

Multiple viewed search engine for an e-journal – A case study on zoological science

Seki, Takahiro

Office for Information of University Evaluation, Kyushu University

Wada, Taiki

Graduate School of Information Science and Electrical Engineering, Kyushu University

Yamada, Yasuhiro

Kyushu University, User Science Institute

Ytow, Nozomi

Graduate School of Life and Environmental Science, University of Tsukuba

他

<https://hdl.handle.net/2324/817826>

出版情報 : Human-Computer Interaction. HCI Applications and Services (Lecture Notes in Computer Science). 4553 (PART 4), pp.989-998, 2007-12-01

バージョン :

権利関係 :



Multiple Viewed Search Engine for an e-Journal - a Case Study on Zoological Science

Takahiro Seki¹, Taiki Wada², Yasuhiro Yamada³, Nozomi Ytow⁴, and
Sachio Hirokawa⁵

¹ Office for Information of University Evaluation, Kyushu University,
Hakozaki 6-10-1, Higashi-ku, Fukuoka, 812-8581, Japan

² Graduate School of Information Science and Electrical Engineering, Kyushu University,
Motooka 744, Nishi-ku, Fukuoka, 819-0395, Japan

³ Kyushu University User Science Institute, Hakozaki 6-10-1, Higashi-ku, Fukuoka,
812-8581, Japan

⁴ Graduate School of Life and Environmental Science, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8572, Japan

⁵ Computing and Communications Center, Kyushu University,
Hakozaki 6-10-1, Higashi-ku, Fukuoka, 812-8581 Japan

¹ tsekiuoc@mbox.nc.kyushu-u.ac.jp, ² t-wada@i.kyushu-u.ac.jp,
³ yamada@usi.kyushu-u.ac.jp, ⁴ nozomi@biol.tsukuba.ac.jp, ⁵ hirokawa@cc.kyushu-u.ac.jp

Abstract. The multiple viewed search engine presented here retrieves documents of an indicated search area and displays a matrix of the distribution of the clustering from two aspects of the retrieval result. The search engine provides a visual and semantic bird's-eye view of the entire retrieval result. In addition, the characteristic words of each cluster are displayed in the matrix, and supports narrowing of the search. Furthermore, it is possible to immediately change the analysis criteria or the number of clusters and to use a zooming function. Thus, various retrieval conditions for a query can be attempted immediately and continuously. As a case study, this paper performs several analyses on the electronic journal *Zoological Science* using a multiple viewed search engine.

Keywords: Multiple viewed search engine, Clustering, Characteristic word extraction, E-journal

1 Introduction

The development of the World Wide Web, explosively increasing the volume of, and extending the availability of information, encourages publishers to move to e-journals from paper one, not only for new articles but also for previously printed papers with the conversion into electronic files. This enables researchers to investigate previous studies with a search engine when writing papers. The current information retrieval is generally expected to support investigation and analysis and therefore require mechanisms not only for mere searches but also for search support and analysis. The current search engines often adopt ranking, clustering, and directories as contrivances

for displaying retrieval results. Since many of the search engines merely return lists containing the search keywords, however, it is difficult to find important matter appearing in the retrieval results and to understand the tendency of these results. It is generally difficult to give appropriate search conditions for obtaining the desired results immediately by a search. Despite narrow or advanced searches under repeatedly changing conditions, the desired results may not always be available easily and accurately. The search engine must solve these problems efficiently in the future.

Many e-journals display such basic information as titles, authors, and abstracts of papers in an HTML page with links to the bodies of the papers. Documents with changeable structures or irregularities are called semistructured documents, including HTML and XML documents. Each substructure of a semistructured document often contains certain aspects. In the search of semistructured documents, it is important to view the retrieval results by utilizing the substructures as the aspects. We cannot say that the retrieval result displayed from a single aspect, adopted by many search engines, can deal with this problem quickly. The general techniques of an XML search can be classified roughly into an XQuery with complicated search requests, and a full-text search for XML documents with simple keywords. Now many proposals have been made about their merger. For example, important subjects are not only the technique of a search from both structures and contents [15] but also an application of ranking technique [6]. However, the merged technique will not be satisfactory in obtaining an entire representation of retrieval results or viewing retrieval results from multiple aspects.

To obtain an entire representation of retrieval results, clustering may be used. For example, KartOO¹ is a search system that compiles related pages of retrieval results and displays the results in two dimensions but cannot always interpret each cluster or its positional relationship easily. Mechanisms for viewing retrieval results from multiple aspects are OLAP and biclustering. OLAP refers to a system or tool that analyzes a database in multiple dimensions and visualizes data. This has a powerful function for numeric data or data that is easy to classify but poses difficulties in creating such data. In biclustering, a matrix consisting of the numeric association between individual data from two aspects is clustered from each aspect and a matrix consisting of a typical value obtained from each cluster by calculation is displayed [1]. For biclustering, however, associations should be calculated in advance and the method of defining a typical value for each cluster should be revised. In addition, people are now considering faceted classification, where several independent categories are set for a narrow search from several aspects. For example, faceted classification is now applied to the navigation of retrieval results [2], [13].

The authors have been proposing a multiple viewed search engine. From the aspects of two partial structures specified by the user as vertical and horizontal axes, this engine clusters the retrieval results of a partial structure from a semistructured document consisting of the common partial structures with respect to the same kind aspects and displays the distribution status in the matrix [7], [11], [12], [14]. The aforementioned OLAP and biclustering are close to the concept of this engine. From OLAP and biclustering, this engine greatly differs in that non-numeric data can be handled dynamically. For the analysis of the retrieval results, this engine outlines the

¹ <http://www.kartoo.com/>

retrieval results both visually and semantically. This engine enables narrowing by determining noted cells by only viewing the characteristic words from rows and columns and also supports searches by obtaining candidate keywords for narrow or advanced searches. For processing after retrieval results are obtained in response to a search request, this engine allows changes of keywords, permitted by ordinary search engines, but also changes of aspects, number of clusters, and zooming. Even with the same search keywords, the possibilities of various search conditions are immediately available.

This paper describes the implementation of the multiple viewed search engine for multiple items on the abstract pages (2720 papers in total) of the e-journal *Zoological Science*² issued by The Zoological Society of Japan.

2 Clustering and extracting characteristic words

For matrix display by the multiple viewed search engine, clustering and characteristic word extractions are important techniques. This engine employs well-known techniques, as described here.

The clustering techniques can roughly be classified into the single linkage method, the complete linkage method, and other hierarchical methods and into the k-means method and other non-hierarchical (partitioning-optimization) method. The hierarchical methods can further be classified into divisive and agglomerative [9]. This engine employs an agglomerative hierarchical method. When generating k clusters from n documents, for example, n clusters consisting of one document only are prepared. Until the number of clusters becomes k by calculating the similarities between clusters, the most similar two clusters are associated into one repeatedly. The series of clustering operations can be drawn by a binary tree called a dendrogram. A dendrogram indicates the status of coupling between clusters in clustering.

Since clustering associates items with similar characteristics, the characteristics of each cluster can be indicated by the extracted characteristic words. In clustering based on a hierarchical method, multiple clusters of weak associations may forcibly be associated into one cluster if the number of generated clusters is small. Therefore, it is appropriate to present several characteristic words instead of only one. To extract characteristic words of clusters, this engine employs statistical indexes used in the field of natural language processing. For example, the statistical indexes are frequency, cosine, pointwise mutual information [4], log-likelihood ratio [5], chi-square value (χ^2) [8], and dice coefficient [10].

² <http://zsj.jstage.jst.go.jp/browse/zsj>

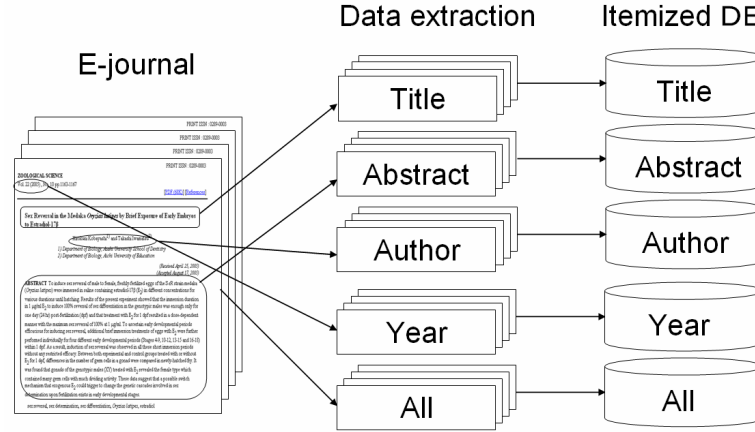


Fig. 1. Data preprocessing

3 Multiple viewed search engine and its implementation

The e-journal *Zoological Science* issued by The Zoological Society of Japan consists of HTML pages describing basic information about papers with HTML pages of linked references and pages of paper bodies. The multiple viewed search engine was implemented for a total of 2720 papers from Volume 1, No. 1 (issued in 1984) to Volume 22, No. 9. The basic information about the papers consists of titles, authors, years of issue, and abstracts. The HTML files of individual papers have a common structure.

The preprocessing is as follows. A file set where each paper constitutes one file is prepared, and the titles, abstracts, authors, years of issue, and all-item data including all items already mentioned are extracted from the set. Then a file of word frequencies is generated from the file sets and itemized indexes are created (Fig. 1). If an item is missing, the data absence is indicated by "Null" or other. Here, words are adopted after all characters are converted into lowercase and stopwords are eliminated.

The multiple viewed search engine is a system that searches a specified search area of an object document for keywords, clusters the results based on two items, and displays the distribution status in the matrix. When the user selects a search area, two items for the matrix and the number of clusters and enters the search keywords through the interface (Fig. 2), a matrix is displayed as the retrieval results (Fig. 3).

For matrix generation, the system internally searches the user-selected search area to generate a retrieval result list. This list is clustered by each of the two user-selected items into the number of clusters to extract characteristic words together. If a retrieval result belongs to cluster i on the vertical axis and cluster j on the horizontal axis, it becomes an element of the (i, j) cell in the matrix. Then a matrix indicating the retrieval result list or the number of retrieval results contained in each cell is

Keywords	water sea	
Connective	☐ AND ☒ OR	
Without the words		
Search area	Abstract ▾	
Row	Abstract ▾	Number of clusters 3 ▾
Column	Title ▾	Number of clusters 3 ▾
Search Reset		

Fig. 2. Sample input screen

Abstract	fish, content, freshwater, adaptation, adapted, salinity, fw, oreochromis, mossambicus, hypoosmoregulatory	0	0	13	13
	seawater, disappeared, eel, eels, atrial, intake, natriuretic, drinking, mucosal, phosphatidylcholine	0	3	17	20
	results, species, found, showed, examined, present, sea, eggs, water, urchin	1	0	296	297
	SUM	1	3	326	330
		steroid, variations, concentrations, sites, tribolodon, river, hakonensis, dace, basin, jinzu	quail, natriuretic, active, coturnix, atrial, intake, excretion, endothelin, parathyroid, biologically	cells, species, sea, effects, japan, urchin, embryos, study, sperm, water	SUM
		Title			

Fig. 3. Sample output screen

displayed (Fig. 4). This engine is implemented by using a CGI program written in Perl. For clustering and characteristic word extraction, the methods described in Section 2 are used. For the index creation and document search, Generic Engine for Transposable Association (GETA)³ is adopted. The library for clustering is the Perl module Algorithm::Cluster of CPAN⁴.

This engine allows the user to change the search area and the vertical-axis and horizontal-axis items on the matrix display freely for a multiple-view search. If the search area, the vertical-axis item and the horizontal-axis items on the matrix, and the number of clusters for each item have five choices each with the search keywords fixed, for example, the total number of choices is $5 \times 5^2 \times 5^2 = 3125$. The presentation of characteristic words in the matrix gives a meaning to each cluster. By a semantic bird's-eye view of the entire retrieval result, the user checks the characteristic words of each cluster and selects the desired cell where a desired cluster on the vertical axis

³ <http://geta.ex.nii.ac.jp/>

⁴ <http://www.cpan.org/>



Fig. 4. Internal processing for matrix generation

and one on the vertical axis intersect each other.

For re-searches under various conditions, the user can change the search keywords, the search area, the vertical-axis and horizontal-axis items in the matrix display, the number of clusters, and the zooming by cell or cluster. Here, zooming refers to a recursive matrix generation function that clusters an entire cell on the basis of two items used for the matrix display immediately before zooming and displays the distribution status in the matrix.

4 Some analyses by the multiple viewed search engine

DNA polymerase chain reaction (PCR) enabled determination of DNA sequence of specific genes. Accumulation of these sequences with improvement in computation power of computers makes molecular phylogenetics, i.e. methods to estimate phylogenetic relationships between various species, as common tool for biology. Phylogeny is also applied to deduce function of genes using knowledge on products of similar sequences. Query to citation databases with phrase containing phylogeny, therefore, may return a set of biological works of these different themes. Here we explain the use of the multiple viewed search engine with focusing on distinction of these themes, i.e. biodiversity and biological function of molecules, in the search result of query words "phylogeny" or "phylogenetic".

Table 1. Results of retrieval with phylogeny *or* phylogenetic(a) 2×2 display of title $\times$ abstract

	A	B	Sum
a	3	37	40
b	0	116	116
Sum	3	153	156

(b) 2×2 display of title $\times$ author

	A	B	Sum
a	1	39	40
b	4	112	116
Sum	5	151	156

Table 2. Characteristic words corresponding to Table 1 (a)

a	gene, mitochondrial, dna, phylogenetic, sequences, relationships, inferred, cytochrome, based, mammalian
b	species, Japanese, sea, ascidian, study, analysis, molecular, phylogenetic, relationships, phylogeny
A	carried, subsequently, fragment, polymorphism, restriction, rflp, dnas, subdivided, subclades, mitotypes
B	results, species, showed, analysis, gene, molecular, phylogenetic, sequences, mitochondrial, relationships

This engine can analyze retrieval results from various aspects through changes of the vertical-axis and horizontal-axis items. For example, characteristic or heterogeneous cells can be found or hidden relationship between the items can be clarified. If there is an intuitive relationship between the vertical-axis and horizontal-axis items and many characteristic words are common to both items, the character is understood to be related to the words. Otherwise, the cell contents will be considered intuitively from the combination of their characteristic words, or the number of hits per cell will be reduced for subdivision.

Table 1 gives the results of searching the specified area for all items with "phylogeny *or* phylogenetic" as the search keywords. The table shows the 2×2 display of title on the vertical axis and abstract on the horizontal axis in (a) and that of title on the vertical axis and author on the horizontal axis in (b).

In Table 1(a), the two cells other than the cell Ab of no hits (hereinafter called an empty cell) and the cell Bb of many hits are easy to interpret. Table 2 lists the characteristic words corresponding to Table 1(a). The characteristic words in the column A, e.g. "subclades" suggesting monophyly or "rflp" expected to be derived from PCR-RFLP (polymerase chain reaction-restriction fragment length polymorphism), a commonly used method in biodiversity analysis based on molecular data, indicate that the cell Aa contains papers on biodiversity. The characteristic words "mitochondrial" and "cytochrome" in the row a suggests that the cell Ba also contains papers on biodiversity because genes of these mitochondrial proteins are commonly used in phylogenetic analysis.

Table 1(b) gives the names of people seemingly belonging to a specific research group as characteristic words and suggests relations between the group and the theme of research.

Table 3. Increase of the number of clusters in Table 1(a): 3×3 display with dendrogram

			A	B	C	Sum
			156			
			3	153		
				49	104	
a		40	3	26	11	40
b	156	116	4	0	0	4
c			12	0	23	89
Sum			3	49	104	156

With this engine, the user cannot only make search conditions restricted by an AND search or NOT search for narrowed retrieval results or detailed analysis, but also promote subdivision by increasing the number of clusters and zooming. In particular, the user may be able to view a single cell or cluster in detail by increasing the number of clusters or zooming. These techniques feature no need to devise search conditions for eliminating other cells or clusters. If characteristic words obviously belonging to different clusters coexist, however, these operations can separate the characteristic words into different clusters.

Since this engine employs clustering based on a hierarchical method, the cluster association status on each axis can be displayed by a dendrogram. If the number of clusters is increased on each of the vertical and horizontal axes of the matrix, one cell is divided into up to four cells. The use of a dendrogram enables the tracking of cell division and the analysis of the association between clusters.

The 2×2 display in Table 1(a) is converted into the 3×3 display by increasing the number of clusters. Table 3 shows the results plus the dendrogram. From this increase in the number of clusters, we see that the row b and the column B in Table 3 are clusters divided from the row b and the column B in Table 1(a). In general, note that a cell of the greatest number of hits is not always divided when the number of clusters is increased by one each on the vertical and horizontal axes.

The characteristic words of clusters at different clustering levels indicate the criteria although we do not show the characteristic words in Table 3. The words "hemoglobin" or "hemoglobins" must be the criterion for the row b because these words are shared by the four papers clustered there but rare in other rows. Some property of mitochondria must be another criterion to distinguish the row B because the row contains 41 times of "mitochondrial" while the others contain 26 in sum.

We should expect that a large matrix may be generated from an enormous number of clusters if the number of clusters is increased repeatedly. In particular, if a cluster of many hits is stout, increasing the number of clusters will merely increase the cells of few hits or empty cells. Once nearly half of the cells have become empty, we should consider a means other than increasing the number of clusters.

For zooming by this engine, the user should merely choose and click a cell, a row, or a column. Increasing the number of clusters is a subdivision based on the general circumstances immediately before, while zooming is exclusive subdivision based on the current circumstances only. With this clear difference in mind, we need to use the

Table 4. Zooming of the Bb cell in Table 1(a)

	A	B	Sum
a	0	4	4
b	4	108	112
Sum	4	112	116

Table 5. Retrieval results of (phylogeny *or* phylogenetic) *not* mitochondrial

	A	B	Sum
a	0	2	2
b	2	83	85
Sum	2	83	87

methods appropriately.

Table 4 gives the results of 116 hits in the cell Bb of Table 1(a) and omits the characteristic words. Since the cell Aa in Table 4 is empty, the cells Ab and Ba of few hits can be identified by viewing the characteristic words of the row a and the column A. The cell Ab is estimated to be of a paper about biodiversity discussing a phylogenetic relationship in a specific area and the cell Ba is estimated to be of a paper about a biological function discussing an amino acid sequence.

In some cases, it may be better to refine the conditions of the AND search and NOT search than to increase the number of clusters or to repeat zooming. This is especially true if there is a rather large and stout cluster. Characteristic words of each cluster in the matrix display may include keywords that are not clear. By using characteristic words, candidate words to be added as search keywords or to be excluded can be found and supported for a re-search. Such words are comparatively easy to find by increasing the number of clusters and zooming.

The word "mitochondrial" among the words in the row a and the column B of Table 2 is not used in an abstract to estimate a molecular biological function from a phylogenetic relationship. Therefore, the word is a candidate to be excluded when the biological function is noted. Table 5 gives the results when the search keywords in the search request of Table 1(a) are changed to "(phylogeny *or* phylogenetic) *not* mitochondrial".

5 Conclusion and further work

A multiple viewed search engine was developed to search the e-journal *Zoological Science* from arbitrary aspects and to display the retrieval results from two arbitrary aspects in a matrix form. This paper discussed the possibility of search support and analysis by using the engine.

Currently, however, the engine is not satisfactory for quality data for which classification is more appropriate than clustering. It is also necessary to discuss various clustering methods and appropriate characteristic word extraction methods. A quantitative evaluation will be performed on the search efficiency of this engine. In

the future, we hope to discuss the implementability of this engine for semistructured documents that do not always have the same substructures even if of the same kind.

Acknowledgments. This study was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 17300071, and Japan Science and Technology Agency's project to develop "Innovative seeds" in Fiscal 2006.

References

1. Cheng, Y., Church, G.M.: Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (2000) 93-103
2. Christophides, V., Plexousakis, D., Scholl, M., Tourtounis, S.: On labeling schemes for the semantic web. WWW2003 (2003) 544-555
3. Chujo, K., Utiyama, M.: Using statistical measures to extract specialized vocabulary from a corpus. English Corpus Studies 12 (2005) 19-35 (in Japanese)
4. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. In Proc. of ACL-89 (1989) 76-83
5. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19 (1) (1993) 61-74
6. Guo, L., Shao, F., Botev, C., Shanmugasundaram, J.: XRANK: ranked keyword search over XML documents. SIGMOD2003 (2003) 16-27
7. Hirokawa, S., Seki, T., Yasumoto, Y., Yamada, Y.: Faceted search engine for activities of academic staff. IPSJ SIG Technical Reports 2005-DBS-137 (2005) 665-672 (in Japanese)
8. Hisamitsu, T., Niwa, Y.: Topic-word selection based on combinatorial probability, in NLPRS-2001 (2001) 289-296
9. Kashima, T.: A survey of recent clustering methods for data mining (part 1) -Try clustering!-. Transactions of the Japanese Society for Artificial Intelligence 18(1) (2003) 59-65 (in Japanese)
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing (1999)
11. Seki, T., Yasumoto, Y., Wada, T., Ytow, N., Hirokawa, S.: A retrieval and analysis of the electrical journal "Zoological Science" by multiple semantic matrix method. Digital Libraries 30 (2006) 72-80 (in Japanese)
12. Seki, T., Wada, T., Yamada, Y., Hirokawa, S.: A faceted search engine for scientific articles. SIG-FPAI-A601, (2006) 13-18 (in Japanese)
13. Tzitzikas, Y., Spyrtatos, N., Constantopoulos, P., Analyti, A.: Extended faceted taxonomies for web catalogs. WISE2002 (2002) 192-204
14. Yasumoto, Y., Wada, T., Seki, T., Hirokawa, S.: Multiple analysis of reputation information of hospitals. SIG-KBS-A501, (2005) 1-4 (in Japanese)
15. Yu, C., Qi, H., Jagadish, H.V.: Integration of IR into an XML Database. INEX Workshop 2002 (2002) 162-169