OfCourse: Web Content Discovery, Classification and Information Extraction for Online Course Materials

Yuhong Xiong, Ping Luo, Yong Zhao, Fen Lin, Shicong Feng, Baoyao Zhou, Liwei Zheng

HP Laboratories
HPL-2010-159

Keyword(s):

Vertical search, online courses, Web classification, Web in- formation extraction

Abstract:

In this paper we present OfCourse, a vertical search engine for online course materials. These materials have the following characteristics: they are scattered very sparsely in the university Web sites; and are generated by the teachers with totally different HMTL templates and layouts. These characteristics impose some challenges for Web Classification (to identify the course materials) and Web Information Extraction (to extract course metadata, such as course title, time and ID) from the identified course homepages. Here, we describe our proposed method to tackle these challenges, and the features of this system. OfCourse, containing over 60,000 courses from the top 50 universities in the US, is currently available for public access.

External Posting Date: October 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: October 21, 2010 [Fulltext]
Published in the 18th ACM Conference on Information and Knowledge Management, Hong Kong (demo paper), November 2-6, 2009

OfCourse: Web Content Discovery, Classification and Information Extraction for Online Course Materials

Yuhong Xiong[†], Ping Luo[†], Yong Zhao[†], Fen Lin[‡], Shicong Feng[†], Baoyao Zhou[†], and Liwei Zheng[†]
[†]Hewlett Packard Labs China, yuhong.xiong@hp.com
[‡]Institute of Computing Technology, CAS

ABSTRACT

In this paper we present OfCourse, a vertical search engine for online course materials. These materials have the following characteristics: they are scattered very sparsely in the university Web sites; and are generated by the teachers with totally different HMTL templates and layouts. These characteristics impose some challenges for Web Classification (to identify the course materials) and Web Information Extraction (to extract course metadata, such as course title, time and ID) from the identified course homepages. Here, we describe our proposed method to tackle these challenges, and the features of this system. OfCourse, containing over 60,000 courses from the top 50 universities in the US, is currently available for public access[1].

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web based services*

General Terms

Algorithms, Experimentation

Keywords

Vertical search, online courses, Web classification, Web information extraction

1. INTRODUCTION

The explosive growth and popularity of the World Wide Web has resulted in a huge amount of unstructured information on the Internet. Except through search, the dominant way to obtain information from the Web, sophisticated Web applications call for flexible techniques that transform information on the Web pages into structured (relational) data [2]. In recent years, progress in this area has moved us closer to this goal [3]. Usually, to build such systems we need a process that is a cascade of two main steps, i.e. Web Classification (WC) to identify the target pages, and Web Information Extraction (WIE) to extract structured metadata from the target pages, after the relevant Web pages are downloaded by *focused crawling*. In this paper we present OfCourse, a vertical search engine for online course materials, to show our work in this area. Different from the other search engines, OfCourse addresses the following challenges:

• The course homepages, as the target Web pages in OfCourse, are scattered very sparsely within the university Web sites. It is totally different from the situations in the other vertical search engines. For example, for product search the online products are often accumulated within several e-commerce sites, such as Amazon, eBay and so on. It indicates that if the crawling starts with a list of e-commerce sites almost all the returned pages contain products. As to our online course portal we have to crawl the pages from a list of university homepages. However, this time only a small portion of the returned pages are course materials. Therefore, the task of WC is more important and difficult in OfCourse if we want to identify all the course materials accurately and completely.

• The course homepages are mostly written freely by course instructors using their own habits. These pages are different from the dynamic Web pages (generated by the database behind), such as product pages, which use the same page template and layout within a site. Due to the variability and heterogeneity in the format and content of online courses materials, accurately extracting metadata from online courses is not a trivial task.

In this paper we describe our progress in focused crawling, WC and WIE, and then details the features of OfCourse, including advanced search by academic discipline and course time, and an *open* search engine for online course materials.

2. METHODS BEHIND OFCOURSE

2.1 Focused Crawling

Focused crawling is the indispensable step before WC and WIE. In the Web site of a typical university, the course pages are less than 5% of all the pages. To discover these course pages without downloading the whole Web site, we need a focused crawler that tries to partially filter out the irrelevant Web pages. We have developed a focused crawling algorithm that is able to quickly find course pages from school web sites. One major problem in focused crawling is determining how to measure the likelihood that a page will lead to the target pages. To address this problem, we developed a measurement of this likelihood, called Navigational Rank (NR). The intuition behind NR is that each page is rewarded by pointing to pages with high relevance and penalized by pointing to pages with low relevance. The experiments show that the resultant focused crawler can download about 90%

[1]http://fusion.hpl.hp.com/OfCourse/

CIKM'09, November 2–6, 2009, Hong Kong, China.
ACM 978-1-60558-512-3/09/11.

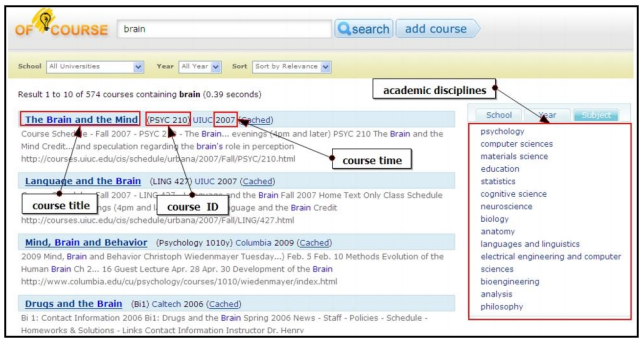

Figure 1: The snapshot of OfCourse

of the target pages by crawling only 30% of the entire Web site.

2.2 Combination of WC and WIE

After the focused crawling we need to accurately identify the course homepages and extract course metadata from them. The traditional approach to reach this goal is to first identify the course homepages by WC, then extract metadata from the homepages by WIE. We notice that it is highly ineffective to use the decoupled strategy, attempting to do WC and WIE in two separate phases, in building a vertical search engine. Specifically, if these two steps are separate, the errors in WC will be propagated to WIC and eventually accumulate to a high level. Thus, the overall performance is upper-bounded by that of WC [2]. We also notice that the mutual dependencies between WC and WIE may provide the improvement room for both these tasks. For the problem in *OfCourse*, on one hand, if a Web page is a course homepage it usually contains some course metadata, such as a course title. On the other hand, the existence of some course metadata, in turn, indicates that the current Web page is a course homepage. This means that there is a forward dependency from WC to WIE, and also a backward dependency from WIE to WC. Therefore, we propose the statistical model to combine WC and WIE, and aim to achieve mutual enhancement between these two steps. As shown by the experimental results, this joint model significantly outperforms the separate models in terms of both precision and recall on course title extraction, and thus addresses the imposed challenges. Please refer to [1] for details on this joint statistical model and the extraction of course time and ID.

3. SYSTEM IMPLEMENTATION AND FEATURE DESCRIPTION

The usage of all the course metadata (including course title, time, and ID) in system implementation is described as follows. 1) Since users often search course materials by the course title, we give more weights to the extracted course titles when indexing the course homepages. This way we aim to achieve better performance of information retrieval. 2) Since course IDs often contain the abbreviations of academic disciplines, course IDs help to group the courses into different disciplines. To do this, we manually map each abbreviation in course IDs to an academic discipline defined by Wikipedia[2]. 3) The course time is used to support the advanced search conducted within certain time duration.

Currently OfCourse contains over 60, 000 courses from the top 50 schools in the US. It has the following features:

[2]http://en.wikipedia.org/wiki/List_of_academic_disciplines

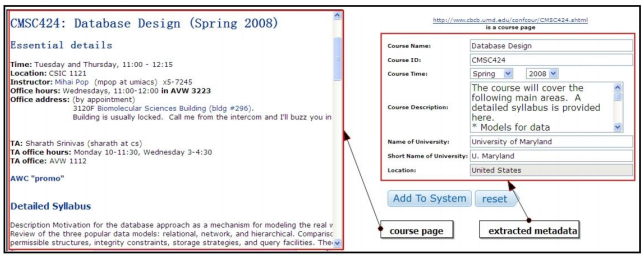

Figure 2: The Web interface for course metadata editing

• Basic keyword search. For each course in the search result, the extracted course title is displayed, together with the school, the extracted year, and an excerpt from the course page containing the search term. The course title links to the course homepage. As shown in Figure 1, the courses returned for the keyword "brain" are listed. The course title, ID and time for the first course in this list are "The Brain and the Mind", "PSYC 210", and "2007" respectively.

• Advanced search. In the advanced search, the search can be conducted within the scope of university, year, and academic disciplines. More interestingly, users can browse the courses according to the academic disciplines they belong to. For example, in Figure 1 the rectangle on the right of the search results contains all the disciplines which the returned courses belong to. It shows that the courses with the keyword "brain" belong to several disciplines, such as biology, psychology, computer science and so on. If we only want to see the courses within the discipline of computer science, we can click the button of "computer science". Then, it will only return the courses offered by the department of computer science for the query "brain".

• An open search engine for online courses materials. OfCourse is open to any user to add new courses into the portal. After submitting the URL of a Web page, the portal automatically judges whether this page is a course homepage and extracts the course metadata if it is. This automatic process is supported by the proposed joint model for WC and WIE. Figure 2 shows the interface for this function, which consists of two parts. The left part shows the user-specified Web page, while the right part shows the course metadata extracted by our model. Users can edit all these course metadata if necessary. After clicking the "Add to System" button, this course can be retrieved from the portal. Additionally, we encourage users to add new courses by giving them high ranking scores.

By *OfCourse* we have seen some interesting searches for online course materials. For example, the search for "Sherlock Holmes" returns a course on "Bestselling Detective Fictions" at MIT, and the search for "hybrid car" (in quotes) also results in some relevant pages.

4. REFERENCES

[1] P. Luo, F. Lin, Y. Xiong, Y. Zhao, and Z. Shi. Towards combining web classification and web information extraction: a case study. In *Proc. of the 15th ACM SIGKDD*, 2009.

[2] A. McCallum. Information extraction: Distilling structured data from unstructured text. *ACM Queue*, 2005.

[3] Z. Nie, J. Wen, and W. Ma. Object-level vertical search. In *Proc. of the Conf. on Innovative Data Systems Research*, 2007.