Information Gathering and Searching Approaches in the Web

Seiji YAMADA

Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori-ku, Yokohama, 226-8502 JAPAN
yamada@ymd.dis.titech.ac.jp

Hiroyuki KAWANO

Department of Systems Science
Graduate School of Informatics
Kyoto University
Sakyou-ku, Yoshida-honmachi, Kyoto, 606-8501 JAPAN
kawano@i.kyoto-u.ac.jp

Received 15 November 2000

Abstract The information accessible through the Internet is increasing explosively as the Web is getting more and more widespread. In this situation, the Web is indispensable information resource for both of information gathering and information searching. Though traditional information retrieval techniques have been applied to information gathering and search in the Web, they are insufficient for this new form of information source. Fortunately some AI techniques can be straightforwardly applicable to such tasks in the Web, and many researchers are trying this approach. In this paper, we attempt to describe the current state of information gathering and searching technologies in the Web, and the application of AI techniques in the fields. Then we point out limitations of these traditional and AI approaches and introduce two approaches: navigation planning and a Mondou search engine for overcoming them. The navigation planning system tries to collect systematic knowledge, rather than Web pages, which are only pieces of knowledge. The Mondou search engine copes with the problems of the query expansion/modification based on the techniques of text/web mining and information visualization.

Keywords The World Wide Web, Information Gathering, Web Search Engine

§1 Introduction

As the Internet and the Web are developing explosively, we can easily utilize them as useful information resources. Since the Web is very huge and constantly updated, information and knowledge in the Web is rich and recent. In this situation, there is a possibility that we can do information gathering and information searching in the Web anytime, anywhere to obtain our target Web pages. However there are significant problems. In general, a user does not know where his/her target Web pages exist. Thus a Web search engine is used to search for the target pages. Unfortunately, since the filtering is not sufficient, a hit list from a search engine may include a large number of non-relevant Web pages. This problem also influences information gathering. If information searching is not correct, the gathered information is also meaningless.

Though traditional information retrieval techniques[28)] have been applied to information gathering and search in the Web, not all of the problems are solved. For example, filtering of a hit list using semantics and the context of Web pages is hard for information retrieval methods and natural language processing is necessary. Fortunately some AI techniques are straightforwardly applicable to such tasks in the Web, and many researchers are trying this approach.

In this paper, first we describe the current state of information gathering and searching in the Web, and survey systems developed as AI applications. Next we point out open problems in the fields. Then we introduce two approaches: navigation planning and a Mondou search engine for overcoming the issues.

§2 Information gathering and searching in the Web

In this section, we survey conventional approaches and implemented systems for information gathering and searching in the Web, then we describe their limitations.

2.1 Traditional systems

[1] *Softbot, Sage and Occam*

There are studies on AI planning to generate procedures for gathering information through the Internet. The *Softbot*[5] provides a planning framework which consists of operators and a planner in the similar way to STRIPS planning. The environment and the operator are described by first-order predicate calculus. An operator represents an action by UNIX commands like `mkdir`, `find` and so on. *Softbot* has user interface for a naive user to describe operators without knowledge on predicate calculus. A complete and sound partial-ordering planner is used in *Softbot*. *Occam*[15] is also a planner for gathering information. It is more efficient and able to reason about the capabilities of different information sources.

Considering a command "send E-mail to Dr. Yamada in TiTech" is given to *Softbot* as a goal description. First it tries to reduce the ambiguity in the command. Concretely *Softbot* identifies a person named Dr. Yamada in TiTech. In this processing, knowledge about the organization of TiTech is used. Next *Softbot* finds applicable operators and generates a plan for achieving the given goal. In this example, the operators describing `finger` and `sendmail` are selected and added into a plan. When a complete plan is obtained, *Softbot* executes it and the goal is satisfied. *Sage*[14] is also developed for integrating planning, execution, replanning and sensing to gather information in distributed resources. The aim of these studies is to generate a plan as a procedure of gathering information, and a plan consists of UNIX commands or database operations.

[2] ARACHNID

ARACHNID[17] (or *InfoSpider*[18]) is an information gathering system which manipulates multiple search agents in the Web. ARACHNID applies an A-Life like approach for controlling the agents. Each agent has energy for describing its activity. If an agent finds a relevant Web page, its energy increases and otherwise teh energy decreases. When an agent gets fully active, it produces children, and if it significantly looses energy, it will be dead. In ARACHNID, since information gathering is done on-line, it can obtain more recent information than a search engine using off-line search.

[3] Meta search engine

While research on a single search engine is active, meta search engines have also been developed. A meta search engine utilizes other search engines for information search in the Web. First it sends a query inputted by a user to several search engines. Next it receives a hit lists from them. Finally it integrates the results and indicates them to a user. MetaClower[23) 24)] is a typical meta search engine. SavvySearch[8)] is able to characterize search engines using user's feedback.

[4] Web robot

Programs called *Spider*, *Clower*, *Web robot* are utilized to gather Web pages for construction of a search engine database[9) 3)]. When start points are given to a Web robot, it begins to trace the links from the current Web pages and gather pages using breadth-first search. In general, the purpose of a Web robot is to collect Web pages as many as possible, and it has no control mechanism for selective information gathering.

[5] Recommender and shopping agent

WebWatcher[10)] is a recommender agent which guides an user in Web browsing. In the similar way to a human guide in a museum, *WebWatcher* can infer and recommend the Web page candidates which a user seems to select next by using his/her browsing history. The knowledge for prediction of the Web pages to which a user goes next is acquired by machine learning methods. Reinforcement learning and instance-based learning are applied to obtain the knowledge. *Letizia*[16)] is also able to recommend to a user the next interesting Web pages.

Some learning systems have been developed for information gathering and browsing in the Web. *ShopBot*[4)] learns the text pattern indicating the price of CD-ROMs, and searches for the cheapest one more efficiently than a human. Fist *ShopBot* collects both of positive and negative training examples by actively investigating on-line shop Web sites. It tries to give different type of input to the site and the returned Web pages are dealt as training examples. Using them, *ShopBot* can utilize an inductive learning method and obtain a description of a vender site.

2.2 Open problems

As mentioned earlier, various AI applications to information gathering and searching in the Web have been developed thus far. Some systems successfully overcome the limitation of traditional information retrieval approach like TFIDF indexing, statistical information and so on. However we can point out open problems in the following.

- *Gathering pieces of knowledge*: In all of conventional systems, the gathered information is only a set of pieces of knowledge. For example, any search engine collects Web pages which mention only surface knowledge on a query. We need more systematic and deep knowledge on a target information described in a query.
- *Appropriate combination of keywords*: We need frequent patterns or associative rules to determine much more suitable combination of keywords in a query. For instance, when we describe a query, we usually consider the distance of keywords in the meaning networks, and evaluate the coverage of expressions.

In the following sections, we introduce two approaches to address the above limitations. Navigation planning and a Mondou search engine are described.

§3 Navigation planning

The Web is very useful for a user who wants to understand a *target concept*. He/she can browse helpful Web pages to understand a target concept. However, in general, this task is very hard because he/she may not know where such Web pages are, and has to search for them over the vast Web search space. A solution to this problem is to use a search engine with the target concept as a query. However, since the retrieved Web pages are not filtered sufficiently, a user has to select useful ones from them. Furthermore, since in most cases the gathered Web pages are pieces of surface knowledge which directly explain the target concept and include concepts that a user does not understand, he/she must search the useful Web pages for them using a search engine again. This task is repeated until a user understands the target concept, and hence wastes time. We consider the task as planning, and propose *navigation planning*[26)][25)] to automatically generate a sequence of Web pages which can guide a user to understand a target concept deeply.

3.1 Web browsing as planning

In this research, *navigation* means a task that indicates a sequence of useful Web pages to a user for guiding his/her concept understanding. The sequence of Web pages is called a *plan*, and *navigation planning* means the automatic generation of the plan. We can summarize the Web browsing task for concept understanding in the following. This procedure is iterated until terminated by the user.

1. Search Web pages using a search engine.
2. Read and understand the pages retrieved by the search engine.
3. Select unknown concepts in the Web pages.
4. Go to *Step*1 with unknown concepts as target concepts.

The procedure above is considered *planning*[7)] under the following correspondence. Using this formalization, we can apply a classical planning framework to navigation planning.

- *Action*: Understanding concepts in a Web page.
- *State*: A user's knowledge state described with a set of words describing concepts which he/she knows.
- *Initial state*: A user's initial knowledge state.
- *Goal state*: A target concept described with a set of words which a user wants to understand.
- *Operator*: *U-Op(URL)* defined by the followings.
  - *Label*: URL of the Web page
  - *Condition*: $C = \{c_1, \cdots, c_i\}$, where C means the *condition words* which are necessary to understand the pages.
  - *Effect*: $E = \{e_1, \cdots, e_j\}$, where E is *effect words* which a user obtains by understanding the page.

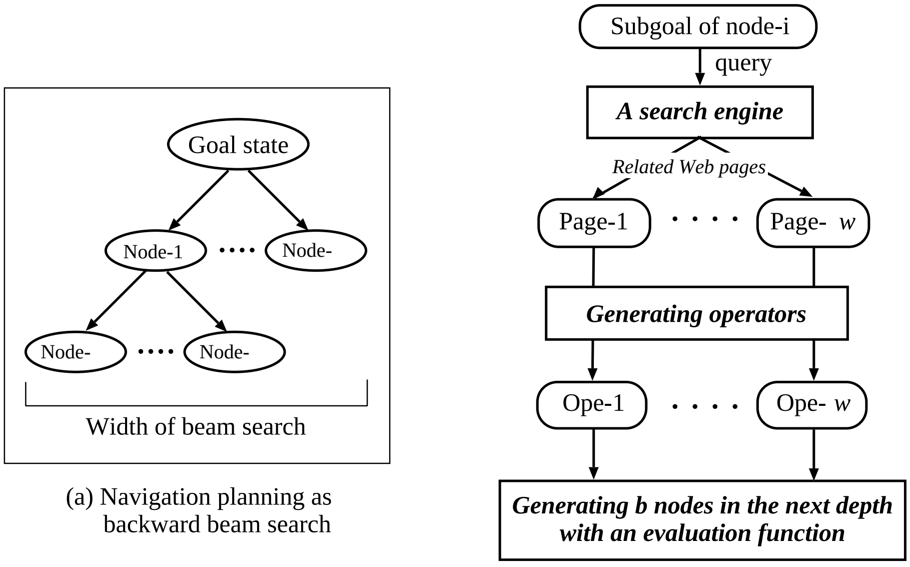
This navigation planning contains a significant problem which has not been presented in planning. It is that the *U-Op(URL)* operators are not given in advance. This is because it is impossible for a human designer to generate the operators from all the Web pages in the Web. Hence the operators need to be automatically generated from Web pages when they are necessary.

3.2 Generating operators from Web pages

[1] Using tag structure in a HTML file

Various methods to extract keywords from text have been studied[22)]. Though most methods are based on the frequency of words, one of the most effective methods is to utilize the structure in text. Since a Web page is described in a HTML format, we can utilize tag structures.

The prime candidates for condition words are those linked to other Web pages, i.e. the words between `<A HREF=URL>` and `</A>`, because this tag is a



(a) Navigation planning as backward beam search

(b) Expansion of a node-i

Fig. 1 Navigation planning.

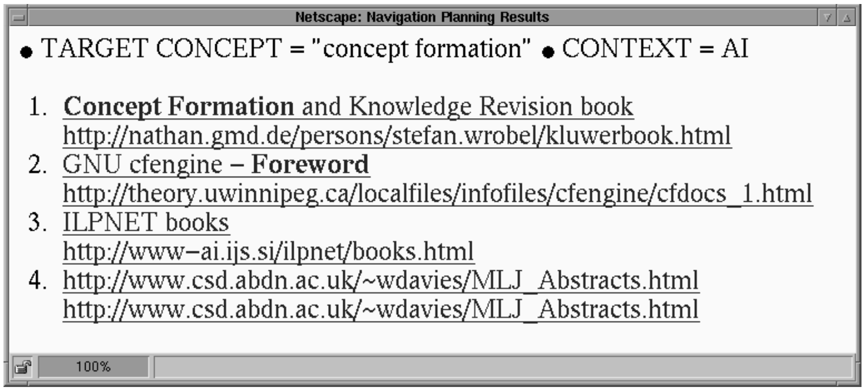
sign of reference to relevant topics, which are important for understanding the current Web page in many cases.

Since the title of the Web page describes words which a user may acquire by reading the page, the words between `<TITLE>` and `</TITLE>` are candidates for the effect words. In the same way, headings describe knowledge which a user may obtain by reading the section. Thus the words between `<Hn>` and `</Hn>` are also candidates for the effect words.

[2] ***KeyGraph*: A keyword extraction method**

The extraction of condition and effect words using the tag structure is not sufficient. All the linked words are not necessarily candidates for condition words, and all the condition words are not necessarily linked. Thus we look for another method to assist it, and *KeyGraph* is found to be appropriate for this task.

KeyGraph is a fast extraction method of keywords representing the asserted core idea in a document[19]. *KeyGraph* composes clusters of terms, based on co-occurrences between terms in the document. Each cluster represents a concept on which the document is based (i.e. condition words), and terms con-



● TARGET CONCEPT = "concept formation" ● CONTEXT = AI

1. **Concept Formation** and Knowledge Revision book
   http://nathan.gmd.de/persons/stefan.wrobel/kluwerbook.html
2. GNU cfengine **– Foreword**
   http://theory.uwinnipeg.ca/localfiles/infofiles/cfengine/cfdocs_1.html
3. ILPNET books
   http://www–ai.ijs.si/ilpnet/books.html
4. http://www.csd.abdn.ac.uk/~wdavies/MLJ_Abstracts.html
   http://www.csd.abdn.ac.uk/~wdavies/MLJ_Abstracts.html

Fig. 2 Plan for “Concept formation”.

necting clusters tightly are obtained as author's assertion (i.e. effect words). Furthermore the likelihood for condition and effect words can be computed by *KeyGraph*, and used for weight of an operator. Another merit of *KeyGraph* is that it does not employ a corpus.

The extraction of condition and effect words using tag structure and *KeyGraph* is integrated. First condition and effect words are obtained with the tag structure. Next *KeyGraph* also extracts condition and effect words and both of them are integrated.

3.3 Planning procedure

We have developed a navigation planning procedure. Fig.1 shows the overview of our navigation planning system, called *NaviPlan*. *NaviPlan* automatically generates a plan for given target concepts. It uses *backward beam search*[21] from a goal state (Fig.1(a)). The expansion of a node (Fig.1(b)) includes the search for relevant Web pages with a search engine and the generation of operators.

We fully implemented *NaviPlan* using Perl. Fig.2 shows a plan (depth = 4) generated by *NaviPlan* with the target concept “concept formation”. Four useful Web pages are indicated with hyperlinked titles and URLs. By reading the Web pages in this order, a user is able to understand “concept formation” deeply.

3.4 Experimental evaluations

Now let us evaluate *NaviPlan* according to the usage of the system as a navigation planner. In this experiment, we employed the search engine MetaCrawler, for expanding nodes in the planning procedure of *NaviPlan*. The system compared with *NaviPlan* here was also MetaCrawler.

For each of 10 subjects (8 graduate school students and 2 university staffs) , we investigated the number of Web pages he/she read for understanding a target concept and the brief explanation on the goal. As results, we found out *NaviPlan* achieved efficient navigations to the goal, due to the strong intension of operators to reach the goal from user's initial state.

In order to investigate the effect of having a plan instead of only one page to read, we also made experiments using more difficult target concepts for the same subjects to understand. Consequently we conclude that planning in *NaviPlan* contributes to increasing user's knowledge until he/she understands a goal concept, by reading multiple but small number of pages.

§4 Intelligent search system

In 1996, We have developped "Mondou" web search engine, which belongs to the first generation of research oriented Japanese web search engines [*1] [12]. We have applied the rapid emerging technologies of data mining to the mechanism of query modification and expansion. We also implemented Java applets in order to support the search steps. In this section, we explain major technical aspects of Mondou systems, such as data cleaning, data mining and information visualization.

4.1 Architecture of Mondou search engine

The Mondou consists of the three primary modules, such as *web robot*, *text database* and *query server*, which are shown in Fig. 3.

The fundamental functions of the *web robot* are described in Section 2.1 [4]. In addition to the basic mechanisms of the web robot, our web robot parses HTML tags and extracts the important keywords from the *title, headings and sub-headings, anchor strings*, and other tagged attributes. Moreover it gives appropriate weight values to derived keywords in order to measure the importance of web pages using the heuristic functions.

[*1] The URL of Mondou web search engine is `http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/index_e.html`. The name of Mondou comes from the term of Zen, which is a sophisticated discussion style in Japan, so the discovery is in the processes of questions and answers.

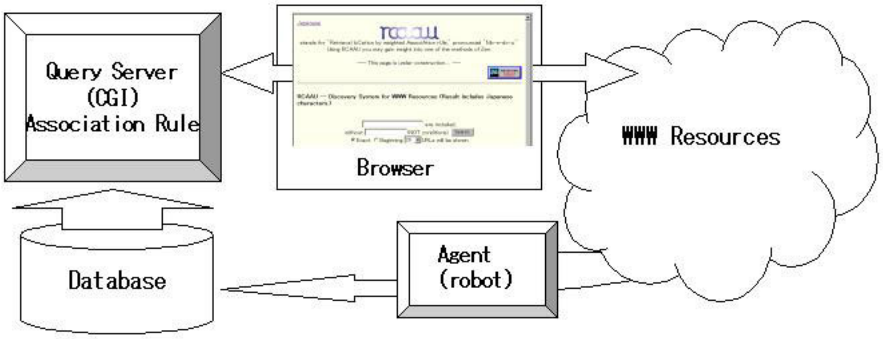


Fig. 3 The structure of Mondou.

Our *text database* handles not only the characters in the web pages, but also the connectivities of hyperlinks. Therefore, the primary database table consists of the following attributes, "*keywords, URLs, hyper links, http servers, IP addresses*", and the system also keeps the control/management data tables for operating the web robots.

The *query server* is the most important program in the Mondou system, we implement typical data mining algorithms to derive appropriate association rules. The associative keywords with some attribute values are visualized on the java applets. Furthermore, in order to adjust the demand of searchers, we tried to apply agent technologies and the integrating mechanism of heterogeneous databases to Mondou systems.

4.2 Data cleaning by web robots

As we mentioned in Section 2.1, the simple web robot visits web servers sequentially in the breadth-first manner. However, we have to consider the performance of gathering the web pages more effectively, then we try to collect meaningful or useful web pages as fast as possible.

It is so hard to discover the meaningful knowledge from a web graph, then we try to extract the interesting features of web pages using the following two types of hyperlinks.

- **Inside link:** the link to other pages on the same web server.
- **Outside link:** the link to other pages on the other web servers.

Actually, when we browse web pages, we consider the connectivities and distance of the web page from/to other web pages. Therefore, it may be helpful to analyze and characterize the connectivities of hyperlinks, in order to fast collect important web pages in the same and different web servers.

Therefore, we proposed a focused crawling algorithm in our paper[12], which is based on the assumption that the important web pages are referred many times from other servers. And it is also helpful to reduce the number of noisy/meaningless web pages from the view point of data cleaning.

Furthermore, in order to reduce the load of the network and the number of web page requests, we are trying to implement cooperative web robots based on a distributed scheduling algorithm[27].

4.3 Data mining for query modification

The research of data mining[6] consists of databases, statistics and various kinds of AI research, such as machine learning, inductive learning, knowledge representation, information visualization and others.

The data mining plays an important role in order to support the discovery procedures in the Mondou system. We improve the typical data mining algorithms of association rules[1], and apply them to the set of web pages.

Shortly speaking, given a set of web documents, where each document consists of various keywords, an association rule is an expression $X \Rightarrow Y$, where X and Y are sets of keywords. The intuitive meaning of such a rule is that documents which contain the keywords X tend to also contain the keywords Y. The support threshold value of the rule $X \Rightarrow Y$ is the percentage of documents that contain both X and Y.

Table 1 shows typical examples of derived keywords, the number of URLs and associative keywords by our algorithm [*2] .

Table 1 Examples of associative keywords.

keywords	No. of URLs	related keywords
Applications	975	windows, speech, computer, internet, helper, ···
Informatics	462	genome, medical, department, center, school, ···
Japan	20,366	japan country, webec, japan, japanese, back
Engine	2,836	lycos, yahoo, dragon, search, creative, japan
Mining	134	akita, geology, college, net, data
Mondou	28	rcaau, search

[*2] A derived keyword, “rcaau”, stands for ”Retrieval loCation by weighted AssociAtion rUle,” which appears in the web page of Mondou.

In order to refine these rules much more by Mondou, it is effective to classify several attributes and use the relationships of different attributes by using semi-structured data. Moreover, our Mondou system derives the different set of association rules from the different databases, such as netnews archives, electronic journals and other digital libralies. Finally our system deduces much more effective combination of keywords from heterogeneous databases by using the difference of derived rules. We also implement the agent oriented communication interface, different Mondou systems can communicate each other by high level query languages like KQML.

4.4 Information visualization

In order to visualize the characteristics of association rules and recognize the features of document clusters, we constructed interactive visual interfaces[13] as Java applets.

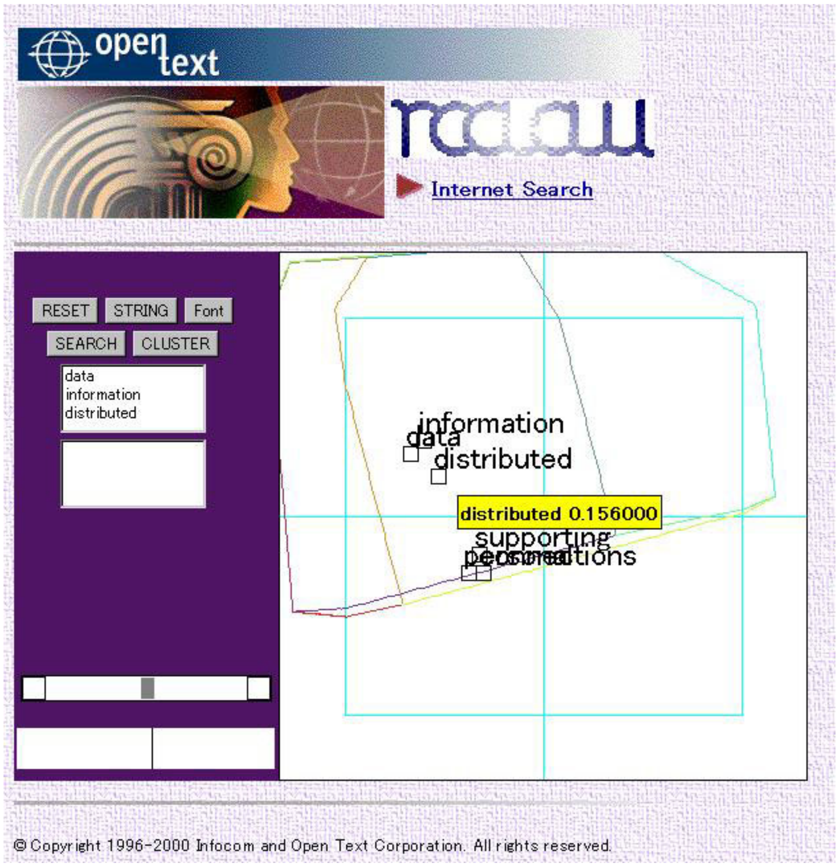
In Fig.4, this 3D graph shows the distributions of derived keywords based on ROC graphs[11, 20]. We also implemented fisheye graph drawing techniques[2]. By the simple mouse operations, users can determine a combination of keywords easily.

For example, in Fig.4, by a query with a keyword “mediator”, Mondou derives associative keywords, such as “information”, “data”, “distributed”, “supporting”, “connections”, “personal” and others. Fig.5 shows an example of visualizing results for an selected query using some of above derived keywords. In the figure, numerical numbers mean the length of queries, and the color of a keyword changes by pointing the number. In the right window, this applet displays the search results which satisfy the combination of black keywords.

Consequently, users can easily select the interesting combination of keywords by data mining and information visualization, which are provided by our Mondou system.

§5 Conclusion

We surveyed the conventional systems of information gathering and searching in the Web. Information gathering systems, *Softbot*, *Sage*, ARACHNID and so on were introduced. Though some of them have successfully overcome the problems which traditional IR systems hardly cope with, there are still open problems. We also pointed out two significant problems, the gathering pieces of knowledge and the appropriate combination of keywords in the fields.



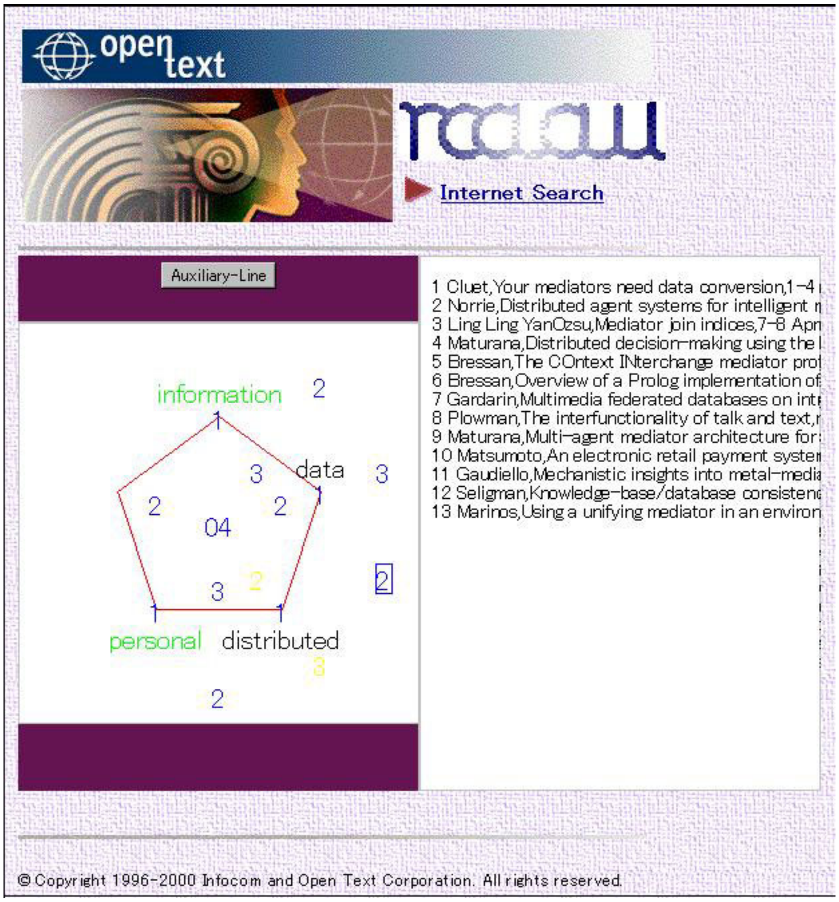
data
information
distributed

© Copyright 1996-2000 Infocom and Open Text Corporation. All rights reserved.

Fig. 4 Visual interface for keyword selection

As an attempt to introduce a promising approach to gather systematic information, not pieces of knowledge, we described navigation planning that generates a plan guiding user to understand a concept in the Web. It has the ability to generate operators during planning from Web pages using keyword extraction methods. The search for useful Web pages for a user to understand goal concepts was formalized using an AI planning framework, and an operator corresponding to the understanding of a Web page was defined with condition and effect knowledge. Then we described the whole planning procedure, and verified the effectiveness experimentally.

In a field of a Web search engine, many researchers are trying to automatically generate the appropriate combination of keywords in a query. A Mondou web search engine was introduced as one of web search engines in the



1 Cluet,Your mediators need data conversion,1−4
2 Norrie,Distributed agent systems for intelligent
3 Ling Ling YanOzsu,Mediator join indices,7−8 Apr
4 Maturana,Distributed decision−making using the
5 Bressan,The COntext INterchange mediator pro
6 Bressan,Overview of a Prolog implementation of
7 Gardarin,Multimedia federated databases on int
8 Plowman,The interfunctionality of talk and text,
9 Maturana,Multi−agent mediator architecture for
10 Matsumoto,An electronic retail payment syste
11 Gaudiello,Mechanistic insights into metal−medi
12 Seligman,Knowledge−base/database consisten
13 Marinos,Using a unifying mediator in an environ

© Copyright 1996-2000 Infocom and Open Text Corporation. All rights reserved.

Fig. 5 Visual interface of document clustering

first generation, which was based on the emerging technologies of data mining, text mining, web mining and information visualization. In particular, we focused on the query modification, which is a typical and important problem in the research of information retrieval. Our implemented functions, the derivation and visualization of associative keywords in order to modify the initial query, are empirically found very effective for novice users.

References

1) R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago, Chile, 1994.

2) C. Chaomei. *Information Visualisation and Virtual Environments.* Springer-Verlag, 1999.

3) F.C. Cheong. *Internet Agents: Spiders, Wanderers, Brokers, and Bots.* New Riders, 1996.

4) R. B. Doorenbos, O. Etzioni, and D. S. Weld. A scalable comparison-shopping agent for the World-Wide Web. In *Proceedings of the First International Conference on Autonomous Agent*, pages 39–48, 1997.

5) O. Etzioni and D. Weld. A SoftBot-based interface to the Internet. *Communication of the ACM*, 37(7):72–76, 1994.

6) R. Feldman. Practical text mining. In *Second Symposium on Principles of Data Mining and Knwoledge Discovery (PKDD-97)*, Nantes, France, 1998.

7) R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208, 1971.

8) A. E. Howe and D. Dreilinger. Savvy search: a metasearch engine that learns which search engines to query. *AI Magazine*, 18(2):19–25, 1997.

9) K. Jamsa, S. Lalani, and S. Weakley. *Web Programming.* Jamsa Press, 1996.

10) T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 770–775, 1997.

11) M. Kawahara and H. Kawano. Performance evaluation of bibliographic navigation system with association rules from roc convex hull method. *Transactions of the IPSJ:Database*, 40(SIG3(TOD1)):105–113, 1999.

12) H. Kawano. Mondou: Web search engine with textual data mining. In *Proc. of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 402–405, 1997.

13) H. Kawano and M. Kawahara. Mondou: Information navigator with visual interface. In *Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000*, pages 425–430, London, UK, Sep. 2000.

14) C. A. Knoblock. Planning, executing, sensing, and replanning for information gathering. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1686–1693, 1995.

15) C. T. Kwok and D. S. Weld. Planning to gather information. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 32–39, 1996.

16) H. Lieberman. Letizia: A agent that assists Web browsing. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 924–929, 1995.

17) F. Menczer. ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 227–235, 1997.

18) F. Menczer and A. E. Monge. Scalable web search by adaptive online agents: An inforspiders case study. In *Intelligent Information Agents*, pages 323–347. Springer, 1999.

19) Y. Ohsawa, N. E. Benson, and M. Yachida. *KeyGraph*: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of IEEE Advanced Digital Library Conference*, pages 12–18, 1998.

20) F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of 3rd Int'l Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48, 1997.

21) S. Russell and P. Norvig. *Artificial Intelligence –A Modern Approach–*. Prentice-Hall, 1995.

22) G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In K. S. Jones and P. Willet, editors, *Readings in Information Retrieval edt.), Morgan Kaufmann*, pages 323–328. Morgan Kaufmann, 1997.

23) E. Selberg and O. Etzioni. Multi-service search and comparison using the metacrawler. In *the 1995 World Wide Web Conference*, 1995.

24) E. Selberg and O. Etzioni. The metacrawler architecture for resource aggregation on the Web. In *IEEE Expert*, volume January-February, pages 11–14. IEEE, 1997.

25) S. Yamada and Y. Ohsawa. Navigation planning to guide concept understanding in the world wide web. In *Proceedings of the Fourth International Conference on Autonomous Agent*, pages 114–115, 2000.

26) S. Yamada and Y. Osawa. Planning to guide concept understanding in the www. In *AAAI 1998 Workshop on AI and Information Integration*, pages 121–126, 1998.

27) H. Yamana, K. Tamura, H. Kawano, S. Kamei, M. Harada, H. Nishimura, I. Asai, H. Kusumoto, Y. Shinoda, and Y. Muraoka. Expreriments of collecting www information using distributed www robots. In *Proc. of SIGIR'98*, pages 379–380, Melbourne, Australia, 1998.

28) R. B. Yates and B. R. Neto. *Modern Information Retrieval.* Addison Wesley, 1999.