Growing Parallel Paths for Entity-Page Discovery

Tim Weninger Fabio Fumarola† Cindy Xide Lin
Rick Barber Jiawei Han Donato Malerba†
University of Illinois at Urbana-Champaign
† Università degli Studi di Bari "Aldo Moro"
weninge1@illinois.edu, ffumarola@di.uniba.it, xidelin2@illinois.edu,
barber5@illinois.edu, hanj@illinois.edu, malerba@di.uniba.it

ABSTRACT

In this paper, we use the structural and relational information on the Web to find entity-pages. Specifically, given a Web site and an entity-page (*e.g.*, department and faculty member homepage) we seek to find all of the entity-pages of the same type (*e.g.*, all faculty members in the department). To do this, we propose a web structure mining method which grows *parallel paths* through the web graph and DOM trees. We show that by utilizing these parallel paths we can efficiently discover all entity-pages of the same type. Finally, we demonstrate the accuracy of our method with a case study on various domains.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

parallel paths, entity pages, semi-structured data, web structure mining

1. INTRODUCTION

Most Web pages contain structures that can be leveraged to extract meaningful information about its content. Originally, wrappers were hand crafted to extract this information, and then automatic wrapper generation approaches were developed and refined in order to automate this process [1]. At the same time, pattern detection algorithms, which operate on the DOM tree of a Web page, allowed for a tag-agnostic way to extract information [3]. While improvements to these approaches remain, we argue that the next step is to extend these methods to operate across multiple Web pages in a Web site.

With that goal in mind, we propose a method for the automatic discovery of *entity-pages* from a Web site. Our approach explores a given Web site by traversing paths through Web structures across multiple Web pages.

Assumption 1. (Based on Blanco *et al.* [1]) Many web sites are deliberately designed in a hierarchical manner. This assumption is true in many cases because Web sites are designed and constructed manually, and in many cases, by professionals and web domain managers. It draws from the observation that as a user navigates into the Web site, the pages tend to be divided into specific subtopics. By this assumption we see that a Web site *conceptually* resembles a tree wherein the Web site's homepage is the root and intra-site links represent the edges.

Assumption 2. (Based on Crescenzi *et al.* [2]) Web page links reflect the regularity of the web page structure. For instance, links that are grouped in collections with a uniform layout and presentation usually lead to similar pages. We call these groups of links *parallel links*.

If these assumptions are valid, then we can deduce that entity-pages of the same semantic type typically share similar page- and DOM-paths through the Web site.

We summarize the task as follows: given (1) a Web site and (2) an example entity-page, we wish to discover all entity-pages having the same *type* as the example. In other words, given a faculty member in an academic department we wish to discover all faculty members in that same department.

The notion of typal-similarity can sometimes be difficult to establish. Given a female associate professor as an example entity-page, do we want to return only the set of female associate professors? Or, given the notion that the example female associate professor is a person-entity, do we want to retrieve all persons from the Web site? Our first assumption dictates that Web site managers deliberately structure their Web site so that the type-semantics of entity-pages appear to be natural and consistent for the domain in question, *i.e.*, the webmaster for a computer science department would not typically segregate the faculty into male and female.

Our approach explores a Web site by traversing paths through Web structures across multiple Web pages. By intelligently navigating these Web structures our algorithm grows parallel paths, and by the intuition above, these parallel paths lead to Web pages that represent entities of the same type.

2. METHODOLOGY

Because of the relational structure of Web pages we represent the Web site as a directed graph G. G may be cyclic and is a rooted directed graph, *i.e.*, there is a node designated as the root, from which there is a path to every other node. Thus, there is guaranteed to be one or more directed click-paths from the root to the example entity-page. The shortest of these click-paths (determined by breadth-first search) on G is called the *example tag path* $\mathcal{P}_e$ (dotted line in Figure 2) because it is the path of hyperlinks along the shortest path from root to example entity-page.

With the information from the example path, the next task is to find paths in parallel with $\mathcal{P}_e$. According to our original assertion, paths which are parallel to the example path should lead to entity-

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

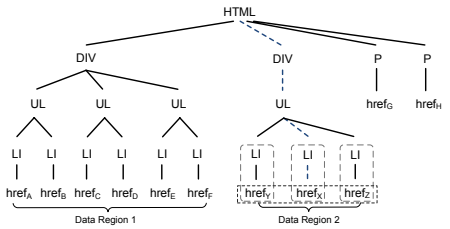

Figure 1: Example finding list of links with augmented MDR. Dotted link represents example link path. Vertical boxes represent parallel data records, and horizontal box represents found parallel links

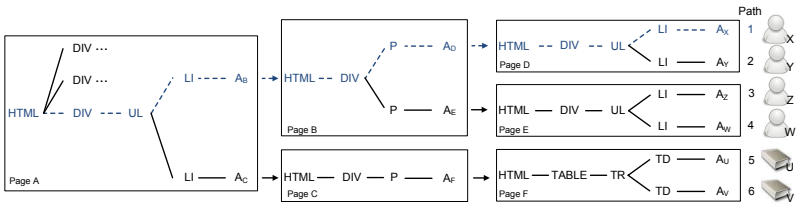

Figure 2: Extended paths over multiple Web pages. Page X is the given example entity-page; therefore, Path 1 is the example path. Paths ending with 1, 2, 3, 4 are in parallel. Paths ending with 5 and 6 are also in parallel.

Table 1: Entity-page discovery results

Entity Type	Reference Page	Count	Prec	Rec.
CS Faculty	`cs.*.edu`	1,410	90.3	87.4
UIUC CS Courses	`cs.illinois.edu`	84	100	100
UIUC CS Groups	`cs.illinois.edu`	36	100	100
Representatives	`house.gov`	441	100	100
Senators	`senate.gov`	100	100	100
House Committees	`house.gov`	45	100	100
Senate Committees	`senate.gov`	40	100	100

pages of the same type. Rather than performing numerous shortest path searches, we, instead, find parallel paths by growing them from the reference page, and splitting at *data regions* as defined by Liu *et al.* in MDR [3] and denoted as dotted boxes in Figures 1–2.

By *growing* parallel paths, we mean that tag-paths continue to grow only so long as it remains in parallel with $\mathcal{P}_e$. As soon as a non-parallel tag is reached ($*$ in Figure 2), the algorithm will cease exploring that tag-path and consider other alternatives. In this way, Web pages and tag-paths are only explored if they contain a path in parallel with the $\mathcal{P}_e$ thereby saving a tremendous amount of computational effort.

Figure 2 illustrates the assumption that tag-paths which are parallel to the example path (dotted line) end with entity-pages of the same type as the example-entity. This processes is repeated for k different example paths, because the initial example path may not have found the complete set of entities.

3. EXPERIMENTS

To test the validity of our assumptions we performed tests on various domains with entity-pages of multiple, diverse types. In our experiments the number of iterations, k, is set to 5. The MDR algorithm also requires two parameters which we set to $K = 5$ and $T = 0.4$. These values were found empirically, and may need to be adjusted for different domains.

Table 1 shows the list of results grouped by entity type.

One side effect of our algorithm is that, because we explore k example-paths, the visitation frequency of found entity-pages show they are related. An actual result of this side effect uses and example entity-page of Tarek Abdelzaher from Illinois. The final result, after all iterations, is that the algorithm visited Professor Abdelzaher 43 times. The 2nd most visited faculty member was Gupta at 42 visits followed by King, Caccamo, Gunter, Lui, Nahrstedt, Godfrey and Kravets at 41 visits each. The total then drops to 32 and so on for the rest of the faculty. After studying the organization of the Computer Science Department at Illinois we found that these professors are all researchers in the systems and networking fields, and therefore share more parallel paths with professor Abdelzaher than a professor in, say, databases does.

A second interesting side effect from the parallel path algorithm is the output of the link tag text. Our initial observation is that the name of the entity at the end of a parallel path is usually included as anchor text. Furthermore, the entity-page's type is usually described by linked text in the parallel paths.

Example. The anchor texts for paths leading to Tarek Abdelzaher illustrate of this side effect. They are: {"People", "Faculty", "Research", "Sensors", "Tarek Abdelzaher", "Personal Site"}.

Clearly, we can infer lots of information about the entities at the end of parallel paths from these labels alone.

4. CONCLUSIONS

In this paper we have introduced the concept of parallel paths and rationalized several properties of these parallel paths in order to discover similarity-typed Web pages. Given a Web site's homepage URL as an entry point and an example Web page representing the *type* of entities to retrieve, our method first finds the shortest path from the Web site homepage to the example entity-page and then finds all paths parallel to the shortest path.

The algorithm exploits the observation that Web site creators deliberately design paths through link structures towards entity-pages, and that these paths (*e.g.*, paths to professor homepages, course pages or group pages) are in parallel.

As alluded to earlier the side effects of finding parallel paths is the extraction of specifically informative information about each entity-page [4]. From our perspective, this research direction is promising given our current results.

5. ACKNOWLEDGMENTS

We would like to thank Jordan Weninger for her help retrieving and labeling data. Research was supported in part by NDSEG Fellowship, the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and by the Strategic Project DIPIS funded by Apulian Region

6. REFERENCES

[1] L. Blanco, V. Crescenzi, and P. Merialdo. Efficiently locating collections of web pages to wrap. In *WEBIST*, pages 247–254. INSTICC Press, 2005.

[2] V. Crescenzi, P. Merialdo, and P. Missier. Clustering web pages based on their structure. *Data Knowl. Eng.*, 54(3):279–299, 2005.

[3] B. Liu, R. Grossman, and Y. Zhai. Mining data records in web pages. In *KDD*, pages 601–606, New York, NY, USA, 2003.

[4] T. Weninger, F. Fumarola, J. Han, and D. Malerba. Mapping web pages to database records via link paths. In *CIKM*, October 2010.