find or interpret it.

The World Wide Web Consortium (W3C) has developed the Extensible Markup Language (XML) which allows information to be more accurately described using tags. However, XML has a limited capability to describe the relationships (schemas or ontologies) with respect to objects. The use of ontologies provides a very powerful way to describe objects and their relationships to other objects. The DAML language is being developed as an extension to XML and the Resource Description Framework (RDF). The latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. (URL - <http://www.daml.org>).

OIL

Ontology Inference Layer, is a proposal for a web-based representation and inference layer for ontologies, which combines the widely used modelling primitives from frame-based languages with the formal semantics and reasoning services provided by description logics. It is compatible with RDF Schema (RDFS), and includes a precise semantics for describing term meanings (and thus also for describing implied information). OIL presents a layered approach to a standard ontology language. Each additional layer adds functionality and complexity to the previous layer. This is done such that agents (humans or machines) who can only process a lower layer can still partially understand ontologies that are expressed in any of the higher layers.

OWL

The W3C Web Ontology Working Group (WebOnt) is tasked with producing a web ontology language extending the reach of XML, RDF, and RDF Schema. This language, called OWL, is based on the DAML+OIL web ontology language. The only substantive changes from DAML+OIL are the removal of qualified number restrictions, the ability to directly state that properties can be symmetric; and the removal of some unusual DAML+OIL constructs, particularly restrictions with extra components. There are also a number of

TQTR-Search02_color.doc	Date	4/10/2003	Page	18 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

minor differences, including a number of changes to the names of the various constructs. These naming changes may indicate potential changes to the preferred names in the concrete syntax for OWL, but the intent of WebOnt is to maintain the DAML+OIL names to the maximum extent reasonable. Current proposal includes 3 subsets of OWL language: OWL light, OWL full and OWL DL (Description Logic). Draft has been available since August 2002. First release is expected in June 2003.

RDF

Resource Description Framework, is a foundation for processing metadata. This W3C standard provides interoperability between applications that exchange machine-understandable information on the Web. RDF uses XML to exchange descriptions of Web resources but the resources being described can be of any type, including XML and non-XML resources. RDF emphasizes facilities to enable automated processing of Web resources.

RDF can be used in a variety of application areas, for example: in resource discovery to provide better search engine capabilities, in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical "document", for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications. Descriptions used by these applications can be modeled as relationships among Web resources.

The RDF data model defines properties and values. RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources. As such, the RDF data model can therefore resemble an entity-relationship diagram. The RDF data model, however, provides no mechanisms for declaring these properties, nor does it provide any mechanisms for defining the relationships between these properties and other resources. That is the role of RDF Schema.

RDF Schema

A specification language that is less expressive, but much simpler to implement, than full predicate calculus languages such as CycL and KIF. Like RDF, the RDF Schema language is also based on metadata research in the Digital Library community. In particular, RDF adopts a modular approach to metadata that can be considered an implementation of the Warwick Framework.

RDF represents an evolution of the Warwick Framework model in that the Warwick Framework allowed each metadata vocabulary to be represented in a different syntax. In RDF, all vocabularies are expressed within a single well defined model. This allows for a finer grained mixing of machine-processable vocabularies, and addresses the need to create metadata in which statements can draw upon multiple vocabularies that are managed in a decentralized fashion by independent communities of expertise.

RDF Schemas might be contrasted with XML Document Type Definitions (DTDs) and XML Schemas. Unlike an XML DTD or Schema, which gives specific constraints on the structure of an XML document, an RDF Schema provides information about the interpretation of the statements given in an RDF data model. While an XML Schema can be used to validate the

TQTR-Search02_color.doc	Date	4/10/2003	Page	19 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

syntax of an RDF/XML expression, a syntactic schema alone is not sufficient for RDF purposes. RDF Schemas may also specify constraints that should be followed by these data models.

TopicMaps







An ISO/IEC standards effort, 13250, entitled, "Topic Maps: Information Technology -- Document Description and Markup Languages [ISO13250]". This International Standard provides a standardized notation for interchangeably representing information about the structure of information resources used to define topics, and the relationships between topics.

A set of one or more interrelated documents that employs the notation defined by this International Standard is called a 'topic map'. In general, the structural information conveyed by topic maps includes: (1) groupings of addressable information objects around topics (occurrences), and (2) relationships between topics (associations). A topic map defines a multidimensional topic space -- a space in which the locations are topics, and in which the distances between topics are measurable in terms of the number of intervening topics which must be visited in order to get from one topic to another, and the kinds of relationships that define the path from one topic to another, if any, through the intervening topics, if any.

Topicmaps.Org is an independent consortium of parties interested in developing the applicability of the Topic Maps Paradigm (first fully described in the ISO/IEC 13250:2000 "Topic Maps" standard [ISO13250], which is SGML and HyTime-based) to the World Wide Web, which is expected to become increasingly XML-based. This work includes the development of an XML grammar for interchanging Web-based Topic Maps, called the **XTM** Specification, which has been recently published by the Topicmaps.Org Authoring Group (the "AG").

The table below shows level of industry adoption for each of the knowledge representation languages:

TQTR-Search02_color.doc	Date	4/10/2003	Page	20 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				

	KIF/OKBC CG/CycL	XML	Topic Maps/XTM	RDF(S)	DAML+ OIL	OWL
Description	<i>Legacy KR Languages</i>	<i>eXtensible Markup Language</i>	<i>Topic Maps/XML Topic Maps</i>	<i>Resource Description Framework</i>	<i>DARPA ML + Ontology Inference</i>	<i>Web Ontology Language</i>
Governance	 and others					
Years since proposed	>10	>5	>5	>3	>2	>1
Commercial Support (KL*)	■	■■■■	■■	■■■■	■■■	■
Open Source Support	Yes	Yes	Yes	Yes	Yes	Coming

2 or less vendors

 10 or less vendors

 5 or less vendors

 > 10 vendors

Table 1: Adoption Level of Knowledge Representation Languages

Metadata

Dublin Core

The [Dublin Core Metadata Initiative](#) (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems. The Dublin Core metadata standard is an element set for describing a wide range of networked resources. The Dublin Core standard comprises fifteen elements, the semantics of which have been established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship.

HTML provides an easily understood format for demonstrating Dublin Core's underlying concepts, but more complex applications using qualification may find that using RDF/XML makes more sense. When considering an appropriate syntax, it is important to note that Dublin Core concepts are equally applicable to virtually any file format, as long as the metadata is in a form suitable for interpretation both by search engines and by human beings.

	Dublin Core Elements
Asset metadata—	Title, Creator, Publisher, Contributor, Date, Type,

The Who, Where and When	Format, Identifier, Source, Language
Subject metadata— The What and Why	Subject, Description, Coverage
Relational metadata— Links between Assets	Relation
Use metadata— <i>How to Monetize Assets</i>	Use

Table 2: Dublin Core metadata example

ISO/IEC 11179

This is a family of ISO standards for informational and organizational structure of metadata registries. ISO 11179 is an international standard for formally expressing semantics of data in a consistent manner. Part 5 of ISO 11179 covers standards for naming conventions.

About TopQuadrant

TopQuadrant is a trusted intermediary for the intelligent application of knowledge technologies. As knowledge system architects, we are assisting leading enterprises to envision, architect, plan and realize knowledge-based solutions. Our consultants have many years of experience in large consulting organizations, for example IBM Global Services, and have a background in AI, Object Technology, Knowledge Management and Methodologies for Knowledge, Software and Systems Engineering.

Using the following unique tools, we address major obstacles to success in building knowledge solutions:

- ❖ **Solution Envisioning**, a scenario-driven approach to experiencing a future system through analogies and examples using a Database of Capability Cases.
- ❖ **Capability Cases**, application solution patterns (e.g., for ontology-based knowledge applications) expressed in a business context with examples of known uses, applicable technologies and leading practices.
- ❖ **TopDrawer™**, a comprehensive knowledge base for storing and dynamically working with Capability Cases.

With a proven track record in the practical application of knowledge technologies, **TopQuadrant** helps clients transition to next generation, semantically integrated systems, while sustaining and optimizing their investments in current systems.

Additional TopQuadrant Technology Briefings are Available

Current:

- Ontology Development Lifecycle and Tool Survey
- Semantic Integration Platforms

Planned:

- Modeling Techniques
- Semantic Solutions for Search and Self Service

To access these papers, please visit our web site at www.topquadrant.com

TQTR-Search02_color.doc	Date	4/10/2003	Page	23 of 23
Copyright © 2002 - 2003 TopQuadrant, Inc. All Rights Reserved. Printed in U.S.A. Confidential, Unpublished Property of TopQuadrant				