Entity Relation Discovery from Web Tables and Links

Cindy Xide Lin[1]   Bo Zhao[1]   Tim Weninger [1]   Jiawei Han [1]   Bing Liu [2]

*[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
[1]{xidelin2, bozhao3, weninge1}@uiuc.edu, [1]hanj@cs.uiuc.edu, [2]liub@cs.uic.edu

ABSTRACT

The World-Wide Web consists not only of a huge number of un-structured texts, but also a vast amount of valuable structured data. *Web tables* [2] are a typical type of structured information that are pervasive on the web, and Web-scale methods that automatically extract web tables have been studied extensively [1]. Many powerful systems (*e.g.*, OCTOPUS [4], Mesa [3]) use extracted web tables as a fundamental component.

In the database vernacular, a table is defined as a set of tuples which have the same attributes. Similarly, a *web table* is defined as a set of rows (corresponding to database tuples) which have the same column headers (corresponding to database attributes). Therefore, to extract a web table is to extract a relation on the web. In databases, tables often contain foreign keys which refer to other tables. Therefore, it follows that hyperlinks inside a web table sometimes function as foreign keys to other relations whose tuples are contained in the hyperlink's target pages. In this paper, we explore this idea by asking: can we discover new attributes for web tables by exploring hyperlinks inside web tables?

This poster proposes a solution that takes a web table as input. Frequent patterns are generated as new candidate relations by following hyperlinks in the web table. The confidence of candidates are evaluated, and trustworthy candidates are selected to become new attributes for the table. Finally, we show the usefulness of our method by performing experiments on a variety of web domains.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms

Keywords

Web Table, Link, Entity Relation Discovery

1. INTRODUCTION

The World Wide Web is traditionally modeled as a collection of unstructured documents, but recently, efforts have been made

to account for the structural and relational nature of the Web. For example, the table shown in Figure 1 consists of 'employees' in an academic department. This table has four columns, each with a domain-specific label and type, wherein the 'name' column contains a group of hyperlinks pointing to the homepage of the listed person. As noted by Cafarella *et al.* [2], this web table essentially is a small relational database, even if it lacks the explicit meta-data traditionally associated with a database.

This poster explores hyperlinks which are contained in web tables for discovering new entity relations. Again consider the table in Figure 1, the professors listed in the table have links to their homepages, and these homepages contain information regarding 'teaching', 'publications', *etc.*, but in slightly different forms with different descriptions. If we can find pieces of common information in these professors' homepages, then we would be able to expand the web table so that each piece of common information becomes a new attribute. Common information could be from contents, hyperlinks, structures and/or metadata of the homepages (however, this poster only considers hyperlink information). In this example, 'teaching' and/or 'acm publication' could be new attributes. Furthermore, by observing which tuples contain the new attributes, 'employees' could further be classified into 'professor' and 'staff'.

Our motivations are: (i) current methods retrieve web tables that are visually expressed in one HTML page, and there is limited experience on discovering attributes across pages; (ii) Due to the fact that a reliable entity group facilitates the discovery of relations, tuples in a web table are (usually) a trustworthy entity group, which supplies guidance for relation discovery; (iii) The discovery of table attributes and relations will mutually help each other.

Name	Title	Office	Email
Deepayan Chakrabarti	Professor	Rm 2123	dee@usa.edu
Anthony K. H. Tung	Professor	Rm 3115	ant@usa.edu
Evaggelia Pitoura	Professor	Rm 4407	eva@usa.edu
Donna Coleman	Staff	Rm 2124	don@usa.edu
Cindy Lin	Staff	Rm 3116	cin@usa.edu
Jordan Vieyra	Staff	Rm 4406	jor@usa.edu

A web table of *employees*

Professor Rm 2123	Professor Rm 3115	Professor Rm 4407
Teaching	Teaching Courses	Teaching Courses
Graph Mining	Data Mining	Database Systems
ACM Publications	My ACM Publications	My ACM Publications
Deepayan Chakrabarti	Anthony K. H. Tung	Evaggelia Pitoura

Figure 1: A Motivating Example

The remaining problem is: can we discover new attributes by traditional methods of relation extraction? The answer is probably

no. There have been extensive studies in this area [6, 7], which have developed techniques that are powerful in many cases. However, in the above example, even though these methods may successfully extract the general 'teaching' relation, they may not extract the less general 'acm publication' relations. Instead, it will probably be omitted because it is a noun phrase or because of data sparseness. The underlying reason is that these methods aim to extract *general* relations, and therefore lack the capabilities to discover relations specific to a table which may not be common to the whole web.
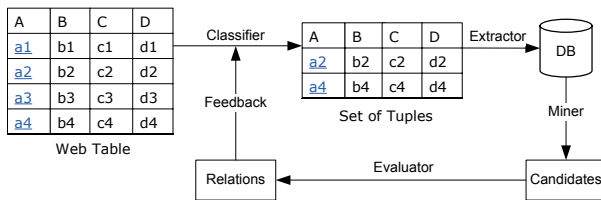
2. THE ALGORITHM

Figure 2: The General Framework

An entity-relation is a triple (e_i, r, e_j), where e_i and e_j denote two entity types and r denotes their relation. Figure 2 depicts the general framework:

> i. Given a web table, a classifier roughly selects tuples that belong to the entity type e_i. Such filtering may be inaccuracy.

> ii. We examine table columns one by one for selected tuples. For a particular table column, we gather the destination pages (abbreviated as P) of hyperlinks in the column, and collect hyperlinks on P to form a transactional database D, where a transaction $d_k \in D$ is a bag of words of any hyperlink-associated information (in this poster we use hyperlink anchor text and hyperlink context).

> iii. We adopt a frequent pattern mining approach to generate frequent itemsets from D, and regard each itemset as a candidate relation.

> iv. For each candidate r, the trustworthiness (denoted as $trust(r)$) is evaluated.

> v. The classifier is updated by adding r into the classifier's feature set.

This procedure repeats iteratively until the trustworthiness converges. Finally, candidates whose trustworthiness are larger than a pre-defined threshold become new relations. In the remainder of this section, we discuss issues of several components.

The **classifier** in component (i). A standard classifier can be adopted here, such as Naive Bayes or SVM. The feature set could be collected from DOM tree structures, pre-defined rules, table contents and/or even discovered relations including r.

The **miner** in component (iii). Pattern-growth-based mining approaches [5] are favored because they are more scalable than, say, Apriori. Pruning techniques can also be used, *e.g.*, require a maximum pattern length, remove patterns with low IDFs, or require a minimum ratio of tuples that point to at least one hyperlink containing r.

The **evaluator** in component (iv). There is more than one way to evaluate the trustworthiness of a candidate relation. In this poster, we employ a classifier C to label hyperlinks as belonging to the relation r or not, and the trustworthiness of r corresponds to the discriminativeness of the classifier.

3. EXPERIMENTS

The experimental datasets are HTML pages crawled from four websites, *i.e.*, $www.cs.uiuc.edu$, $cis.ksu.edu$, $esteelauder.com$ and $senate.gov$, downloaded in Jan. 2010, among which UIUC and KSU are academic department sites, ESTEE is a cosmetics site, and SENATE is a government site. We select the four datasets to demonstrate that our method works on diverse web domains.

In all, the four datasets contain 65,452 pages, 1,018,510 hyperlinks and 104,596 web tables, where 44.09% of the tables contain hyperlinks. These statistics empirically confirm our motivation: the World-Wide Web has a vast amount of web tables which are valuable structured data, and hyperlinks widely exist in web tables.
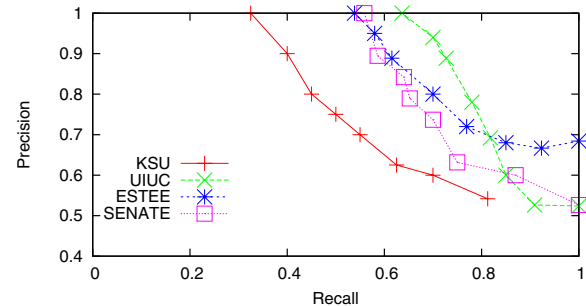

Figure 3: The Precision-Recall Curve

The average numbers of generated candidates for each web table in UIUC, KSU, ESTEE and SENATE are 108.2, 92.5, 20.1 and 66.0, respectively. The reason ESTEE has a smaller number of candidates is because of the relatively limited vocabulary of cosmetics.

A web table is selected from each of the 4 datasets, from which the gold standard is created by manually extracting new attributes. We rank candidates of each web table according to their trustworthiness, and show the precision-recall performance in Figure 3. We observe that the precisions of all datasets are generally high, which confirms our hypothesis that a reliable entity group facilitates the discovery of relations. In terms of recall, some relations are missed because page authors sometimes express the same meaning using different words. We may be able to improve recall if prior knowledge on word correlations is given.

4. REFERENCES

[1] G Miao, J. Tatemura, W-P Hsiung, A.Sawires and L.E. Moser, *Extracting data records from the web using tag path clustering* In *WWW*, p981-990, 2009.

[2] M.J. Cafarella, A.Y. Halevy, D.Z. Wang, E. Wu and Y. Zhang, *WebTables: exploring the power of tables on the web*, In *VLDB*, p538-549, 2008.

[3] S. Mergen, J. Freire and C. Heuser *Mesa: A Search Engine for Querying Web Tables*, In *SBBD*, demo, 2008.

[4] M.J. Cafarella, A.Y. Halevy and N. Khoussainova, *Data Integration for the Relational Web*, *VLDB*, p1090-1101, 2009.

[5] J. Han and J. Pei, *Mining Frequent Patterns by Pattern-Growth: Methodology and Implications*, In SIGKDD Exploration, p13-20, 2000

[6] A. Yates, M. Banko, M. Broadhead, M.J. Cafarella, O. Etzioni and S. Soderland, *TextRunner: Open Information Extraction on the Web*, In HLT-NAACL, p25-26, 2007.

[7] A. Culotta, A. McCallum and J. Betz, *Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text*, In HLT-NAACL, 2006.