A Report on Metadata for Web Databases(Web Data Mining)

NAKATOH, TETSUYA
Computing and Communications Center, Kyushu University

OHMORI, KEISUKE
Graduate School of Information Science and Electrical Engineering, Kyushu University

HIROKAWA, SACHIO
Computing and Communications Center, Kyushu University

出版情報：情報処理学会研究報告. ICS, [知能と複雑系]. 2004 (125), pp.95-98, 2004-12-04. 一般社団法人情報処理学会
バージョン：
権利関係：

2004－ICS－138　(17)
2004／12／5

A Report on Metadata for Web Databases

Tetsuya Nakatoh ,† Keisuke Ohmori ††
and Sachio Hirokawa †

Increasing number of Web Databases are available on the Web. The query for such databases is not just a keyword but a complex query and that search results are not a listing of general Web pages but a listing of records generated dynamically from the database behind the Web interface. This paper is a survey of 2,880 Web databases listed in Dnavi (Japanese National Diet Library). The first part analyses the form components TextInputFields, SelectMenu, RadioButtons and CheckBoxes for 880 "book search" DBs in Dnavi. The second part reports a strong connection between the input form and output data by analyzing the attribute names for 100 Web DBs chosen randomly from 1,541 DBs which accepts complex query.

1. Introduction

Increasing number of search engines are available on the Web besides general search engines such as Google. They are databases with Web interface. We can obtain high quality information from these databases.

Information on those databases cannot be indexed by search engines and cannot be referred directly because it is referred to only by the page that is formed dynamically from the database according to the user's query. Because of that, these are called as Invisible Web[5),6)], Deep Web[1)] and Hidden Web[2),3)]. These sites often provide information of height quality limited to the specific theme.

We are developing the system DAISEN[9)] which performs meta-search for such databases on the Web. Conventional meta-search engines integrate a fixed number of particular general search engines. On the other hand, the goal of DAISEN is dynamic integration of arbitrary set of databases on the Web.

There is a new trend of databases on the Web, where a user can send complex query. The query he sends is not just simple keywords but are the keywords which specify each field of the records he wants to retrieve from the database.

For example, Amazon.com returns a list of book information which consists of the author, the title, the publisher, the price, ISBN and so on. kakaku.com returns a list of prices of PCs and other electric products. Travelocity.com returns a list of hotel information in specified area.

To aggregate such databases on the Web, we need to know what kind of information are provided in each DB in advance. More precisely, we need to know what kind of input query are required and what kind of record are returned as search results. In other words, we need "metadata" of each database.

If they are usual databases and Web Services, we can assume that a data schema would be given explicitly. But this is not the case for databases on the Web. They only provide interface for Web browsers.

Therefore, the data schema behind the Web site should be extracted and guessed from the "form information" of the site and from the HTML files of search results.

A large listing of such databases is known at Dnavi, a database navigation service provided by the National Diet Library☆. Note that not all of sites in Dnavi are dynamic databases. Some of them are nothing but a listing. We confirmed that there are 2,800 search engine in Dnavi and proposed a method to estimate the query form automatically[4)]. A key problem of automatic estimation of data schema is in obtaining attribute names for query (input metadata) and search results (output metadata). The problem is not easy for input as well as for output, however, information for input would help guessing metadata for output. This paper reports the current situation of Web

† Computing and Communications Center, Kyushu University
†† Graduate School of Information Science and Electrical Engineering, Kyushu University

☆ http://dnavi.ndl.go.jp/

databases with complex query and a possibility to guess input metadata from output metadata.

2. Components of Complex Query

A "books search" is a typical search site with complex query form. An example of a typical books search site is shown in Fig. 1. There are 880 such sites among 2,880 listed in Dnavi. In this section, we report the numbers of such sites that uses form components TextInputFields, SelectMenu, RadioButtons and CheckBoxes to describe complex query.

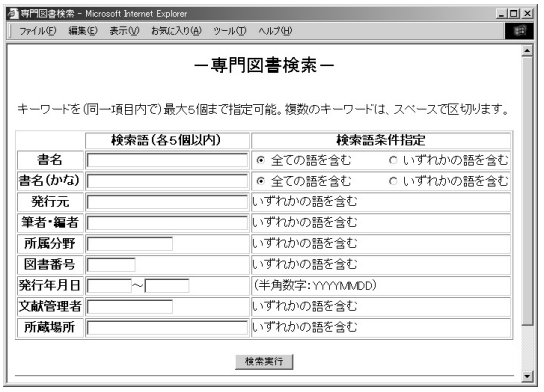

Fig. 1 A sample of books search sites

The number of text input fields in books search sites is shown in Fig. 2. 160 sites uses only one TextInputField, but most other sites uses multiple TextFields. Menu components with Select elements are typically used to specify the attribute of the TextInputField.

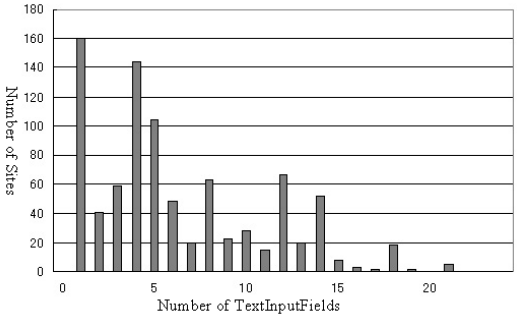

Fig. 2 The number of TextInputFields in books search sites

Fig. 3 shows the distribution of sites according to the number of Menu in the form. With such Menus, a user chooses a database and specifies the format of search result, or describe the meaning, e.g., "author", "title" or "publisher", of a keyword set in the TextInputField.

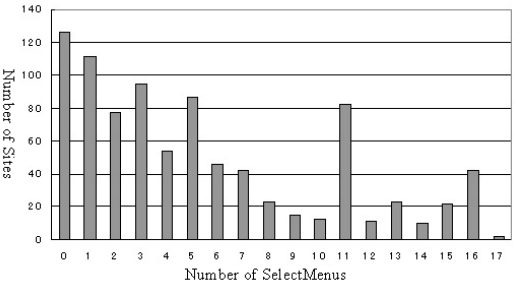

Fig. 3 The number of Menus in books search sites

RadioButtons (Fig. 4) are used in most of the sites to set up the search condition. Similar to Menus, the number of RadioButtons is not large when it is used to specify the attribute of the text input field.

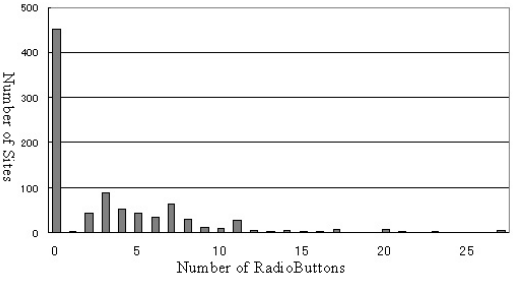

Fig. 4 The number of RadioButtons in books search sites

The checkboxes (Fig. 5) are used to select the databases for search at almost all sites.

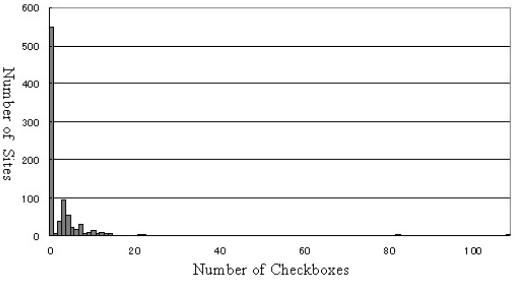

Fig. 5 The number of checkboxes in books search sites

3. Attributes of Complex Query

Most of the search engines have one text input field. They search for text files and return a listing of links to the retrieved files with brief information of the files.

On the other hand, databases on the Web have more than one input fields which determine the record a user wants. Fig. 6 is a typical example of such Web DB with complex query. We can use the site for searching books by specifying the author, the title etc. The search result is not just a list of links but a list of book records which consist of the similar data given as query (Fig. 7). In Fig. 6, we can see that the names of the input fields appear in the search result.

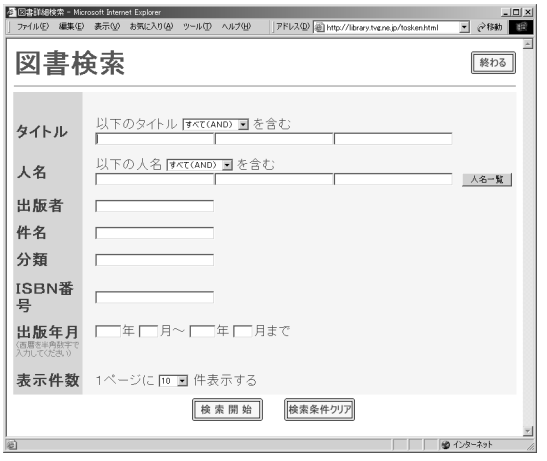

Fig. 6 a Web DB with complex query

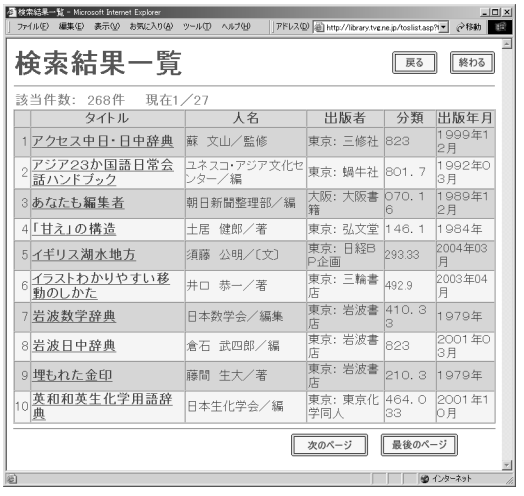

Fig. 7 a search result page

There are 1,541 sites in 2,880 Web DBs in Dnavi. We randomly chose 100 DBs from 1,541 DBs and analyzed the data scheme of query form and that of search result.

We manually compiled the attribute names for query and search result as follows. As for the attribute names of query, we extracted the string which locates near the text box by hand. When there is a pull down menu near to the text box, we assume that all the items of the menu are candidates of the attribute name.

As for the attribute names of search result, we used the following two heuristics. The first heuristics is that the attribute names appear at the first line. The second heuristics is that the attribute names appear at the first column. When a search result consists of just a list of links, and snippets with no attribute name, we obtained the name from the linked pages.

We analyzed the correspondence between the query metadata and output metadata. We used the ratio R of the number of common fields of the input page and the output page compared with the number of fields in the input. If a site has a high ratio, most of the attributes a user send appear in the result page.

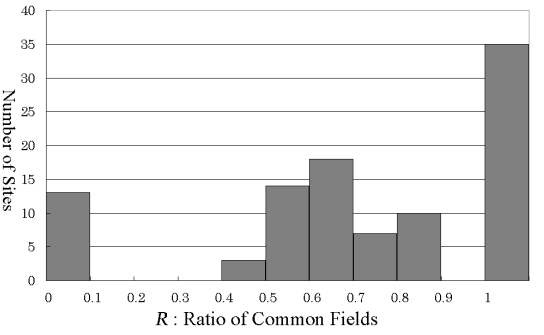

Fig. 8

A result of investigation is shown in the histogram of the Fig. 8. At 35% of the sites, all the fields of the input item appear in the field of a search result. For these sites, we can guess input metadata from output data. At 84% of the site, more than half of input items appear in the field of a search result. The actual ratio of the 13 sites in the range of 0.0 to 0.1 is exactly equal to 0, which means that query metadata cannot be estimated from output metadata. The reason of this failure, for 11 sites, is that the name

of query attribute is "keyword" for these sites. We should ignore this case because "keyword" is a general word which cannot be a particular field of some data type.

Most of success ratio, except for 0.0 and 1.0, locate in the range of 0.4 to 0.9. "Keyword", "Contents" and similar words are the main reason of poor ratio, which we already mentioned. Another reason of poor ratio is the ambiguity of words, e.g. "ISBN" and "book ID". There would be case, which we did not see among 100 DBs in this report, that attribute names are omitted because it is trivial for human. In such a case we cannot obtain attribute information from search result and hence would yield low success ration.

4. Summary

We analyzed the components of query form used in 880 books search sites in Dnavi. It revealed that Web databases are moving toward to use complex query. We confirmed a strong connection between input metadata and output metadata by analyzing the attribute names for 100 Web DBs chosen randomly from 1,541 DBs which accepts complex query.

More investigation is necessarily for larger number of sites. Automatic estimation of attribute names for query and search result are further work to be investigated.

Acknowledgments

This research was partially supported by the Japan Society for the Promotion of Science Grant-in-Aid for Exploratory Research No. 16650030, 2004, Grant-in-Aid for Young Scientists (B) No. 16700106, 2004, and by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Research on Priority Areas No. 16015267, 2004.

References

1) BrightPlanet, The Deep Web: Surfacing Hidden Value, BrightPlanet White Paper, 2000.
2) P. Ipeirotis, L. Gravano and M. Sahami, PERSIVAL Demo: Categorizing Hidden-Web Resources, JCDL2001, 2001.
3) P.Ipeirotis, L.Gravano and M.Sahami, Probe, Count, and Classify: Categorizing Hidden-Web Databases, ACM SIGMOD 2001, 2001.
4) T. Nakatoh, K. Ohmori, Y. Yamada and S. Hirokawa, *COMPLEX QUERY AND METADATA*, Proc. ISEE2003, pp. 291-294, 2003.
5) P. Pedley, The invisible web, ASLIB, 2001.
6) C. Sherman and G. Pric, The Invisible Web, Information Today, Inc., Medfore, New Jersey, 2001.
7) T. Taguchi, Y. Koga and S. Hirokawa, Integration of Search Sites of the World Wide Web, Proc. of *the International Forum cum Conference on Information Technology and Communication*, Vol. 2, pp. 25–32, 2000.
8) S. Thakkar, C. A. Knoblock, J. Ambite and C. Shahabi, Dynamically Composing Web Services from On-line Sources, Proc. of 2002 AAAI Workshop on Intelligent Service Integration, Edmonton, Alberta, Canada.
9) Directory Architecture for Integrated Search Engines, http://daisen.cc.kyushu-u.ac.jp/