Regular Paper

Combining Page Group Structure and Content for Roughly Filtering Researchers' Homepages with High Recall

Yuxin Wang[†] and Keizo Oyama[†,††]

This paper proposes a method for gathering researchers' homepages (or entry pages) by applying new simple and effective page group models for exploiting the mutual relations between the structure and content of a page group, aiming at narrowing down the candidates with a very high recall. First, 12 property-based keyword lists that correspond to researchers' common properties are created and are assigned either organization-related or other. Next, several page group models (PGMs) are introduced taking into consideration the link structure and URL hierarchy. Although the application of PGMs generally causes a lot of noises, modified PGMs with two original techniques are introduced to reduce these noises. Then, based on the PGMs, the keywords are propagated to a potential entry page from its surrounding pages, composing a virtual entry page. Finally, the virtual entry pages that score at least a threshold number are selected. The effectiveness of the method is shown by comparing it to a single-page-based method through experiments using a 100 GB web data set and a manually created sample data set.

1. Introduction

Collecting and utilizing information from the Web is becoming increasingly crucial activity now since the World Wide Web has become the primary source of knowledge for people all over the world. Research information has been one of the typical examples since its beginning.

Some search engines and database services provide information on research papers through search functions using keywords, author names, citations, etc. However, they do not support the identification of the authors. This lack of support indicates that a high-quality information resource regarding researchers is necessary.

Information on researchers themselves is also a useful research resource. For instance, some application systems [1),2)] would be more effective if information on all the researchers were made available. Thus, researchers' homepages on the web would be a useful resource. It is, however, very difficult to comprehensively collect them using existing methods.

One reason is that, in the web, information of an entity can be presented on a single page or a set of pages that constitutes a logical page group. Therefore we are investigating a system for building a collection of researchers' homepages that can handle them not only in single pages but also in logical page groups.

Although some methods that take into consideration only the contents of single pages perform to a certain degree in the case of collecting homepages in single pages, they perform rather poorly in the case of collecting homepages in logical page groups. On the other hand, although methods that take into consideration the global link structure (e.g., in/outlink references, and anchor texts across web sites) are quite effective for collecting popular researchers' homepages, they are almost of no use in the cases of homepages rarely referred to by others. Therefore, in addition to the contents of the pages, the local structure among the pages must be considered.

Since there has been no way to guarantee the completeness of web-based information, up till now almost all the related applications have only been best-effort-types, especially in terms of recall. Therefore we must stress high recall in the investigated system, since it is crucial especially for guarantee-type applications utilizing its output.

In this paper, we propose a rough filtering method for gathering researchers' homepages with a very high recall by applying page group models for combining the structure and content among the pages in a page group. As will be described in Subsection 3.1, the method is applied to a component of the investigated system to work as the first processing step. The method is simple enough to narrow down the candidate

† The Graduate University of Advanced Studies (SOKENDAI)
†† National Institute of Informatics (NII)

pages from a huge amount of web pages with a very high recall. Its output is then further processed by the following steps using much more complicated and costly methods to accurately collect the target pages.

The effectiveness of this proposed method compared to a single-page-based method is shown through experiments using a 100 GB web data set and a manually created sample data set with various parameters.

The rest of this paper is organized as follows. Related work is introduced in Section 2, and then a task description of this work, including the background, goal, and the related items, is given in Section 3. Next, the construction of the proposed method, property-based keyword lists, and page group models with the considered parameters are described in Sections 4, 5 and 6 respectively. Section 7 shows the experimental results and considerations on the performances, and Section 8 finally summarizes the present work and discusses our future work.

2. Related Work

As will be described in Subsection 3.2 and Section 4, we try to gather all the probable researchers' entry pages by exploiting the link structure and URL hierarchy, together with types of the property-based keyword lists, from a huge amount of the web pages. Our method proposed in this paper could be deemed as a web page search method as well as a web page classification method, both use a set of property-based keywords and take into consideration web page groups. There are many related research works and those mentioned below are just examples.

Regarding research on web page search methods, Oyama et al. described a method for searching the web pages of particular categories by adding domain-specific keywords called "keyword spices" to the input query and forwarding them to a general-purpose search engine [3]. In a sense, our idea of using property-based keywords is similar to their keyword spices; however, their purpose is to reduce the number of irrelevant pages on rather specific topics, whereas our purpose is to gather all the possible pages in one category. Accordingly, our keyword selection policy and usage of keywords differ.

Matsuda et al. introduced a method for task-oriented web information retrieval utilizing document classification by various page characteristics, such as content, out-link, URL, and so on, and succeeded in reducing task-irrelevant pages [4]. However, all of the characteristics they use are extracted from individual pages and no page group structure is utilized.

Rosell analyzed methods for web information retrieval through the link structure of the web graph and concluded that analyzing the link structure can among other things, improve search engine results, and find clusters of pages with similar topics, etc [5]. Li et al. presented an algorithm to efficiently retrieve information units that can perform progressive query processing by considering both semantic similarity and link structures [6]. Each of these works tries to utilize some measures obtained by analysis of the structures among the web pages for reducing the number of irrelevant pages from the search results. Contrarily, our method uses the local web structure among the pages to collect useful keywords for comprehensively gathering candidate pages, while keeping the increase in irrelevant pages at a relatively low level.

Regarding research on web page classification, Calado et al. reported that global link information, such as co-citation, can achieve a high performance in regards to web classification [7], and Wang et al. showed that in-link reinforcement and anchor window can improve the quality of web page clustering by effectively utilizing the global link structure [8]. However, as mentioned in Section 1, the effectiveness of global information for achieving high recall is limited.

Sun et al. proposed an iterative web unit mining method for finding and classifying web units of web pages [9]. Each of the iterations includes two steps: merging sub-units and classifying the key page of the unit. However, the page content is only used for classifying individual pages, and is indirectly combined with structure-related information. Therefore the method has only a limited effectiveness for the improvement of the recall.

To use page group structure to collect information distributed in several related pages is a rather popular idea, and the applications of such ideas have been attempted for web search techniques and others. All the works referred to in this section, except the first two, exploit structural information (i.e., global/local links, anchor text and URL hierarchy) in some way together with page content information; however, none of them tries to apply the above-

mentioned idea. We know of no work right now that has succeeded in making the idea work for general cases, because of the difficulty in excluding irrelevant information, or noises.

Nevertheless, our method does apply the idea and works to a certain degree. The reason is that, as will be shown in Section 4, our method exploits the mutual relations between the page content and relative page location in generating virtual pages by propagating keywords. The key is our original techniques introduced in modifying the simple PGMs as will be described in Section 6.

3. Task Description

3.1 Background

Our final goal is to obtain a high quality collection of researchers' homepages that can be used for extending a guarantee-type information service, such as CiNii☆, not only as link targets but also as reference data for record linkage [10]. Considering the applications, we set the overall target performance of the investigated system at 95% recall and 99% precision.

We did a preliminary experiment to classify researchers' homepages using the manually created positive and negative samples that are described in Subsection 3.3, using SVM-Light☆☆ with features based only on the content words of individual pages, resulting in 59.6% recall and 90.15% precision.

Starting from this point, we will have many things to do before achieving our overall goal. It is generally effective to use features exploiting link structure, directory structure, document tag structure and document semantic structure, among others. However, they are too costly to process for all the web pages. Therefore we split the process into two steps: rough filtering for efficiently narrowing down the page amount with a very high recall and accurate classification for achieving both high recall and high precision using rich features. If necessary, human assessment may further be involved in the process to achieve the performance goal. We will only discuss the rough filtering in this paper, and the other parts will be presented in the future.

3.2 Goal

To achieve the overall target performance given in the previous subsection, the rough filtering is required to achieve a recall of at least 98% and desirably 99%. Precision does not matter so much but the amount of output pages should be as small as possible as long as the recall is satisfied.

Since not only a single page but also an entry page of a logical page group can be a homepage, the rough filtering is for gathering both. In addition, since it is impossible to determine the extent of a logical page group in general cases, we will not try to identify the logical page groups in the current work.

The basic concepts necessary for clarifying our goal of the task are given below.

(1) Researchers' homepages

Although various styles and presentations are used to provide researchers' information, they contain several basic elements in common. We define a researcher's homepage as a web page whose main subject is the researcher and which includes his/her personal attributes such as the name, affiliation and address, and research activities such as the research topics, publications, majors, and academic societies. The information may be provided either in a single page or in a logical page group. In the latter case, the entry page is regarded as the homepage. The information may be lacking in some part, but must be sufficient to identify the researcher and to outline his/her research activity. A homepage is usually created by the researcher him/herself or by the organization he/she belongs to. The current work handles researchers' homepages written only in Japanese.

(2) Logical page groups

A logical page group is a group of web pages that consists of an entry page and one or more component pages, where none of them solely contains sufficient information on the topic and only where the entry page, together with the component pages, logically contains sufficient information. The component pages are not necessarily accessible from the entry page directly or indirectly. For example, if an entry page does not include the organization's name while a site top page includes it, then the site top page is considered to be a component page, even if there is no path from the entry page to the site top page. In the current work, a component page has to be either an in-linked page, an out-linked page, or a directory entry page in the directory path of the entry page and to exist in the same site.

☆ CiNii is a research paper navigation service provided by the National Institute of Informatics. For details, see http://ci.nii.ac.jp/.

☆☆ SVM-Light Support Vector Machine, http://svmlight.joachims.org/.

3.3 Data

The rough filtering takes all the web pages as input when it is in real operation. In the current work, however, we use the fixed set of data described below.

(1) Web data

For the experiments, we used a corpus of 100-gigabyte web document data, NW100G-01[☆], which was gathered from the '.jp' domain for WEB Tasks[☆☆] [11),12)] of the Third and Fourth NTCIR Workshops [13),14)]. It contains 11,038,720 web pages. We used the link list attached to the document data and the full-text index of the document data generated by "namazu[☆☆☆]". Note that when the keyword lists described later are fixed, we can use a string matching algorithm like Aho-Corasick's integrated into the web crawling process instead of "namazu" without loss of efficiency.

(2) Sample data

We prepared sample data from the NW100G-01 document set. We first collected 113,380 pages containing some typical Japanese family names and randomly selected 11,338 pages, 10% from them (hereinafter we call this set of 11,338 pages as **Jname data**). Each of the pages was then manually assessed by the authors according to its content and, if necessary, the contents of the in/out-linked pages. Consequently, we obtained 426 positive samples and 10,912 negative samples.

It should be noted that since the rough filtering stresses recall, the sample data must be prepared with sufficient care as not to introduce biases in any aspects. If we use precision-preferred methods to collect possible pages that are to be manually assessed, rather "easy" pages are collected more and "difficult" pages less. The final recall would thus be inaccurate and overestimated. Since we could find no other bias-free way to efficiently collect positive sample pages, we used only typical Japanese family names, under a hypothesis that the presence of a person's name is independent from other characteristics of a researcher's homepage.

It should also be noted that since the sample data are manually assessed, some entry pages of the logical page groups that should be judged as positive might be overlooked because they include very few or even no clue words. Indeed, existences of such pages have been confirmed, as mentioned in Subsection 7.2.

On the other hand, we randomly selected 1,103 pages (0.01%) from the corpus, checked if each page contained any personal names, and found just 400 pages. We can thus estimate that the total number of pages in the corpus that contain a person's name is 4,000,000 (i.e., 400/0.01%). Since the size of Jname data is 113,380 pages, its sample size is approximately 0.283% (i.e., 113,380/4,000,000).

4. Construction of the Method

The proposed method uses property-based keyword lists and several kinds of page group models. **Figure 1** illustrates the simplified construction of the method.

Each web page is first mapped to a **document vector** consisting of binary values, each of which corresponds to a keyword list and represents if any of the keywords in the keyword

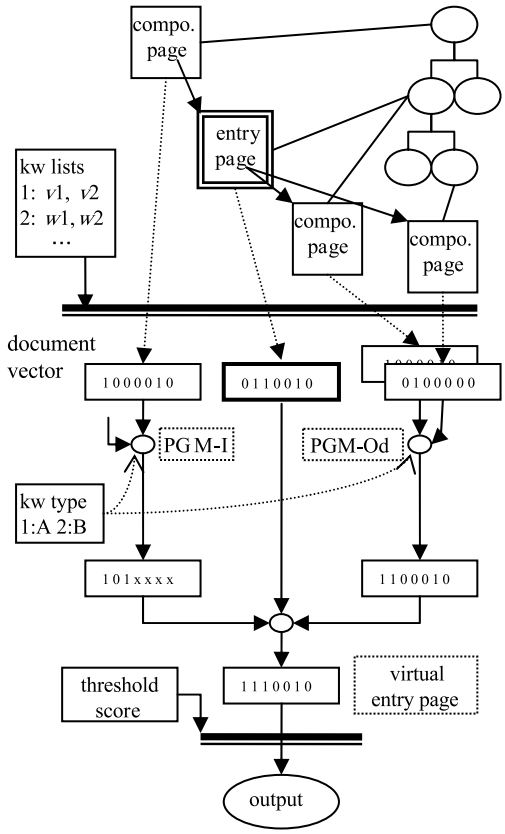

Fig. 1 Construction of the proposed method.

☆ NW100G-01 is available for research purposes from the National Institute of Informatics. See http://research.nii.ac.jp/ntcir/permission/.

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

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

list are found in the web page.

Next, for each of the page group models, the document vectors are merged by making a logical sum of each vector element. In this process, only the elements corresponding to the keyword lists of suitable types for the page group model are considered (in the figure, ignored elements are indicated with 'x' at the output from PGM-I). They are further merged to the page's document vector in the same way to compose a final document vector. Here, a conceptual document represented by the final document vector is called a **virtual entry page**, and the process to merge the document vectors of the individual pages to that of a virtual entry page is called **keyword propagation**.

Finally, **scores** of virtual entry pages are obtained by counting the number of 1's in the document vector, and those that scored more than or equal to a **threshold score** are output. The threshold score will be set to as high as possible by considering the evaluation results so that the recall is satisfactory for the supposed applications, e.g., 99.0%.

It should be noted that we do not use popular IR techniques, such as TF-IDF or probabilistic models, for two main reasons: term frequency makes no sense for property-based keywords, only the presence does; and property-based keywords are essentially popular terms and their specificities are of no use.

5. Property-based Keyword Lists

We cannot use specific search terms such as the person names and organization names because specific search targets are not provided. The search engine techniques used in the "Homepage Finding Task" [15] of TREC Web Track are therefore not applicable for gathering researchers' homepages. However, although the styles and structures of homepages tend to differ greatly, they usually contain several basic information elements in common. Therefore, we use property-based keywords representing properties common to a certain number of researchers and expected to be included in their homepages.

Content-based keywords are usually extracted from sample data by means of statistical techniques. However, if we try to extract property-based keywords from the sample data, the following problems would be encountered: (1) collecting a complete set of appropriate keywords is difficult because preparing a sample data set of a sufficient size is very expensive, especially for positive data; (2) the logical relationship between the keywords and the definition of a target data is unclear; and (3) tuning for improving the potential recall of unseen types of target data is impossible.

To overcome these problems, we use several sets (or lists) of manually selected words and phrases (or terms) characterizing the researchers' properties. We mainly use property-name-related terms. Property-value-related terms are used only when they can be enumerated within a small number, otherwise their maintenance would require a lot of effort.

The conceptual procedure for creating the keyword lists is shown below. All the steps are manually executed and thus it is only a conceptual one.

A. Create keyword lists corresponding to researchers' properties commonly described in their homepages by studying the contents in positive sample data and the database structure of ReaD☆, etc.

B. Select a small number of keywords for each keyword list from the contents of the positive sample data.

C. Add synonyms associated to the newly selected keywords using a general Japanese dictionary.

D. Apply each of the keyword lists to the positive samples; check the contents of the pages that include no keywords in the keyword list, and select new keywords from the pages, if there are.

E. Add the new keywords to the existing keyword list. Create a new keyword list if there isn't an appropriate one. Split an existing keyword list into two if it starts to include keywords that are too diverse.

F. Repeat steps C through E until no more keywords remain to be selected.

Using the above-mentioned procedure, we created nine keyword lists in step A and consequently obtained 12 keyword lists containing 86 keywords in the current work.

Each of the keyword lists is then designated as either organization-related or non-organization-related. Keyword lists corresponding to the properties common to the members in the same organization are desig-

☆ ReaD is a research information service including information of researchers in Japan. For details, see http://read.jst.go.jp/.

Table 1 Property-based keyword lists and keyword samples.

Type	Keyword list	Keyword example*
non-organization-related	general word	research
	research topic	research topic, theme, etc.
	title	doctor, professor, etc.
	position	present position, duty, etc.
	history	biography, personal history, etc.
	achievement	paper, bibliography, etc.
	lecture	course, seminar, etc.
	academic society	academic society, regular member, etc.
organization-related	major	major, specialty, research field, etc.
	member	staff, member, etc.
	organization	university, institute, school, etc.
	section	section, department, etc.

*Original keywords are in Japanese.

nated as organization-related, while keyword lists corresponding to individual researcher's properties are designated as non-organization-related. The types and meanings of these 12 keyword lists are listed in **Table 1** along with some keyword examples. Note that the actual keywords are in Japanese.

In order to confirm their quality roughly, we conducted a T-test to evaluate the effectiveness of each selected keyword. The result indicated that 76 out of 86 keywords (83.7%) are statistically significant at 95% confidence level.

In addition, using a statistical technique we tried to extract other useful keywords from both the positive and negative sample data. However, many of the extracted words are university names, era names, places, department names, and the like; only eight of the words may be the appropriate candidates. We do not use university names, for instance, as keywords since a complete list of the university names is hard to maintain over a long time because they change rather frequently.

6. Page Group Models

To achieve high recall in gathering researchers' homepages, we need to take into consideration the structure in individual logical page groups. We propose four simple page group models (PGMs) using in-links, out-links, and relative URL hierarchy between an entry page and a component page respectively, as well as using directory entry pages in the URL directory path.

In this section, we present concepts and definitions of the PGMs. Related notations are listed in **Table 2**, and the definitions of PGMs are listed in **Table 3**. It should be noted that in all PGMs, only the pages in the same site are considered.

Single page model (**SPM**) is a baseline that uses keywords in individual pages only. Single site model (**SSM**) is a baseline of the simplest PGM that uses keywords in all out-linked pages in the same site. They are compared to the proposed PGMs in order to evaluate the effectiveness of proposed PGMs.

PGM-Od is intended to exploit all kinds of keywords in out-linked component pages in the lower levels of the directory subtree. **PGM-Ou**, **-I**, and **-U** are intended to exploit organization-related keywords from component pages in the upper levels of the directory path, PGM-Ou for out-linked pages, PGM-I for in-linked pages, and PGM-U for directory entry pages and site top pages, respectively.

Parameters used in each PGM are also listed in Table 3. Here, s and l specify the ranges of the directory levels to propagate the keywords from. The upper bounds of l for PGM-Od, -Ou, -I, and -U are 2, 4, 3, and 8 respectively, since in the corpus all the pages in the lower and upper directories that are out-linked from any of the positive samples are within 2 and 4, all the in-linked pages of any of the positive samples are within 3, and the uppermost level of directory entry pages of all the positive samples is 8.

PGMs generally propagate many keywords irrelevant to the researcher to the virtual entry page and consequently include many noise pages. Modified PGMs, which are our originalities and key techniques, are thus introduced to reduce such noises, whereas to keep useful keywords propagated.

PGM-Od@θ is a modified PGM derived from PGM-Od with the intention of excluding irrelevant pages introduced by PGM-Od, based on the observation that one of the noise sources is groups of many pages mutually linked within a directory, and that an entry page having many

Table 2 Notations.

Notation	Definition
$P_{\mathrm{out-link}}(r)$	set of pages having link from page r (r's out-linked pages) in the same site
$P_{\mathrm{in-link}}(r)$	set of pages having link to page r (r's in-linked pages) in the same site
$P_{\mathrm{down}}(r,s,l)$	set of pages in directories s to l levels lower in the directory subtree of page r
$P_{\mathrm{up}}(r,s,l)$	set of pages in directories s to l levels upper in the directory path of page r
$P_{\mathrm{dir-ent}}(r,s,l)$	set of entry pages of directories s to l levels upper in the directory path of page r
$P_{\mathrm{site-top}}(r)$	set of site top page(s) of page r
$N_{\mathrm{Lod}}(r)$	number of links from page r to the pages in the same directory and in directories lower in the directory subtree of page r

Note: The level of the same directory is defined as 0.

Table 3 Definitions for PGMs and parameters.

Models		Description	Propagated pages	Parameters
SPM (baseline)		Single page model; no keyword propagation is used.	—	
SSM (baseline)		Reference page group model; all out-linked pages in the same site are used.	$P_{\mathrm{out-link}}(r)$	
Simple PGM	Od(s,l)	A PGM based on out-links downward; out-linked pages in the directory subtree are used.	$P_{\mathrm{out-link}}(r) \cap P_{\mathrm{down}}(r,s,l)$	$s=0,1$, $l=s..2$
	Ou(s,l)	A PGM based on out-links upward; out-linked pages in the directory path are used.	$P_{\mathrm{out-link}}(r) \cap P_{\mathrm{up}}(r,s,l)$	$s=0,1$, $l=s..4$
	I(s,l)	A PGM based on in-links upward; in-linked pages in the directory path are used.	$P_{\mathrm{in-link}}(r) \cap P_{\mathrm{up}}(r,s,l)$	$s=0,1$, $l=s..3$
	U(s,l)	A PGM based on directory entry pages; site top and directory entry pages of the directory path are used.	$P_{\mathrm{site-top}}(r) \cup P_{\mathrm{dir-ent}}(r,s,l)$	$s=0,1$, $l=s..8$
Modified PGM	Od@θ	Od with additional conditions on the number of out-links; if there are too many out-links, Od is not used.	If $N_{\mathrm{Lod}}(r) \leq \theta$, same as Od; otherwise, same as SPM.	$\theta = 5, 10, 20$
	Ou#,I#,U#	Ou, I, and U, each propagating organization-related keywords only.	Same as Ou, I, and U for organization-related keywords; for others, same as SPM.	

Note: r is a possible entry page; s and l specify the ranges of the directory levels.

out-links within the directory subtree always contains sufficient keywords in itself.

PGM-Ou#, **-I#**, and **-U#** are modified PGMs derived from PGM-Ou, -I, and -U respectively, with the intention of excluding irrelevant keywords based on an observation that non-organization-related keywords are not included in the upper directory hierarchies. Therefore only organization-related keywords are propagated within these PGMs.

7. Experiments and Considerations

Using the data described in Subsection 3.3, we experimented on PGMs with various parameters. In Subsection 7.1, it is shown that despite the use of PGMs, our method can reduce the page amount to an allowable level when compared to SPM. In Subsection 7.2, the effectiveness of our method over SPM is shown by performing additional assessment on the candidate pages that are output by the proposed method but had been judged as negative in the first manual assessment. Then, in Subsection 7.3, considerations on the overall performance of the proposed method are given.

7.1 Performance on Manually Assessed Samples

In this subsection, the results from various experiments are shown in several graphs. All the graphs include SPM and SSM plots for comparison. The x-axes are the page amount $n_c(i)(1 \leq i \leq 12)$, namely, the number of pages in the corpus that scored at least i. The y-axes are recall defined by $n_p(i)/N_p$, where N_p is the total number of positive sample data, $n_p(i)$ is the number of positive sample data that scored at least i.

For each plotted curves, the uppermost and

rightmost data corresponds to threshold score 1, and every next one corresponds to a threshold score incremented by 1. In general, a higher recall and a less page amount indicate better performance; however, we put priority on the recall.

Since the amount of positive sample data is not large, in order to grasp the accuracy of observed recall values and to clarify their targets, their confidence intervals should be considered in the following experiments. With the sample size fixed to 426, the confidence intervals for the observed recalls are shown in **Table 4**. In other terms, in order to assure at least 98% recall with 90% confidence, the observed recall must be a little higher than 99%☆.

(1) Comparison of simple PGMs

First, we experimented on individual simple PGMs with typical parameters in order to understand their basic performances. The results for PGM-Od(0,2), -Ou(0,3), -I(0,3), and -U(0,3) are shown in **Fig. 2**. It shows that all the simple PGMs deteriorate in their page amounts than SPM since a lot of noises are introduced by the keyword propagation, but with fewer noises than when compared to SSM.

(2) Effects of modified PGM-Od

Next, we experimented on simple and modified PGM-Od's with several s, l, and θ values. Some of the results are shown in **Fig. 3**. Only for cases where $l = 2$ are shown because the changes of l made almost no difference for each combination of s and θ. For $s = 1$, only PGM-Od(1,2) is shown since the modified PGM-Od results for every θ were almost the same as for simple PGM-Od.

The figure indicates that if we select $s = 1$, the page amount is almost at the same level as SPM, whereas the recall value for each threshold score is completely the same for SPM. This implies that almost all non-organization-related keywords are collected from within the same directory. Thus, $s = 1$ is useless for PGM-Od. On the other hand, focusing on the recall area of around 99%, if we select $s = 0$, the page amount increases by 80% over SPM with simple PGM-Od whereas modified PGM-Od can reduce the increase down to 50%.

Although smaller θ tends to result in a lesser page amount, since PGM-Od is the only PGM that propagates non-organization-related keywords, the parameters should be carefully selected and will be further investigated in the next subsection.

(3) Effects of modified PGM-Ou, -I, and -U

Then, we experimented on modified PGM-Ou, -I, and -U, and compared them with corresponding simple PGMs with typical parameters.

The results of PGM-Ou and -I are shown in **Fig. 4** and **Fig. 5**, respectively. Only cases where $l = 3$ are shown because the changes of l made almost no difference for each s. The re-

Table 4 Confidence intervals of observed recalls.

Confidence	95%			90%		
Observed recall	97.0	98.0	99.0	97.0	98.0	99.0
Upper bound	98.2	98.9	99.6	98.0	98.7	99.5
Lower bound	94.8	96.0	97.6	95.2	96.3	97.9

Note: Recalls are in percents.

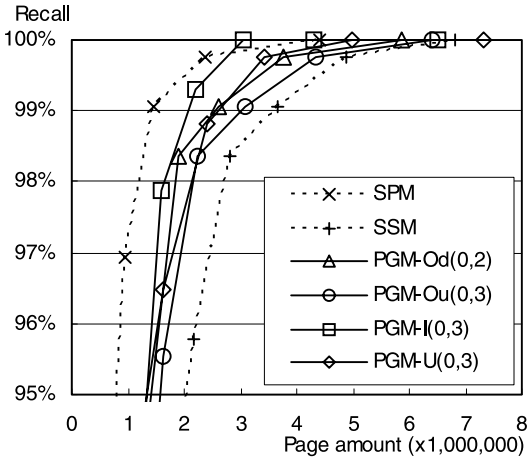

Fig. 2 Performance of simple PGMs.

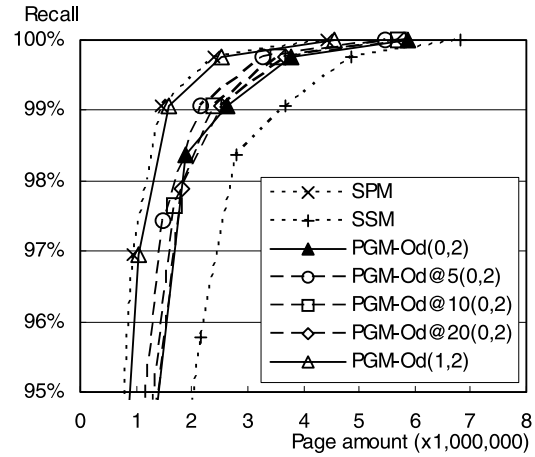

Fig. 3 Performance of simple and modified PGM-Od's.

☆ More details are described in Appendix A.1.

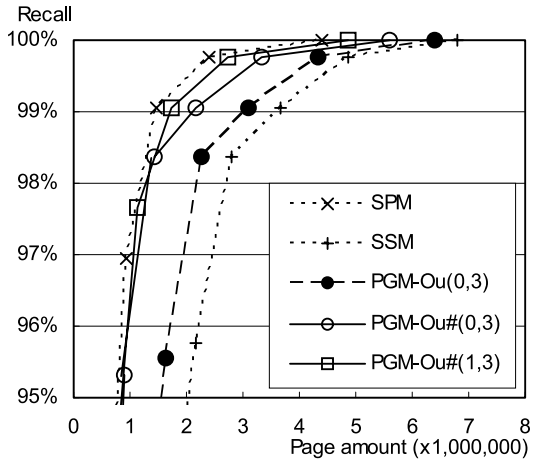

Fig. 4 Performance of simple and modified PGM-Ou's.

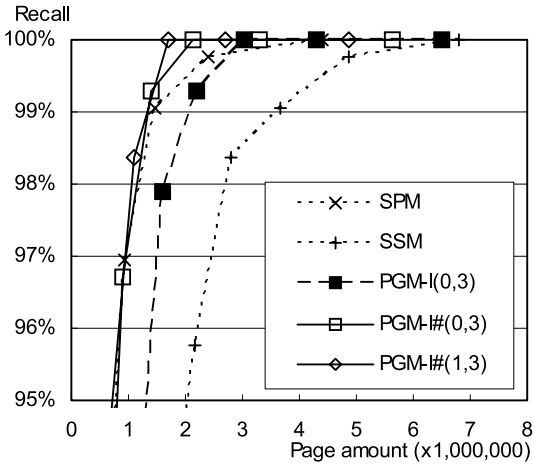

Fig. 5 Performance of simple and modified PGM-I's.

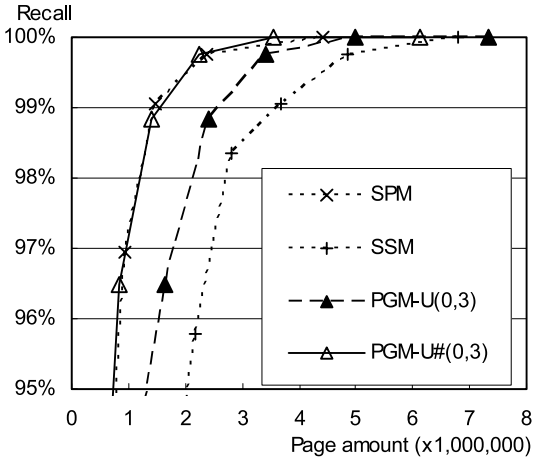

Fig. 6 Performance of simple and modified PGM-U's.

sults of PGM-U are shown in **Fig. 6**. Only for cases where $s = 0$ and $l = 3$ are shown because the changes of l made almost no difference for each s and the changes of s made only small shifts along the corresponding plots.

Focusing on the recall area of around 99%, although the page amount increases by 40% to 120% over SPM with each simple PGM, modified PGMs can reduce the increase to almost the same level as SPM.

(4) Effects of combining PGMs

Finally, we experimented on combinations of PGMs with several promising parameter sets.

We selected parameters for each PGM based on the following policy: if the difference of page amounts in the recall area of around 99% is small between two parameter sets, then the one that collects keywords from more pages should be selected.

Furthermore, for PGM-Od, we select $s = 0$ because of the reason mentioned in item (2), although modified PGM-Od still collects a rather large amount of noise pages.

Since pages in the same directory have many chances to be used as keyword sources, combinations of $s = 1$ as well as $s = 0$ were tested for the other PGMs.

Figure 7 shows the run results of three combinations of PGMs selected from those that performed relatively well. We will refer to each of them hereinafter as follows:

PGM-C1: PGM-Od@5(0,2),Ou#(1,3), I#(0,3),U#(0,3)
PGM-C2: PGM-Od@10(0,2),Ou#(1,3), I#(0,3),U#(0,3)
PGM-C3: PGM-Od@20(0,2),Ou#(1,3), I#(0,3),U#(0,3)

Note the scale of the x-axis is twice larger than the other figures.

Each of them uses all four modified PGMs with the same parameters except for θ of PGM-Od. As $s = 0$ is used for PGM-Od, $s = 1$ is selected for PGM-Ou. All the other parameters were eventually the same for all combinations.

The results show that even the best performed run PGM-C1 is inferior to SPM in all the recall ranges except for 100%. On the other hand, it is also shown that the proposed method reduced the page amount to a certain degree despite its use of PGMs.

However, since some true positive data are overlooked in the manual assessment as men-

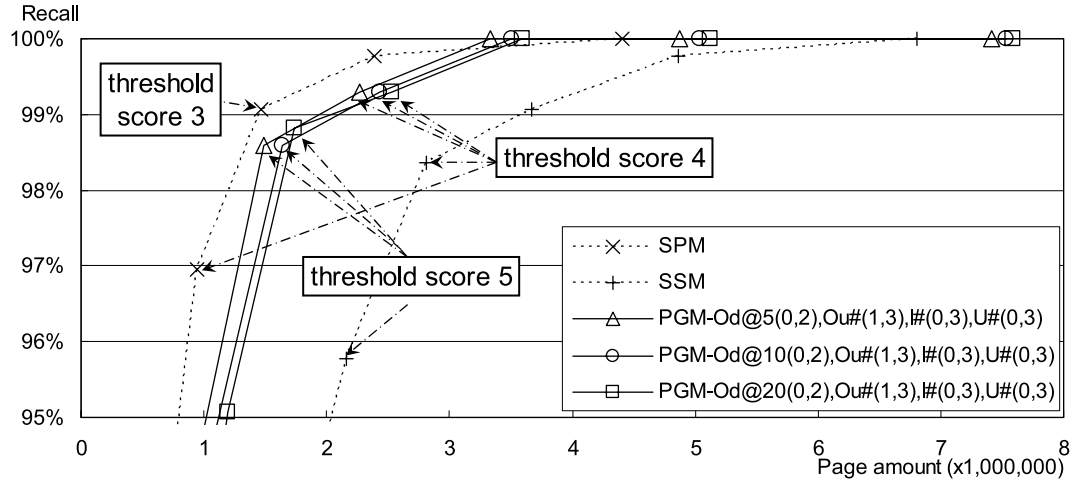

Fig. 7 Performance of selected PGM combinations.

Table 5 Overlooked positive pages per score.

		Score for SPM			
		0	1	2	Total
Scores for PGM-C1 through -C3	4	1	2	0	3
	5	2(1)	5	0	7(6)
	6-12	1	1	1	3
	Total	4(3)	8	1	13(12)

Note: Each cell indicates the number of positive pages. All the numbers are same for PGM-C2 and -C3. The numbers in parentheses are that for PGM-C1 only.

tioned in Subsection 3.3 and their effects are not shown in Fig. 7, the effectiveness of the proposed method is not clearly shown with the results. Thus, we will further investigate this issue in the next subsection.

7.2 Ability to Find Overlooked Homepages

In this subsection, we evaluate the ability of the proposed method to find positive pages that were overlooked in the manual assessment using additional assessments to confirm the effectiveness of the proposed method.

First, the pages that were contained in Jname data and scored less than 3 with SPM but scored at least 4 with PGM-C3 were assessed and 13 new positive data were found. Then, for each of these newly found pages, we checked the scores for SPM and PGM-C1 through -C3 respectively. The page numbers for each score are shown in **Table 5**.

This partly shows the ability of the proposed method to find positive pages that cannot be gathered by SPM even if we select the threshold score 3 so that the recall is more than 99% for the manually assessed positive samples.

7.3 Considerations

First, as mentioned in Subsection 3.3, Jname data had a sample rate of 0.283%. Then, the total number of positive data in the corpus can be estimated as 426/0.00283=150,530 pages. With this estimate, the precision is calculated by dividing 150,530 by each page amount.

Table 6 summarizes the results for PGM-C1, -C2, and -C3. The precision values are rather low when compared to the state-of-the-art web page classification methods. However, considering the very high recall, the performance is considered to be fairly good. The "Reduction ratio" is the proportion of the page amount to the corpus size. They are not satisfactory but show that the processing cost of the following processes can be greatly reduced.

Second, when taking into account the new 13 positive data shown in Table 5, SPM's recalls at threshold scores 2 and 3 should be corrected from 99.8% (425/426) to 97.0% (426/439) and from 99.1% (422/426) to 96.1% (422/439), respectively. By comparing these values with the recalls of the proposed methods at threshold scores 4 and 5 respectively, it is obvious that the proposed methods outperform SPM with 95% confidence. Furthermore, four positive pages cannot be gathered with SPM even if the threshold score is set to 1. This implies SPM can hardly achieve the goal recall with a feasible page amount.

Third, a failure analysis on all three pages that scored only 3 with PGM-C1 through -C3 revealed the following facts. All of them are researchers' introduction pages from the same site that is officially provided by a uni-

Table 6 Summary of experiment results.

PGM	Threshold score	Recall	Precision*	Page amount	Reduction ratio
-C1	4	99.3%	6.6%	2,265,478	20.5%
	5	98.6%	10.2%	1,482,980	13.4%
-C2	4	99.3%	6.2%	2,429,250	22.0%
	5	98.6%	9.2%	1,635,703	14.8%
-C3	4	99.3%	5.9%	2,530,850	22.9%
	5	98.8%	8.7%	1,738,404	15.7%

* Precisions are estimates.

versity department. Their page styles are similar and contain minimum information, scoring only 2 with SPM. Although they have hyperlinks to the researchers' personal homepages, our method cannot exploit them because they exist in separate sites. However, they are expected to be gathered with our method instead, although they were not actually gathered because they were not included in the corpus. The facts support that for applications where only an informative homepage suffices when multiple homepages exist for a researcher, the proposed method must have worked if their personal homepages had been crawled.

Finally, as there are trade-offs between the recall and page amount, in general it is difficult to say which of PGM-C1, -C2, and -C3 is the best. In order to guarantee that the overall recall will be more than 98%, according to the discussion in the first part of Subsection 7.1, we should set the threshold score to 4. We will eventually select PGM-C2 as the most appropriate for the current goal, because the recall at threshold score 4 is the same for PGM-C2 and -C3.

8. Conclusion and Future Work

We described a method for comprehensively gathering probable researchers' homepages from the web within as few pages as possible. We proposed a method of using property-based keyword lists combined with four page group models. Two original key techniques were introduced to reduce irrelevant keywords to be propagated by exploiting the mutual relations between the content and structure among pages in a logical page group.

We evaluated the method by comparing to a single-page-based method through experiments using a 100 GB web data set and a manually created sample data set with various parameters. It could successfully reduce the increase of gathered page amount to an allowable level despite its use of a page-group-based method, which generally causes many noises. Then, the proposed method was shown to be able to gather a significant number of positive pages that could not be gathered with a single-page-based method.

The proposed method is considered to have fulfilled the goal set for the rough filtering. When the overall system described in Subsection 3.1 is completed, its outcome can be utilized to realize guarantee-type applications.

Our future research may include:

- Find a more systematic way for modifying the property-based keywords and the set of properties.
- Try combinations of individual keyword lists with other possible keyword types in a systematic way.
- Implement a full-operational system for the method and apply it to the real web data.
- Attempt to apply the method to other categories. We expect the method is applicable if only a single entity is described on each entry page and if specific properties and corresponding keyword lists can be prepared.

To tackle the diversity of web data pursuing a very high recall is a challenging problem. The approach presented in this paper is just the first step. As there are various issues remaining to be solved, a great deal of challenges should be continuously made to solve them.

Acknowledgments We used NW100G-01 document data under permission from the National Institute of Informatics. We would like to thank Professors Akiko Aizawa and Atsuhiro Takasu of NII for their very helpful advice and technical support.

References

1) Mori, J., Matsuo, Y., Ishizuka, M. and Faltings, B.: Keyword Extraction from the Web for FOAF Metadata, *Workshop Notes of 1st Workshop on Friend of a Friend, Social Networking and Semantic Web*, Galway, Ireland, pp.1–8 (2004).
2) Matsuo, Y., Tomobe, H., Hasida, K. and Ishizuka, M.: Mining Social Network of Con-

ference Participants from the Web, *Proc. 2004 IEEE/WIC/ACM International Conference on Web Intelligence* (*WI 2003*), IEEE, p.190 (2003).
3) Oyama, S., Kokubo, T. and Ishida, T.: Domain-Specific Web Search with Keyword Spices, *IEEE Transactions on Knowledge and Data Engineering*, Vol.16, No.1, pp.17–27 (2004).
4) Matsuda, K. and Fukushima, T.: Task-oriented World Wide Web Retrieval by Document Type Classification, *Proc. 8th International Conference on Information and Knowledge Management* (*CIKM'99*), Missouri, United States, pp.109–113 (1999).
5) Rosell, M.: A Brief Introduction to Information Retrieval through the Link Structure of the Web Graph. http://www.nada.kth.se/rosell/courses/ia/term˙paper.pdf
6) Li, W., Candan, K., Vu, Q. and Agrawal, D.: Retrieving and Organizing Web Pages by "Information Unit", *Proc. International WWW Conference* (*10*) (*WWW 2001*), pp.230–244 (2001).
7) Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B. and Goncalves, M.A.: Combining Link-Based and Content-based Methods for Web Document Classification, *Proc. 12th International Conference on Information and Knowledge management* (*CIKM'03*), New Orleans, Louisiana, USA, pp.394–401 (2003).
8) Wang, Y. and Kitsuregawa, M.: Enhancing Contents-Link Coupled Web Clustering and Its Evaluation, *Proc. Data Engineering Workshop 2004* (*DEWS2004*), pp.5–B–05 (2004).
9) Sun, A. and Lim, E.: Web Unit Mining: Finding and Classifying Subgraphs of Web Pages, *Proc. 12th International Conference on Information and Knowledge management* (*CIKM'03*), New Orleans, Louisiana, USA, pp.108–115 (2003).
10) Aizawa, A. and Oyama, K.: A Fast Linkage Detection Scheme for Multi-Source Information Integration, *Proc. International Workshop on Challenges in Web Information Retrieval and Integration* (*WIRI 2005*), Tokyo, pp.30–39 (2005).
11) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop, Technical Report NII-2003-002E, National Institute of Informatics (2003).
12) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure, *IEICE Trans. Inf. Syst.*, Vol.E86-D, No.9, pp.1804–1813 (2003).
13) Oyama, K., Ishida, E. and Kando, N.(eds.): *Proc. Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering* (*Sep. 2001 – Oct. 2002*), Tokyo, National Institute of Informatics (2003). (online, available from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html)
14) Kando, N. and Ishikawa, H.(eds.): *Proc.Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization* (*Apr. 2003 – Jun. 2004*), Tokyo, National Institute of Informatics (2005). (online, available from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/index.html)
15) Hawking, D. and Craswell, N.: Overview of the TREC-2001 Web Track, *Proc. 10th Text REtrieval Conference* (*TREC 2001*), Gaithersburg, Maryland, NIST, pp.61–67 (2002). (available from http://trec.nist.gov/pubs/trec10/papers/web2001.ps.gz)

Appendix

A.1 Confidence Interval of Observed Recalls

Although the sample size 426 is large enough, since confidence intervals are considered at a very high recall range, a normal distribution cannot be applied, and therefore a binomial distribution should be used. **Figure 8** shows the confidence intervals at 80%, 90%, and 95%, exactly calculated with a binomial distribution for the observed recall value x.

Note that it is accurate on condition that the samples are randomly selected from the population, and that it does not take into ac-

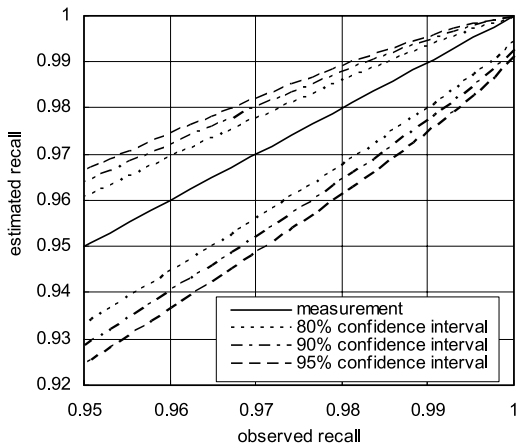

Fig. 8 Confidence intervals of recall for 426 samples.

count manual assessment errors, false-positives or false-negatives.

(Received September 19, 2005)
(Accepted March 13, 2006)

(Editor in Charge: *Hiroshi Ishikawa, Masayoshi Aritsugi, Kaoru Katayama, Yutaka Kidawara, Masashi Tsuchida*)

Yuxin Wang is a Ph.D. student of the Department of Informatics at The Graduate University of Advanced Studies (SOKENDAI), Tokyo, Japan. She received the B.E and M.E degrees from East China Normal University, Shanghai, China in 1990 and 1993, respectively. Her research interests include web information utilization, information retrieval and natural language processing. She is a student member of IEICE and DBSJ.

Keizo Oyama received the B.E., M.E. and Dr.Eng. from the University of Tokyo in 1980, 1982 and 1985, respectively. He is a professor of National Institute of Informatics (NII), Japan and the Graduate University for Advanced Studies (SOKENDAI). His research interests are structured text processing, web retrieval systems and full-text search technologies. He is a member of IPSJ, IEICE, JSIMS and DBSJ.