Web Data Management

Michael J. Cafarella
University of Michigan
Ann Arbor, MI 48109-2121
michjc@umich.edu

Alon Y. Halevy
Google, Inc.
Mountain View, CA 94043
halevy@google.com

1. INTRODUCTION

Web Data Management (or WDM) refers to a body of work concerned with leveraging the large collections of structured data that can be extracted from the Web. Over the past few years, several research and commercial efforts have explored these collections of data with the goal of improving Web search and developing mechanisms for surfacing different kinds of search answers. This work has leveraged (1) collections of structured data such as HTML tables, lists and forms, (2) recent ontologies and knowledge bases created by crowd-sourcing, such as Wikipedia and its derivatives, DBPedia, YAGO and Freebase, and (3) the collection of text documents from the Web, from which facts could be extracted in a domain-independent fashion.

The promise of this line of work is based on the observation that new kinds of results can be obtained by leveraging a huge collection of independently created fragments of data, and typically in ways that are wholly unrelated to the authors' original intent. For example, we might use many database schemas to compute a schema thesaurus. Or we might examine many spreadsheets of scientific data that reveal the aggregate practice of an entire scientific field. As such, WDM is tightly linked to Web-enabled collaboration, even (or especially) if the collaborators are unwitting ones.

We will cover the key techniques, principles and insights obtained so far in the area of Web Data Management.

Categories and Subject Descriptors: H.3.5 [Information Storage and Retrieval] Online Information Services, Data sharing, Web-based services

General Terms: Algorithms, Design, Experimentation, Theory

2. TUTORIAL OUTLINE

The tutorial covers the following WDM topics. The references mentioned here are only representative of the ones that will be covered in the tutorial and are not intended to be comprehensive.

2.1 Domain-Independent Extraction from Text

There is a growing body of work in open information extraction, or Web-centric extraction. These are extraction systems that are able to effectively construct relational databases out of very large document corpora. Because they operate at very large scale, the traditional approach of designing domain-specific extractors, or of collecting domain-specific training data, is not feasible.

There have been a large number of recent and relevant academic projects in this area. The KnowItAll [7] and TextRunner [1] projects extract fact triples from natural language Web text. The WebTables system [5, 4] constructs a large corpus of databases from HTML tables on the Web. Researchers at IIT Bombay [10] have attempted to annotate tabular data elements with extra semantic information (*e.g.,* the type of a column). Suchanek, *et al.'s* YAGO system [13] extracted a large number of tuples from Wikipedia in several dozen semantic categories; it used a relatively large amount of domain-specific knowledge than the above projects, and also obtained unusually high precision and recall.

We will discuss the various techniques used by these systems, and their relative advantages and disadvantages. In particular, we will discuss ways in which they might be combined in the future to obtain large numbers of output sets with very high recall and precision.

2.2 Online Data Communities

Of course, it is possible to pursue the goal listed above - a large collection of multi-topic structured datasets - without using extraction at all. Instead, one could employ large numbers of human editors to generate and clean their datasets. The best-known example of such a system is Wikipedia, which has added a substantial amount of structured data to its text-centric encyclopedia pages. This structured data comes mainly in the form of "infoboxes" that accompany many pages.

However, Wikipedia is not the only relevant work on this topic. Freebase [3] is a community-constructed graph-oriented database. Although it was designed to be a primarily socially-driven site, it only met moderate success in appealing to a large number of contributors and eventually obtained much of its structured data from Wikipedia's infoboxes. DBPedia [2] is an effort to unify several online structured databases. Wikipedia is the highest-profile dataset, but there are many others: MusicBrainz for music, Geonames for geographic information, Drugbank for pharmaceutical information, and so on. Its data size has been growing rapidly in recent years. Finally, research into Wikipedia has yielded a large number of other secondary data products and applications, including improvements to information extraction and to interesting mixed-initiative interfaces. One example is the Kylin project [15], which bootstraps an ontology using extracted resources.

SIGMOD'11 June 12–16, 2011, Athens, Greece.
ACM 978-1-4503-0661-4/11/06.

We will cover similarities and differences among these socially-driven data creation systems, and discuss the ways in which Wikipedia data has become a critical standard in most socially-driven data work.

2.3 Social Data Management Tools

A new generation of online tools have arisen to address data management tasks that arise specifically in Web-style settings. These tools share some qualities with traditional relational data systems, but also deeply embed social activities into their design.

For example, FusionTables [8] is a Google tool that enables socially-driven creation of tabular datasets (it also has some light information extraction features). IBM's ManyEyes [14] site allows groups of people to discuss and visualize data sets. In a roughly similar vein, Socrata [12] offers tools for mashing up and visualizing uploaded datasets, in particular governmental data. The DBLife [6] system allows groups of people to easily design a topic-specific website that collects much of its data from external sources. Although describing the initial site can be somewhat time-consuming, updates to the site take place automatically via information extractors. The system has explicit support for users trying to design these extractors.

We will cover techniques and features common to all of these systems. We will also discuss technical challenges with these systems, and which new domains may lend themselves to a novel tool.

2.4 The Deep Web

The Deep Web is the collection of databases with Web front-ends, containing data that can only be accessed via submitted Web forms. Many estimates of the Deep Web put its size at several times the data that can be accessed via the traditional Web. Several researchers [9, 11] have looked into techniques that make Deep Web data more accessible, often by automatically formulating appropriate Web form queries. Of course, because the number of Deep Web sources is enormous, it is not feasible for humans to create these queries by hand.

2.5 General Themes

There are several challenges faced by Web Data Management that are larger than any of the specific topics mentioned above. These themes include the following:

- The interplay between structured and unstructured data. How patterns extracted from unstructured data can help recover the semantics of structured data
- Expansion of large ontologies, and using them to find additional facts on the Web
- Using Web data to answer factual queries on the Web
- Creating data communities on the Web

3. RELATED WORK

It is important to put progress in WDM in the context of other related areas of work.

Unlike **content management**, WDM is interested in the data products of many different people, not a single person or organization. WDM is not fundamentally concerned with providing versioning or storage reliability - it assumes such services are available. WDM exposes collaboration among data-authors; facilitating that collaboration is a second-order priority.

Data mining is occupied with techniques for obtaining high-quality predictive or other statistical results from examining datasets. WDM uses data mining approaches extensively, but is also interested in data models and practices that lie outside traditional data mining concerns.

Information extraction focuses on obtaining a refined version of data from an unstructured source. Unlike many extraction projects, WDM often operates on objects that are already nearly-completely-structured: databases, spreadsheets, social data structures, etc. In addition, WDM often generates aggregate information from the source data, whereas information extraction is generally focused on acquiring the relevant dataset.

Dataspaces are concerned with managing collections of loosely coupled heterogeneous data sources and providing best-effort answers on them. While inheriting some of the main ideas of dataspaces, WDM is concerned with the special challenges of the Web as an extreme dataspace.

4. REFERENCES

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *IJCAI*, pages 2670–2676, 2007.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009.

[3] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250, 2008.

[4] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. WebTables: Exploring the Power of Tables on the Web. *PVLDB*, 1(1):538–549, 2008.

[5] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the Relational Web. In *WebDB*, 2008.

[6] P. DeRose, W. Shen, F. Chen, Y. Lee, D. Burdick, A. Doan, and R. Ramakrishnan. Dblife: A community information management platform for the database research community (demo). In *CIDR*, pages 169–172, 2007.

[7] O. Etzioni, M. J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.*, 165(1):91–134, 2005.

[8] H. Gonzalez, A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google fusion tables: data management, integration and collaboration in the cloud. In *SoCC*, pages 175–180, 2010.

[9] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the deep web. *Commun. ACM*, 50(5):94–101, 2007.

[10] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. *PVLDB*, 3(1):1338–1347, 2010.

[11] J. Madhavan, L. Afanasiev, L. Antova, and A. Y. Halevy. Harnessing the deep web: Present and future. In *CIDR*, 2009.

[12] Socrata. `http://www.socrata.com/`.

[13] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[14] F. B. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1121–1128, 2007.

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