Building Knowledge-bases from the Web

Srinivasan H Sengamedu

Yahoo! Labs
Bangalore
shs@yahoo-inc.com

Abstract

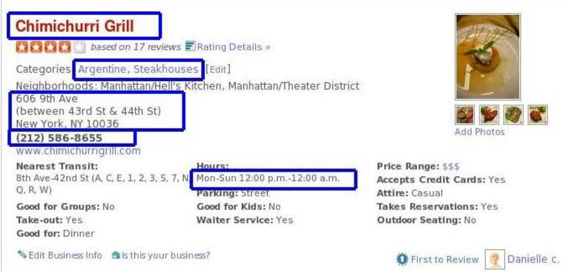
The web is a vast repository of information. Most of the information on the web is meant for human consumption. Extracting structured information from the web can enable several applications like advanced ranking, semantic search, etc. In this talk, we first list different types of content available on the web, survey known techniques for extracting information from them, present the architecture of Vertex information extraction system developed at Yahoo, and discuss in detail a new technique for information extraction leveraging content redundancy.

1 Introduction

Web pages contain rich information about real-world objects. Figure 1 shows a restaurant web page. Some of the information of interest is highlighted in the figure. This information can be leveraged to enable applications like Enhanced Search Results (in which additional information like business phone number is presented in the Search Result Page), ranking by distance/price/rating, etc. Unfortunately this information is not readily available to computers and has to be explicitly extracted. Web Information Extraction typically exploits the structure of the web pages in addition to the nature of content extracted. Section 2 lists different types of web content and Section 3 mentions techniques which are effective for different types of content. Finally, in Sections 4 and 5, we present details of the Vertex information extraction system developed at Yahoo and a new technique for unsupervised information extraction by exploiting another aspect of web content – *content redundancy.*

2 Web Content

Template-generated pages: A vast majority of web pages are generated automatically by populating fixed positions (XPaths) in web pages with values from a database. Hence values for different attributes (like restaurant name) occur at fixed XPaths *in a given site.*



Chimichurri Grill
★★★★☆ *based on 17 reviews* Rating Details »
Categories Argentine, Steakhouses [Edit]
Neighborhoods: Manhattan/Hell's Kitchen, Manhattan/Theater District
606 9th Ave
(between 43rd St & 44th St)
New York, NY 10036
(212) 586-8655
www.chimichurrigrill.com

Nearest Transit: 8th Ave-42nd St (A, C, E, 1, 2, 3, S, 7, N, Q, R, W)
Good for Groups: No
Take-out: Yes
Good for: Dinner

Hours: Mon-Sun 12:00 p.m.-12:00 a.m.
Parking: Street
Good for Kids: No
Waiter Service: Yes

Price Range: $$$
Accepts Credit Cards: Yes
Attire: Casual
Takes Reservations: Yes
Outdoor Seating: No

Figure 1: A restaurant web page.

List pages: List pages contain information on a set of entities in a single page. Here there is structural repetition in a page.

Tables: Table also contain information about multiple entities but the formatting of information is two-dimensional.

Free text: Either entire pages or sections of pages (like "Product Description") can contain unformatted text.

Deep Web: Deep web content is accessed by filling forms and not through following links. Automatic form-filling is a challenging task.

3 Extraction Techniques

Wrapper Induction: Wrapper-based techniques primarily leverage the web page structure. A research challenge is coming up with robust XPaths for extraction.

Regex: Regexs primarily leverage content format, e.g., date have a specific format. Automatic regex learning is an interesting research problem.

Sequential Models: One can consider web pages as sequence of tokens/section and sequential models like Hidden Markov Models, Conditional Random

International Conference on Management of Data
COMAD 2010, Nagpur, India, December 8–10, 2010

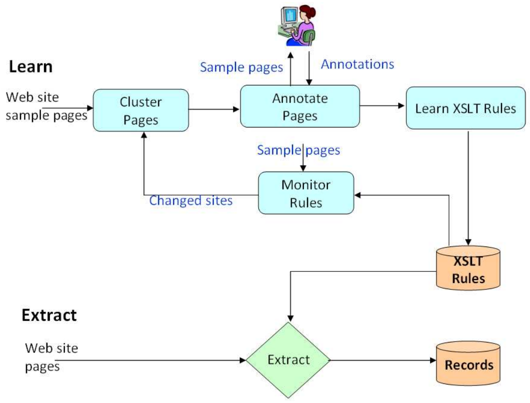


Figure 2: Vertex Architecture.

Fields, etc. exploit the sequential structure in addition to other cues.

Relational Models: Relational learning techniques like Markov Logic Networks describe relational constraints through weighted first order logic formulas and hence are more powerful.

Linguistic techniques: Extraction of information from free text leverages either linguistic patterns (like Hearst patterns) or linguistic context (such as distributional similarity-based techniques).

4 Vertex

Vertex is a *Wrapper Induction* system developed at Yahoo! for extracting structured records from template-based Web pages. It has been deployed in a production environment within Yahoo! to extract data with high precision from hundreds of millions of crawled pages from thousands of Web sites. To operate at Web scale, Vertex employs a host of novel algorithms for (1) Grouping similar structured pages in a Web site, (2) Picking the appropriate sample pages for wrapper inference, (3) Learning XPath-based extraction rules that are robust to variations in page structure (4) Detecting site changes by monitoring sample pages, and (5) Optimizing editorial costs by merging clusters, reusing rules, etc. Figure 2 shows the architecture of Vertex. To the best of our knowledge, Vertex is the first system to do high-precision information extraction at Web scale.

5 Leveraging Content Redundancy

Wrapper-based extraction, while being very precise, requires site-specific training data. Hence the editorial costs can be large. It is easy to see that multiple sides belonging to a given vertical like "Shopping" have some *content redundancy*. In other words, any two sites have pages about a small number of common products. It is possible to leverage content redundancy on the web to extract structured data from *template-based* web sites. We start by populating a seed database with records extracted from a few initial sites. We then identify values within the pages of each new site that match attribute values contained in the seed set of records. To match attribute values with diverse representations across sites, we define a new similarity metric that leverages the templatized structure of attribute content. Specifically, our metric discovers the matching pattern between attribute values from two sites, and uses this to ignore extraneous portions of attribute values when computing similarity scores. Further, to filter out noisy attribute value matches, we exploit the fact that attribute values occur at fixed positions within template-based sites. We develop an efficient Apriori-style algorithm to systematically enumerate attribute position configurations with sufficient matching values across pages. Once we have identified a few pages within a site with redundant content matching record values in the seed set, we exploit the structural similarity among the pages of a template-based site to extract records from the remaining pages of the site. The newly extracted records from a site are added to the seed set, and enable further extractions (due to content redundancy) from additional sites. Finally, we conduct an extensive experimental study with real-life web data to demonstrate the effectiveness of our extraction approach. See "Exploiting Content Redundancy for Web Information Extraction" by P Gulhane, R Rastogi, S H Sengamedu, and A Tengli, VLDB, 2010.

6 Conclusions

Information extraction is an enabling technology for several web applications. While techniques like wrapper induction have high editorial costs for high-precision extraction, we can leverage content redundancy in the web for high-precision extraction from template-based sites without high editorial costs.

Acknowledgements: Vertex resulted from the work of several people at Yahoo Labs, Bangalore – notably Ajesh Babu, Pankaj Gulhane, S R Jeyashankher, Amit Madaan, Rajeev Rastogi, and Charu Tiwari.