Mondou: Web search engine with textual data mining

Hiroyuki Kawano
Kyoto University

ABSTRACT: It is too difficult to discover useful documents on web without rich background knowledge. In this paper, with applying techniques of the textual data mining to the web resource discovery, we try to derive effective association rules in order to submit more effective queries. From 1995, we have been developing the resource discovery system, Mondou, which derives associative keywords from collected and parsed Japanese web pages. This paper includes a brief evaluation of our system and java applet in order to visualize search results with multi-dimensional measurements.

1 Introduction

The volume of web documents is increasing exponentially, it makes difficult to find cool URLs for various users. Especially, increasing of web documents makes it difficult to find out the adequate URLs including invaluable information, and to discover the relationship among world wide web sites[3].

Many search engines, Lycos, AltaVista and many others in the web, make it possible to retrieve web documents by the technology of textual database as a legacy system using typical boolean expressions. However, without rich background knowledge about the relations or structures of keywords, it is impossible to judge whether interesting documents really include the combination of several keywords. Therefore, in order to describe effective queries, it is important to grasp the suitable combination or descriptions of keywords, which exist in the web as textual database.

We develop the search engine with applying techniques of textual data mining to the web resource discovery[7]. Our developed search engine provides the search results including associative keywords, users are easily able to modify initial query with boolean expressions of keywords.

In this paper, we explain the ability of our developing search system, which is named as Mondou, with applying our proposed mining algorithm to the collected Japanese HTML documents. RCAAU means the "retrieval location by weighted association rule" in the digital "monde", and R-C-A-A-U also spells Mondou. Using RCAAU, we may gain insight into one of the methods of Zen. The URL[1] is also referred from many search pages in Japan.

Our Mondou executed more than one million queries required by anonymous users from February in 1996. We analyze the part of log, and evaluate the effectiveness of our developed search strategies.

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

0-7803-3905-3/97/$10.00

2 Approach to web mining

2.1 Structure of web space

We consider web space as a typical hyper graph, we try to analyze the structure of web space by following two types of links.

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

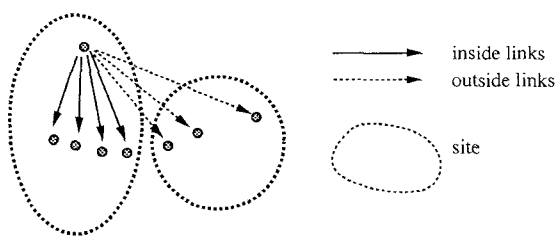

Figure 1: Inside links and outside links

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 1. Then, each child documents C_k are referred by parent P, we can also describe (s_k, i_k, o_k) for C_k.

We define the following cost function p_k in order to evaluate the quality of contents C_k.

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

We define W_i and W_o, as the weighted values for inside and outside links respectively. Thus, we could evaluate p_k as the approximate cost of the several attributes for web documents. Figure 2 shows average utilized points p_k for the part of the web space, that are evaluated for 6,621 parent documents including 18,397 links in jp-domain with $W_i = 1$ and $W_o = 30$. From these figures, it must be too difficult for us to discover the interesting contents C_k by the function of (s_P, i_P, o_P) based on only links.

Therefore, in this paper, we regard the web space as a simple textual database with hyper links, which doesn't have strongly integrated data model.

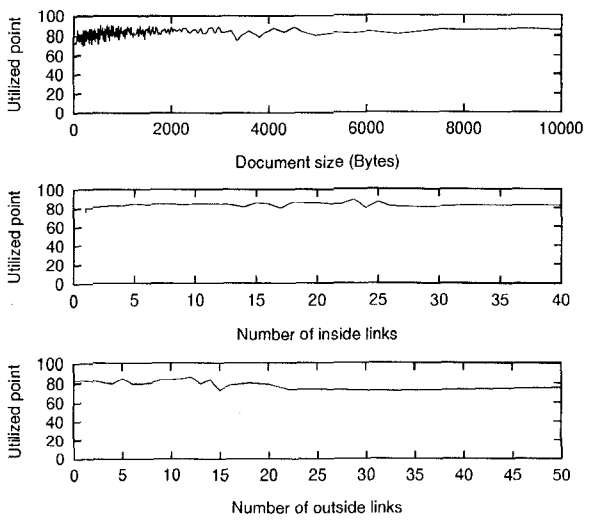

Figure 2: Utilized point vs. size, inside links, outside links

2.2 Textual data mining

At present, data mining is noteworthy field to be studied based on various kinds of researches, such as machine learning, inductive learning and knowledge representation, with considering characteristic features of database[4, 2, 6, 5].

In order to make it possible to retrieve web documents more sophisticatedly by derived keywords, We extend the typical mining algorithm to derive association rule[1].

Weighted association rule

We have a brief sketch for mining weighted association rule[7]. Basically, we extend the original mining algorithm to handle weighted keywords for markup language, especially for tags in HTML.

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

We create all rules whose *confidence* is equal to or greater than *minconf*. If rules, { program } and { program, database }, are derived from the set of keywords, we discover the rule "program ⇒ database," where the values of *support* and *confidence* are important measurement for the strongness of rules.

In addition to the conditions of threshold values, in the case of retrieving keyword k_j, if k_j appears w_{ij} times in documents T_i, we consider that k_j has the *weight* of w_{ij}, and define $a_{ij} = (k_j, w_{ij})$.

Thus, we select documents $T_i (i \in I)$ including set of keywords $K = \{k_j | j \in J\}$ from all documents set $\mathcal{T}$. Given T_i including $a_{ij} = (k_j, w_{ij})$, $sup(K) = \frac{N(K)}{N_0}$ can be defined by the following equations, where K includes any combination of keywords from all keywords $\mathcal{K}$ related with retrieving documents.

$$N_0 = \sum_{\mathcal{T}} \max_{\mathcal{K}} w_{ij}, \qquad N(K) = \sum_{i \in I} \min_{j \in J} w_{ij}.$$

3 Mondou

3.1 Structure of Mondou

Mondou consists of the following three main modules, *(1) agent, (2) database, (3) query server*, which are shown in Figure 3.

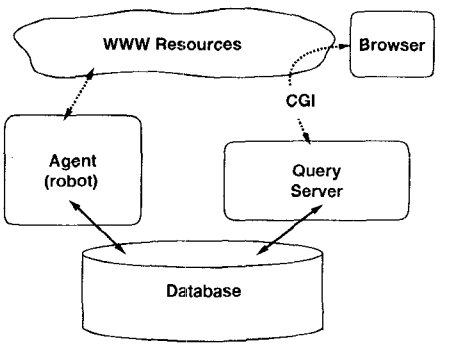

Figure 3: The structure of Mondou

Generally, the first module is called as the robot, spider or agent[9], and the robot collects web documents in the net and store them into the textual database. In addition to the standard functions 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 (children) are referred many times from other web pages (outside parent pages).

The *database* stores huge numbers of attribute values not only about keywords, but also date, size of documents and the number of links from other URLs.

The CGI (Common Gate Interface) module of *query server* is the search program, and it provides search results and mining association rules.

The input form of Mondou is shown in Figure 4, it is possible to enter combination of search keywords using AND and NOT boolean expressions in each empty box.

When one submitted initial keyword *knowledge*, Mondou provided several keywords of association, "engineering, systems, 知識 (knowledge in Japanese), acquisition" and so on, which is shown in Figure 5. Consequently, even by applying our proposed algorithm without taxonomies, conceptual trees or ontologies, we are able to grasp the association of important keywords in their interesting URLs.

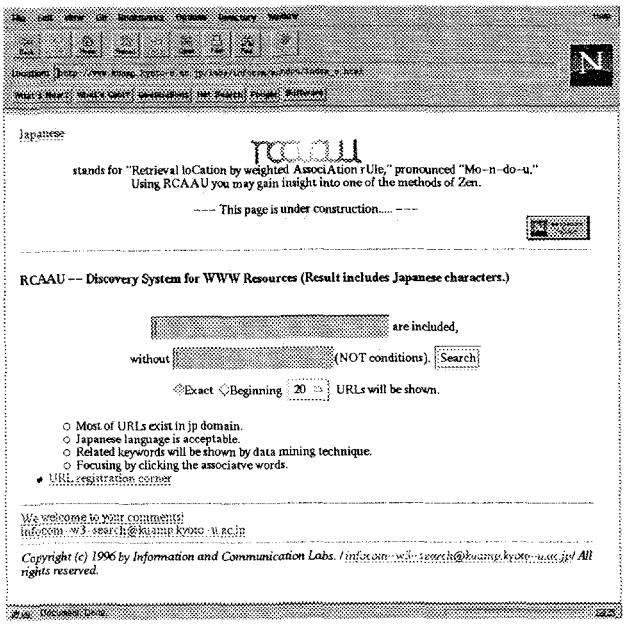

Figure 4: Input form of Mondou

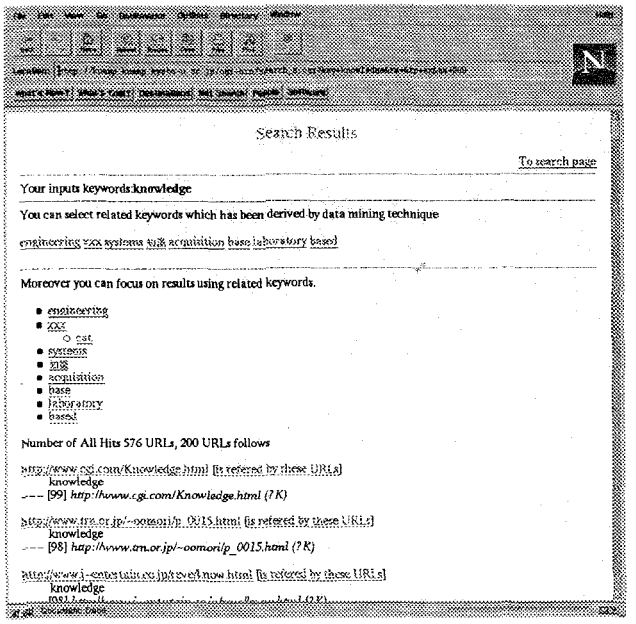

Figure 5: Output results from Mondou

3.2 Quality of associative keywords

At present, we have been operating Mondou in the net, and we try to examine the quality of search queries, patterns of combination of keywords and so on.

Table 1 shows typical examples of derived keywords, the number of URLs and related keywords derived by our algorithm.

Thus, we can easily get interesting combination of the keywords that can be treated as several meanings in web documents. Most of beginners could easily grasp the structure of web documents via association of keywords.

Moreover, from February to October in 1996, Mondou executed 931,537 queries submitted from the netters. 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.

The number of query patterns is 338,535, and 51,510 patterns (15.2%) are described by only one keyword, 287,025 (84.8%) patterns are described using more than two keywords. Most of users can submit various queries using the derived keywords. Furthermore, keywords including all executed queries covers 129,445 words (40.5%) for all 319,426 keywords, which are stored in our database. Thus, our Mondou can provide the rich combination of keywords in order to modify initial submitted query.

As a result, by using textual data mining algorithm in Mondou, web surfers could get much more information about the relationship of keywords in the natural way. Even if users don't know well about web documents to be find out, it is possible to discover the interesting URLs quickly.

3.3 Visualization: results of web mining

It is also necessary to develop a visual interface that can express information with multi-dimensional metric including network environment on client side, and we can focus on interesting sets with derived association rule. Then, we have been developing an interactive search interface by Java in Figure 6, we show one example of search results with the attributes as shape, area size, brink interval and so on.

In current implementation, we represent the cost of access time from users to the web server, and relevance of the URLs including given keywords. Moreover, the size of documents as area, support sets by derived keywords as color, and structure of URLs as arrows with quantity are also presented.

4 Conclusions and future works

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 documents, which are collected by agents. Many users can easily focus on interesting web resources without the knowledge of taxonomy, or intelligent data dictionary given by database administrators. We can conclude that our proposed algorithm works very effectively in searching textual data in the web.

We developed the first prototype of Mondou as a centralized system, but we should improve it as a distributed system in order to keep much more URLs and to focus on URLs more effectively since web grows very rapidly. We are also improving the visual interface for effective search on web browsers

Table 1: Examples of textual data mining

keywords	number of URLs	related keywords
applied	1,168	mathematics, mechanics, analysis, physics, media, geochemistry, superconductivity, optics, geology
engine	876	honda, search, stirling, dragon, エンジン (engine in Japanese), behavior, similarities, parts
analysis	1,904	fujita, numerical, behavior, method, top, plan, multidimensional, applied
simulation	634	software, シミュレーション (simulation in Japanese), computer, results, numerical, sciences, conference, noise, advanced
travel guide	77	center, singapore, internet, usa, thailand, las vegas, leisure

Acknowledgment

This work was supported in part by a Grant in Aid for Science Research 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. We also thank students, Mr. Hideki Nishimura (Sharp Corporation) and Mr. Koichiro Ito (The Goldman Sachs Group), as excellent programmers.

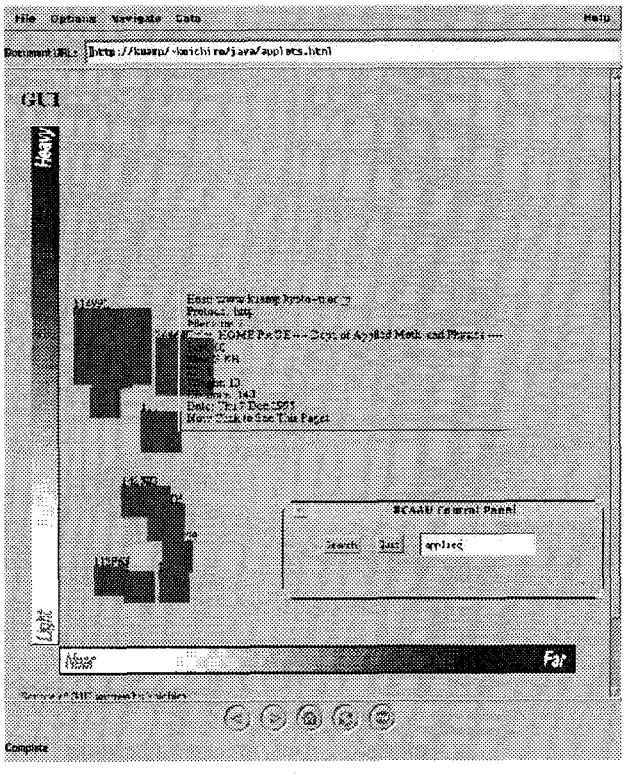

Figure 6: Visual interface for Mondou

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.