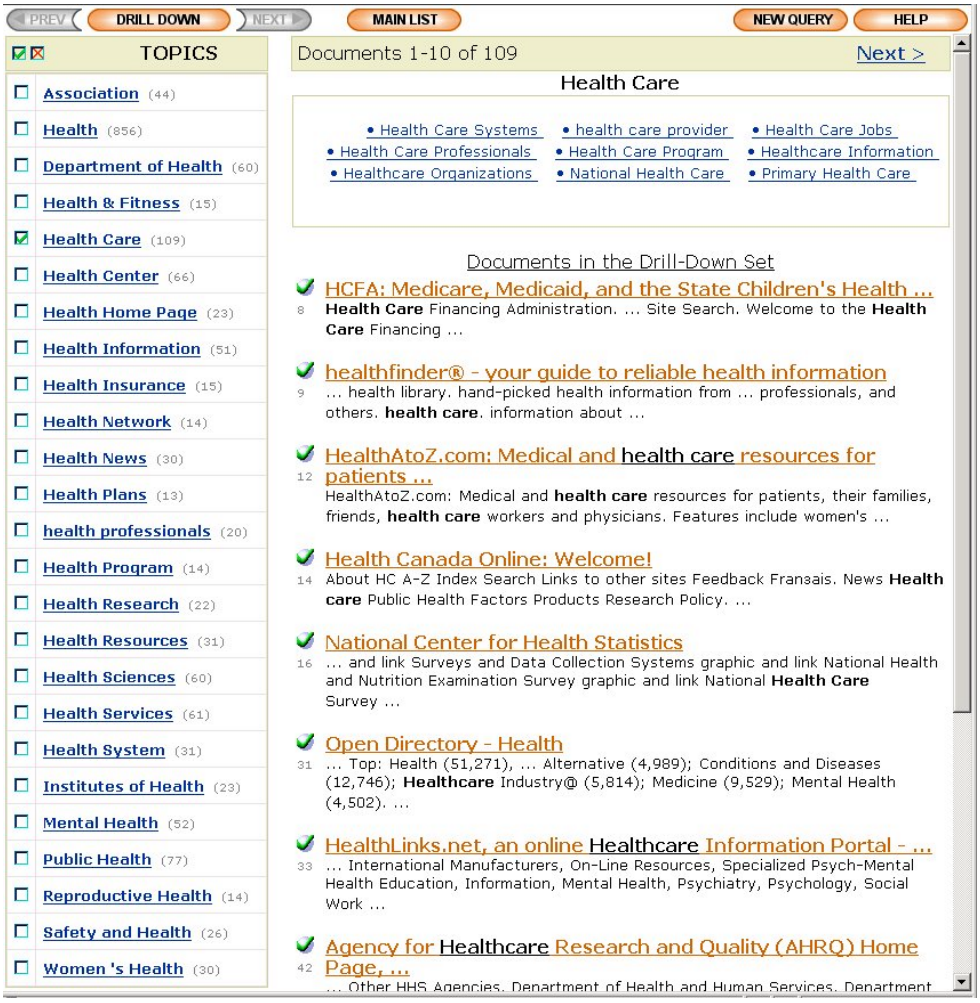
RealTerm
Dynamic Content Management Solution
(White Paper)

Use RealTerm at http://www.infonetware.com



infogistics

power through knowledge

40 Maritime Street
Edinburgh EH6 6SA

Tel: +44 (0)131 554 6461

Email info@infogistics.com Web: http://www.infogistics.com/

Table of Contents

The world is drowning in information

Ever-expanding data in corporate Intranets, World Wide Web, local networks and even on personal computers (all those emails, word documents, spreadsheets...) means that information on everything you can imagine seems to be available at your fingertips. Searching is as easy as pressing a key. Or is it?

Searching is easy...

State-of-the-art search technology such as corporate and Internet search engines let you search for what you want. You type in what you are looking for, press the key and bingo! there it is. Searching is a solved problem.

...finding is not

But what about finding what you are looking for? Search engines are easy to use, but how about search results? Usually if you don't see exactly what you are looking for in the first 5-10 listed search results, you re-formulate your query and hope for better luck next time. After a few failed attempts you make do with what you have got, or you probably just give up.

We want content structured…

No matter whether you are dealing with your personal emails or with documents in your corporate Intranet you want similar documents to be grouped together so you can easily access them when needed. Information directories and portals such as Yahoo offer electronic content classified into areas or topics such as "News & Media", "Recreation & Sports", "Health", etc.

*… and we want **our** content …*

But they classify their content not yours. Yahoo employs multiple hundreds of human analysts to classify documents using *their* database into *their* hierarchy. This is not an easy task. Try to manually classify your emails! What about your corporate Intranet with hundreds of thousands of documents that is growing and changing all the time?

… and structured flexibly

Also the way they classify documents not always corresponds to how you would do it. How many times you were lost in the wrong branches of their hierarchy. Moreover, you know that the same document can belong to several different topics, and topics themselves cannot be fixed. Every new task can create a new view to your content and therefore produce a new range of topics, which requires reclassifying your documents accordingly.

We want personalized and dynamically structured content!!!

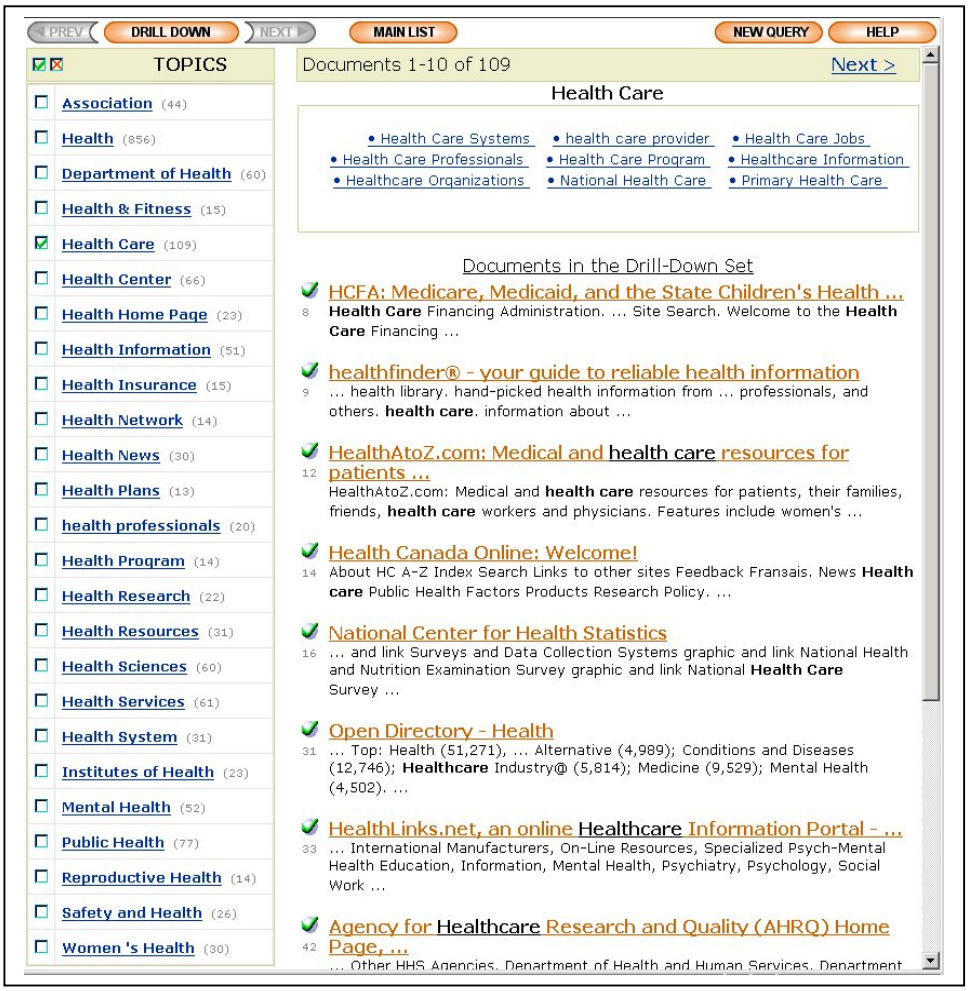
So what we really want is a way to structure our content not according to some static rules predefined by some analysts hierarchy, but dynamically, according to our intuition and needs. At the same time we cannot afford to spend multiple hours in building our personal information space.

So is there an answer to this? Yes, recent advances in Artificial Intelligence and Natural Language Processing allowed Infogistics to develop the first personalized and dynamic content management solution – RealTerm.

RealTerm dynamic content structuring

Infogistics' RealTerm is a topic identification and clustering technology. It analyses results returned by a search, organizes these results into a hierarchy of topics and presents them to you for browsing and exploring. For example, if you search for "*chocolate cake recipe*" on the Internet, RealTerm analyses abstracts of the retrieved documents and creates topics such as "*chocolate slab cake recipe*", "*chocolate pudding cake recipe*", "*chocolate Zucchini cake recipe*", "*chocolate cake recipe from Cuisine Magazine*", etc. Clicking on a topic retrieves all and only relevant documents no matter whether they were listed in the first ten or the last hundred of the original search.

The following is an example of searching the Internet for the word "health" and drilling down on "*Health Care*" topic identified automatically by RealTerm.



The original search returned over a thousand documents containing the word "health".

RealTerm analyzed the document titles and abstracts and identified a number of major topics: "*Health & Fitness*", "*Health Care*", "*Health Center*", "*Health Insurance*", "*Women's Health*", etc. shown in the "TOPICS" box

In the document box we see the results of clicking on the "*Health Care*" topic.

RealTerm aggregated all the documents (109 out of 1037) belonging to the topic *"Health Care"* and then listed subtopics ("*Health Care Systems*", "*Health Care Jobs*", "*Healthcare Information*", etc.). Drilling down further on the "*Health Care Jobs*" returns only three documents. In our example these were documents numbered 611, 614 and 789 in the original return set of 1037 documents. Documents you would never have been able to scroll down to.

There are two main features that make RealTerm content delivery unique:

- RealTerm **automatically** creates topics and subtopics; they do not come from an existing static hierarchy like Yahoo or Yellow Pages but are **dynamically constructed** to closely model the content of every specific document collection and even every specific search. This means that information structuring is very highly personalized to reflect each individual search.
- Topics and subtopics are identified and structured into hierarchical relations **in real-time**. A matter of seconds after the search engine has returned its results RealTerm explores the results and identifies the key pieces of information to build the topic list. This means that RealTerm can be installed on top of any existing search engine, database, corporate Information Management System, legacy systems etc. Therefore any existing corporate information structure can be turned into an advanced knowledge discovery system in a matter of seconds.

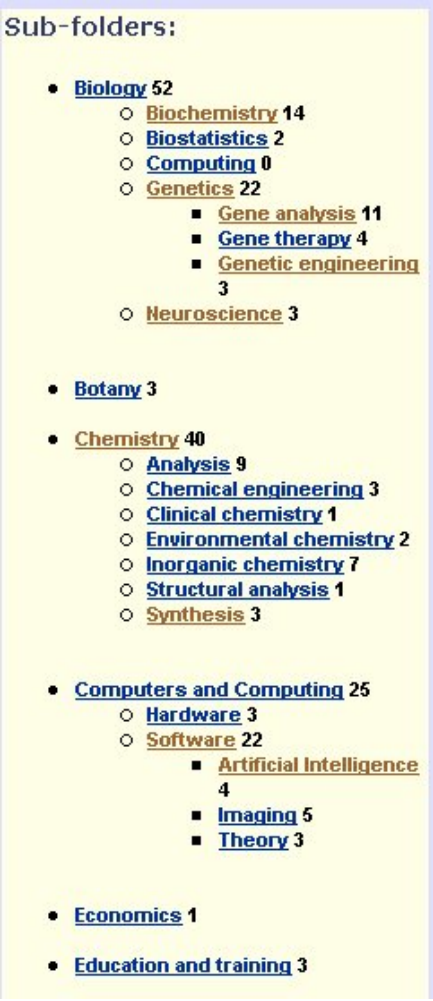
RealTerm not only allows you to find and aggregate all relevant content from a variety of data sources, it also allows you to explore new information to see what is there. Rather than require the user to guess exactly the right query, RealTerm relies on our ability to "know it when we see it". Therefore a broad and general request such as "search for all documents which mention health" will identify specific areas of interest and suggest areas for expansion of the search. RealTerm constantly monitors the user's actions to continually tailor the topics, accurately reflecting the user's interest, helping to find that needle in the haystack.

In short, RealTerm solves a long-standing requirement of the information overflow economy: i.e. the highly personalized delivery of information structured in a sifted hierarchy of topics. Since RealTerm can be easily integrated with various information sources in real-time it can aggregate information coming from these diverse sources in a consistent way. So for instance all documents on a specific subject scattered across multiple departments of an organization can be aggregated and made available.

RealTerm also comes with its own search engine, which can index and make searchable all electronic content in your organization ranging from emails and MS-Word documents to Lotus Notes and Oracle based information. This is only needed if a client does not already have an installed Information Management structure to which RealTerm can be dynamically integrated.

RealTerm Content Management

RealTerm content delivery functionality is complemented by its content management capabilities. RealTerm allows you to build your own content structures similar to Yahoo or Yellow Pages but for your own data and with no additional effort. You can create folders and subfolders within RealTerm to store the documents you aggregate through RealTerm topics browsing.



For example, you can issue a query on "microelectronics" and drill down to the "*Semiconductors Industry*" topic identified by RealTerm and save all these documents in an appropriate folder. You can then drill down further to "*Digital CMOS*", a subtopic of "*Semiconductors Industry*", and create a new subfolder for these documents.

On the left is an example of working folders in the application of a RealTerm solution to Intellectual Property Information from a consortium of Scottish Universities. These folders have been created and partially populated by a user of the system during various searches through the underlying documents. Highlighted folders have been populated during the last search and browsing session.

Unlike currently adopted approaches, RealTerm does not force you to define your information hierarchy upfront and then spend hours of classifying your documents. On the contrary, using RealTerm you build your information hierarchy along as you use the system and in line with your current working patters. RealTerm will identify topics and you can decide to store or not to store documents belonging to one or more topics in a particular folder. RealTerm allows the user to create, populate and manage only the folders they want at no extra hassle.

One document can belong to several different topics and several different folders, reflecting multiple ways of looking at the same information, based either in the document's content or different projects. Since RealTerm stores not the documents themselves but rather pointers this does not waste resources or take addition space. It further means that changes made to a document will be propagated throughout the entire hierarchy.

After a folder is populated with documents RealTerm can be told to automatically create an agent that will monitor new documents entered into the system and notify you when a document similar to ones you store in a specific folder appears. You can even ask this agent to go and spider other information sources, including the Internet and accessible by your databases, for required information.

Benefits of RealTerm

- **Hierarchical**

RealTerm is able to determine and identify hierarchies of information within large collections of documents. Collecting topics into groups and sub-groups enables topic hierarchies to be created. This makes information more easily searchable and therefore makes information easier to find.

- **It finds related topics**

By exploring the information contained within the documents returned from a search, RealTerm is able to identify additional topics and terms. By offering additional and related terms users are presented with a far more complete set of returned documents as well as being given the opportunity to expand and extend their search.

- **It sorts out spelling errors**

RealTerm recognises typographical errors, spelling mistakes and ambiguous terms that exist to frustrate search most engines. By not excluding these terms and the documents that contain them the user can be infinitely more confident that the document collection will contain the documents and information that they require and that this is not lost simply because of a slip of a finger.

- **It makes suggestions of things to look at**

The additional terms and topics that RealTerm identifies and lists will provide the user with options for additional search terms that they may not have considered. This is particularly useful when the search subject or criteria is not immediately obvious.

- **It normalises information**

The algorithms used within RealTerm aggregate terms and topics identifying this even where these terms may not explicitly occur, but are referred to by understanding. ("Aborigines are the indigenous occupants of Australia – *they* were *there* many years before the Europeans." RealTerm knows that '*they*' and '*there*' mean '*Aborigine'* and '*Australia*'.)

- **It uses existing information sources**

By using existing databases, dictionaries, thesauri and lists RealTerm is able to use sophisticated lookup references to find related terms and topics. This means that in fields where specialised terms may be used (science, engineering and technology), the information is not open to just the few specialists that work and understand the language that is spoken.

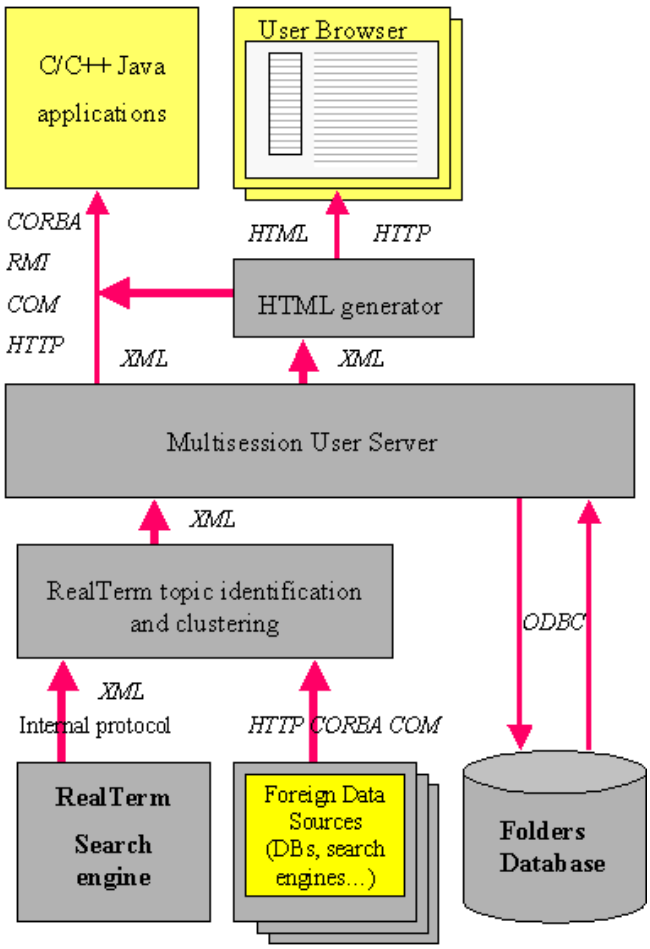
- **It shows where the words are in the document**

Traditional search engines will present the first couple of lines of the document in a list of search returns. Where the search term or phrase appears in the document and the context in which it is used is often an indicator of the relevance of the document in the research being undertaken. RealTerm will present the returned documents and will show all the instances of the search term in the context in which it is used allowing the user to instantly judge its relevance.

RealTerm Functionality at a Glance

- real-time automatic topic detection and active browsing through information;
- persistent storage of relevant documents in user-defined folders: dynamic creation of hierarchical information directories (portals);
- automatic content monitoring and spidering; automatic information harvesting;
- full text retrieval and Natural Language Processing;
- ability to be connected to a wide variety of existing databases and search engines;
- support of domain-specific knowledge bases, thesauri and word-lists;
- support of meta information stored in the databases;
- single point of access;
- full Internet enablement;
- compatibility with currently used browsers;
- highly intuitive, user friendly and customizable interface;
- automatic spidering for new information from predefined sources;
- automatic user profiling;
- automatic information routing;
- virtually no administrative overheads on information input and update;
- easy integration with C/C++ and Java applications on all major platforms;

RealTerm Architecture



RealTerm is implemented as a server application or through an ASP, which connects to thin clients (web browsers) or other applications through HTTP or other network protocols passing XML or HTML. The *Multi-User Server* monitors connections with multiple users coordinating requests and data traffic. It generates XML which then is transformed into HTML (if necessary) using XSLT.

RealTerm can query its own *Search Engine* as well as many external search engines, databases, and other information sources. Retrieved information is then processed by the *Topic Identification and Clustering Engine* that creates a semantic network of topics linked to documents an each other.

Persistent information classification is supported through a database that keeps documents associated with user-defined folders.

RealTerm Technical Specification

RealTerm Distribution

- RealTerm multi-user server
- RealTerm real-time topic extraction and clustering engine;
- RealTerm advanced search engine;
- Bridge to connect to ODBC database to store user folders;
- Bridges to other search engines, databases and legacy systems;
- Libraries for integration to C/C++ and Java applications;
- User Documentation;

Topic Extraction and Clustering Engine

The RealTerm Topic Extraction and Clustering Engine intuitively blends the flexibility of a query-based search with guided browsing capabilities of Information directories. It achieves this by using advanced statistical, linguistic and conceptual analysis to automatically identify major topics and their sub-topics that are contained in a document collection. This allows the user within a few mouse clicks to identify and aggregate all and only relevant documents regardless of whether they have been listed in the first ten or the last hundred in the return set.

Currently RealTerm technology is available for

- English (all variations)
- German
- Spanish
- French
- Italian
- Portuguese

It will shortly be available for a number of Eastern European and Scandinavian languages.

The RealTerm Topic and Clustering Engine is capable of processing 15,000 words a second on Pentium 1GH computer. This speed allows it to generate the topic index for a typical return set of 1,000 documents in less than 3 seconds.

RealTerm Search Engine

RealTerm comes with its own advanced search engine that supports

- **KeyWord in Context document summaries.** RealTerm not only finds relevant documents and ranks them according to the user's query, it also extract specific passages from these documents that contain required information. This functionality, also call Key-Word In Context (KWIC), enables the user to validate the importance of the document at a glance without opening and scanning through the document itself.
- **Concept based search.** RealTerm not only searches for keywords specified in the user query but it also dynamically detects synonyms and can extend its search to them after the user validates which synonyms to use. For instance, in a search for "heart disease"

RealTerm detects that "cardiac arrest" or "cardiology" are often mentioned and will suggest these terms to extend the search.

- **Keyword and Natural Language queries.** Coming from a team of developers with a collective experience of over 30 years in Natural Language Processing and Information Retrieval in both academia and industry, naturally RealTerm can handle both keyword queries and queries expressed in Natural Language.
- **Multiple file format support.** RealTerm indexing engine can index over 30 file formats including Microsoft Word, PDF, Postscript, HTML, Lotus, etc.

RealTerm external compatibility

RealTerm can be integrated with any mainstream search engine or database and will add advanced knowledge management capabilities to traditional Information Management software without extensively re-engineering the existing deployment platform. Here is a partial list of most popular external systems accessible from RealTerm:

- Internet Search Engines - Yahoo, AltaVista, FAST, Google...
- Oracle Context and Intermedia
- Inktomi
- Verity
- SQL database search through ODBC
- legacy indexing systems
- ... and many others including Lotus Notes and Domino

Hardware/Software Platforms

RealTerm is designed to operate as a server that communicates with thin clients (usually Internet browsers) through HTTP protocol. Therefore there is no limitation on the client configurations, provided that they are capable of running Internet Explorer 4 and above or Netscape Navigator 4 and above.

RealTerm server runs on all major platforms including Microsoft NT/2000, Linux and Solaris.

Integration

RealTerm server can be directly accessed from other applications through special developers libraries. Currently we support integration with C/C++ applications on various platforms, COM integration, Java integration and EJB (stateful session) integration.

User Interface

The RealTerm user interface is easily customizable in terms of corporate look-and-feel and information layout. RealTerm server communicates with the interface generation module by outputting XML structures with required information. The interface-building module applies special templates and XSLT scripts to generate HTML, which then can be displayed by the browsers. Both templates and scripts can be easily modified to reflect specific user and application requirements.

RealTerm: How it Works

RealTerm applies advanced statistical, linguistic and conceptual analysis to automatically identify major topics and their sub-topics in electronic texts. Extracted terms are arranged into a semantic network which links them to documents and other terms. This network then supports RealTerm topic browsing functionality which allows the user within a few mouse clicks to identify and aggregate all and only relevant documents regardless whether they have been listed in the first ten or the last hundred in the return set.

Term Extraction

First, RealTerm scans document summaries returned by a search engine and identifies the most important **words and phrases (terms)** which characterize these documents. Since documents might come from a variety of diverse sources, RealTerm applies different kinds of spelling correction and phrase unification algorithms which allow it to unify differently (mis)spelled variants of the same phrase e.g. "Mono Lisa" and "Mona Lisa", "Gregory" and "Grigory", etc.

It also applies morphological and syntactic transformations to unify phrases according to a Language grammar. For example, "linguistic and statistical method" and "statistical and linguistic methods", "information retrieval" and "retrieval of information" can be unified according to syntactic rules of English.

Document Clustering

At the next stage documents, together with the identified terms, are **clustered** into groups of related topics according to the content of their summaries. This is done by applying statistical algorithms that evaluate the strength of co-occurrence between terms and documents. Therefore groups of documents form topic clusters which can be described by the terms identified in these documents. RealTerm looks at the documents beyond individual words appearing there and groups them into topics on the basis of the information contained in them.

Term Relations Identification

Finally, terms are arranged into **hierarchical relationships** of more general with more specific topics. This is done in two ways. Lexical analysis of term structure can indicate that one term is more general than another, as for example, "Manchester United Football Club" is a more specialised term than "Football Club". The other way of uncovering term relations is distributional statistical analysis which can reveal that, for example, "myocardial infarction" often co-occurs with "heart disease" and therefore can be treated as its specialization.

Knowledge-Base Support

When applied to a specific domain RealTerm can be directed to make use of existing knowledge bases, thesauri and word-lists. For example, in the medical domain RealTerm can be used in conjunction with the MEDLINE Meta-Thesaurus where relations between many medical terms are already established.

Frequently Asked Questions

Is RealTerm a search engine?
No. RealTerm is a system that serves as a mediator between a search engine and the user. It analyses results returned by a search, organizes these results into a hierarchy of topics and presents them for browsing. It comes with its own search engine but it can work with other search engines and databases.

Is RealTerm a meta-search engine?
RealTerm can act as a meta-search engine because it can combine output from multiple information sources such as search engines and databases. However, this is just one of the features of RealTerm while its main functionality is in the analysis of the search results and structuring retrieved documents into topics.

Are the topics assigned by human annotators?
No! This is the main point! Topics are identified automatically using statistical, linguistic and conceptual analysis.

Do topics come from a pre-existing list?
No. There is no static list of topics. RealTerm creates topics dynamically to reflect information content of each individual document collection or search and is refreshed for each search.

Is every document assigned to a single topic?
No. Topics are not constraints, but an aid to finding the information you want. Documents can belong to multiple topics reflecting different view to the same information.

Does RealTerm analyze the entire text of the documents?
No. It would take a long time to download all the documents. RealTerm analyses titles and document abstracts returned by a search engine.

How fast is RealTerm topic structuring?
The engine is capable of processing 15,000 words a second on Pentium 1GH computer. This allows it to generate the topic index for a return set of 1,000 documents in 2-3 seconds.

How many topics RealTerm identifies?
This depends on a particular document collection but on average there are 3-4 topics per document. So for a set of 1000 documents RealTerm will create about 4000 topics.

How can I digest all these topics?
You don't need to. Since these topics are organized into a hierarchy the user always sees only a fraction of them. Initially only the most general topics are presented. RealTerm monitors the user's actions to continually tailor the topics to reflect the user's interest and suggests only the relevant topics.

Where RealTerm has already been applied?
RealTerm is being extensively used to browse through job seekers resumes in online recruitment agencies. It has been applied to browsing through a large database of patent and IP documentation in a context of a consortium of five major Scottish Universities. It has been applied to browsing through products and offerings in a context of an online music shop. It is available for general use as a Web search tool at www.infonetware.com. Try it!

What does it take to add RealTerm functionality to my corporate intranet?
Probably 15 minutes if your intranet already has a search engine you are happy with. Alternatively, we will index all documents in your Intranet with RealTerm internal search engine, which is guaranteed to give excellent performance. This might take a couple of hours.

How do you license your technology?
We are very flexible. We license it to the end users on a server bases, information-load basis or using ASP model. We license it to technology providers using OEM agreements.

About Infogistics

Infogistics Limited is an Edinburgh-based company founded by internationally recognized experts in the fields of text-mining and document retrieval. Infogistics software products are based on proprietary technology that originated in Edinburgh University, where the founders of Infogistics' were based for a period in the 1990s.

Infogistics is a leading provider of text-analysis and content delivery solutions across multiple markets including Human Resource, Law Enforcement, Knowledge Management and CRM. Combining award-winning software applications with patent-pending technology, Infogistics helps organizations locate and link key pieces of information within their vast databases.

Recently Infogistics have won a prestigious SMART award run by Scottish Enterprise on behalf of the UK Department of Trade and Industry (DTI). The award marks the recognition of the groundbreaking work carried out by Infogistics in developing technologies that allow the extraction and search of electronic documents using Natural Language Processing

40 Maritime Street
Edinburgh EH6 6SA

Tel: +44 (131) 554 6461
Email info@infogistics.com Web: http://www.infogistics.com/