A Survey of Concept-based Information Retrieval Tools on the Web

Hele-Mai Haav

Tanel-Lauri Lubi

Institute of Cybernetics at Tallinn Technical University
Akadeemia tee 21, 12618 Tallinn
helemai@cs.ioc.ee
lauri@cc.ioc.ee

Abstract. In order to solve the problem of information overkill on the web current information retrieval tools need to be improved. Much more "intelligence" should be embedded to search tools to manage effectively search, retrieval, filtering and presenting relevant information. This can be done by concept-based (or ontology driven) information retrieval, which is considered as one of the high-impact technologies for the next ten years. Nevertheless, most of commercial products of search and retrieval category do not report about concept-based search features. The paper provides an overview of concept-based information retrieval techniques and software tools currently available as prototypes or commercial products. Tools are evaluated using feature classification, which incorporates general characteristics of tools and their information retrieval features.

1 Introduction and Motivation

Current information retrieval tools mostly use keyword search, which is unsatisfactory option because of its low precision and recall. In this paper, we consider concept-based information retrieval model as a new and promising way of improving search on the web. Informally, concept-based information retrieval is search for information objects based on their meaning rather than on the presence of the keywords in the object.

In the last 5 years, concept-based information retrieval tools have been created and used mostly in academic and industrial research environments [Guarino et al 1999, Woods 1998]. For example, in the survey of information retrieval vendors by R. J. Kuhns [Kuhns 1996] only 4 vendors from 23 surveyed vendors delivered concept-based information retrieval tools. This survey did not cover web search tools. Currently we an in the situation, where new commercial and experimental concept-based information retrieval tools are rapidly emerging. Most of these tools offer search facilities for the web. Nevertheless, according to Internet Product Watch

[Internet Product Watch], only about 10 from 116 commercial products of search and retrieval category have reported about concept-based search features.

Our survey covers 13 concept-based information retrieval tools, which according to our knowledge is the most relevant set of tools with respect to exhibiting concept-based retrieval features. Information about commercial tools is gathered mostly from the vendors' homepages and white papers. Research prototypes are described in corresponding research papers and projects. A status of the product development of the tools is different. Among 13 tools 4 are research prototypes and others are commercial products. As one of the goals of the paper is to provide an overview of existing concept-based information retrieval techniques, then choosing tools with different production status is motivated. Another goal of the survey is to identify important features to study and evaluate concept-based information retrieval tools for the web.

The rest of the paper is organised as follows. Section 2 discusses existing information retrieval models and provides an overview of the concept-based information retrieval techniques. Section 3 presents a methodology of a survey and a review of considered tools using the methodology. Section 4 draws conclusions from a survey and evaluates the tools.

2 Concept-based Information Retrieval

This section serves as an introduction to the field of concept-based information retrieval on the web giving background knowledge for the survey methodology used.

It is necessary for information retrieval that information objects have a description of their contents. Matching their descriptions against a user's query can then retrieve information objects. Text can serve as a universal description of any type of information source, including images, audio and video. This is wellknown and well-utilised in most of search tools. We distinguish two main information retrieval models as described in the following subsections.

2.1 Keyword-based Information Retrieval Model

Information retrieval model commonly used in commercial search engines is based on keyword indexing systems (manual or automatic) and Boolean logic queries that are sometimes equipped with statistical methods (e.g. frequency of occurrence of a keyword is taken into account or some proximity constraints are used). We call this model keyword-based information retrieval model.

In this model, keyword lists are used to describe contents of information objects. Keyword list is a description that does not say anything about semantic relationships between keywords. One could easily choose a valid synonymous word that is not in any textual objects and therefor fail the search.

Principal problem with this kind of information retrieval model is that it does not take into account meaning of the word or phrase. A word for this model is only a sequence of binary codes representing a word. Even if some linguistic search systems use word stemming and phrase dictionaries, this does not mean that they use a different information retrieval model.

2.2 Concept-based Information Retrieval Model

In the cognitive view of the world, there exists the presumption that the meaning of a text (word) depends on conceptual relationships to objects in the world rather than to linguistic or contextual relations found in texts or dictionaries. A new generation information retrieval model is drawn from this view. We call it concept-based information retrieval model. Sets of words, names, noun-phrases, terms, etc. will be mapped to the concepts they encode.

Generally, a content of an information object is described by a set of concepts in this model. Concepts can be extracted from the text by categorisation. Crucial in this model is existence of a conceptual structure for mapping descriptions of information objects to concepts used in a query. If keywords or noun-phrases are used, then they should be mapped to concepts in a conceptual structure.

Conceptual structures can be general or domain specific, they can be created manually or automatically, they can differ in the forms of representation and ways of constructing relationships between the concepts. Naturally, the tools considered in this paper differ in this respect.

In this section, we concentrate to description of fundamental features of concept-based search tools: conceptual structure and its usage for improving search. Additional ordinary search methods are not discussed here but only given in the tables 1-3 presented in appendix.

Types of conceptual structures. For establishing definitions of concepts it is necessary first to identify concepts inside the text and then classify found concepts according to the given conceptual structure. There are several ways of identification of concepts present in the text. This process is called categorisation. Texts explicitly contain words rather than concepts. As concepts are expressed by natural language, then it is possible to identify them in the text by analysing phrases. In many concept-based information retrieval systems (tools) Natural Language Processing (NLP) is used to analyse syntax and semantics of the text for categorisation [Adi et al 1999, DioWeb 2001, LexiGuide, MetaMorph, etc].

Concepts can be identified also by using fuzzy reasoning about the cues (terms) found in the text for calculating likelihood of a concept present in the text [Loh, etc 2000].

After the concept is categorised, it can be given the definition by a classification process. Classification is determining where in the conceptual structure a new concept belongs. For this purpose, either an existing conceptual structure (like dictionary, thesaurus or ontology) or automatically generated one can be used. It is reported in many papers [Loh, etc 2000], [Guarino 1998] that pre-existing dictionaries often do not meet the user's needs for interesting concepts, or ontology like WordNet [Miller 1995] does not include proper nouns.

Conceptual structure can be automatically generated by learning process. In the case of unsupervised learning this process is called conceptual clustering, which organises information objects into groups or categories, where each category represents a relevant concept interpreted in the problem domain (context).

The main types of conceptual structures used in concept-based information retrieval systems are described below.

Conceptual taxonomy. Conceptual taxonomy is a hierarchical organisation of concept descriptions according to generalisation relationship. Each concept in taxonomy has link to its most specific subsumers ("parents" or superconcepts) and links to its most general subsumees ("children" or subconcepts) in a taxonomy.

Usually, conceptual taxonomies are constructed manually by deciding where in the taxonomy each concept should be located. Conceptual taxonomies can be constructed automatically using special conceptual indexing technique as proposed in the project by Sun Microsystems [Woods 1997].

Formal or domain ontology. Ontology is a conceptual representation of the entities, events, and their relationships that compose a specific domain. Two primary relationships are abstraction (subsumption) and composition ("part-of" relationship) [Guarino and Giaretta 1995]. It is said in [Gruber 1995] "Ontologies are agreements about shared conceptualisation".

Depending on the subject of the conceptualisation, some authors [Van Heijst 1997] distinguish between *application ontologies, domain ontologies, generic ontologies* and *representation ontologies.*

According to them, top-level ontologies describe very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain. On the other hand, domain ontologies and task ontologies describe, respectively, the vocabulary related to a generic domain (like medicine, or automobiles) or a generic task or activity (like diagnosing or selling), by specializing the terms introduced in the top-level ontology.

In concept-based information retrieval systems, an ontology can serve as a resource description and can be used for query formulation. However, there are large linguistic ontologies available like WordNet and EuroWordNet [Miller 1995], many researches report about a lack of ontological information adaptable for knowledge retrieval purposes [Guarino 1999]. The problem relates to that those ontologies are built based on linguistic criteria and as such they are hard to use for non-linguistic applications. A solution can be in construction of formal ontologies or restructuring linguistic ontologies by using methods for formal ontology design.

For example, in OntoSeek project [Guarino et al 1999], Sensus ontology was used for concept-based retrieval in yellow pages and product catalogs. Conceptual graphs obtained from queries have been linked to ontology by using lexical conceptual graphs. In [Haav and Nilsson 2000] two approaches of using lattices as formal basis for ontology construction are considered and proposed for utilising in OntoQuery project [OntoQuery].

Two interesting projects SHOE [Luke and Helfin 1997, Helfin and Hendler 2000] and ONTOBROKER [Decker, et al 1999, Erdmann and Studer 1998] are based on an idea to annotate HTML pages with ontologies for concept-based retrieval purposes. A special annotation language is used to annotate HTML pages with ontological information. In SHOE, description logic is used for ontology description. ONTOBROKER relies on Frame Logic that supports more powerful inference mechanism than description logic.

Semantic linguistic network of concepts. In many commercial concept-based information retrieval tools [LexiGuide, RetrievalWare, MetaMorph, DioWeb, Webinator] NLP is used for creation of conceptual structure in some form of semantic network. Usually, in these systems a user can submit queries in natural language.

For example, in Excalibur RetrievalWare [RetrievalWare], natural language concept search is based on special semantic network. It supports for over twenty languages. Cyc [Lenat 1998] NLP converts text to formal language CycL for inclusion to Cyc Knowledge Base (KB). The KB consists of terms constituting the vocabulary of CycL, and assertions, which relate those terms. These assertions include both simple ground assertions and rules. Approximately 3,000 terms capturing the most general concepts are referred to as the "upper Cyc® ontology" that is made publicly available.

Thesaurus. Thesaurus is a collection of words or phrases linked through a set of relationships including synonymy, antonymy, and “isa” relationship. Thesaurus provides automatic semantic term expansion of queries in information retrieval systems [MetaMorph, Webinator]. Thesaurus building is manual work and as such very time-consuming.

Predictive model. Predictive models like neural networks can be used for concept-based information retrieval. HNC Software Inc. [HNC Software 2000] uses Context Vector™ technology for encoding textual information. Using special training algorithms, context vectors are assigned to objects in such a way that vectors for related objects will be closer together than vectors for unrelated objects. Thus, finding vectors that are closest to each other solves the problem of associating similar objects based on a textual description. Traditional query and retrieval is just finding documents that are similar to the query. Combined with a "self-organising" neural network technique Context Vectors actually "learn" the meaning of content - whether it is text, symbols, or images. They make it possible to eliminate the need for costly and time-consuming human work.

3 Methodology of a Survey

In this section we present feature classification scheme developed to study concept-based information retrieval software tools on the Web. Features relevant to study are grouped into 3 groups as follows: general features, features of conceptual structures and additional search features. The classification is applied to 13 concept-based information retrieval tools and project prototypes. The results of the study are presented in 3 different tables found in the appendix of this paper. The following subsections show groups of features together with their explanations.

3.1 General Features of Tools (Table 1 in appendix)

The following features are considered as general characteristics of the systems:

1. Product name and vendor, home page location on the Web
2. Purpose and functionality
3. Production and legal status: commercial (C), research prototype (RP)

Legal Status: Freeware (F), Commercial (C)
4. Demo: demo version available for download on the Net, demo available on request, unknown (-)
5. Network and system architecture: Intranet, Internet, Extranet, agent-based, client/server

3.2 Features of Conceptual Structures (Table 2 in appendix)

Conceptual structures used for concept-based information retrieval are characterised by the following features:
1. Type of a conceptual structure: concept taxonomy, domain ontology, top ontology, linguistic ontology, semantic linguistic network, predictive model, thesaurus, dictionary
2. Form of representation of a conceptual structure: tree, semantic network, context vectors, conceptual graphs, rule-based language, and logic language, etc.
3. Relationships supported by a conceptual structure: subsumption, a kind-of, a part-of, associations, and relations, etc.
4. Way of creation of a conceptual structure: manual creation, automatic learning, and NLP

3.3 Additional Search Features (Table 3 in appendix)

Most of commercial concept-based retrieval systems offer wide spectrum of ordinary advanced search methods in addition to the concept-based search features. We grouped these features as follows:
1. Additional search: Boolean, statistical, thesaurus-based fuzzy search, stemming, terms weighting, pattern matching (PM), and Natural Language Querying (NLQ)
2. Indexing methods: keyword indexing, conceptual indexing, category based indexing
3. Data types: databases, HTML, XML, text, PDF, images, video, audio, etc.

When gathering the information about commercial concept-searching tools we faced a problem that some vendors do not will to publish technical characteristics of their search features. Also, in many cases different vendors use slightly different terms to denote the same search feature. In the tables 1-3 we tried to use as common notions as possible to denote search features.

4 Conclusions

From the tables 1-3 (see appendix) we can draw some important conclusions. First of all, the table 1 shows that ontology-driven search is the goal of research projects rather than commercial information retrieval tools. Companies delivering or at least advertising their concept-based search tools are not so ambitious in their purposes, even if they have realised that concept-based approach is valuable for improving precision of the search. Also agent technology and machine learning are not yet widely used in these systems.

It is interesting to observe (in the table 2) that commercial products tend to use less complex and less formal conceptual structures than research projects. Commercially available concept-based search tools are mostly built on the basis of semantic linguistic networks or thesauri, and as such usually support NL queries in multiple languages. Commercial companies have developed very large rule bases capturing words, phrases and their relationships. Nevertheless, these knowledge bases are represented in different languages and as forming vendor's assets are not publicly distributed. Available linguistic ontologies like WordNet are not used by most of commercial companies. Research projects and sophisticated commercial products use conceptual taxonomies or ontologies to enable mapping query terms to concepts. In these projects effort is made to automation of creation of conceptual structure using a kind of automatic extraction of knowledge from information sources.

From the table 3 we can see that very wide range of data types are supported by concept-based information retrieval tools. All commercial tools provide also wide range of additional power search features, in contrast to research prototypes, which are concentrated on concept searching. Keyword indexing is used in most of commercial tools in contrast to research prototypes, where the goal is to get free of indexes.

In conclusion, we draw some future trends in concept-based information retrieval on the Web. One direction is to replace (or at least assist) human intervention in ontology construction by some inductive learning technique or text mining method. Automatic construction of domain ontologies will probably lead to their usage in commercial products. Another interesting trend is in merging domain ontologies and XML for information integration. The latter is related to agent communication, where agents need to share knowledge.

Acknowledgements

We thank Estonian Research Foundation for supporting this work partly by the grants 2772 and 4705.

References

[Adi and Ewell 1987] T. Adi and Mr. O. K. Ewell, "Letter Semantics in Arabic Morphology: A Discovery About Human Languages," pp. 21-52, Jul. 1987, Stanford University.

[Adi, et al 1999] T. Adi, O. K. Ewell and P. Adi, High Selectivity and Accuracy with READWARE's Automated System of Knowledge Organisation, 1999, (available online at www.readware.com)

[Decker, et al 1999] Stefan Decker, Michael Erdmann, Dieter Fensel, and Rudi Studer: Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In R. Meersman et al. (eds.): Semantic Issues in Multimedia Systems. Proceedings of DS-8. Kluwer Academic Publisher, Boston, 1999, 351-369.

[DioWeb 2001] Diogene Technology White Paper (available at www.dioweb.com)

[Erdmann and Studer 1998], Michael Erdmann, Rudi Studer: Ontologies as Conceptual Models for XML Documents. In: Proceedings of the 12th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'99), Banff, Canada, October 1998.

[Gruber 1995] Gruber T, Toward Principles for Design of Ontologies Used for Knowledge Sharing, International Journal of Human and Computer Studies, 43 (5/6): 907-928

[Guarino 1998] N. Guarino, Formal Ontology and Information Systems, In N. Guarino (Ed), Formal Ontology in Information Systems, Proc. Of the 1st International Conference, Trento, Italy, June 1998, IOS Press, Amsterdam, pp 3-15

[Guarino et al 1999] N. Guarino, C. Masolo, G Vetere, OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, May/June 1999, pp 70-80

[Haav and Nilsson 2000] Haav H-M and Nilsson J. F., Approaches to Concept Based Exploration of Information Resources, W. Abramowicz and J. Zurada (Eds), Knowledge Discovery for Business Information Systems, Kluwer Academic Publishers, 2000, ch 4, pp 89-111

[Helfin and Hendler 2000] Helfin J., and Hendler J. Dynamic Ontologies on the Web, In Proceedings of American Association for AI Conference (AAAI-2000), Menlo Park, Califirnia, AAAI Press 2000

[HNC Software 2000] Intelligent Response-The Right Answer, HNC Software White Paper, August 2000 (available online at www.hnc.com)

[Internet Product Watch] Internet Product Watch at http://ipw.internet.com

[Kuhns 1996] Kuhns, Robert J., "A Survey of Information Retrieval Vendors," Technical Report SMLI TR-96-56, Sun Microsystems Laboratories, Mountain View, CA, October 1996. (Available online at: http://www.sun.com/research/techrep/1996/abstract-56.html).

[Lenat 1998] Lenat D., The Dimensions of Context-Space, Cycorp, 1998 (available online at www.cyc.com)

[LexiGuide] LexiQuest homepage at http://www.lexiquest.com

[Loh, et al 2000] S. Loh, L K Wives, and J P M de Oliveira, Concept-Based Knowledge Discovery in Texts Extracted from the Web, SIGKDD Explorations, ACM SIGKDD, July 2000, Vol 1, Issue 1, 29-39

[Luke and Helfin 1997] SHOE 1.0 proposed specification (available at http://www.cs.umd.edu/projects/plus/SHOE/spec.html

[MetaMorph] Thunderstone homepage http://www.thunderstone.com

[Miller 95] Miller, G. A., WORDNET: A Lexical Database for English, Communications of ACM (11): 39-41

[OntoQuery] OntoQuery Project, www.ontoquery.dk

[RetrievalWare] Excalibur homepage at http://www.excalib.com

[Van Heijst 1997] Van Heijst, G., Schreiber, A. T., and Wielinga, B. J., Using Explicit Ontologies in KBS Development. *International Journal of Human and Computer Studies, 1997*

[Verity Search 2000] Verity Search- The Advantage of Advanced Information Retrieval, Verity White Paper, Nov. 2000 (available online at www.verity.com)

[Webinator] Thunderstone homepage http://www.thunderstone.com

[Woods 1997] Woods, W. A., "Conceptual Indexing: a better way to organize knowledge," Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April 1997. (Available online at: http://www.sun.com/research/techrep/1997/abstract-61.html).

Appendix: Features of the Concept-based Retrieval Systems

Table 1. General features of systems

	Product name, vendor and homepage	Purpose and functionality	Status	Demo	System and network architecture
1.	OntoSeek, LADSEB-CNR, IBM http://www.ladseb.pd.cnr.it/infor/ontology	Content-based IR from yellow pages and product catalogs	RP	-	Internet Client-server
2	Ontobroker, University of Karlsruhe http://ontobroker.aifb.uni-karlsruhe.de	Ontology-based search and answering	RP, F	Available	Internet, Intranet
3.	SHOE,University of Maryland http://www.cs.umd.edu/projects/plus/SHOE/	Ontology description for HTML search on the web	RP, F	Available	Internet Agent-based
4	Sun Microsystems Conceptual Indexing http://www.sun.com	Solving "paraphrase problem" in IR	RP	-	Internet
5	Cyc Knowledge server, Cycorp., http://www.cyc.com	Concept-based search	C	-	Internet Client-server
6.	Verity Search, Verity, Inc, http://www.verity.com	Full-text search for business portals.	C	-	Intranet, Internet Databases
7.	HNC Software Inc., Mindwave, Context Vectors, http://www.hncs.com	Content-based retrieval	C	-	Intranet, Extranet Internet
8.	LexiGuide, LexiQuest http://www.lexiquest.com	Intranet search and Web search.	C	Available on request	Intranet, Internet Agent-based learning
9	Excalibur RetrievalWare, Excalibur Technologies, http://www.excalib.com	Intranet search. Supports Web-based GUI.	C	Available	Intranet, Internet Agent-based learning
10	Readware ConSearch, Management Information Technologies, Inc. http://www.readware.com	Intranet and Internet search	C	Available	Intranet, Internet Non-agent-based
11	Thunderstone MetaMorph,Thunderstone http://www.thunderstone.com	Site's internal search	C	Available	Internet Non-agent-based
12	Thunderstone Webinator,Thunderstone http://www.thunderstone.com	Intranet document search	C, F	-	Expandable Intranet
13	DioWeb Search, Diogene TechnologyTM www.dioweb.com	Contextual search	C	-	Intranet, Internet

Table 2. Features of Conceptual Structures (CS)

	Product name	Type and form of representation of CS	Relationships in CS and way of creation of CS
1.	OntoSeek [Guarino 1999]	Ontology, linguistic ontologies, conceptual graph. Ontolingua syntax adopted for ontology representation	Subsumption Using existing linguistic ontologies WordNet, Sensus
2	Ontobroker [Decker, et al 1999]	Domain specific ontology, Frame Logic [Kifer at al 1995] is used for ontology description	Subclassing, instance of, part-of, relations, attributes Partly automatic extraction
3.	SHOE [Heflin and Hendler 2000]	Domain Ontology, ontology description language (description logic) as HTML extension	Subsumption between categories, inference rules, relationships between domain ontologies Manual mapping concepts to ontology's vocabulary
4	Sun Microsystems Conceptual Indexing. [Woods 1997]	Conceptual taxonomy, KL-One like knowledge base, 30000 rules	Intensional subsumption, a kind-of relationship Automatic conceptual indexing system
5	Cyc Knowledge server [Lenat 1998]	Context space, Cyc Top Ontology Knowledge base of assertions (rules) given in CycL,	12 classes of dimensions of context-space (e.g. time) Manual entering of rules
6.	Verity Search [Verity Search 2000]	Classification, rules in Verity's representation language, weighted semantic graph	Relationships between words, weights for relationships Manual building of rules by experts
7	Mindwave, Context Vectors. [HNC Software 2000]	Predictive model. Context Vector™ technology combined with a "self-organising" neural networks	Similarity associations between context vectors Automatic learning and categorisation
8.	LexiGuide [LexiGuide]	Semantic linguistic network, dictionary Semantic network	Links between words and semantic concepts, NLP, multiple languages
9	Excalibur RetrievalWare, [RetrievalWare]	Semantic network of concepts. built-in Knowledge Base of 50000 word meanings and over 1,6 million relationships	Links between concepts, links to dictionaries and thesaurus, NLP
10	Readware ConSearch [Adi, et al 1999]	Ontology of word roots (natural concepts). Semantic network, concept base of knowledge types. Rules called Letter Semantics	Matrix theoretical concept-relations Automatic text analysis
11	Thunderstone MetaMorph, [MetaMorph]	Thesaurus Linguistic network	Word and concept associations. logical operations with concepts, NLP
12	Thunderstone Webinator [Webinator]	Thesaurus Linguistic network	Word and phrase associations, NLP Concept expansion
13	DioWeb Search [DioWeb 2001]	Semantic linguistic network Morphological-contextual linguistic network	Word and concept associations NLP

Table 3. Additional search features

	Product name	Additional search methods	Indexing methods	Data types supported
1.	OntoSeek [Guarino 1999]	Graphical browsing of an ontology	Not used	Web pages, Yellow pages, product catalogs
2	Ontobroker [Decker, et al 1999]	Queries in Frame Logic, Inference in Horn Logic	Index is used	Web pages
3.	SHOE [Heflin and Hendler 2000]	Creation of ordinary queries for search engines	Ontology and category based indexing	Web pages
4	Sun Microsystems Conceptual Indexing, [Woods 1997]	-	Conceptual indexing	Web pages, text
5	Cyc Knowledge server [Lenat 1998]	NLQ front end to CycL query	-	Web, text and images. Databases
6	Verity Search [Verity Search 2000]	Boolean, proximity, frequency, fuzzy search	Indexing documents into Verity Collections	Web pages, e-commerce sites, text, PDF, etc
7.	Mindwave, Context Vectors. [HNC Software 2000]	NLQ, text and audio search, compound search	-	Different types of information sources
8	LexiGuide [LexiGuide]	NLQ, Integration with other web search engines	-	Web pages, PDF, MS Office documents, text, etc
9	Excalibur RetrievalWare, [RetrievalWare]	Adaptive Pattern Recognition Processing, Boolean, statistical	Indexing across the network	Web pages, PDF, over 200 in total
10	Readware ConSearch [Adi, et al 1999]	Word search, concept search Super-concept search	-	Text , text databases
11	Thunderstone MetaMorph [MetaMorph]	PM Algorithms (numeric PM, approximate PM, etc), wildcard search, NLQ	Not used	Text, also embedded into other formats, API for databases
12	Thunderstone Webinator [Webinator]	NLQ, set logic, fuzzy pattern matching	Common index for multiple sites	Over 100 data formats SQL query interface
13	DioWeb Search [DioWeb 2001]	Relevance order ranking	Indexing is widely used	225 different file formats