A Method for Creating a High Quality Collection of Researchers' Homepages from the Web

Yuxin Wang[†] Keizo Oyama[‡]

[†‡] School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

[‡] National Institute of Informatics, Research Organization of Information and Systems

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8403 Japan

Email: [†]mini_wang@grad.nii.ac.jp, [‡] oyama@nii.ac.jp

Abstract This paper proposes a method for creating a high quality collection of researchers' homepages. The proposed method consists of three phases: rough filtering of the possible web pages, accurate evaluation of the web pages and precise selection of the correct homepages. For the rough filtering, the authors first define content-based keyword-lists, then generate filtering rules and relax the rules with heuristics. For the evaluation and the selection, they use a support vector machine with the feature sets derived from the content words of the web pages and propose an approach utilizing web-specific properties for improving the measures.

Keyword Web Mining, Web Information Retrieval, Machine Learning, Web Page Classification

1. Introduction

Various kinds of information are provided on the web and collecting and utilizing information from the web are becoming increasingly crucial. Researchers' information is one of such typical examples.

As for information of research papers, CiteSeer[1] is one of the well known search engine for the web-based resources. There are also many database services of research papers. They provide not only sufficient information on research papers, but also search functions using author names and citations. However, because no personal identification is supported in those systems, the users are often annoyed by the noises caused by other persons with the same name. In order to solve the problem, we need a high quality information resource on researchers that are available for the identification of authors.

There are some databases on researchers which might be available. For instance, ReaD[2] is one of database services on Japanese research activities and it contains information on researchers. However, it provides about 200,000 researchers' information, only 20%-30% of the whole researchers in Japan.

There are many practical systems which use researchers' information for various applications, but most of them are using the existing information which is available at hand. While some systems are aiming at collecting homepages, they usually focus on specific domains and seek for high precision [3,4,5], paying little attention to the low recall.

In addition, if a list of researchers' names is available, it is not difficult to search for each researcher through conventional search engine such as Google. However, they are not suited for gathering homepages of many and/or unspecified researchers. Therefore, it is meaningful to collect as many researchers' homepages as possible from the web, both for compiling and maintaining researchers' database and for further applying them to other practical systems.

In this paper, we propose a method to collect unspecified researchers' homepages with high recall as well as high precision. We begin with the background in section 2 and introduce the overall structure of the proposed method in section 3. Section 4 shows detailed explanation of the methods and the related experiment results for rough filtering of the possible web pages and Section 5 for accurate evaluation of the web pages and precise selection of the correct homepages. We then discuss the future work in section 6 and conclude in section 7.

2. Background

There are several kinds of researchers' information and there may be more than one homepages for a researcher. Here we focus on collecting unspecified researchers' homepages in Japanese that contain information on their research activities such as research topics, publications and projects.

Unlike for a specified researcher, commonly used methods can not be used for searching for an unspecified researcher. A researcher’s name, for example, can not be used as a query for a search engine. Hence, the properties of researchers’ homepages must be learned from sample pages, such as content-based keywords, link structures, URL patterns, and so on.

As the working data, we use 100-gigabyte web document data NW100G-01[1][6,7,8], which were mainly gathered from ‘.jp’ domain for WEB Tasks[9] at the Third and Fourth NTCIR Workshops[10]. The number of web pages in NW100G-01 is 11,038,720. We use both the link list attached to the document data and the full-text index of the document data generated by “namazu”[11].

We prepared sample data of researchers’ homepages from NW100G-01 document set. 113,380 pages were gathered with some typical Japanese family names as keywords, and 10 percent of them were judged by one of the authors one by one and classified to positive and negative classes. In this paper, we call them **sample data**. The composition of the sample data is shown in Table 1. They are divided into two parts in order to meet the different goal in different processing afterwards.

Table 1 Information of sample data

Total		First half		Second half	
#positive data	#negative data	#positive data	#negative data	#positive data	#negative data
403	10,936	214	5,456	189	5,480

As we mentioned above, human names can not be used as queries directly in collecting unspecified researchers’ homepages. However, the sample data collected with some human names can be used for training and evaluating our method which does not use human names in any way.

3. Overall structure of the method

The proposed method aims at creating a high quality collection of researchers’ homepages. It consists of three phases, through which the researchers’ homepages are gathered with high recall as well as high precision from the vast amount of the web pages. Figure 1 shows the outline of the process.

(1) Rough filtering of the possible web pages

This phase is to collect web pages which are possibly researchers’ homepages. The result is mainly evaluated in terms of the recall and relatively low precision is allowed. Because whole web pages are to be processed, efficiency is highly required. Thus, the goal is to filter out obviously irrelevant web pages fast enough with a simple processing.

(2) Accurate evaluation of the web pages

The web pages gathered in the previous phase are accurately evaluated in terms of likelihood of researcher’s homepage. Relatively heavy processing is allowed in this phase. The goal is to calculate a score which represents the likelihood appropriately.

(3) Precise selection of the correct homepages

The web pages scored in the previous phase are labeled as “positive”, “negative”, or “possibly positive”. The “possibly positive” pages are further to be judged by hand. Ideally all the pages should be labeled as either “positive” or “negative”. However, since there are unpredictable varieties among the web pages, such a pragmatic approach may be necessary to guarantee both recall and precision at given levels. Thus, the goal is to find best thresholds, one between positive and possibly positive ones and the other between possibly positive and negative ones.

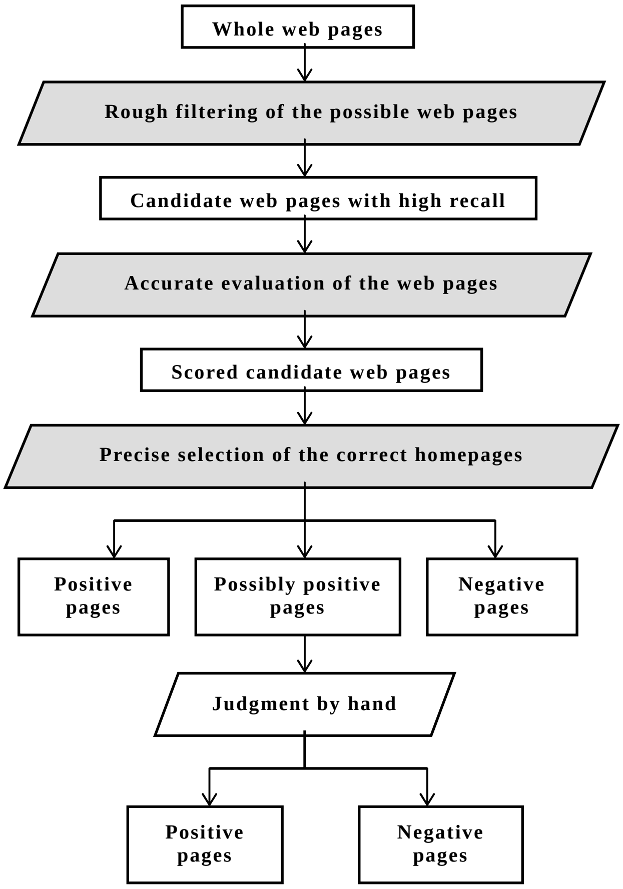
Whole web pages
Rough filtering of the possible web pages
Candidate web pages with high recall
Accurate evaluation of the web pages
Scored candidate web pages
Precise selection of the correct homepages
Positive pages
Possibly positive pages
Negative pages
Judgment by hand
Positive pages
Negative pages

Figure 1 Overview of the proposed method.

[1] NW100G-01 is available for research purposes from National Institute of Informatics. See <http://research.nii.ac.jp/ntcir/permission/perm-en.html>.

Since we are still at the preliminary investigation stage for the second and the third phases, we handle them as a single step referred to as “Evaluation and selection of the correct homepages” hereinafter.

4. Rough filtering of the possible web pages

4.1 Content-based keyword-lists

For different kinds of homepage, the contents in and the styles of the kind of homepage tend to be quite different. In the meanwhile, for different researchers, the situation is the same. To find a simple and general method for gathering possible researchers’ homepages, first of all the contents in the first half of sample pages are studied. Besides the words of number, symbols, particles, dates (“平成”, “月”, etc.) and place names, all the words appeared in both positive and negative pages are classified into two sets. One set consists of “property words” that support the properties of general researchers and that researchers’ homepages should contain logically. The other set consists of general words which can appear in many kinds of general web pages. The appearance of the words segmented by “Chasen” are counted and all the property words and the examples of general words in the top 100 frequent words in both positive and negative pages are listed in Table 2. The statistical result of the word frequency shows that property words may be contained in both positive and negative pages and the most frequent words may be both property words and general words. Therefore, it is difficult to use the statistical method directly to select and weight the keywords for collecting candidate researchers’ homepages.

Table 2　Frequent words in the sample data

	Property words of researchers’ homepages (number)	Examples of general words
Positive pages	研究，学会，論文，教授，大学院，報告，大学，協会，学部，担当，研究所，分野，博士，共著，専門，所属，講座，テーマ，助教授 (18)	経済，社会，国際，科学，文化，情報…
Negative pages	教授，論文，研究所，出版，講座，発表，専門，担当，課題，専攻，講義，演習 (12)	経済，構造，学習，基礎，方法，人間，管理…

In addition, by observing on the content of first half of sample data, we found there are two kinds of keywords that can be used for collecting candidate researchers’ homepages, “on-topic” keywords and “off-topic" keywords. The on-topic keywords are almost same as the property keywords. The off-topic keywords are the words that do not appear in researchers’ homepages. We did the experiment to extract the words that appeared in negative pages but not in positive pages, and found out several off-topic keywords such as “案内”, “サービス”,”お願い” etc. Though the off-topic keywords did not appear in the positive sample data, we decided just to use the on-topic keywords but not the off-topic keywords for two reasons. One is the number of sample pages is limited and we cannot say the statistical reliability is high enough. The other reason is there is the risk to miss some researchers’ homepages if we exclude the pages that contain the off-topic keywords, since any off-topic keywords can appears in the title part of the publications.

As the result, 12 properties are selected and the corresponding content-based keyword-lists are defined using only the on-topic keywords. The keyword-lists are shown in List 1 and the line under each keyword-list is its explanation. Each keyword-list may contain several synonyms corresponding to a researcher’s properties expected to be included in his/her homepage, and consists not only keywords extracted from the sample data but also keywords obtained with some dictionaries and some existing code tables. In the 12 keyword-lists, keyword-list 0 is quite different from the others, in that the word(s) in “keyword-list 0” is (are) the word(s) on general concepts that might appear in a researcher’s homepage while the words in the other keyword-lists are the specific words that represent the researcher’s properties. The reason for using keyword-list 0 as a separate keyword-list will be explained in the latter part of Section 4.2.

We collected the pages from all the working data using all keywords in each keyword-list. We use “namazu” as a tool. The keywords are used as partial match to the index. For example, all the web pages that contain the words including the character string “研究” in any place are retrieved using the following command:

```
namazu -a -s -l '*研究*' index-file
  > output-file
```

This is the reason for there is the keyword “教授” but not “助教授” in the keyword-list 3. By this means, we can not only decrease the amount of keywords in each keyword-list but also avoid missing the potential keywords. Then we got 12 result files in which the document IDs corresponding to the web pages are listed.

List 1　Keyword-lists for rough filtering

Keyword-list 0: 研究

General words characterizing a researcher;

Keyword-list 1: 研究内容，研究対象，キーワード，課題，テーマ，研究活動，研究紹介

Words related to research topics;

Keyword-list 2: 専門，分野，学科，専攻

Words related to the major of researcher;

Keyword-list 3: 博士，教授，講師，助手，技官，研究員，リサーチアソシエート，ポスドク，ポストドクター

Words related to specific position of researcher;

Keyword-list 4: 資格，職名，役職，現職

Words related to general position of researcher;

Keyword-list 5: メンバー，教官，スタッフ，職員，研究者，教員，構成員

Words related to general word as a member;

Keyword-list 6: 略歴，学歴，職歴，学位，履歴，卒業，経歴，プロフィール, profile，自己紹介，出身

Words related to general words used for researchers' history;

Keyword-list 7: 大学，研究所，研究センター

Words related to the organization that the researcher belongs to;

Keyword-list 8: 論文，報告，著書，共著，出版，発表，賞，業績，プロジェクト，研究成果，フィールドワーク，特許，フィルドワーク, paper

Words related to researcher's achievement;

Keyword-list 9: 所属，部門，講座，研究室，実験室, lab, laboratory，学群，学系，学部，研究科

Words related the concrete section that the researcher works at;

Keyword-list 10: 講義，ゼミ，授業，クラス，科目，演習

Words related to education activities the researcher is going for;

Keyword-list 11: 学会，会員，協会，研究会

Words related to the societies the researcher joined to.

List 2　Statistical results of first half of sample data

#keyword list	#pages out of working data	#pages out of 214 (#increased)	percentage of pages out of 214 (%increased)
12	3754	21(21)	9.81%(9.81%)
11	16747	49(28)	22.90%(13.08%)
10	38567	86(37)	40.19%(17.29%)
9	75302	125(39)	58.41%(18.22%)
8	135602	156(31)	72.90%(14.49%)
7	229376	173(17)	80.84%(7.94%)
6	373504	197(24)	92.06%(11.21%)
5	591111	208(11)	97.20%(5.14%)
4	915546	212(4)	99.07%(1.87%)
3	1420452	213(1)	99.53%(0.47%)
2	2311853	214(1)	100.00%(0.47%)
1	4303023	214(0)	100.00%(0.00%)

We calculated the number of the pages that are contained in N or more files (N=1..12), i.e. the pages containing keywords from N or more different keyword-lists. The statistical results are given in List 2.

The statistical results show that 100% training data contain at least 2 keywords from different keyword-lists and more than 99% training data contain at least 4. It means the definition of keyword-lists is feasible.

4.2 Rules for retrieving candidate pages

Our object is to gather all the possible pages but the least amount of irrelevant web pages with a rule set as simple as possible. In the current work, C4.5 [12] is used as a tool for finding out the rules that are hidden in the sample data. The first half of sample data is used as training data for learning rules by C4.5 and the second half of sample data is used as tuning data for relaxing the rules to cover all the positive samples. For the training data, positive and negative data constitute two classes used in C4.5 and the presence of the keywords in each keyword-list (whether at least one keyword in each keyword-list appears or not) is used as the features of each page.

When generating rules by C4.5, the rate of false negative (positive data that are classified into negative data) must be kept very low whereas the rate of false positive (negative data that are classified into positive data) may be relatively high, because it is required to keep the recall very high in the meanwhile to decrease the number of irrelevant pages in the gathered candidate pages. In the experiments, the positive data are weighted 160 times larger than the negative data and the classification purity is set as 98 percent when running C4.5 so that the false negative is kept as zero and the generated rules are simple.

After studying the result tree generated by C4.5, the rules can be reduced based on the class of leave nodes. The procedure to reduce the rules is as below (In the procedure, "C*n*" represents the condition of containing at least one keyword from keyword-list *n*).

1. choose all the leaves that are classified as 1(positive class);
2. write down the path for every leaf chosen in step 1 with intermediate nodes that are to include the keywords, for example, the rule of (C0 & C1 &C2);
3. ignore the rules which are included in other rules, for example, the rule of (C0 & C1 & C2 & C3) is ignored in condition the rule of (C0 & C1 &C2) exists;

4. apply the rules to the tuning data for testing the recall. If the recall is 100%, then go to step 7;
5. if the recall is improved than before then refresh the rule set, otherwise recover the rule set as before;
6. relax the rules with the following steps then go to 4:
    (1) find a group of rules which have the most common conditions, for example, (C0 & C1 & C2 & C3) and (C0 & C1 & C2 & C4) constitute a group because they have the most common conditions of (C0 & C1 & C2);
    (2) make a candidate rule consisting of the common conditions and replace the group of rules with it;
7. fix the rule set.

Applying the procedure to the rules generated from the result tree of C4.5 using the tuning data, a rule set containing 10 rules were finally obtained and the recall of the tuning data was 100 percent. The final rule set is listed in List 3.

List 3 The final rule set

C0 & C1 & C2,
C0 & C1 & C3,
C0 & C1 & C6,
C0 & C1 & C9,
C0 & C2 & C6,
C0 & C2 & C8,
C0 & C7 & C9,
C2 & C10,
C3 & C2 & C6,
C3 & C8 & C11.

All the rules in the final rule set satisfy the criterion showed in the statistical result in List 2 on the end of Section 4.1, in that 9 rules out of 10 contain three keywords in different keyword-lists and only one rule contains two keywords.

Next, the final rule set is applied to the whole working data and 661,454 possible researchers' homepages, only 5.99% of total amount of working data, are collected. Because the number of collected pages is less than 915,546 (8.29%), the number of pages that contain keywords from at least four different keyword-lists and is only little more than 591,111 (5.35%), the number of pages that contains keywords from at least five different keyword-lists, the rule set is considered to be effective.

In addition, we experimented on generating and relaxing rules using 11 keyword-lists, all the keyword-lists given in List 1 except keyword-list 0, in the same way as the previous one. The results of the two experiments are compared and shown in Table 3.

Table 3 Results comparison of using 12 keyword-lists to using 11 keyword-lists

Case	Result of 12 keyword-lists	Result of 11 keyword-list
#Rules	10	8
#Rules with 2 conditions	1	6
#Rules with 3 conditions	9	2
#collected pages	661,454	820,180
Percentage out of total pages	5.99%	7.43%

The results comparison shows it is necessary to use "keyword-list 0" as a separate keyword-list. It is the words of general concepts that remedy the insufficiency of just using the property words and make it possible to achieve our goal that to generate rules based on the keyword-lists for gathering all the possible researchers' homepages in the least amount of web pages.

The exact evaluation on the recall of the proposed method is a future issue. As the number of the positive sample data is small and we used all of them for training and tuning, we need to prepare more sample data for evaluation later.

5. Method of evaluation and selection

5.1 Baseline: content-based classifier

For evaluation and selection of the collected candidate researchers' homepages, SVM light is used as the tool both for classifier model learning and for classifying. The classifier model is learned from the first half of the sample data as the training data, and the second half of the sample data is used for evaluation as the testing data. Positive and negative data of both training data and testing data are labeled as "1" and "-1" respectively. The feature sets are generated in two ways: one is to use only nouns included in the page, which is called **feature set 1**; and the other is to use nouns, word bi-gram of nouns and the occurrence counts of the other types of part of speech, which is called **feature set 2**. Here the morphological analyzer 'ChaSen' [14] is used as a tool for segmenting Japanese text and labeling each word a part of speech. The value of a feature is the presence of the corresponding word (or word bi-gram) in a page, that is to say the value is 1 if the feature word is present in the page else the value is 0. The experiment results on content-based classifiers are listed in Table 4.

Table 4 The experiment results on the sample data

feature set ID	feature set 1	feature set 2
precision	79.88%	83.87%
recall	71.43%	68.78%

Both the precisions and the recalls are not high enough and could not satisfy the ideal requirement. We assume the reasons are that the positive sample data are all single page data and content-based only features are used for evaluating the web pages. It is insufficient to use only the content words as features in the single pages.

5.2 Utilization of logical page group

Since Web pages are more complicated than usual documents, in addition to the content-based features, properties specific to the web should be taken into account to improve the precision of the classifier [15]. To extract logical page groups proposed in [16] is an approach because a logical homepage may consists of multiple physical pages as a group, which is called a logical page group. It is expected to be effective to consider logical page groups by utilizing the web-based features, such as link structures, URL patterns, anchor texts, etc. between the top page and the other component pages in a logical page group.

There are many cases in that, although no physical pages in the logical page group contain the sufficient content-based keywords on researchers, one of the physical pages can be deemed as the top page (or the entry page) of the logical page group and the top page can be regarded to logically contain all the contents of all the physical pages in the logical page group. Although it may be easy to achieve high recall and high precision only by using the content-based features in the case that the homepage consists of only a single page, it often does not hold for the case of logical page group.

Because it is difficult to judge the extent of logical web pages for general cases, a method to improve precision and the recall at the same time by finding the possible top page first, then propagating the content-based keywords in each component physical page to the possible top page. In this way, a virtual page of the logical page group can be formed and can be evaluated with the same method presented in Section 5.1. This method will improve both the precision and the recall.

We are currently doing the related experiment, hoping that the precision can be improved with the proposed method compared to the baseline content-based classifier. After a satisfactory classification performance is achieved, we will investigate on the methods of the accurate evaluation of the web pages and the precise selection of the correct homepages separately.

6. Future work

For the rough filtering of possible web pages, though the method used for relaxing rules in section 4.1 is simple, we have to verify it from an objective view. We will do an experiment using the second half of the sample data as the training data and the first half of sample data as the tuning data and relaxing the rules in the same way. Then, if the result rule set is the same as that shown in List 3, we can say that the rule set is stably obtained with the proposed method. Otherwise, a more appropriate and objective algorithm should be proposed for relaxing the rules.

For accurate evaluation of the web pages, we use SVM light as the classifier in the current work. The classifier other than SVM light will be considered later for comparing the measures. The method for improving the precision of classifier we proposed in section 5.2 remains to be verified.

Besides this method, other methods should be investigated too. For example, to analyze the link structures including member list pages might be an effective way to improve the likelihood of each web page. In addition, some pages contain contents of more than one researcher's information and they can be treated as more than one researchers' homepages logically. Therefore if we can investigate the method to confirm this kind of pages, it will be a supplementary method to improve the precision of the classification.

For precise selection of the correct homepages, we have to investigate a method for classifying the candidate pages into three classes, which is to decide both the threshold between positive pages and possibly positive ones and the threshold between possibly positive pages and negative ones.

After we have achieved satisfactory results with the proposed method on the researchers' homepages, we will try to apply the method to other page categories. We hope it would also be applicable to page categories under conditions that only a single entity is described in each page and that specific property words can be defined.

7. Conclusions

The method we proposed in this paper is for collecting unspecified researchers' homepages from a vast amount of the web data with high recall as well as high precision, aiming at creating a high quality collection of researchers' homepages. The proposed method consists of three phases: (1) rough filtering of the possible web pages; (2) accurate evaluation of the web pages; and (3) precise selection of the correct homepages.

The key techniques of the rough filtering for collecting unspecified researchers' homepages with high recall within relatively small amount of web pages are the definition of appropriate content-based keyword-lists, the usage of a decision tree for generating the rules and the relaxation of the rules based on heuristics.

The key point of the accurate evaluation would be that the virtual page of the logical page group which is composed by merging the content words in the component physical pages with the top page is appropriately evaluated.

Though we have not finished all the experiments on the proposed approach yet, we hope it could be a practical method for collecting large amount information from the web for further utilization.

Acknowledgements

We used NW100G-01 document data as our working data under permission from the National Institute of Informatics. We would like to thank Professor Akiko Aizawa of NII for precious advices and technical supports.

References

[1] NEC and Penn State, Computer and Information Science Papers CiteSeer Publications ResearchIndex, http://citeseer.ist.psu.edu/cs

[2] Japan Science and Technology Agency, ReaD Directory Database of Research and Development Activities:, http://read.jst.go.jp/EN/

[3] Y. Yang, S. Slattery and R. Ghani, A Study of Approaches to Hypertext Categorization, Journal of Intelligent Information Systems, Vol.18, Issue 2-3, pp. 219-241, March-May 2002.

[4] K. Matsuda and T. Fukushima, Task-oriented World Wide Web Retrieval by Document Type Classification, Proc. 8th international conference on Information and knowledge management, pp. 109-113, Kansas City, Missouri, United States, 1999.

[5] T. Kuno, T. Agata, E. Ishida and S. Ueda, The Judge Method of Web Page Type, http://www.slis.keio.as.jp/~urda/webir/index-j.html

[6] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama, Overview of the Web Retrieval Task at the Third NTCIR Workshop, NII Technical Report, No.NII-2003-002E (Jan. 2003).

[7] K. Eguchi, K. Oyama, E. Ishida, N. Kando, K. Kuriyama, Overview of the Web Retrieval Task at the Third NTCIR Workshop, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan (Mar. 2003).

[8] K. Eguchi, K. Oyama, E. Ishida, N. Kando, K. Kuriyama, Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure, IEICE Transactions on Information and Systems, Vol.E86-D, No.9, pp.1804-1813 (Sep. 2003).

[9] NTCIR WEB Task (NTCIR-WEB), http://research.nii.ac.jp/ntcweb/

[10] NII-Test Collection for IR Home Page, http://research.nii.ac.jp/ntcir/

[11] Namazu: a Full-Text Search Engine, http://www.namazu.org/index.html.en

[12] Machine Learning-Decision Trees-C4.5 Tutorial, http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/c4.5.html

[13] ChaSen's Wiki - FrontPage, http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[14] SVM-Light Support Vector Machine, http://www.cs.cornell.edu/People/tj/svm_light/

[15] P. C. Marco, C. Edleno, M. Nivio, Z. Berthier, R. Marcos and A. Gonçalves, Combining Link-based and Content-based Methods for Web Document Classification, Proc. 12th International Conference on Information and Knowledge Management, pp. 394-401, New Orleans, LA, USA, 2003.

[16] K. Tajima, K. Hatano, T. Matsukura, R. Sano and K. Tanaka, Retrieving and Organizing Web Pages by "Information Unit", Proc. 10th International Conference on World Wide Web, pp. 230-244, Hong Kong, 2001.