Mondou: Interface with Text Data Mining for Web Search Engine

Hiroyuki Kawano
Department of Applied Systems Science, Kyoto University
kawano@kuamp.kyoto-u.ac.jp

Toshiharu Hasegawa
Department of Applied Systems Science, Kyoto University
hasegawa@kuamp.kyoto-u.ac.jp

Abstract: *In order to submit queries to web search engines, we have to carefully choose the suitable combination of keywords. Without rich background knowledge about keywords in web documents, it is too difficult to find out invaluable URLs by search engines. In this paper, with applying techniques of text data mining to the web resource discovery, we try to derive associative keywords by extended association algorithm. We explain our developed interface of web resource discovery system using association rules, which are derived from the cluster of Japanese HTML pages in the text database. This paper also includes a brief discussion of evaluation of "Mondou" system and java applet in order to visualize search results with multi-dimensional measurements.*

1 Introduction

Recently, in the internet, digital documents, programs, sounds, images, movies and various other types of resources organize the huge information space. Especially, it is very difficult to find out the adequate URLs including invaluable information and to discover the relationship among world wide web sites[3].

Lycos, AltaVista and many other search engines in Table 1 have been developed and they provide us informative URLs by search queries with boolean expressions[13, 5].

However, without rich background knowledge about the web space, it is too hard to judge whether interesting documents really include the combination of a few keywords, especially in the case that those keywords may have different meanings in other domains. Therefore, in order to describe adequate queries, it is important to grasp the suitable combination or association of keywords, which exist in the database.

At present, we have been developing the search engine with applying techniques of text data mining to the web resource discovery[7]. The algorithms of data mining or KDD (Knowledge Discovery in Databases)[4, 2, 6] derive some constraints, rules or patterns in various types of databases. Our developing search engine provides association rules by the techniques of text data mining, and it is possible to modify initial submitted query with boolean expressions including a few keywords.

In this paper, we explain the ability of our developing search system, which is named as *Mondou*, with applying our proposed algorithm for weighted association rules to huge number of web documents. Moreover, since collecting documents are written in Japanese, we also use Japanese morphological analysis and other heuristic operations to derive Japanese keywords.

RCAAU stands for "retrieval location by weighted association rule" in the digital "monde", and pronounced "Mo-n-do-u." Using associatve keywords provided by Mondou, we may gain insight into one of the methods of Zen. The URL[1] is also listed in the Japanese net search page of Netscape Navigator.

Our Mondou executed more than one million queries required by anonymous users from February in 1996. We analyze the part of httpd log, which has been recorded by the apache server, and evaluate the effectiveness of our proposed search strategies.

2 Structure of web space

In this section, we characterize the structure of web space, in order to discover important clusters of web documents. For example, when we have to check a few web documents, it is very useful to get the relation between the access page and other popular web pages. Actually, we often make the list of cool web pages based on the voting of internet users. Then, we analyze web space as a typical structure of loosely connected hyper graph in Figure 1, and we try to evaluate the features of the interesting web pages by following two types of hyper links.

[1] `http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/index_e.html`

1060-3425/98 $10.00 © 1998 IEEE
Proc. 31st Annual Hawaii International Conference on System Sciences

Table 1: List of popular search engines.

Search Engine (URL)	Types	Features
WebCrawler (http://webcrawler.com/)	Centralized	Full text database and weighted keywords are used.
Lycos (http://www.lycos.com/)	Centralized	Titles, headings and other important attributes are mainly stored.
Harvest (http://harvest.transarc.com/)	Distributed	The structure of search engine is based on distributed broker database.
InfoSeek (http://www.infoseek.com/)	Centralized	Robot program collects documents for this commercial search engine.
AltaVista (http://altavista.digital.com/)	Centralized	Advertising search engine for the computing power.
goo (http://www.goo.ne.jp/)	Centralized	Technologeies of NOW and morphological analysis for Japanese documents.

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

We estimate the hyper links in the web by the values of (s_P, i_P, o_P), which is the combination of the document size, the number of inside links, and the number of outside links for the parent document P in Figure 2. Then, each child document C_k is referred by parent P, we can also describe (s_k, i_k, o_k) for C_k.

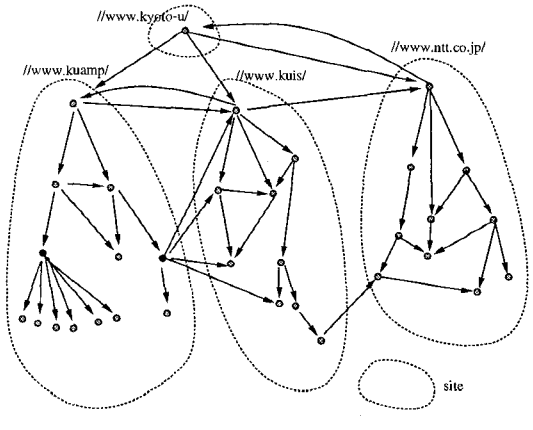

Figure 1: Hyper graph presentation for Web space.

In order to evaluate the quality of contents C_k, we define the following cost function p_k as a reference score.

$$p_k = (s_k - i_k \cdot W_i - o_k \cdot W_o)/s_k. \quad (1)$$

In the above equation 1, we must define W_i and W_o, as the weighted values for inside and outside links respectively. Thus, we can caliculate p_k, as the approximate reference score based on the several attribute values with web documents.

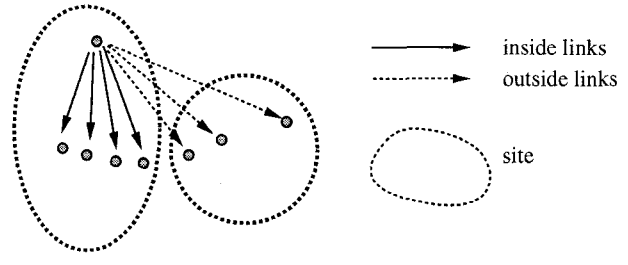

Figure 2: Features of hyper links, inside/outside links.

Figure 3 shows typical graphs of average reference score p_k, that are evaluated for 6,621 parent documents including 18,397 links in jp-domain with $W_i(= 10)$ and $W_o(= 30)$.

At first, from these simple graphs, we want to extract important clusters of documents from web space. However, it is too hard to discover the relation between the cluster of interesting web pages C_k and attribute values (s_P, i_P, o_P), document size, the number of inside links, or the number of outside links.

Therefore, in this paper, we regard the web space as a simple text database without document clusters, which doesn't have any integrated data model. In the following section, we try to derive other rules from the huge amount of text data stored in the internet by several kinds of algorithms for text data mining.

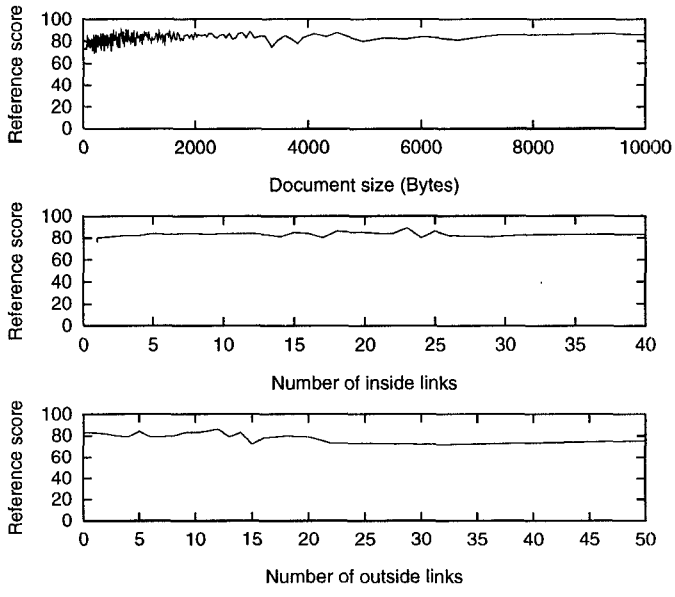

Figure 3: Reference score vs. size, inside links, outside links.

3 Text data mining in the web space

In recent years data mining is noteworthy field to be studied based on various kinds of researches, such as machine learning, inductive learning, search and knowledge representation, with considering characteristic features of database[12, 4, 6]. Many mining algorithms, such as *tuple-oriented algorithm* with dependency between tuples or *attribute-oriented induction algorithm* which generalizes attributes with taxonomy information, provide us useful knowledge as rules, constraints, regularities and others.

In this section, we try to extend the algorithm to derive association rule[11, 1], which is one of tuple-oriented algorithms.

3.1 Association rule

We have a brief introduction to mining association rule.

For example, the rule of "program ⇒ database" from the sets of { program } and { program, database } will be found in databases by the following typical threshold values.

Create all the rules whose *confidence* is equal to or greater than *minconf*, which is the threshold value to determine the confidence-rated rules. If { program } and { program, database } are found from the set of keywords, we create the rule "program ⇒ database," where *support* of the rule is the value of { program, database }, and *confidence* of the rule is given by dividing the *support* of former by the later support value.

3.2 Weighted association rule

We extend the previous mining algorithm to handle weighted keywords for markup language, especially for tags in HTML.

Definition 1 *In the case of retrieving keyword k_j, if k_j appears w_{ij} times in tuple T_i, we say that k_j has the weight of w_{ij}, and define $a_{ij} = (k_j, w_{ij})$.* □

Next, we select tuples $T_i (i \in I)$ including set of keywords $K = \{k_j | j \in J\}$ from all the tuple set $\mathcal{T}$. Given T_i including $a_{ij} = (k_j, w_{ij})$, we have to evaluate the support $sup(K)$. $sup(K)$ can be defined by the following equations, where K includes any combination of keywords from all the keywords $\mathcal{K}$ related with retrieving.

$$N_0 = \sum_{\mathcal{T}} \max_{\mathcal{K}} w_{ij},$$

$$N(K) = \sum_{i \in I} \min_{j \in J} w_{ij},$$

$$sup(K) = \frac{N(K)}{N_0}.$$

Besides, the next inequality shows that *support* is appropriately defined for set of keywords K_1 and K_2 that satisfy $K_1 \subset K_2$.

$$sup(K_1) - sup(K_2) = \frac{N(K_1) - N(K_2)}{N_0} \geq 0$$

Table 2: URL and list of (keyword, weight value).

URL	keyword-set
URL_1	{ (management, 3), (technology, 2), (economics, 1) }
URL_2	{ (data, 4), (management, 2), (program, 2) }
URL_3	{ (route, 2), (management, 2), (program, 1) }
URL_4	{ (sort, 3), (program, 2), (file, 1) }
URL_5	{ (hash, 2), (program, 1), (file, 1) }
...	...

Table 3: Results including "management".

URL_1 { (management, 3), (technology, 2), (economics, 1) },
URL_2 { (data, 4), (management, 2), (program, 2) },
URL_3 { (route, 2), (management, 2), (program, 1) },
....

Example of weighted association rules:

We will show several tables in the steps of mining rules by applying our proposed algorithm.

The pair of URL and weighted keywords are stored in the database in Table 2. Let's assume that we try to search tuples including keyword "management" and more than two other words. The result is shown in Table 3.

And next step, we can execute mining algorithm to derive association rule. By applying our proposed algorithm with *minsup* 1/5 in section 3.1, all *frequent* keyword-sets are shown as Table 4.

Table 4: Frequent keyword sets.

keywords	*support*
{management}	7/9
{data}	4/9
{program}	3/9
{route}	2/9
{management, data}	3/9
{management, program}	3/9
{management, technology}	2/9
{management, route}	2/9
{data, program}	2/9
{management, data, program}	2/9

Then, we have to focus on the tuples including specific keywords, and we can remove *verbose* keyword-sets without "*management*", and we derive set in Table 5.

As a result, the association rule tree, for example, with *minconf* 1/4 can be calculated as Table 6. Consequently, we will get the associative keywords of "program," "technology," "route" and "data" in order to modify keyword in initial submitted query.

Table 5: *Verbose*-removed keyword sets.

keywords	*support*
{management}	7/9
{management, data}	3/9
{management, program}	3/9
{management, technology}	2/9
{management, route}	2/9
{management, data, program}	2/9

Table 6: Example of association rules.

rule	*conf*	*support*
management⇒program	3/7	3/9
management⇒technology	2/7	3/9
management⇒route	2/7	2/9
management⇒{data, program}	2/7	2/9
{management, data}⇒ program	2/3	2/9
{management, program}⇒ data	2/3	2/9

In the following section, we propose the search system to use derived rules as associative keywords. Especially, we make it possible to retrieve web documents more sophisticatedly by using feedback of associative keywords. Actually, many internet users can easily focus on interesting web resources without taxonomy, or intelligent data dictionary given by database administrators.

4 Mondou: web search engine with mining algorithm

4.1 Structure of Mondou

Our Mondou system consists of the following three main modules, **agent**, **database**, **query server**, which are shown in Figure 4.

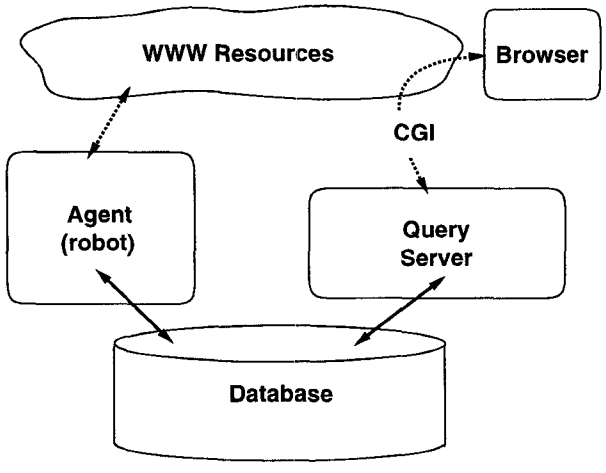

Figure 4: The structure of Mondou system.

The first module is often called as the robot program, spider or *agent*[9], and this program collects web pages in the internet and store them into the text database. In addition to the normal function of the robot, our intelligent agent parses collected documents by several methods including natural language processing[10] for Japanese documents. Moreover, in order to collect more interesting documents, our agent often visits to special URLs as interesting web pages, if they are referred many times from other web pages.

The *database* stores data not only about keywords, but also the number of links from other URLs. We have to prepare the several tables for several different attributes, keywords, URLs, hyper links, http servers, stop words, IP addresses and control/management data for Mondou system.

The *query server* is the search program, which is executed by CGI (Common Gate Interface), and it provides search results and mining association rules to the users. The mining algorithm is described in the previous subsection 3.2.

The input web page of Mondou system is shown in Figure 6, it is possible to enter any suitable combination of search keywords using AND, NOT boolean expressions in each empty box.

For examle, when he/she submitted initial keyword *knowledge*, Mondou provided him/her several rules as associative keywords, "engineering, systems, "knowledge" (in Japanese), acquisition" and so on, which is shown in Figure 7. Consequently, even by applying our proposed algorithm without taxonomies, conceptual trees or ontologies, he/she can grasp the association or relation among important keywords in those interesting web pages.

4.2 Extension of Mondou in heterogeneous databases

At present, we try to extend the implementation of Mondou as a full text search system, and also to derive much more effective rules from heterogeneous databases.

Firstly, we use full text retrieval systems as basic database systems in Figure 5. We combine the programs of Mondou system and the interface program, which is the agents for the full text database, and try to provide typical communication interface using KQML language.

Secondly, we implement the CGI programs and the mining program to derive different set of association rules from heterogeneous databases.

Using the search results and the association rules from heterogeneous databases, improved Mondou system can provide more effective keywords to users. In Figure 9, after selecting of electrical news "fj.rec.autos" as one of databases, we enter the keyword "oil" (in Japanese) into improved Mondou system. In the results, improved Mondou system provide us more interesting associative keywords which are tightly related to the domain of "autos".

4.3 Evaluation of Mondou

We are operating Mondou system in the internet, and we examine the quality of many search queries, patterns of combination of keywords and other features. Table 7 shows typical examples of derived keywords, the number of URLs and associative keywords derived by mining algorithm.

By using associative keywords, we can easily get interesting combination of keywords that can be treated as several meanings in web documents. It is helpful for many users to grasp the structure of web documents by associative keywords.

From February to October in 1996, Mondou executed 931,537 queries submitted from the users. Surprisingly, there are only 20,734 queries (2.23%) with NOT expression, it is too difficult for most of users to describe the query with adequate NOT keywords.

Table 7: Example of associative keywords.

keywords	number of URLs	associative keywords
analysis	2,027	fujita, numerical, behavior, method, infomration, top, plan, multidimensional, applied
applied	1,244	mathematics, mechanics, geochemistry, analysis, physics, media, superconductivity, optics, geology
engine	964	honda, search, stirling, dragon, "engine" (in Japanese), behavior, similarities, parts, serch
simulation	661	software, "simulation" (in Japanese), computer results, numerical, sciences, conference, advanced, gam
travel	1,393	"travel" (in Japanese), "trip" (in Japanese), guide, asia, tourism, resources, hokkaido, access
guide	5,808	infoseek, nagoya,"guide" (in Japanese katakana), yahoo, map, user, library, "guide" (in Japanese), event
travel guide	76	center, tokyo, japan, singapore, thailand, internet, usa

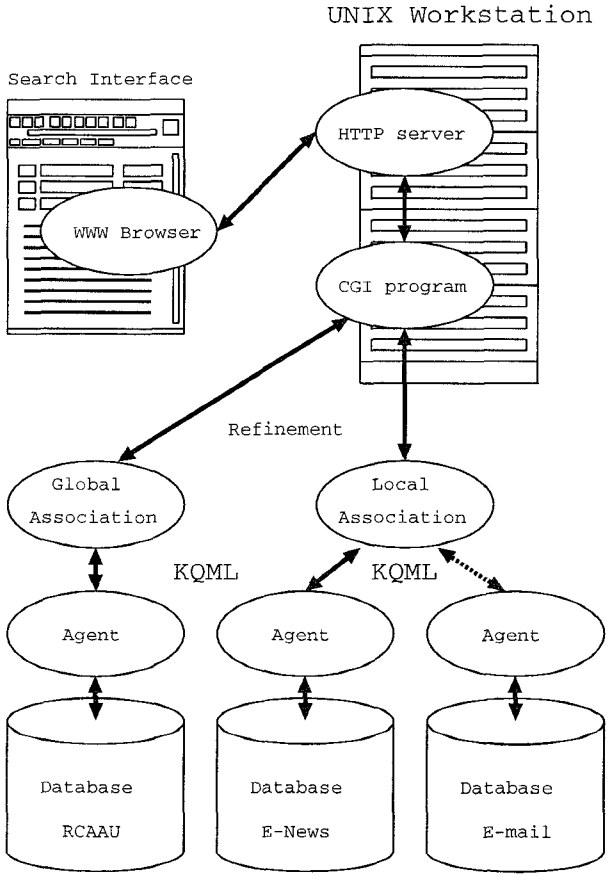

Figure 5: Structure of search engine in heterogeneous environment.

Moreover, the number of unique query patterns is 338,535. Most of users used only one keyword in the query, 51,510 patterns (15.2%) are described by one word. On the other hand, 287,025 (84.8%) patterns are described using combination of more than two keywords. Generally speaking, it is too difficult to describe various combination of keywords.

Furthermore, keywords including all executed queries covers 129,445 words (40.5%) for all 319,426 keywords, which are stored in the database. Therefore, according to the statistics of Mondou system, most of users seem to submit modified query using the derived associative keywords.

As a result, Mondou provides the rich combination of keywords in order to modify initial submitted query. Even if users don't know well about web documents to be find out, it should be possible to reach the interesting URLs. By using text data mining algorithm in Mondou, web users can get much more information about the relationship of keywords in the natural way.

4.4 Visual interface for Mondou

It is necessary to improve the web browsing interface by visualization, especially for the results of search engine we often want to check the accesibility of the documents and the status of networks from our site. Furthermore, it is easy to discover the interesting association rules by the drawing of graphs. Therefore, in order to visualize several attibute values and rules we have been developing an interactive search interface in Java language.

In Figure 8, in order to display the search results on

the browser, we use several drawing functions, *shapes, colors, interval of blinking time, arrows* and other attributes. Especially, we have to pay attention to the access cost to the URL from users and the relevance score of documents satisfying search query. By the implementation of present Java programs, each URL is shown with several attributes, the access cost, the score of documents, color and hyper links. Figure 8 shows one of visual examples for search results with the keyword "applied".

5 Conclusions and future works

The volume of text data in the web is increasing exponentially, it makes difficult to find the useful or suitable URLs providing rich contents. Several search engines in the web makes it possible to retrieve web documents by usual text database. However, users may not judge easily whether the documents have useful information, especially in the case that given keywords have wide concept.

In this paper, in order to retrieve efficiently web documents by suitable queries, we applied the algorithm of mining association rules, which is extended to handle weighted keywords in HTML, to the web document collected by robots. As a result, users can focus on the URLs appropriately by applying our rules without depending conceptual tree, or meta knowledge. We could confirm that our proposed algorithm works very effectively in searching text data in the web. We also developed the visual interface in Java for efficient searching on web browsers.

Still now, we have been using centralized systems for Mondou, but they have to be distributed in order to keep much more URLs and to focus on URLs more effectively since web grows very rapidly.

Acknowledgment

This work was supported in part by a Grant in Aid for Science Research (08244103) from the Ministry of Education, Science, and Culture of Japan. And the part of this work was also supported by the educational grant from Mitsubishi Electric Corporation and Sharp Corporation.

References

[1] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp.487–489, 1994.

[2] M.-S. Chen, J. Han and P. S. Yu, "Data Mining: An Overview from a Database Perspective," IEEE Trans. on Knowledge and Data Engineering, Vol.8, No.6, pp.866–883, 1996.

[3] O. Etzioni, "The World-Wide Web: Quagmire or Gold Mine?," Communications of the ACM, Vol.39, No.11, pp. 65–68, Nov 1996.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," AAAI/MIT Press, 1996.

[5] T. Honkela, S. Kaski, K. Lagus and T. Kohonen, "Newsgroup Exploration with WEBSOM Method and Brousing Interface," Technical Report A32, Laboratory of Computer and Information Science, Helsinki University of Thechnology, 1996.

[6] H. Kawano, S. Nishio, J. Han and T. Hasegawa, "How Does Knowledge Discovery Cooperate with Active Database Techniques in Controlling Dynamic Environment?," Proc. 5th International Conference on DEXA, Athens, Greece, pp.370–379, 1994.

[7] H. Kawano and T. Hasegawa, "Textual Data Mining for Intelligent Search Engine in WWW information space," Advanced Database Symposium '96, Tokyo, pp.27-34, 1996. (In Japanese)

[8] D. A. Keim and H.-P. Kriegel, "Visualization Techniques for Mining Large Databases: A Comparison," IEEE Trans. on Knowledge and Data Engineering, Vol.8, No.6, pp.923–938, 1996.

[9] M. Koster, "Guidelines for Robot Writers," http://info.webcrawler.com/mak/projects/robots/guidelines.html.

[10] Y. Matsumoto, S. Kurohashi, T. Utsuro, Y. Myoki and M. Nagao, "Japanese Morphological Analysis System JUMAN Manual, version 1.0," Nara Institute of Science and Technology, 1993.

[11] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. of the 21st VLDB, U. Dayal, P. M. D. Gray and S. Nishio (Eds.), Zurich, Switzerland, pp.407–419, 1995.

[12] M. Stonebraker, R. Agrawal, U. Dayal, E. Neuhold and A. Reuter, "DBMS research at a crossroads: The Vienna update," Proc. of the 19th International Conference on Very Large Data Bases, pp. 688–692, Dublin, Ireland, Aug. 1993.

[13] O. R. Zaiane and J. Han, "Resource and Knowledge Discovery in Global Information Systems: A Preliminary Design and Experiment," Proc. 1st International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, pp.331–336, 1995.

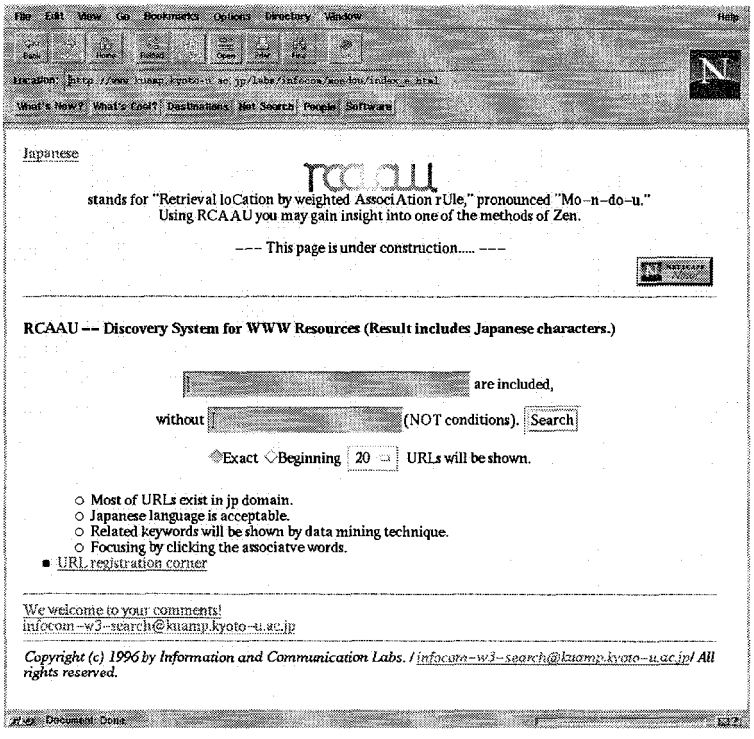

Figure 6: Input web page for Mondou system.

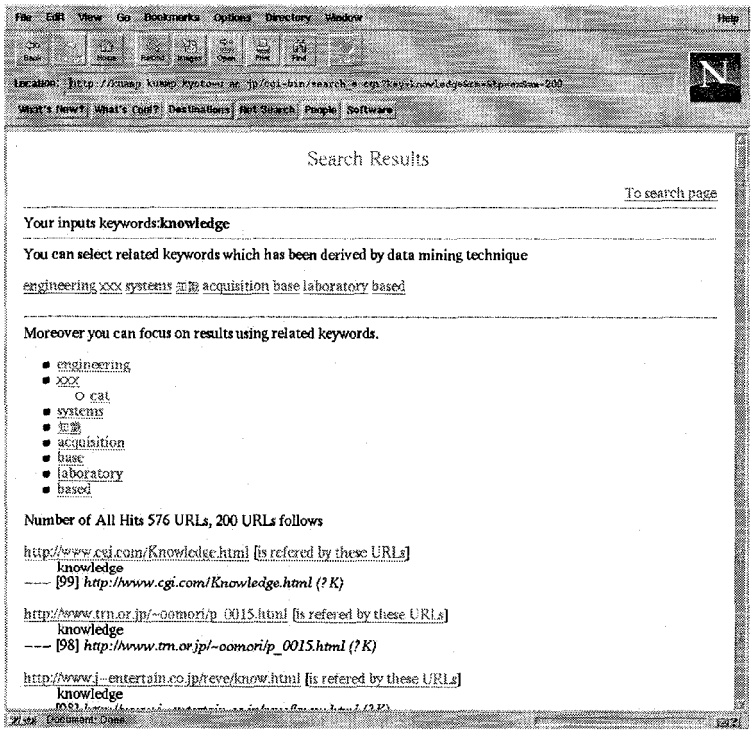

Figure 7: Output results from Mondou system.

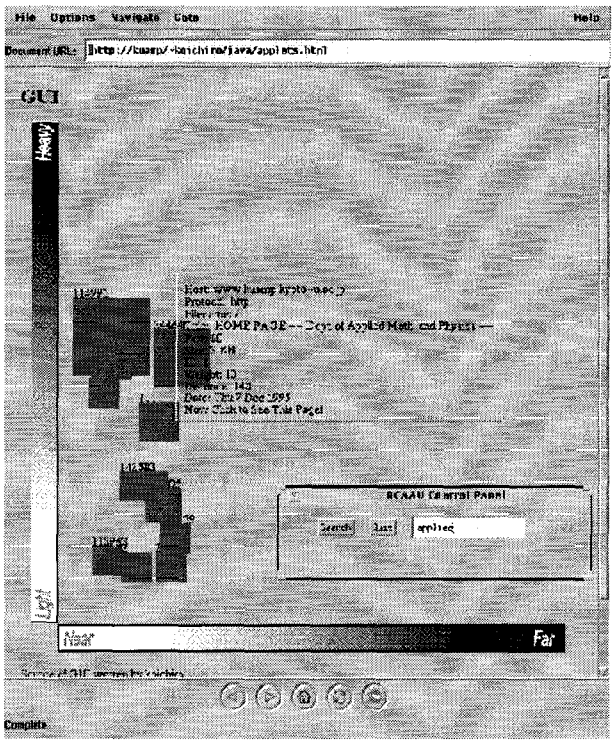

Figure 8: Visual interface for Mondou system.

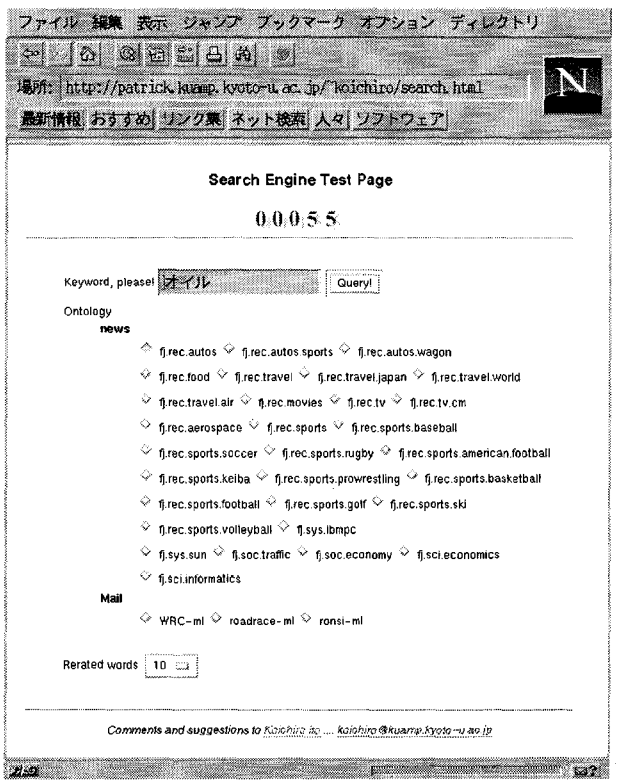

Figure 9: Prototype system in heterogeneous environment.