Semantic search

with the Infolution product and solution suite

An evaluation of the functional and technical features

R. W. van der Pol, © Thoughtwell 2007

Contents

0. Introduction

Ever since the introduction of the desktop computer in the early 1980s and the advent of the World Wide Web (WWW) in the mid-1990s, the amount of information available has increased progressively.

As a result, a new challenge arose: the ability to retrieve (finding and collecting) relevant information
efficiently and effectively, eliminating all non-relevant information. These demands are not met today, not even close. Relevant information remains unfound, and irrelevant information is found. Search is time-consuming and mentally tiresome and costly.

For instance, a person searching the WWW usually finds hundreds of thousands of documents (when using a simple query of a single term in a search engine such as Alta Vista or Google). Some of these documents are relevant, but not necessarily at top of the result list and will thus remain undiscovered. Moreover, even though the result list is long, it may not include all relevant documents. The obvious way forward, by formulating a more specific query, proofs too difficult a task for most people. Even a skilled person runs the risk of excluding relevant documents from the result list, leaving valuable information undiscovered.

Since approximately 2000, several search engines have entered the marketplace, varying in technical sophistication and user-friendliness. One of these tools, introduced recently, is the **Infolution** product and solution suite.

In this report, we compare this particular suite to the most relevant competing products. For example we aim to answer the question: "What distinguishes the **Infolution** suite positively from its competitors?" This report aims to be of use to potential buyers/users and investors.

In this study we compared information on various manufacturer websites, added our knowledge on search technology, and compared some initial experiences with the Infolution suite and some of the other products. A detailed empirical comparison lies outside of our scope since this is not a straightforward task because the most important features are not only search performance in terms of relevant documents, but also mental load and user comfort or user friendliness. All of these are difficult to measure.

The information available on the WWW lists a number of features in most of the investigated search tools. Therefore, it is not a question of available program features, but more one of how these features are implemented and what precisely they entail. However, this detailed view was not always apparent. We discuss these details for the most important features of the Infolution suite. The most exciting feature is by far the semantic technology (which not only means using syntax information but also extracting the actual 'meaning' of the words), hence the report title.

The report is structured as follows. In Chapter 1, the terminology of search processes is explained. Chapter 2 provides the architecture of the suite in functional terms and for some details of technical terms. Chapter 3 compares the suite to some competing products. Finally, Chapter 4 discusses the comparison and draws conclusions.

1. Introduction to search processes

The existing search methods in unstructured or semi-structured documents can always be defined either as searching with a query, or explorative search, or a combination thereof.

Query-based document-search processes are according to the schema of Figure 1. Here, the following functions are recognized:

- indexing of the information resources, i.e. documents on the desktop, intranet or WWW,
- formulating a query, that describes what the information (document) must be about,
- matching, i.e. calculating the similarity of query to indexes,
- presenting the results list,
- viewing/inspecting some of the results - revising the query, matching and re-viewing until satisfied or out of time/energy (this is called 'relevance feedback', which may be partly automated by adding terms from the top ten relevant documents).

Thus, the main components are:

1. indexer
2. matcher
3. user interface

The latter serves as an interaction between the search and retrieval/access program and the user. Its functions include: displaying the query screen, the result list (or: result screen) and/or the metadata as suggested by the query terms.

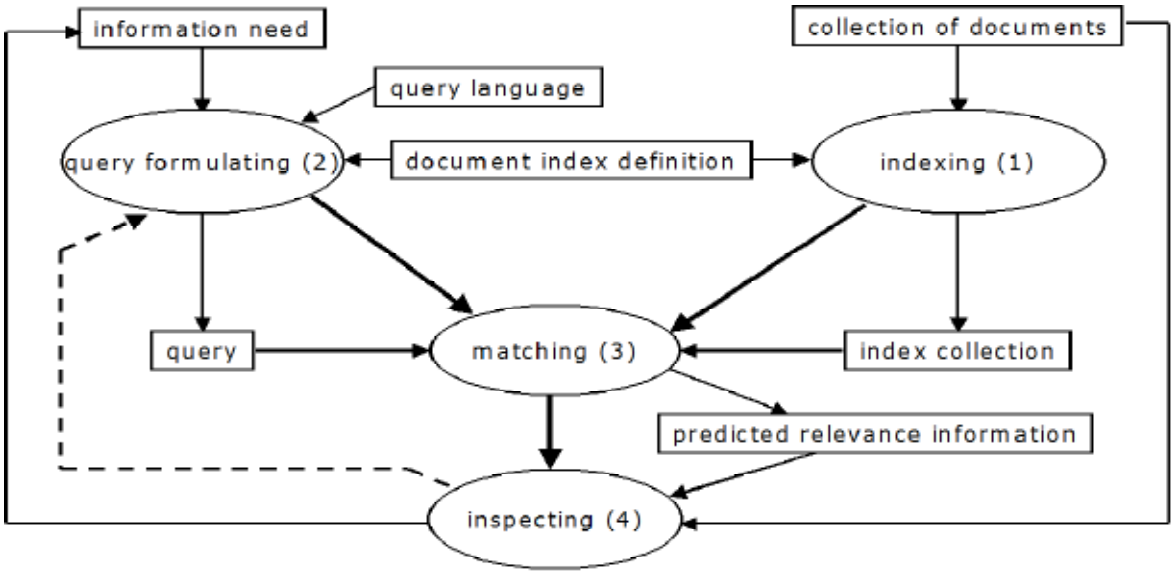

Figure 1: Schematic overview of query-based document-search processes.

The simplest indexes are full-text indexes – using the words within a document. Such indexes are still most commonly used. The queries are often "Boolean", meaning that the user types in words and connects them by using Boolean operators (such as AND, OR and NOT, sometimes also wildcards and others). This process is simple to implement. The result list yields a value for each document: either 'relevant' or 'irrelevant'. Thereafter, a relevance ranking of the result list is possible (yielding a value on a continuous scale of 0% to 100%). Extended Boolean models are available – these are more advanced in the sense that they allow for weighting the terms in a query and providing output listed by relevance; yet a fast implementation remains difficult.

Another advanced approach is the vector model; this is a model in which a query only consists of words and no operators. Matching is done by a multiplication of a query vector and an index vector. The result list is displayed on a continuous scale ranging from 0% to 100% and is thus ordered.
Advantages: easy query formulation, most relevant results on top, and fast matching performance.

We also mention the existence of probabilistic, or Bayesian, models. These allow for ranked output and are well suited for integration with hyper documents. However, it is not easy to obtain a fast performing implementation. Among these various matching models, there is no clear winner today. Each model has its own advantages.

Sometimes a dictionary is used, or even a thesaurus (including terms and relations such as 'broader than', narrower than' and 'synonymous to') can be used for suggesting query terms and for filtering/cleaning-up the indexes (removing typos and stop words in the articles). Sometimes, queries are (automatically) expanded using a thesaurus, adding for example synonyms and spelling variants. Adding terms, that have other relationships to the query term, is likely to improve the recall (number of relevant documents found), but will decrease precision (i.e., more irrelevant documents will be included) unless expansion is done in a smart way.

The dictionary may be taken from the document collection, or may have been predefined. In libraries, they are usually predefined (and structured as a thesaurus), in desktop, intranet and web search they are usually automatically extracted from the document collection.

A further extension, beyond the thesaurus, is a semantic network. This is a 'network' of binary relations between (pairs of) concepts. Some of these networks just connect the concepts, weighting them and thereby indicating the 'strength' of the connection. In addition other networks also contain relation type names. For example, usually, the relation type 'is a' (equivalent to 'is a subtype of') is present, which allows for representing taxonomies, i.e., type hierarchies (such as 'Trojan horse as a horse' and 'Trojan horse as a computer virus').

Semantic networks may be used for associating search terms, in order to find closely related – and therefore also potentially interesting – terms in queries or in indexes. Additionally, semantic networks may be used for clustering the search results into groups of related documents. For instance, clusters may be made by author or by the meanings of an ambiguous/homonymous search term (such as 'Jaguar' – meaning a car brand, an animal species and/or a computer game). Semantic networks are thus helpful in searching, but they slow down the speed of matching.

The user interface usually shows a query box and a result list. The latter showing titles of documents and 3-4 lines of text thereof, usually those containing the query terms. In some cases a list of alternative terms is available which the user may add to the query. These terms are taken from the thesaurus, the semantic network and/or from the document collection.

Much research has been focusing on the so-called Semantic Web (or Web 2.0). Here, WWW-pages are provided with richer metadata than before, visa vie the important concepts/classes in hyper documents and some relationships among those concepts (mainly hierarchical relations, thus "is a"). This is expected to improve search quality (especially with regards to relevance).

2. Architecture of Infolution product and solution suite

2.1 Functional description: what it does

In Figure 2, the Infolution architecture is described and to be viewed bottom up. Here, the information resources can be seen. These data sources provide structured and semi-structured data to the Infolution suite. A wide variety of data sources (over 200 types) are recognized, including relational database files, various word processor files, spreadsheets, hypertext documents (e.g. from intranets), RSS feeds, and so on.

Figure 2 shows that the indexer and semantic network builder use the data sources as described in section 1. The engine consists of the following four functional units:

Harvest - Rosetta
This unit converts the data from the various sources into an internal format for Infolution.

Semantic Analysis - Strannix
Here, the converted data (including documents) is analyzed, resulting in seman tic networks. For each document, a network and an index are created. Then, the individual networks are integrated. Thereby, common clustering techniques are used in order to recognize the various meanings a particular concept may have.

Semantic Store & Retrieve - Nebula
This stores the semantic networks in relational database format, retrieves data, and performs fast and precise matching of queries to documents.
Known marker-passing techniques promote high speed and high precision.

Acces Control - Pretorian
Some documents have limited access in the sense that only designated persons are allowed access. Infolution maintains these limitations by means of an access control unit. It supports a number of restriction mechanisms such as access structure to Windows folders.

The Infolution engine works continuously in the background, performing all four functions, updating its files as soon as documents are added or modified. It prepares data for the three main functions, as shown in Figure 2 in the section above the engine. The three main functions include:

- **Search**
- **Classify**
- **Summarize**

'Search' and 'classify' may be regarded as a matching function. In 'search', queries are matched with document indexes; in 'classification' classes are matched with document indexes.

In text 'summarization" unimportant sentences and phrases of a text document are eliminated, leaving only the most relevant sentences. What is relevant is determined by a 'query', which indicates the context (including purpose) of the summarization. Summarization here resembles matching in the sense that the index terms of each sentence are matched with the query/description of the context.

These three main functions access the engine via the API (Application Programming Interphase). Other, future, functions may do the same, which simplifies future extensions of the suite.
These three main functions can be used by application programs via a web server which makes it possible to use these programs anywhere in the world where there is an internet connection.

The application (mainly search) programs available are listed at the top in Figure 2. They include an intuitive GUI (Graphical User Interface) developed by Infolution. The user communicates with the suite via the GUI thus inspects documents, (re)-formulates queries, etc.

2.2 Technical description: how it works

In order to distinguish the Infolution suite from its competitors, we provide information of how it works regarding the most interesting areas.

1. Semantic networking
Without a doubt the core technology of the Infolution suite is pivoted around the semantic networks. The networks themselves are well-known, though not standard technology today. Figure 3 shows an example network.

Such networks are usually created in a partly automated process, followed by manually cleaning-up and fine-tuning. Infolution does this fully automated. This also holds for the maintenance of the networks when documents are added. As a benefit result, either maintenance costs are decreased (compared with manual fine-tuning) or the search performance is improved (compared with low-quality automated creation and maintenance).

The understanding of natural language by computers has been a research issue for a few decades now and progress is slow. The holy grail of natural language processing is a full understanding of an arbitrary text. This goal has by far not yet been met. However, some understanding is possible and Infolution claims to have implemented this. They claim that their software can extract a large part of the concepts and their interrelations from a text. For those who insist on hearing numbers, this pertains to 80% of a piece of text.

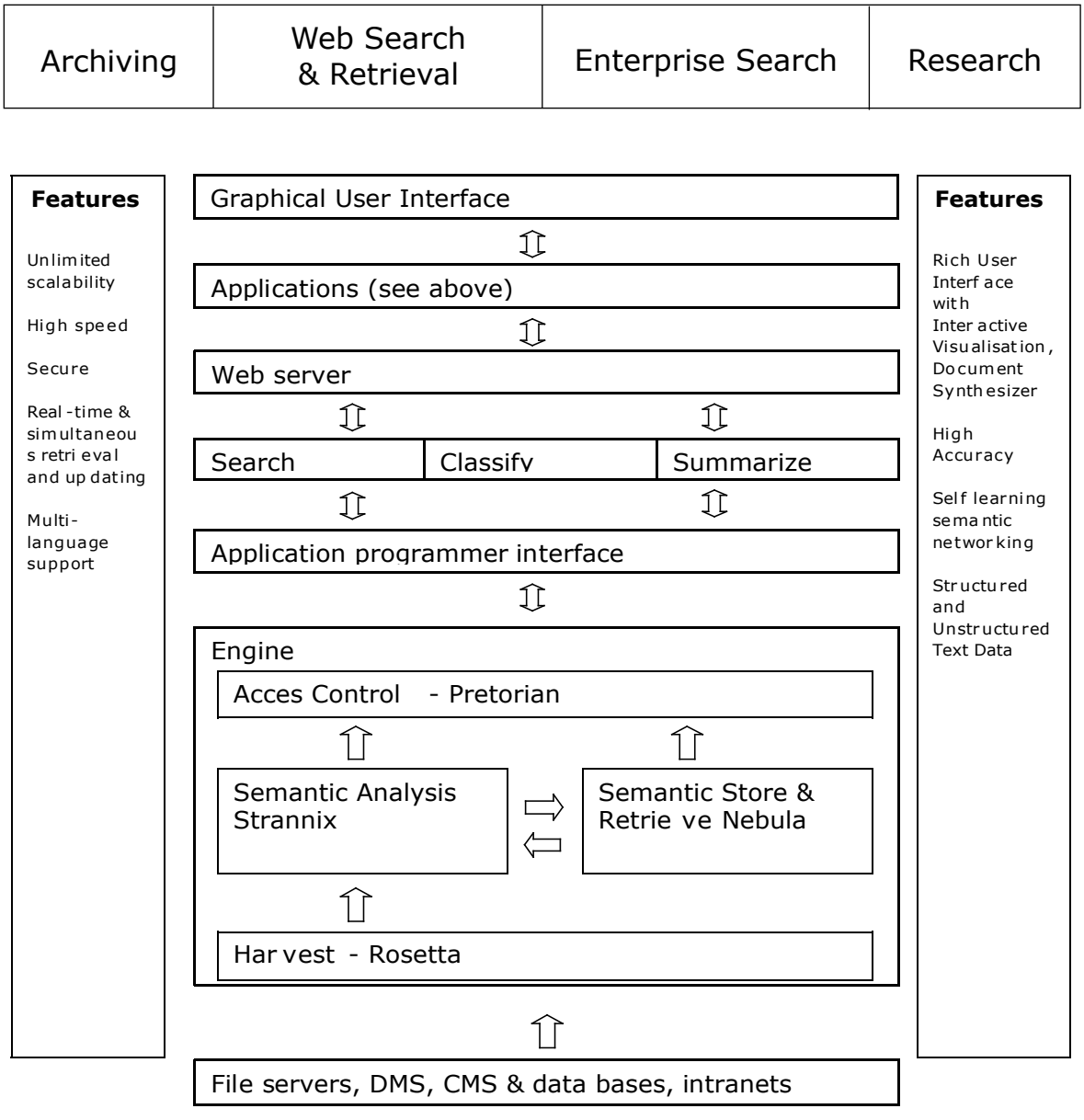

Figure 2: The architecture of the Infolution product suite

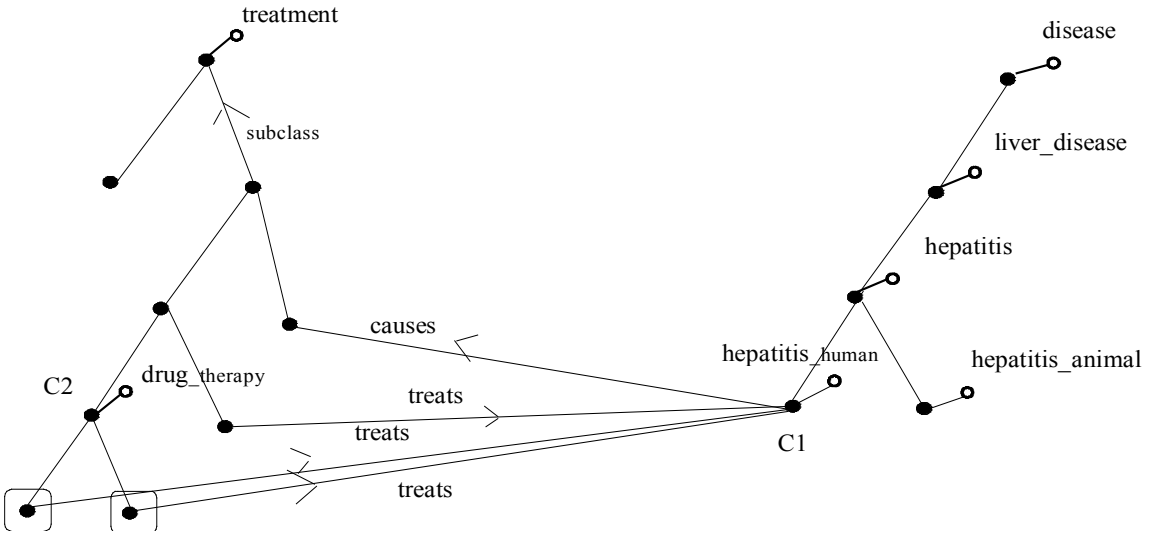

Figure 3: Semantic network example

Needless to say, Infolution is not keen on exposing their technology, but they are prepared to offer some insight. The automated process uses heuristic rules as well as pre-defined vocabularies to extract the concepts and relations. The rules make use of knowledge that is already present, therefore the more knowledge is gathered, the more they make use of it – in that sense they are self-learning. Just like people, who, the more knowledge they already have, the more they can do with new information. The types of relations gathered automatically are mainly hierarchical relations ("Is a") and sometimes other types e.g. 'causes'. Although the technology is sufficiently mature, Infolution continues to improve it.

2. Use of context information
Furthermore, the search task is treated as part of another task and makes use of the context thus provided. During a search information of the context is enclosed in the query and as such it improves the precision of the search process. In the application programs, the context information and the subject of each of the applications is formed.

3. Text summarisation
Text summarisation, also available as a separate function, is incorporated in the search interface. It allows the user to grasp the essence of a document quickly. The essence depends on a person's interests. The Infolution solution takes into account – which is unique – the interest by providing context information to the summarization function. The latter function then essentially focuses on the pieces of text that contain the concepts of the context information and closely related concepts and erases the other pieces by means of the semantic network. There's more to it than this, but this is the essence.

4. UI including graphical browser
Another feature of the Infolution suite is the user interface (UI) as visible in Figure 4 below. The main components of the UI are:

Query box
The query box is straightforward and meant to type a list of words. A Boolean query is also allowed for. For less experienced users, some help on query language would we advisable. This is available in a long list but would be more convenient as a context sensitive help feature. Again, this should not be difficult to realize.

Result list
The result list is a list of documents matched. For each document, it shows the degree of relevance to the query, a few pieces of text with the query terms, and an icon that – upon activation – also shows a summary of the corresponding document (focused on the query). The author would prefer to see the summarisations permanently instead of the other pieces of text – the former seem more useful.

Figure 4a: Infolution Graphical User Interface

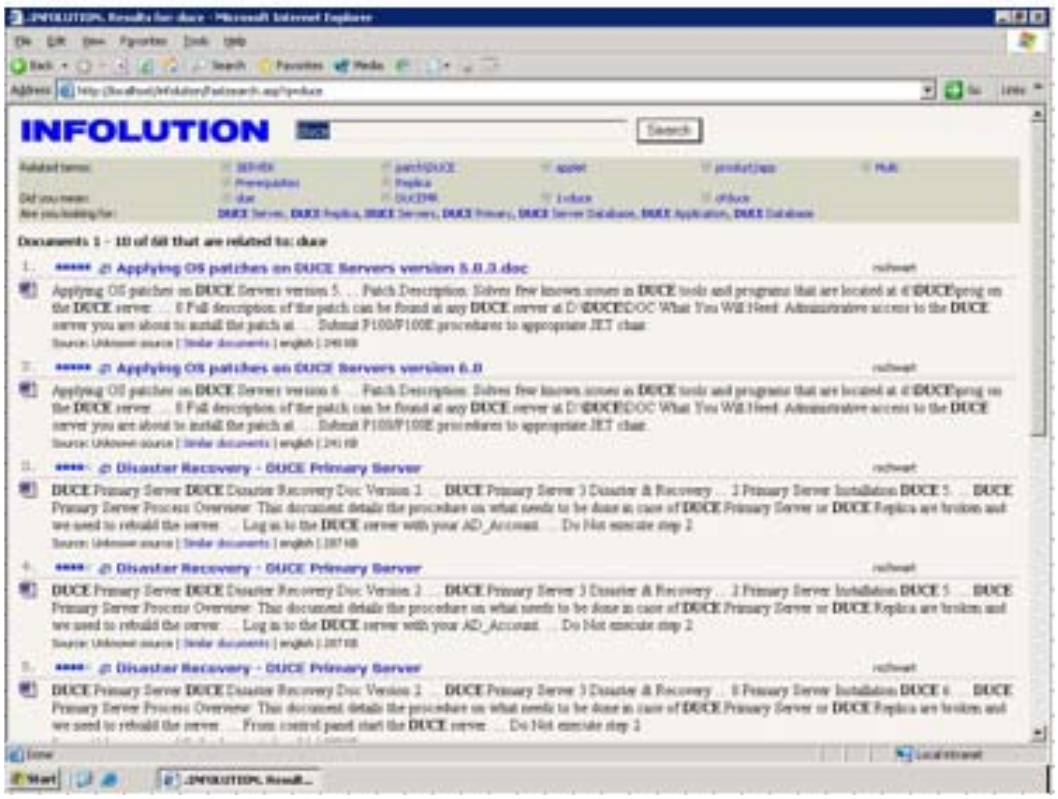

Figure 4b: Infolution list User Interface

Term screen: graphical browser or list
After a query has been entered and matched, the graphical browser shows a part of the semantic network, containing query terms and related terms in a graphical manner. Closely related terms are shown close to each other, and important terms are shown larger than relatively unimportant terms. The relationships are not shown; this is deliberate as not to confused the user with information-overload.

Although this type of interface was developed some ten years ago (among others) by the predecessor of Infolution and thus may not seem new at first glance, it is incorporated in a useful manner: it shows information from the search results. Moreover, marker-passing technology was added to make it faster.

Alternatively, the related terms are shown as a list. The user may click on a +-sign, or a minus sign to add or remove a term to the query. In the list presentation no relations among the terms are shown. Other search products exist in which a term hierarchy is shown, which seems highly useful. Adding this to the suite shouldn't be difficult.

The browser offers something similar: by clicking a term in the browser, that term will be put in the query box. By right clicking, another term may be added.

The absence of a backward navigation button in the browser seems an omission: typically, browsers are used for exploring a domain and this incorporates going forth and back. This possibility should be added.

To add another critical note: for the past ten years the author has used the same graphical browsers in the interface and prefers the versions in which the relations and relation type names are shown. One reason for this preference is that this information is useful for exploring less-well-known subject areas (this was found in scientific research). Moreover, a clear distinction between hierarchical relations and other kinds of relations would be welcome, but so far, no known graphical browser offers this feature. It seems that Infolution can easily provide the relation type names, since these are known – they are in the semantic network. This will greatly improve user friendliness and search performance.

Settings/preferences box
Ordering by relevance or date, and choice between vector query (term list) and Boolean query. The author would always prefer an exact match combined with relevance ranking. In the suite, this is obtained by Boolean matching followed by ranking.

The whole GUI can be configured to a great extent to meet user demands. The interface is embedded in the application programs.

5. User comfort
With regards to usability, the author needs to rely on some impressions from his own experiences with the suite. Overall, the product makes a balanced impression in the underlying technology and the graphical user interface. The various components work neatly together (fast matching, summarization, clear user interface), allowing for a relaxed search experience by the user. In other words, it offers a high user comfort. This will improve even more if the above-mentioned improvements are made.

3. Comparison to competing products

The Infolution suite is compared to competitive products. We divide the products into two main groups: those without semantic technologies, and those with.

We gathered our information from the respective websites and from relevant white papers. For an understanding of the terms used, we refer to Chapter 1.

Group I

Search tools, non semantic

- Google desktop
- Microsoft Windows XP/vista search & Sharepoint Server 2007
- Copernic 2.0
- Yahoo! Desktop Search 1.1 Beta
- Wizetech Archivarius 3000 3.14
- Open Text: a document management package with built in search

These products offer search, but not semantic search. They all have a full-text index containing words (not concepts) and use Boolean matching. All above products use a box as query input and offer a list as output; no graphical interface. They all handle a large number of file types. Conditional access is not, or hardly, supported. Performance is usually fast. Installation is usually easy, maintenance automatic. Prices vary: from 0 to some 40 Euros.

Some of the products in this group are used by millions of users. These tools offer low performance yet they are better than no tools at all. As such, they make the user familiar with search tools – possibly preparing them for a step up in performance.

Group II

Semantic tools

All these products have some form of semantic search, as opposed to those of group I. Semantic may just mean that concepts are recognized (to some degree) in the documents when indexing, or it may mean that also relationships among concepts are recognized – the latter in varying degree.

Below we show a table with features for each of the products. It is noted that many more features are present in all of these products or can be easily included on demand. For instance, if the product creates a semantic network, it is simple to include a predefined semantic network with the vocabulary of a specific group of users as well. We therefore decided to omit some obvious features and to focus on more essential ones.

The table covers the most relevant competing products. A myriad of new products exists based on natural language processing and semantic technology. Most of these are still beta-versions and
immature. On the basis of public information it is difficult to determine how advanced their technology is. Due to the aforementioned difficulties in understanding natural language by a computer, any breakthrough would probably have been mentioned on the website of the proud owner.
To mention some of these newer search products: Twine, Adaptive blue, Hakia.com, Talis, Trueknowledge (claims to do question-answering; which will work only for a few types of questions and can be easily integrated in all semantic products), Tripit, Clearforest, and Spock.

Comments to the Table
Autonomy searches multi-media data: text and also images, video, and sounds. It has additional functions such as subdividing audio and video files and adding hyperlinks to the sections.

Infolution has a patent application pending on the automatic detection and correction of spelling errors, using context information and another one on the automatic creation of stochastic semantic networks. When integrated in an application program, Infolution obtains context information from that application. Otherwise, it guesses the context from the query.

The Aduna graphical 'Autofocus' interface shows the result lists of a query and of its terms as connected spheres. Each sphere represents a cluster, e.g. the hits of a single query term. This is meant to provide the user insight into the effects of adding and removing and combining query terms. The graphical interface is beautiful. However in search, it didn't seem helpful when tested. The clusters may provide some additional overview, but most of it is already visible in the normal result list in list format. It wasn't clear how the spheres help to do the next selection-step i.e. finding the relevant documents in the result list. Thus, at the moment, the Autofocus interface adds little value; further development is needed. Aduna is offered for free; maintenance and support are not free.

Fast ESP has a ® on the use of context information. It guesses the search goal (as context) from the query. Clustering in Fast is used to organize the result list into groups/clusters. Furthermore, Fast ESP uses context information to show only specific parts of documents. And it recognizes specific entities in documents such as names and phone numbers.

Eidetica still runs a website which has not been renewed over the last four years and it does not promote its products; they no longer appear to be active in business.

	Infolution Suite	Fast ESP	Intellisearch	Autonomy K2v7	Irion/21
Main functions					
Search					
Enterprise Search	+	+	+	+	+
Web search	+	+	+	+	+
Desktop search	+	+	+	+	+
Classification	+	-	+	-	+
Summarization	+	-	-	-	+
Search Effectiveness, efficiency & comfort					
Indexing					
File types supported	> 200	> 400	> 200	> 300	many
Index type	conceptual	conceptual	conceptual	conceptual	
Use of hyper structure	-	-	-	-	
Multilingual	+	+ (81)	+	+	+
Clustering	+	+	+	+	+
Correction of typos etc	+	+	+	+	+
Matching					
Query language	T, B	B	B	NL, B	T, B
Automated query expansion	+	+	+	+	+
Use of context information	+	+	-	-	-
Use of sem networks	+	tax, syn	+	+	+
Relevance ranking	+	+	+	+	+
Autom. creation of sem. nets	+	+	+	+	+
Speed	+	+	+	+	+
UI					
Graphical	+	-	-	-	-
List	+	+	+	+	+
Summarization	+	-	-	-	-
Various qualities					
Support of conditional access	+	-	+	+	-
Query management	+	+	++	+	-
System maintenance	+	?	+	+	-
Scalable document collection	+	+	+	+	?
Open source	+	-	-	-	-
Downloadable via WWW	+	-	-	-	-
Integration via API	+	+	+	+	-
Pricing	on inquiry	on inquiry	on inquiry	> $ 80k inquiry	on

T = term list
B = Boolean
NL = natural language
Tax = taxonomy (= type hierarchy)
Syn = synonyms

	Carp Sinope	Google Search Appliance	Eidetica	Aduna
Main functions				
Search				
Enterprise Search	-	+	+	+
Web search	+	+	+	+
Desktop search	+	+	-	+
Classification	+	-	+	-
Summarization	+	-	-	-
Search Effectiveness, efficiency & comfort				
Indexing				
File types supported	many	many	many	several
Index type	conceptual	conceptual	conceptual	conceptual
Use of hyper structure	-*	-	-	-
Multilingual	+	+	+	?
Clustering	+	+	+	+
Correction of typos etc	+	+	+	?
Matching				
Query language	NL, T, B	B	T, B	B
Automated query expansion	+	+	-	-
Use of context information	-	-	-	-
Use of sem networks	tax	tax	tax, syn	tax
Relevance ranking	+	+	+	-
Autom. creation of sem. nets	+	+	+	+
Speed	+	+	+	+
UI				
Graphical	-	-	-	+
List	+	+	+	+
Summarization	+	-	-	-
Various qualities				
Support of conditional access	-	+	-	-
Query management	-	+	-	+/-
System maintenance	-	+	-	-
Scalable document collection	+	+	+	+
Open source	-	-	-	+
Downloadable via WWW	-	-	-	+
Integration via API	-	+	-	-
Pricing	on inquiry	see below	on inquiry	free or paid

* Google web search uses the page rank algorithm that utilized hyperlink structure of the WWW. Google desktop search and Search Appliance do not use the hyper link structure.

Google pricing:

Google Hosted search up to 5000 docs	> $ 100 per year
Google Mini Search Appliance	> $ 2k
Google Search Appliance	$ 30k – 500k

The search appliance uses its own specific hardware.

4. Discussion and conclusion

Discussion
In the previous Chapter, we distinguished two groups. The first, search products without semantic technologies, which is considered an entry market that offers search at low cost. This group is useful for users with low demands on search performance.

The second group, to which the Infolution suite belongs, offers semantic search in varying degrees. Unfortunately, all producers are vague on how precisely their semantic search works. Experimental data is also absent, especially quantitative data.

The user interfaces can be evaluated somewhat easier by 'playing' with them. However, this can only be done with respect to the user friendliness, not in terms of the search performance. Regarding the two products with a graphical user interface, Infolution and Aduna, the following comments are made:

Regarding the Infolution user interface, we discussed and criticized this already in Chapter 2. It is a nice, useful interface that still needs some fine-tuning. The graphical browser in particular seems useful but might become more so, when it shows more of the already available information (eg. relation names).

Regarding the Aduna interface, we concluded in Chapter 3 that it is absolutely beautiful. Usability seems to be almost absent, though.

In search of a positive distinction of Infolution from the competitors, we note:

- The suite has all the essential features. Yet, there are competitors that also have most of these: Intellisearch, Autonomy and Fast.
- First distinguishing feature: the 'semanticity' of Infolution seems to go beyond that of competitors, esp. regarding the automatic creation of advanced semantic networks. We say seems, because it is stated by Infolution and appears true in our own experimental exploration but it was not scientifically proven.
- Second distinguishing feature: the user interface is the only one that combines a graphical part, a list and summarization. Due to the underlying technology that offers high performance search (effective and fast response of the interface), it is comfortable to work with. This will be even more so when the aforementioned fine-tuning is completed.

Conclusion
The answer to our initial question, "What distinguishes the Infolution suite from its competitors?" is:

- advanced semantic technology, and
- a comfortable, user-friendly interface

These two features seem to lead to a higher search performance in terms of effectiveness, efficiency and user comfort (lower mental load).

Additionally, the summarisation and classification are superior. In search, the use of context information is also done by Fast, summarisation also in Sinope by Carp and by Irion. However the latter two do not use context information to focus the summarisation.

About the author.

The author works as an independent consultant in computer science and systems engineering. He is an engineer and has a PhD in knowledge-based (i.e. semantic) information retrieval.

About the review

The review requested by Infolution. The author was given free hand to be critical and the opinions given are sincere and his own. The author has tried to provide accurate factual information, however, accepts no liability for any mistakes.