A WEB SYLLABUS CRAWLER AND ITS EFFICIENCY EVALUATION

Matsunaga, Yoshihiro
Graduate School of Information Science and Electrical Engineering, Kyushu University

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

Ito, Eisuke
Computing and Communications Center, Kyushu University

Hirokawa, Sachio
Computing and Communications Center, Kyushu University

出版情報：Proceedings of International Symposium on Information Science and Electrical Engineering. 2003, pp.565-568, 2003-11
バージョン：
権利関係：

A WEB SYLLABUS CRAWLER AND ITS EFFICIENCY EVALUATION

Y. Matsunaga, S. Yamada

Kyushu Univ
Graduate School of ISEE
Hakozaki 6-10-1, Fukuoka 812-8581, JAPAN

E. Ito, S. Hirokawa

Kyushu Univ
Computing and Communication Center
Hakozaki 6-10-1, Fukuoka 812-8581, JAPAN

ABSTRACT

Many educational material are available on the web. Typical examples are syllabus pages of universities. A university and a department of the university usually provides a complete list of the courses which they offer. Therefore, syllabus data are not only typical but outstanding educational material with respect to quality and quantity. We are developing a syllabus database of Japanese universities. As the first step of the project, we developed a crawler that gathers syllabus pages on the Web. This paper describes a crawling mechanism which utilizes the characteristic keywords of syllabus pages. The efficiency evaluation is presented compared with general purpose crawlers.

1. INTRODUCTION

The progress of information technology and the infrastructure of telecommunications caused a spread of network and computers in many fields of the society. Education and training are not exceptions of this phenomenon. There are many educational material available on the web. Syllabus data of universities are typical examples. A university or a department of the university usually provides a complete list of the courses which are linked to syllabus pages. A syllabus page is a kind of contract document between students and the university, so that the quality is guaranteed. Syllabus data are not only typical but outstanding educational material with respect to quality and quantity. We are developing a syllabus database of Japanese universities. We have a plan to realize the database in the following four steps.

1. Feature analysis of syllabus pages and meta data construction
2. Construction of syllabus crawler and compilation of syllabus pages
3. Extraction of records from syllabus pages and integration into a database
4. Query system of syllabus data.

The first step is a preparation for constructing a crawler. We analyzed many syllabus pages and constructed a metadata for syllabus, which is represented as a list of keywords [1, 2].

A crawler is a program which gathers web pages following hyperlinks. There are too many web pages and we cannot waste computer resource and network. Therefore, the crawler has to avoid collecting irrelevant pages. A crawler for a specific target is called an intelligent crawler [3] or focused crawler [4].

This paper describes a focused crawler for syllabus pages. Efficiency is realized by a decision tree to recognize syllabus pages. This knowledge is used to control the behavior of the crawler to follow the hyperlinks.

2. CHARACTERISTICS OF SYLLABUS PAGES

There are many educational material on the web. As the first step of the analysis, we checked keywords that appear in syllabus pages of 25 Japanese university sites. NIAD (National Institution for Academic Degrees and University Evaluation) proposed an XML schema for syllabus data [1]. But there are many variation of the terms used in each syllabus pages, because each university can describes the web pages as they like. For example, keywords such as “担当教官”, “担当”, and “教官名” used to represent “teacher” of the metadata (Table 1). In spite of these variation, these keywords are not generic but specific to Web syllabus pages. So, we used these keywords to recognize syllabus pages.

As the second step to analyze the relationship between syllabus pages, we accumulated more pages from university sites. We collected 649 URLs which is derived form Google with the keyword “シラバス (syllabus)”. Then we ran a tiny program which follows hyperlinks from one of the 649 pages up to depth 5. The result is a list of 80446 pages from 452 sites. Out of these pages, we checked 4273 web pages, whose hostname begins with “www.a”, in detail. This analysis made it clear that 2738 web pages out of all 4273 pages are syllabus pages and that there are two kinds web pages related to syllabus data. One kind is a syllabus page itself. The other is a listing page of lectures which links to each syllabus page. We call the former “B-type” and the latter “A-type” in this paper.

共通項目名	対応項目名
担当教官 (teacher)	担当教官，担当，担当者，教官名，担当教員
授業科目名 (name of course)	授業科目名，授業科目，テーマ，研究主題，講義科目，科目名
概要 (outline)	概要，内容，授業目的，概要と目標，計画，講義の狙い
教材 (materials)	教材，教科書，参考図書，テキスト，関連ホームページ
関連科目 (related subject)	関連科目，予備知識，必要知識，受講条件，履修しておくべき科目，先履条件
キーワード (keyword)	キーワード，キー
授業コード (code)	授業コード，コード番号，ID
授業学期 (term)	授業学期，開講学期，学期
単位数 (credit)	単位数，単位
曜日と時間 (period)	日時，開講日
評価方法 (evaluation)	評価方法，評価，成績

Table 1. A Meta data for syllabus and variation of terms

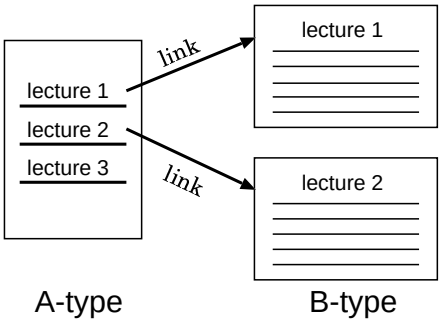

Fig. 1. A link structure of most syllabus sites: they have two types of pages. “A-type” is a list page which links to many syllabus page explaining each lecture. “B-type” is a syllabus page itself.

3. A WEB SYLLABUS CRAWLER

3.1. A Decision Tree for Web Syllabus

A page of “A-type” links to many syllabus pages. Therefore, we can efficiently collect syllabus pages by following links from a page of A-type. Thus, A-types are preferable to non A-types to accumulate web syllabus effectively. To realize such a strategy, we need a judgment method of high precision to tell whether a page is of A-type or not. A low precision may cause the explosion of traverse and the efficiency of the strategy might be worse than that of a random crawler.

We used the feature keywords of syllabus in Table 1 and 4273 web pages mentioned previously to constructed decision trees for A-type and B-type. The decision tree for A-type has the accuracy of 89% and the recall of 87%. The decision tree for B-type has the accuracy of 99% and the recall of 99%. This indicates that the precision of these decision tree is high enough.

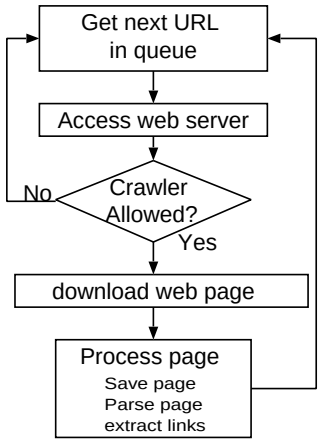

Fig. 2. Work flow of a typical Web Crawler

3.2. A Web Syllabus Crawler

We review a general framework of crawler before we explain our crawler that follows hyperlinks selectively. A Web crawler, also known as a robot or a spider, is usually implemented as a graph search algorithm, where nodes are web pages and edges are links between web pages. Each time a web page is downloaded, the page is parsed in order to find URLs which may be accessed in the next time. Figure 2 shows this work flow.

A Web crawler requires two lists of URLs. One contains URLs which are found during crawling but not collected yet. The other contains URLs which it has already collected. The second list prevents to download the same page that was downloaded before. Most crawlers use a FIFO queue for the first list of URLs. Hence the such crawlers follows the breadth first search. This behavior may be adequate to collect many web pages of wide variety and may be appropriate for general search engines. It provides not specific but general and broad search services with high coverage of the web by having a large index. For instance, a crawling strategy of Google, whose goal is to crawling important, or popular pages faster, finds mention in [5]. But it is not suitable for compiling pages of a specific topic. Our syllabus crawler teats the list as a priority queue, whose intention is to compile web pages linked from A-types. The goal of our syllabus crawler is to compile many web syllabus pages faster.

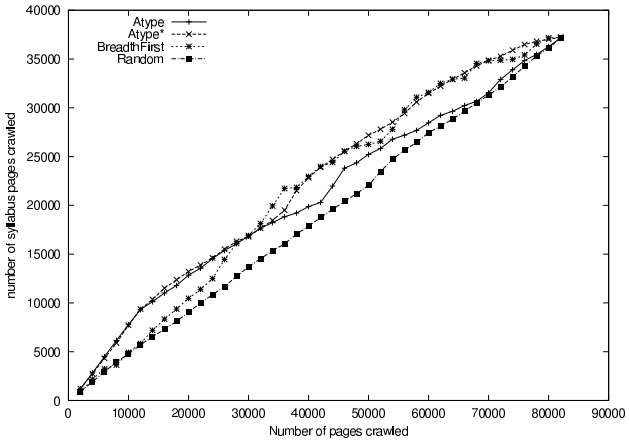

Fig. 3. A comparison of crawlers using number of syllabus pages crawled

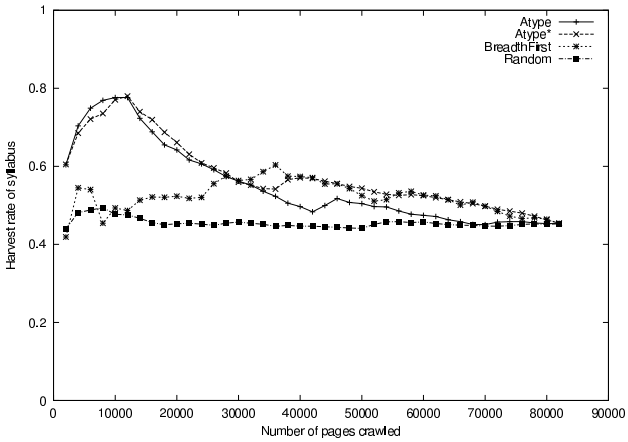

Fig. 4. A comparison of crawlers using curve of harvest rate: Our syllabus crawler(Atype) is more efficient than random crawler(Random) at all the time and most efficient at the beginning of crawling.

4. EVALUATION SYLLABUS CRAWLER

We ran experiments in order to evaluate the efficiency of the syllabus crawler. We used 80446 sample pages that we have collected for analysis of characteristics of syllabus pages as mentioned in the section 2. The collected data contain both syllabus pages and non-syllabus pages. According to the decision tree for B-type, we know that about 50% are syllabus pages. In the experiments, the crawler traversed hyperlinks between these web pages in local disk.

A focused crawler should collect many syllabus pages preferentially by avoiding non-syllabus pages. We evaluated efficiency of crawlers by the number of syllabus pages and by the harvest rate compared with the number of web pages collected. The harvest rate is the percentage of the syllabus pages over all web pages collected during crawling. We compared the efficiency of our crawler with that of the following crawlers. The “Breadth first crawler” chooses the next URL to be crawled from the FIFO queue of uncollected URLs. The “Random crawler” chooses the next target at random. The “modified A-type crawler” combines the methods of our syllabus proposed in section 3 and the breadth first crawler. It accesses the next URL which is linked from a A-type page if any and an oldest pushed URL if there is no such URL. These crawlers begin crawling from the start URLs of 649 web pages in the section 2. The number of syllabus pages is counted by the decision tree stated in the section 2.

Figure 3 and 4 illustrate the result of this experiments. Figure 3 displays the number of syllabus pages. Figure 4 plots the harvest rate, which is the percentage of syllabus pages over the total pages collected by the crawler. Notice that the crawler (A-type) is more efficient than that of random crawler (Random).

One measure of the performance is the number of web pages which is necessary to be crawled in order to compile 10,000 web syllabus pages. Our syllabus crawler accomplished it at 15,000 web pages crawled, however, the breadth first crawler (BreadthFirst) must crawl about 20,000 web pages and the random crawler must crawl about 24,000 web pages.

Another measure is the harvest rate when the crawler compiled 16,000 web pages. For our crawler, about 70% pages are syllabus pages at this point. In the case of breadth first crawler about 50% pages are syllabus, and only about 45% are syllabus pages for the random crawler.

Thus, our syllabus crawler reaches first to accumulate desired number of web syllabus pages and has the highest precision within the specified web pages crawled. In both respect, our crawler is more efficient than the bread first crawler and than the random crawler.

It is worthwhile to note that, in Figure 4, our crawler has lower harvest rate near 30,000 web pages compared with that of the breadth first crawler. This is because the decision tree fails to recognize A-types in some web sites. Contrary to this failure, the breadth first crawler steadily compile syllabus pages. Therefore, we improved our syllabus crawler as “modified A-type crawler(A-type*)”, which is mentioned above. This crawler is more robust and efficient than the others at any stage of crawling.

5. CONCLUSIONS AND FUTURE WORK

Many educational material are being published on the web. We focus on web syllabus pages of Japanese universities. An intelligent crawler for syllabus pages is the first step to

realize a syllabus database.

In this paper, we proposed a web syllabus crawler and showed an evaluation of its efficiency. The crawler uses two characteristics of syllabus pages.

This first characteristics is that there are particular keywords which distinguishes syllabus pages from other web pages. The second characteristics is the link structure between syllabus pages. It is often the case that an individual syllabus page (B-type), is linked from a listing of courses (A-type). We collected 4273 web pages for analysis and judged whether each page is A-type or not and is B-type or not. We used these data as training data and constructed decision trees which decide for A-type and B-type. The decision trees is used to control the traverse of our syllabus crawler.

Finally, we showed an efficiency evaluation of the crawler in terms of the harvest rate, which is the percentage of the number syllabus pages compared with total number of collected pages. We performed the evaluation with 80,446 web pages collected in local disk in advance and confirmed that our crawler is more efficient than the other crawlers.

In this experiment, the crawler started from a list of URLs obtained by searching keyword "シラバス (syllabus)" toward Google. In future, we plan to make our crawler starting from top pages of Japanese universities. Another plan is to improve the precision of the decision trees. The efficiency of our syllabus crawler depends on the precision of the decision tree for A-type. We expect that the crawler can modify the decision tree while it is crawling.

The present paper concerned only collecting web pages. But, extraction and integration of syllabus data is our final goal. Table 1 may be a metadata for syllabus. These keywords may be used as attribute names of syllabus records in integrating them.

6. REFERENCES

[1] Masaaki Ida, Kazuteru Miyazaki, Fuyuki Yoshikane, and Hajime Kita, "A study on constructing syllabus database with XML," *65th Meeting of the Information Processing Society of Japan*, 2003.

[2] Eisuke Itoh, Yoshihiro Matsunaga, Shintaro Yamada, and Sachio Hirokawa, "Auto db construction from web syllabus," *17th Annual Conference of the Japanese Society for Artificial Intelligence*, 2003.

[3] Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu, "Intelligent crawling on the world wide web with arbitrary predicates," *World Wide Web*, pp. 96–105, 2001.

[4] Soumen Chakrabarti, Kunal Punera, and Mallela Subramanyam, "Accelerated focused crawling through online relevance feedback," *WWW2002, Hawaii*, 2002.

[5] Junghoo Cho, Hector García-Molina, and Lawrence Page, "Efficient crawling through URL ordering," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 161–172, 1998.