Querying for relations from the semi-structured Web

Sunita Sarawagi

IIT Bombay
Mumbai
India
sunita@iitb.ac.in

Abstract

We present a class of web queries whose result is a multi-column relation instead of a collection of unstructured documents as in standard web search. The user specifies the query either via a few example records, or a text description of columns of the relation. Starting from this seed, we show how to compile the result from several, possibly overlapping, tables and lists on the web. Many challenges arise in the process. First, we need to be able to extract structured records from HTML pages with little user supervision. We present algorithms for jointly aligning arbitrary record sets on the web with the query table. We adapt state of the art extraction models like Conditional Random Fields to exploit inter and intra source regularity in a unified framework. Second, we need to be able to consolidate the results from several sources in the face of missing columns, noisy extractions, and zero human supervision. We show how a suitably designed Bayesian networks allows us to compose a resolver from a library of type-specific similarity functions and table statistics. Finally, we discuss the problem of ranking the result rows by their estimated membership in the hidden target relation.

Motivation

Many research and commercial efforts are underway to add structure to the current, predominantly keyword based web searches. These can be categorized broadly into two types:

1. The first approach, exemplified by recent commercial search engines like Wolfram Alpha [1], and True Knowledge [2], is to create separate repositories of structured data in the form of Ontologies. Such Ontologies are typically constructed through long and painstaking manual effort. However, many ongoing projects are attempting to harvest structured databases from sources such as Wikipedia [18, 21], or the general web [4, 6, 19, 2].

2. The second approach is to annotate web documents with entity and relationship labels from a well-known catalog like Wikipedia as in [10, 17, 15]. This allows the enrichment of keyword queries with structured primitives such as a type specifier for entities.

The first approach has the advantage of providing high quality structured answers but suffers from low recall. The web is too huge and diverse to be captured in a structured database, at least with the existing technology. The second approach of bringing structured annotations to web documents, while more recall-friendly, throws in severe challenges of information extraction, particularly on largely text oriented pages. As of this writing, automatic domain-independent extractions on arbitrary HTML pages continues to be a highly error prone process.

Most of these extraction tasks view the input web documents as a sequence of tokens as in classical named entity recognition from natural language text. However, a typical web document contains a variety of other formatted data in the form of lists, tables, and heirarchies that are already partially structured. In this talk, we show that such sources can serve as a compact and high quality source of structured information, that has been ignored until recently [11, 5, 13, 3]. Most prior work on extracting record-like structures from regular web pages, has been localized to specific vertical applications like Shopping, advertisements, and other catalog structures [1, 7, 22, 16, 9, 11]. The excitement of late is in exploiting these resources for domain independent web searches.

In this talk, we will present one such paradigm of horizontal structured search that returns as answers, a ranked list of multi-column records instead of unstructured documents. The user poses the query by specifying the query in one of the following formats:

1. A seed set of records in the answer set. For example, the user might specify a seed set of two rows

15th International Conference on Management of Data COMAD 2009, Mysore, India, December 9–12, 2009

[1] http://www.wolframalpha.com/, October 2009
[2] http://www.trueknowledge.com/, October 2009

such as {(Codd, Relational Algebra),(Alan Turing, Turing Machine)} in trying to compile a list of other (inventors, invention) pairs in computer science.

2. A textual description of the columns of the answer set: such as "Computer scientist" and "Invention".

Several regular sources exist on the web that contain answers to such queries. However, finding the relevant sources, extracting structured records from them, integrating, and ranking them present several challenges. In [13] we present our approach to some of these problems when the source is restricted to lists on the web.

We are working on expanding our sources from lists to arbitrary multi-dimensional tables, or regular structures that are expressed as neither tables or lists. In addition, we plan to exploit existing large Ontologies such as Wikipedia and Yago [18]. We are developing algorithms to automatically annotate tables with type and relationship links of an Ontology. This will enable us to understand a table in relation with other tables with overlapping content and with the user query. In response to a user query, a challenging problem is selecting the set of sources from which the answer will be constructed. This problem can be cast as a novel form of query planning involving soft joins [8] and soft unions [14] across tables. Finally, ranking of records based on their membership in overlapping tables, and in the presence of extraction and linkage errors, continues to be an important unsolved problem in spite of much research [12, 20, 19].

Our ongoing work on these problems can be tracked at `http://www.cse.iitb.ac.in/ sunita/wwt`.

References

[1] M. Álvarez, A. Pan, J. Raposo, F. Bellas, and F. Cacheda. Extracting lists of data records from semi-structured web pages. *Data Knowl. Eng.*, 64(2):491–509, 2008.

[2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007.

[3] M. Cafarella, N. Khoussainova, D. Wang, E. Wu, Y. Zhang, and A. Halevy. Uncovering the relational web. In *WebDB*, 2008.

[4] M. J. Cafarella, D. Downey, S. Soderland, and O. Etzioni. Knowitnow: Fast, scalable information extraction from the web. In *HLT/EMNLP*, 2005.

[5] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Webtables: Exploring the power of tables on the web. In *VLDB*, 2008.

[6] M. J. Cafarella, C. Re, D. Suciu, and O. Etzioni. Structured querying of web text data: A technical challenge. In *CIDR*, pages 225–234, 2007.

[7] C. Chang. and S. Lui. Iepad: Information extraction based on pattern discovery. In *WWW*, pages 681–688, 2001.

[8] W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems*, 18:288–321, 2000.

[9] W. W. Cohen, M. Hurst, and L. S. Jensen. A flexible learning system for wrapping tables and lists in html documents. In *WWW*, 2002.

[10] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL 2007*, pages 708–716, 2007.

[11] H. Elmeleegy, J. Madhavan, and A. Halevy. Harvesting relational tables from lists on the web. In *VLDB*, 2009.

[12] Z. Ghahramani and K. A. Heller. Bayesian sets. In *NIPS*, 2005.

[13] R. Gupta and S. Sarawagi. Answering table augmentation queries from unstructured lists on the web. In *Proc. of the 35th Int'l Conference on Very Large Databases (VLDB)*, 2009.

[14] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: Similarity measures and algorithms. Tutorial at SIGMOD, 2006.

[15] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466, New York, NY, USA, 2009. ACM.

[16] K. Lerman, C. Knoblock, and S. Minton. Automatic data extraction from lists and tables in web sources. In *Workshop on Advances in Text Extraction and Mining (ATEM)*, 2001.

[17] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, 2008.

[18] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[19] P. P. Talukdar, J. Reisinger, M. Pasca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In *EMNLP*, 2008.

[20] R. C. Wang, N. Schlaefer, W. W. Cohen, and E. Nyberg. Automatic set expansion for list question answering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008.

[21] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *CIKM*, pages 41–50, 2007.

[22] Y. Zhai and B. Liu. Web data extraction based on partial tree alignment. In *WWW*, pages 76–85, 2005.