Knowledge Discovery over the Deep Web, Semantic Web and XML

Aparna Varde[1], Fabian Suchanek[2], Richi Nayak[3] and Pierre Senellart[4]

1. Department of Computer Science, Montclair State University, Montclair, NJ, USA

2. Databases and Information Systems, Max Planck Institute for Informatics, Saarbrucken, Germany

3. Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia

4. Department of Computer Science and Networking, Telecom Paristech, Paris, France

(vardea@mail.montclair.edu, suchanek@mpi-inf.mpg.de, r.nayak@qut.edu.au, pierre@senellart.com)

Abstract. In this tutorial we provide an insight into Web Mining, i.e., discovering knowledge from the World Wide Web, especially with reference to the latest developments in Web technology. The topics covered are: the Deep Web, also known as the Hidden Web or Invisible Web; the Semantic Web including standards such as RDFS and OWL; the eXtensible Markup Language XML, a widespread communication medium for the Web; and domain-specific markup languages defined within the context of XML We explain how each of these developments support knowledge discovery from data stored over the Web, thereby assisting several real-world applications.

Keywords: Information Retrieval, Standards, Web Mining

1 Introduction

This tutorial focuses on knowledge discovery from the Web with particular emphasis on the Deep Web, Semantic Web and XML including domain-specific markup languages. The vast amount of data stored and exchanged over the World Wide Web is a huge source of knowledge that can be useful in various potential applications.

Among the recent advances in Web technology, we have the Deep Web over which stored information is not obvious but needs to be inferred, for example from queries through forms. The Semantic Web encompasses standards such as RDFS and OWL which often serve as the basis for defining ontology with reference to context.

XML, the eXtensible Markup Language has become a widely accepted means of communication with its descriptive tag sets that can be extended to add semantics to the data stored on the Web. This also facilitates the development of domain-specific markup languages that can be accepted as the lingua franca for communication in their respective fields.

All these developments provide great potential for mining the Web, i.e., discovering knowledge from the stored data. The data available on the Web is typically in a semi-structured format which presents additional challenges in knowledge discovery as opposed to data stored in traditional relational databases. In this tutorial we address various aspects of knowledge discovery from the Web with respect to these developments. We give an overview of the Deep Web, Semantic

Web, XML and domain-specific markup languages in terms of their fundamental concepts and explain how each of these enable knowledge discovery. Suitable examples are provided at relevant points in the tutorial. Interesting real-world applications are also described. The tutorial is thus divided into four parts as described in the following four sections.

2 The Deep Web

A large part of the information present in the World Wide Web is *hidden* to current-day search engines, because it is not accessible through hyperlinks but lies in databases queried through forms. This *Deep* Web (or *Hidden* Web, or *Invisible* Web) has been estimated to contain 500 times as much data as the *Surface Web*. If such precise measures are debatable, this order of magnitude has been confirmed by recent work, and it is unquestionable that with information of the best quality (e.g., *Yellow Pages* services, U.S. *Census Bureau,* library catalogs, bibliography), the hidden Web is not only an invaluable source of information, but is also, due to its semi-structured, template nature, a rich source for knowledge discovery.

Access to content of the Deep Web requires filling in and submitting (HTML) forms, in order to retrieve some response pages, typically structured as lists or table records. Two approaches coexist for benefiting of the data hidden behind forms. The first one, the most straightforward, which has been advocated and experimented with by Google is an *extensional* one: response pages generated by the deep Web service are just stored as regular Web pages, that can be queried and retrieved as usual. The second approach, exemplified by the METAQUERIER system, is *intensional*: the goal is not to store response pages, but to understand the structure of both forms and response pages, and thus to know the semantics of this service, that can then be called as needed, depending on a user query. In either case, some schema matching and text mining techniques are used to associate form fields with concepts, in order to generate corresponding response pages. In the intensional case, understanding the structure of a response page means discovering the template this page was created from, either by unsupervised techniques such as ROADRUNNER, or by (semi-)supervised techniques.

This part of the tutorial presents different approaches for accessing the Deep Web, including but not limited to our own work, and shows how relevant data and information can be discovered and extracted from it.

3 The Semantic Web

The Semantic Web project envisions that people will publish semantic information in a computer-processable formalism that allows the information to be globally interlinked. For this purpose, the World Wide Web Consortium (W3C) has developed the knowledge representation formalisms RDFS and OWL. These formalisms are based on XML, but go beyond it by specifying semantic relationships between entities and even logical constraints on them. A collection of world knowledge in these formalisms is commonly called an *ontology*.

In this section of the tutorial, we first explain the vision and the applications of the Semantic Web project. We then give an introduction to semantic knowledge representations and ontologies in general. We also explain the knowledge representation formalisms RDFS and OWL, their syntax and semantics. We show where the Semantic Web has already taken off: Several large-scale ontologies are available online and are interlinked in the spirit of the Semantic Web. We explain how this information was gathered from different sources and how it can be queried using the SPARQL query language. Furthermore, we emphasize how this enhances knowledge discovery.

4 XML, the eXtensible Markup Language

The eXtensible Markup Language (XML) has become a standard language for data representation on the Web. With the continuous growth in XML based Web data sources, the ability to manage collections of XML documents and discover knowledge from them for decision support is increasingly important.

Mining of XML documents significantly differs from structured data mining and text mining. XML allows the representation of semi-structured and hierarchal data containing not only the values of individual items but also the relationships between data items. Element tags and their nesting therein dictate the structure of an XML document. Due to the inherent flexibility of XML, in both structure and semantics, discovering knowledge from XML data is faced with new challenges as well as benefits. Mining of structure along with content provides new insights and means into the knowledge discovery.

Recognizing the increasing interest in XML mining, this portion of the tutorial aims to discuss challenges that occur while mining the XML based Web data along with their solutions. We also provide issues and directions for research and development work in the future.

5 Domain-Specific Markup Languages

A significant expansion in the area of the Web and XML is the development of domain-specific markup languages. Such languages typically encompass the syntax of XML and capture the semantics of the given domains. Storing and exchanging data in this format, i.e., using XML based markups greatly boosts knowledge discovery. In this section of the tutorial we provide an overview of domain-specific markup languages with some real-world examples. We consider state-of-the-art markups, e.g., MML, the medical markup language, MatML, the Materials Markup Language and a few more.

We briefly outline the steps involved in the development of markup languages, the desired features of the languages, the use of XML constraints in preserving domain semantics and additional requirements, and the retrieval of information from the markups using XQuery, XPath and others in the XML family. We explain how data storage using such markup languages can assist data mining with classical techniques

such as association rules. We also stress on the fact that in addition to a having adequate schemas for the markups, relevant ontological developments using standards in the literature can further assist the discovery of knowledge from data in the respective domains. A summary of our own research as well as related work by others in the area is discussed.

In conclusion, we summarize the most important points in the tutorial and briefly touch upon the potential for future work in the area of Web Mining.

6 References

1. Chang, C., Kayed, M. Girgis, M.R. and Shaalan, K.F., A survey of Web information extraction systems. IEEE Transactions on Knowledge and Data Engineering, 18(10):1411-1428, Oct 2006.
2. Crescenzi, V., Mecca, G. and Merialdo, P., Roadrunner: Towards automatic data extraction from large Web sites. In VLDB, Rome, Italy, Sep 2001.
3. He, B., Patel, M., Zhang, Z. and Chang, K.C., Accessing the deep Web: A survey. Communications of the ACM, 50(2):94–101 May 2007.
4. Madhavan, J., Halevy, A.Y., Cohen, S., Dong, X., Jeffery, S.R., Ko, D. and Yu, C. Structured data meets the Web: A few observations. IEEE Data Engineering Bullerin, 29(4):19–26, Dec 2006.
5. Senellart, P. Mittal, A., Muschick, D., Gilleron, R. andTommasi, M., Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge. In WIDM, pp. 9–16, Napa, USA, Oct 2008.
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z.G., Dbpedia: A nucleus for a Web of open data. In ISWC 2007.
7. Lenat, D., and Guha, R.V., Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, 1989.
8. Staab, S. and Studer, R., editors. Handbook on Ontologies, 2nd edition. Springer, 2008.
9. Suchanek, F.M., Kasneci, G. and Weikum, G., YAGO: A Core of Semantic Knowledge. In WWW 2007.
10. Word Wide Web Consortium. OWL Web Ontology Language (W3C Recommendation 2004-02-10). http://www.w3.org/TR/owl-features/
11. Li, H., Shan, F. and Lee, S.Y., Online mining of frequent query trees over XML data streams. In *15th international conference on World Wide Web* (pp. 959-960). Edinburgh, Scotland: ACM Press, 2008.
12. Kutty, S. and Nayak, R., Frequent Pattern Mining on XML documents, Chapter 14 in "Handbook of Research on Text and Web Mining Technologies", Ed: Min Song and Yi-Fang Wu. Publisher: Idea Group Inc., USA. pp. 227 -248, 2008.
13. Nayak, R., Fast and Effective Clustering of XML Data Utilizing their Structural Information. *Knowledge and Information Systems* (KAIS). Volume 14, No. 2, pp. 197-215, Feb 2008.
14. Rusu, L. I., Rahayu, W., and Taniar, D. Mining Association Rules from XML Documents. In A. Vakali & G. Pallis (Eds.), *Web Data Management Practices, 2007.*
15. Wan, J., Mining Association rules from XML data mining query. *Research and practice in Information Technology, 32*, 169-174, 2004.
16. Boag, S., Fernandez, M., Florescu, D., Robie J. and Simeon, J.: XQuery 1.0: An XML Query Language. W3C Working Draft, November 2003.
17. Clark, J. and DeRose, S.: XML Path Language (XPath) Version 1.0. W3C Recommendation, Nov 1999.
18. Davidson, S., Fan, W., Hara, C. and Qin, J.: Propagating XML Constraints to Relations. In International Conference on Data Engineering, March 2003.
19. Guo, J., Araki, K., Tanaka, K., Sato, J., Suzuki, M., Takada, A., Suzuki, T., Nakashima, Y. and Yoshihara, H.: The Latest MML (Medical Markup Language) —XML based Standard for Medical Data Exchange / Storage. In: Journal of Medical Systems, Vol. 27, No. 4, pp. 357 – 366, Aug 2003.
20. Varde, A., Rundensteiner, E. and Fahrenholz, S.: XML Based Markup Languages for Specific Domains, Book Chapter, In Web Based Support Systems", Springer, 2008.

(Please note that the references have been cited as they are used in the respective sections, i.e., references 1 through 5 are for Section 2, references 6 through 10 are for Section 3, references 11 through 15 are for Section 4 and references 16 through 20 are for Section 5.)