PLF: A Publication List Web Page Finder for Researchers

Kai-Hsiang Yang, Jen-Ming Chung, Jan-Ming Ho
Institute of Information Science,
Academia Sinica, Taipei, Taiwan
{khyang, jenming, hoho}@iis.sinica.edu.tw

Abstract

Finding and keeping track of other researchers' publication lists is an essential activity for every researcher, because they often contain citations not found elsewhere and may provide access to information, such as slides and talks, which can help other researchers keep abreast of state-of-the-art knowledge and technology. There are many different ways to generate publication list web pages, and a researcher may have several different versions of a publication list on the Web because he holds different positions. So it is difficult to find the correct publication list web page from the top results retrieved from search engines, especially when we only know the name of the researcher. Very few works have addressed the problem. In this paper, we propose a system called the "Publication List Web Page Finder" (PLF), which can automatically find the publication list web pages for a given researcher's name. The PLF system is an automatic and language-independent system, and its main idea is that publication list web pages often contain many citations about a specific researcher, so the system uses those citations as clues to find out publication list web pages. Our experimental results show that the PLF system outperforms other approaches, especially when a researcher has multiple publication list web pages.

1. Introduction

Researchers usually create their homepages on the Internet for various reasons, such as describing their research and contributions, or providing material for their new courses. A researcher's homepage usually contains his/her biography, course materials and research achievements, including all publications, projects and patents. Hence, finding and keeping track of other researchers' publications is an essential activity for every researcher, because in this way, it is possible to learn about state-of-the-art knowledge and technology from other researchers' publication lists.

Although this problem is really important, it is not easy to develop an automatic system to find out publication list web pages for a specific researcher because of the following reasons. (1) Researchers might be professors in universities, or work in high-tech companies. They may have publication lists in different formats, as well as multiple publication pages. (2) A researcher may have several versions of a publication list because of job changes. (3) A person's name appearing on a web page may be ambiguous because, in the real world, many people have the same name.

The most common way for researchers to generate their publication list web pages is to design their own homepages in web spaces provided by their institutes. They are responsible for maintaining the content, and usually provide a hyperlink connecting to their publication list web pages. For a number of reasons, this kind of list is usually more accurate and informative than those provided by digital libraries, such as Cite-Seer, ACM, and Google Scholar. First, some latest papers that have been accepted only appear on their own publication list web pages, and such information can not be found in digital libraries. Second, many researchers provide useful information about their works, such as related software libraries and PowerPoint files. Third, the publication lists in digital libraries often suffer from the name ambiguity problem, However, the problem also suffers from the name ambiguity problem, which means all the publications of researchers with the same name are grouped together. This problem has been addressed in [1-4].

To date, relatively little research has been conducted on the problem of finding publication list web pages. In contrast, several works have addressed the issue of homepage finding [5-9]. However, they can not solve the problem completely. Even though they can find the homepage correctly, they still need a method to identify which hyperlink that connects to the publication list web page. Hence, it is very difficult to develop an automatic and language-independent system. On the other hand, the most common way we use to find a publication list web page is to input a person's name to a search engine, such as Google, and then manually check the results one by one. However, the search performance is not very well, especially when we

only know the name of the researcher.

In this paper, we first define the problem of finding the publication list web pages for a researcher, and then propose a system, called the "Publication List Web Page Finder" (PLF), which finds a researcher's publication list web pages by inputting his/her name only. The key idea of PLF is that the publication list web pages often contain many citations about the specific researcher that we are interested in. The PLF uses citation records as clues for finding the publication list web pages. Moreover, since the method only depends on the results returned by search engines and digital libraries, it is language-independent. We conducted many experiments to evaluate the search performance of the proposed system. Compared with other two approaches that searching by Google, PLF achieves 79.2% for the recall metric when the parameters are set to $n = 5$ and $m = 40$. The results show that PLF is easy to implement and performs quite well.

The remainder of this paper is organized as follows. In Section 2, we defines the problem of finding the publication list web page, and in Section 3, we describe the system architecture of PLF in detail. Section 4 discusses our experiment methodology and results. Finally, in Section 5, we present our conclusions.

2. Problem definition

In this paper, we are particularity interested in the publication list web pages that researchers themselves are responsible for maintaining. So we provide the following definitions so that we can define the problem without any ambiguity.

The first item we consider is a citation string (CS), which is defined as follows.

Definition 1. Citation String (CS): a citation string refers to a structured record or semi-structured sentence that contains metadata about a publication.

Definition 2. Publication List: a publication list is a list of citation strings.

Definition 3. Affiliated Personal Publication List Web Page (APPL): Given a specific person name, a web page is called an "affiliated personal publication list web page" when it belongs to a researcher in real world and belongs to the affiliated web site that the researcher works in.

In this paper, we only focus on the set of ***APPL***. Note that, according to definition of ***APPL***, a publication list web page provided by an institute that a researcher has worked for also belongs to the ***APPL***.

3. System architecture

Figure 1 illustrates the system architecture of PLF, which comprises three components: (1) a citation string search component that collects the citation strings from digital libraries; (2) a web page search component that collects the hyperlinks of web pages from search engines by using the collected citation strings as queries; and (3) a ranking function that analyses the statistics of all the collected hyperlinks of web pages, and reports the results to the user.

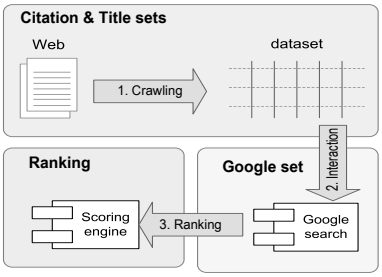


Figure 1. System architecture of PLF

Given a person's name, PN, the first component tries to collect the citation strings from digital library web sites, such as Google Scholar. Let $scholar(PN, m) = \{c_1, c_2, \ldots, c_m\}$ be the set of top m citation strings returned by querying the person's name PN to Google Scholar.

After the citation strings have been gathered, the second component collects all the hyperlinks of pages by querying the title of each citation string to a search engine. For each citation string c_i, let t_i denote the paper title of c_i and the set of t_i be denoted by $T(PN, m)$. For each t_i in $T(PN, m)$, the second component sends the t_i with double quotation marks as a query to a search engine, such as Google, and then retrieves the top n results. Let $Google(t_i, n)$ be the result set, and $L(PN, m, n)$ be the union of all $Google(t_i, n)$ which is defined as follows:

$$L(PN, m, n) = \bigcup_{t_i \epsilon T(N,m)} Google(t_i, n)$$

In the third step, the PLF system ranks each URL by the number of its appearances in $L(PN, m, n)$, and returns the most likely URLs to the user.

4. Experiments

To evaluate the performance of the PLF system, we chose the committee of the WWW2006 conference as our dataset by randomly collecting 200 names from the WWW2006 Conference Committee website. Each person in the dataset has several attributes, such as the institute or university he/she belongs to and research interests. We manually gathered all the publication list web pages, and divided the dataset into two groups according to the number

of publication list web pages. Researchers with more than web page, were classified as ***multi-group***, while those with only a publication list web page were classified as ***single-group***.

Table 1. Dataset distribution

APPL Types	#APPL	#people	%population
$others$	0	22	11%
$single-group$	1	120	60%
$multi-group$	2	35	17.5%
	3	16	8%
	4	7	3.5%

Table 1 shows the statistics of the dataset. 60% of researchers had a single publication list web page, while 29% had more than one publication list web page. We could not find publication lists for the remaining 11% by manual searching. Therefore, the data can be divided into three groups. There were 58 people in the ***multi-group***, and 120 people in the ***single-group***.

Two Google-based approaches are compared in our experiments. The first approach, called "GSE", simply inputs a person's name as a query to Google. The second approach, called "GESK", inputs the person's name and the word "Publication" as a query to Google. The two approaches are designed to simulate the search behavior of a normal user and an advanced user. In other words, a normal user only inputs the person name to a search engine, but an advanced user may input the person's name with the additional word "publication". To evaluate the performance of each approach, we consider the top-5 results and focus on the top-5 recall metric.

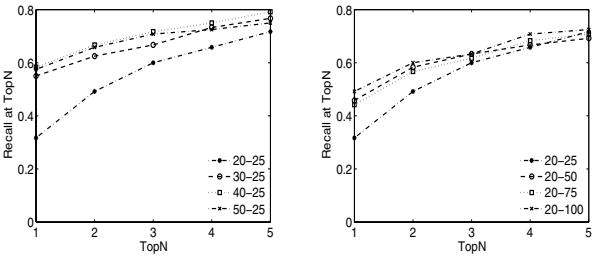
4.1. Experiment results

In this section, we first study the effect of parameters used in the PLF system, and then choose the best set of parameters to evaluate system performance. All the experiments are conducted on the two datasets, ***single-group*** and ***multi-group***.

4.1.1 Parameter analysis

We use two parameters in PLF: m, which refers to the number of citations gathered from Google Scholar; and n, which refers to the number of hyperlinks gathered from Google by using the title of each citation.

We first fixed $n = 25$, and changed m to evaluate the influence on the ***single-group*** dataset. Figure 2(a) shows the results. We observe that, when m increases, the recall also increases. In particular, when $m = 40$, the top-5 recall reaches about 80%, but when $m = 20$, the recall is only is 70%. The results show a clear and significant relationship between the number of citation strings crawled from Google Scholar and the system performance. It also verifies the assumption that a publication list web page usually contains a lot of citation strings. In our results, the best recall (about 79.2%) for top-5 results occurs when we set $n = 25$ and $m = 40$. When m increases to 50, the top-5 recall starts to decrease. This is because too much noise is gathered when using larger m.



(a) Fixed n with different m (b) Fixed m with different n

Figure 2. Parameter effect for $single - group$

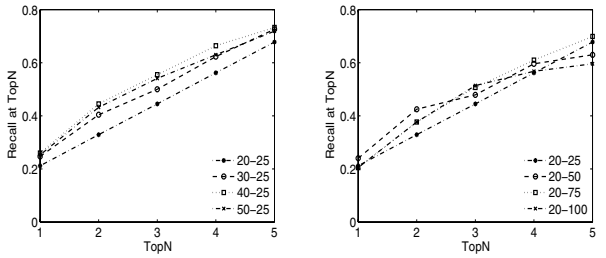
Again, we start by evaluating the effect of parameter n while fixing $m = 20$. Figure 2(b) shows the results. We observe that all the recall scores are around 70% for the top-5 results. This shows that if important citation strings have already been gathered, retrieving the top 25 results from Google is enough to find the correct publication list web pages. This result is very important for us when designing an on-line system because it saves processing time.

¿From the above experiments, we found that m significantly affects the recall metric for the ***single-group*** dataset. Specifically, when $m = 40$ and $n = 25$, the top-5 recall reaches 79.2%. The experiment results show that the impact of m is much higher than n. Hence, the PLF system is very efficient because it only needs to obtain enough citation strings from Google Scholar in order to retrieve a few results from the Google search engine.

We performed the same experiment on the ***multi-group*** dataset with $n = 25$. Figure 3(a) shows the results, and the performance while $m = 40$ is always better than the other settings. The reason is the same as for the results in ***single-group***. Again, we fix $m = 20$ to observe the effect of n on the ***multi-group*** dataset, as in Figure 3(b). The best performance (recall=70%) occurs when $n = 75$.

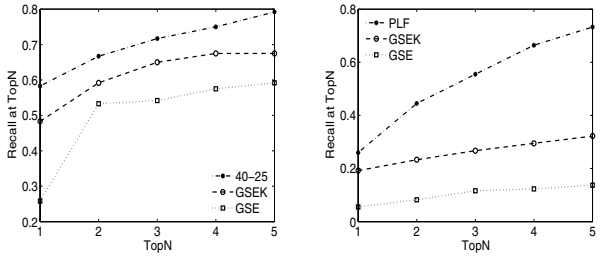
4.1.2 Performance evaluations

We now compare the results of PLF with those of the other two approaches for the different datasets. For the ***single-group*** dataset, we use the best setting of the parameters $(40, 25)$ for PLF. From the results in Figure 4(a), we observe that the PLF system always achieves better recall (the highest recall is 79.2% for the top-5 results). The GSE ap-



(a) Fixed n mix of different scale m (b) Fixed m mix of different scale n

Figure 3. Parameter effect for $multi - group$



(a) Performance of approaches in $single - group$ (b) Performance of different ways in $multi - group$

Figure 4. Performance evaluations

proach returns the worst results (recall is always less than 60%), while the GSEK approach achieves approximately 67.5%.

For the ***multi-group*** dataset, we use the best setting of the parameters $(20, 75)$ for PLF. From the results in Figure 4(b), we observe that PLF can achieve much better top-5 recall (around 73%) than other two approaches. In addition, it is interesting to note that the GSEK approach only achieves 32% top-5 recall, which means that using a search engine is not an effective way to find a publication list web page, even with an additional word.

We summarize our findings as follows. (1) The parameter m has a strong influence on the system's performance, but an oversized m may degrade the performance. (2) The parameter n has little influence on the system's performance. (3) The PLF system outperforms the other two approaches on both the ***single-group*** and the ***multi-group*** datasets.

5. Conclusion

Finding and keeping track of other researchers' publication lists is an essential activity for every researcher, because it is important to keep abreast of state-of-the-art knowledge and technology. The main contribution of this paper is that we propose a system, called the "Publication List Web Page Finder" (PLF), which can find publication list web pages by using person names only. The PLF system is automatic, language-independent, and most of all, easy to implement. Our experimental results show that it outperforms other approaches, especially when researchers have many publication list web pages. There are still two issues needed to be addressed in our future research. The first is how to deal with the name ambiguity problem, and the second is how to merge the multiple publication list web pages for a specific person into a single page.

Acknowledgment

This work was supported by the project of NSC 95-2221-E-001-021-MY3.

References

[1] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. *in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, 2001, pp. 250-257.

[2] B.-W. O. Dongwon Lee, Jaewoo Kang, Sanghyun Park. Effective and scalable solutions for mixed and split citation problems in digital libraries. *in IQIS*, Baltimore, Maryland, 2005, pp. 69-76.

[3] K.-H. Yang, J.-Y. Jiang, H.-M. Lee and J.-M. Ho. Extracting Citation Relationships from Web Documents for Author Disambiguation. *Technical Report (TR-IIS-06-017)*, Institute of Information Science, Academia Sinica, 2006.

[4] K.-H. Yang, K.-Y. Chiou, H.-M. Lee and J.-M. Ho. Web Appearance Disambiguation of Personal Names Based on Network Motif. *in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, Hong Kong, December, 2006.

[5] W. Kraaij, T. Westerveld, and D. Hiemstra. The Importance of Prior Probabilities for Entry Page Search. *in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, 2002, pp. 27-34.

[6] P. Ogilvie, and Jamie Callan. Combining Structural Information and the Use of Priors in Mixed Named-Page and Homepage Finding. *in Twelfth Text Retrieval Conference (TREC-12)*, 2003.

[7] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, vol. 21, pp. 286-313, 2003.

[8] V. N. Anh and A. moffat. Homepage Finding and Topic Distillation using a Common Retrieval Strategy. *in Eleventh Text Retrieval Conference (TREC-11)*, 2002.

[9] W. Xi, Edward A. Fox, Roy P. Tan, J. Shu. Machine Learning Approach for Homepage Finding Task. *in String Processing and Information Retrieval (SPIRE)*, Lisbon, Portugal, 2002, pp. 145-159.