

COMPLEX QUERY AND METADATA

Nakatoh, Tetsuya
九州大学情報基盤センター

Omori, Keisuke
九州大学大学院システム情報科学府

Yamada, Yasuhiro
Graduate School of Information Science and Electrical Engineering, Kyushu University

Hirokawa, Sachio
Computing and Communications Center, Kyushu University

<https://hdl.handle.net/2324/2980>

出版情報 : Proceedings of International Symposium on Information Science and Electrical Engineering. 2003, pp.291-294, 2003-11

バージョン :

権利関係 :



COMPLEX QUERY AND METADATA

T. Nakatoh[†], K. Ohmori[‡], Y. Yamada[‡] and S. Hirokawa[†]

[†] Computing and Communications Center, Kyushu University
6-10-1 Hakozaki, Higashi-ku Fukuoka 813-8581, JAPAN

[‡] Graduate School of Information Science and Electrical Engineering, Kyushu University
6-10-1 Hakozaki, Higashi-ku Fukuoka 813-8581, JAPAN

ABSTRACT

We are developing a search system DAISEn which integrates multiple search engines and generates a metasearch engine automatically. The target search engines of DAISEn are not general search engines, but are search engines specialized in some area. Integration of such engines yields efficiency and quality. There are search engines of new type which accept complex query and return structured data. Integration of such search engines is much harder than that of simple search engines which accept keywords and return a list of URLs.

This paper reports the current situation of complex forms by analyzing 2,880 specialized search engines and demonstrates a possibility to construct metadata for complex query.

1. INTRODUCTION

The flood of information on the Internet is a serious problem for people and companies. Search engines are keys to get rid of this flood of information. We use general search engines, e.g., Yahoo, Alta-vista and google. One of the problems of general search engines is the quality of search result. The search results tend to contain irrelevant pages. Many companies and organizations provide their information with their own search engines[12]. We call such web sites “Search Sites” compared with general search engines. A search site of a company focuses on their authorized information. CompletePlanet¹ estimates that there are more than 100,000 searchable databases available on the Web. Integration of such search sites, i.e. metasearch engine, is a solution to obtain good quality search and there are several difficulties to generate a metasearch engine in general. DAISEn² is the system we are developing. It analyzes search sites and generates metasearch engine automatically. Effectiveness of DAISEn is based on the following key technology.

- Generation of Query Parameters [10]
- Estimation of Logical Expression [9]
- Generation of Extractor [13]
- Feature Extraction [2, 8]

The first three technologies enable automatic generation of a wrapper which conceals the difference of interfaces. Another core technology is automatic feature extraction of search sites. The extractor is applied to eliminate advertisements and decoration that surround the search result. Thus we obtain purified feature of the database.

There is a new direction of search engines, where complex query is performed and specific items are returned instead of a simple list of URLs. Amazon.com returns a list of books. Kakaku.com returns a list of PCs together with their prices. Travelocity.com returns a list of hotels of specified area. They are not a simple list of URLs but are lists of items consisting of several small components of information. The input form for such a specialized search engine is much complex than that for a general search engine. They require several keywords to specify the query and each keyword represents different attribute. The departure and the arrival station, time and date are required for the transfer search engine Jorudan (<http://www.jorudan.co.jp>). The check-in date, the check-out date, the number of people, the number of rooms, the price range and the region are required for Mytrip (<http://www.mytrip.net>), a search engine for hotels.

In [15], we reported that there are more than 1,500 search sites, in Dnavi (Database Navigation Service of the National Diet Library)³, which contain complex query form with multiple inputs. There are 2,931 sites that are retrieved by the keyword “search” in Dnavi⁴. They are candidates of search sites. We checked the sites and classified by hands in the following three groups.

¹<http://www.completeplanet.com/>

²<http://daisen.cc.kyushu-u.ac.jp/>

³<http://dnavi.ndl.go.jp/netnv/>

⁴Dnavi contains more than 7,000 databases. But it is not a metasearch engine.

- search sites (1,845)
- link pages to search sites (624)
- simple listings without search (462)

We obtained 1,035 search sites following the link pages. Hence, 2,880 search sites, which we collected in this way, are the target of our analysis.

We report the current situation of complex forms by analyzing these 2,880 search engines in detail. We also demonstrate a possibility to construct metadata for complex query. Search sites with complex query return lists of records which vary site to site. Integration of such various records is not as easy as integration of lists of URLs which are returned from simple search sites. Metadata of records and matching between fields are keys to aggregation of search sites of complex query. Though the target is limited to search sites in this paper, we think that a similar technique can be applied to the more general Web services [14].

2. SIMPLE QUERY VS. COMPLEX QUERY

The number of text input fields is the main difference of simple search engines and complex search engines. Google is a typical example of sites with simple query with single text input field (Fig. 1).

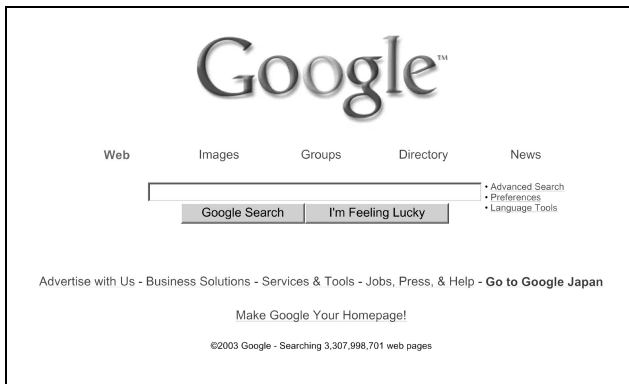


Fig. 1. Simple Query Site

The analysis of the number of text input fields is shown in Table 1. A text input field is formed by an HTML tag $\langle \text{INPUT type=text} \rangle$ and $\langle \text{textarea} \rangle$. There are many $\langle \text{INPUT type=edit} \rangle$ in the last investigation [15]. But "type=edit" is considered as the omission of type attribute. It is not in the standard of HTML. Default omission is "type=text".

In general search engines, a search keyword (query) is set in text input field. There are 559 sites (19.4%) that have single text input field. There are 1,541 sites (53.5%) which use multiple text input fields. This figure is much large than

Table 1. Number of Text Input Fields

number of textbox	number of fields	number of textbox	number of fields
0	780	14	92
1	559	15	14
2	168	16	6
3	236	17	2
4	252	18	18
5	163	19	1
6	127	20	1
7	92	21	5
8	82	23	2
9	50	24	1
10	42	26	1
11	56	74	1
12	78	100	1
13	50		

Table 2. Search Site without text input field

pull-down menu, checkbox and radio button	48
HTML frame document	547
refresh, jump	24
script	27
connection error	134
total	780

we expected before. We can say that more than half search sites are providing complex query.

Typical examples are search sites for books where the title of the book and the name of the author is specified in multiple text input fields as attributes for the book (Fig. 2).

There are 780 exceptional search sites (27.1%) where no text input field is found. Table 2 shows the detailed analysis of the sites. 48 sites (1.7%) are providing a search function with pull-down menus, checkboxes and radio buttons. We were not able to analyze some pages. Typical reasons are that the tool used for the experiment cannot handle pages with frame, jump and scripts (javascript, VBscript), that the site uses a cookie or the referer and that the page does not exist.

3. QUERY OPTIONS

Even in a search site with single text input field, optional information are augmented by pull-down menus, check boxes and radio buttons. The number of sites with designated number of components is shown in Table 3. Pull-down menus are contained in 57.4% sites. There is a search site with 35

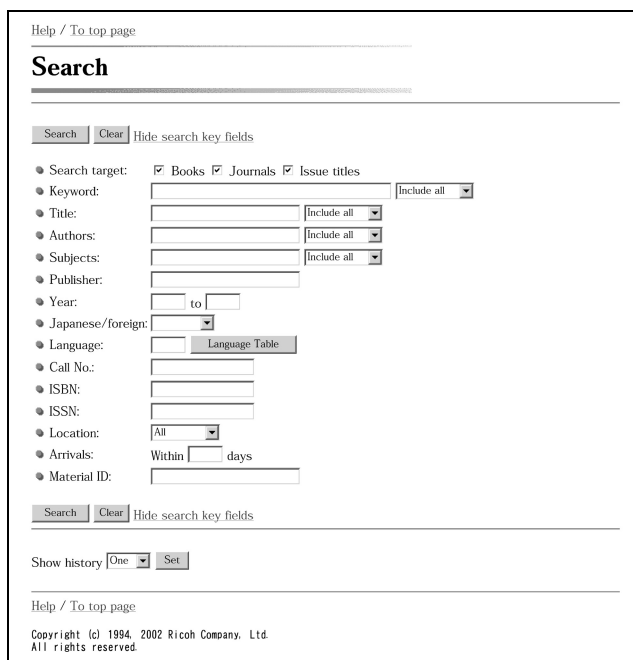


Fig. 2. Complex Query Site

pull-down menus. The average of the number of pull-down menus is 5.3 and the averages of the number of the choices contained in a pull-down menu are 12.3. The most complex pull-down menu contains 1,495 choices. 21.4% sites have checkboxes. The average of the number of checkboxes in a site is 11.5. There is a search site with 206 checkboxes. 29.4% sites have radio buttons. The average of the number of radio buttons is 5.9. There is a search site with 113 radio buttons.

Most of search sites with multiple text input fields provide searches for specific records. But there are some exceptions. Library of Yamanashi University⁵ has multiple text input fields. But it is to construct logical query with multiple keywords (Fig. 3). There is another kind of exception, where different search services are provided in the same page.

4. CONSTRUCTION OF METADATA FROM COMPLEX QUERY FORM

We classify search sites into three groups according to the number of text input fields and the attributes they represent.

(A) Sites with one text input field. A text search by the keyword is possible at this kind of sites. They are targets of present DAISEn.

⁵<http://licfather.lib.yamanashi.ac.jp/oudan/index.html>

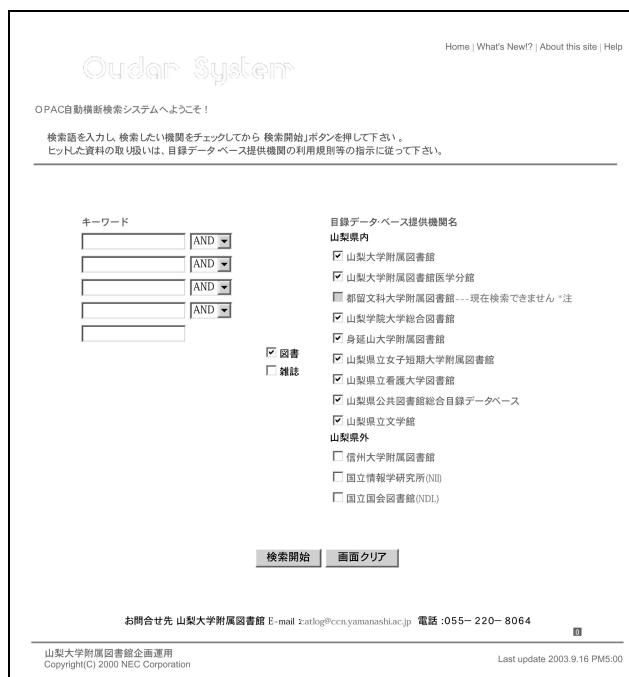


Fig. 3. Logical Query

(B) Sites with multiple text input fields of the same attributes. A text search with logical operation (AND and OR) between the keywords is possible at this kind of sites.

(C) Sites with multiple text input fields of different attributes. Each keyword specifies the field of record that is being searched. The sites of this group are targets of this paper.

Most sites have pull-down menus, checkboxes and radio buttons as well. An indication order, a number of indications, a target of databases, a range of data and a field name are specified by these query options.

The search sites of the type (C) are classified and shown in the following.

- A field name is associated to each text input field. It is displayed in the left or the top of the field.
- A pull-down menu is associated to each text input field. It is displayed in the left or the top of the field. That is the name of the field to choose.

In both cases, metadata can be formed by collecting these field names and classifying them. The following is an example of such a metadata which we constructed from complex search sites of libraries.

Table 3. # of Components

# of components	# of sites		
	pull-down menu	checkbox	radio button
0	1,227	2,263	2,036
1	310	49	5
2	221	52	169
3	271	111	156
4	144	81	117
5	163	26	78
6	88	34	70
7	61	77	80
8	63	24	44
9	32	13	25
10	21	20	15
>10	279	130	88

text name 書名, 書名/タイトル, 書名 (叢書名), 書名等, 書名読み, 図書名, 双書名, 叢書名, タイトル, TITLE, 雑誌名, 誌名, 書籍・雑誌名, ...

author name 著者, 著者名, 著作者, 著作者名, 著者・作者, 著者・作者名, 原著名, 主要著者名, 書籍著者名, 著訳者, AUTHOR, 編著者名, ...

5. CONCLUSION

In this paper, we reported an analysis of 2,880 search sites and their search forms. It turned out that more than half of sites have complex query form with multiple text input fields. Most of such sites have menus, check boxes and radio buttons as query options. It is pointed out that metadata will be constructed by extracting the names for text input fields.

Metasearch engines are expected as next generation search engines of high quality search result. Each individual search engine is evolving toward complex query. Therefore, automatic generation of metadata from complex query is important future work.

6. REFERENCES

- [1] D.Embley, S.Jiang, Y.-K. Ng, Record-boundary discovery in Web documents, Proceedings of 1999 ACM SIGMOD International Conference on Management of Data, pp. 467-478, 1999.
- [2] S. Hirokawa, S. Watanabe, Y. Koga and T. Taguchi, *Automatic Feature Extraction of Search Sites*, Proc. SSGRR2001(CD-ROM).
- [3] A. E. Howe and D. Dreilinger, Savvy Search A Metasearch Engine That Learns Which Search Engines to Query, AI Magazine Vol. 18, No. 2, pp. 19-25, 1997.
- [4] P. Ipeirotis, L. Gravano and M. Sahami, *Automatic Classification of Text Databases through Query Probing*, Proc. of the ACM SIGMOD Workshop on the Web and Databases (WebDB'00), 2000.
- [5] Y. Koga, T. Taguchi and S. Hirokawa, *Wrapper Generation for Search Sites Integration(in Japanese)*, Proc. DEWS'01, 2001.
- [6] N. Kushmerick, *Wrapper induction: Efficiency and expressiveness*, Artificial Intelligence, Vol.118, No.1-2, pp.15-68, 2000.
- [7] W. Meng, C. Yu, K. Liu. Building Efficient and Effective Metasearch Engines. ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp.48-89.
- [8] T. Nakatoh, Y. Koga and S. Hirokawa, *Automatic Classification of Search Sites(in Japanese)*, Proc. DB-Web2001, pp. 225-228, 2001.
- [9] T. Nakatoh, M. Sakai, S. Hirokawa, Automatic Estimation of Set Operator for Search Sites, Forum on Information Technology 2002,pp.9-10, 2002 (in Japanese).
- [10] T. Nakatoh, M. Sakai, Y. Koga and S. Hirokawa, *Generation of Query URL for Search Sites*, Proc. SSGRR2002w(CD-ROM).
- [11] E.Selberg, O.Etzioni, The MetaCrawler architecture for resource aggregation on the Web, IEEE Expert 12(1), pp. 11-14, 1997.
- [12] C. Sherman and G. Price, The Invisible Web, Information Today, Inc, Medfore, New Jersey, 2001.
- [13] T. Taguchi, Y. Koga and S. Hirokawa, Integration of Search Sites of the World Wide Web, Proc. of International Forum cum Conference on Information Technology and Communication, Vol. 2, pp. 25-32, 2000.
- [14] Andreas Hess, Nicholas Kushmerick, Automatically attaching semantic metadata to Web services, Proc. IJCAI-03 Workshop on Information Integration on the Web, pp. 111-116, 2003.
- [15] Y. Yamada, Y. Matsunaga, M. Noguchi, T. Nakatoh, S. Hirokawa, Automatic Wrapper Generation and its Evaluation on Meta-Search System DAISEn, Summer Database Workshop 2003. (in Japanese)