Social Semantic Web Mining

Synthesis Lectures on the Semantic Web: Theory and Technology

Synthesis Lectures on the Semantic Web: Theory and Application is edited by Ying Ding of Indiana University and Paul Groth of VU University Amsterdam. Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives.
Topics to be included:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging "Web of Data" and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

Social Semantic Web Mining
Tope Omitola, Sebastián A. Ríos, and John G. Breslin
2015

Semantic Breakthrough in Drug Discovery
Bin Chen, Huijun Wang, Ying Ding, and David Wild
2014

Semantics in Mobile Sensing
Zhixian Yan and Dipanjan Chakraborty
2014

Provenance: An Introduction to PROV
Luc Moreau and Paul Groth
2013

Resource-Oriented Architecture Patterns for Webs of Data
Brian Sletten
2013

Aaron Swartz's A Programmable Web: An Unfinished Work
Aaron Swartz
2013

Incentive-Centric Semantic Web Application Engineering
Elena Simperl, Roberta Cuel, and Martin Stein
2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web
Eero Hyvönen
2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery
Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding
2012

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer
2011

Social Semantic Web Mining

Tope Omitola, Sebastián A. Ríos, and John G. Breslin

www.morganclaypool.com

ISBN: 9781627053983 paperback
ISBN: 9781627053990 ebook

DOI 10.2200/S00623ED1V01Y201412WBE010

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY

Lecture #10
Series Editors: Ying Ding, *Indiana University*
Paul Groth, *VU University Amsterdam*
Founding Editor Emeritus: James Hendler, *Rensselaer Polytechnic Institute*
Series ISSN
Print 2160-4711 Electronic 2160-472X

Social Semantic Web Mining

Tope Omitola
University of Southampton

Sebastián A. Ríos
University of Chile

John G. Breslin
National University of Ireland Galway

SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY #10

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

The past ten years have seen a rapid growth in the numbers of people signing up to use Web-based social networks (hundreds of millions of new members are now joining the main services each year) with a large amount of content being shared on these networks (tens of billions of content items are shared each month). With this growth in usage and data being generated, there are many opportunities to discover the knowledge that is often inherent but somewhat hidden in these networks. Web mining techniques are being used to derive this hidden knowledge. In addition, the Semantic Web, including the Linked Data initiative to connect previously disconnected datasets, is making it possible to connect data from across various social spaces through common representations and agreed upon terms for people, content items, etc.

In this book, we detail some current research being carried out to semantically represent the implicit and explicit structures on the Social Web, along with the techniques being used to elicit relevant knowledge from these structures, and we present the mechanisms that can be used to intelligently mesh these semantic representations with intelligent knowledge discovery processes. We begin this book with an overview of the origins of the Web, and then show how web intelligence can be derived from a combination of web and Social Web mining. We give an overview of the Social and Semantic Webs, followed by a description of the combined Social Semantic Web (along with some of the possibilities it affords), and the various semantic representation formats for the data created in social networks and on social media sites.

Provenance and provenance mining is an important aspect here, especially when data is combined from multiple services. We will expand on the subject of provenance and especially its importance in relation to social data. We will describe extensions to social semantic vocabularies specifically designed for community mining purposes (SIOCM). In the last three chapters, we describe how the combination of web intelligence and social semantic data can be used to derive knowledge from the Social Web, starting at the community level (macro), and then moving through group mining (meso) to user profile mining (micro).

KEYWORDS

World Wide Web, Web Mining, Web Intelligence, Semantic Web, Social Web, Social Semantic Web, Provenance, Knowledge Discovery, Knowledge Management

We would like to dedicate this book to our families,
who have supported us and been patient with us throughout
the writing of this book.

Contents

Acknowledgments

Tope Omitola would like to express gratitude for the managerial support of Professor Sir Nigel Shadbolt of the Web and Internet Science Group, University of Southampton and of Dr. John Davies of British Telecommunications during the writing of this book. He would also like to acknowledge the support of EPSRC (Grant Number EP/K503770/1) during this time.

Sebastián Ríos would like to thank the continuous support from the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FBO16) and the research grant for Initiation into Research (FONDECYT), project code 11090188, entitled "Semantic Web Mining Techniques to Study Enhancements of Virtual Communities."

John Breslin would like to acknowledge input from his collaborators Dr. Fabrizio Orlandi and Dr. Alexandre Passant. He would also like to acknowledge that his work on this publication has emanated from research supported in part by research grants from Science Foundation Ireland (SFI) under Grant Number SFI/08/CE/I1380 (DERI Líon 2) and also under Grant Number SFI/12/RC/2289 (Insight).

Tope Omitola, Sebastián A. Ríos, and John G. Breslin
January 2015

Grant Aid

This publication was grant-aided by the Publications Fund of
National University of Ireland Galway

Rinneadh maoiniú ar an bhfoilseacháin seo trí Chiste Foilseacháin
Ollscoil na hÉireann Gaillimh

CHAPTER 1

Introduction and the Web

1.1 INTRODUCTION

The past ten years have seen a rapid growth in the numbers of people signing up to use Web-based social networks (hundreds of millions of new members are now joining the main services each year) with a large amount of content being shared on these networks (tens of billions of content items are shared each month).

With this growth in usage and data being generated, there are many opportunities to discover the knowledge that is often inherent but somewhat hidden in these networks. Web mining techniques are being used to derive this hidden knowledge.

On the other hand, the Semantic Web, including the Linked Data initiative to connect previously disconnected datasets, is making it possible to connect data from across various social spaces through common representations and agreed-upon terms for people, content items, etc.

In this book, we detail some current research being carried out to semantically represent the implicit and explicit structures on the Social Web, along with the techniques being used to elicit relevant knowledge from these structures, and we present the mechanisms that can be used to intelligently mesh these semantic representations with intelligent knowledge discovery processes.

We will begin this book with an overview of the origins of the Web (Chapter 1), and then show how web intelligence can be derived from a combination of web and Social Web mining (Chapter 2). We give an overview of the Social and Semantic Webs (Chapters 3 and 4), followed by a description of the combined Social Semantic Web (along with some of the possibilities it affords) (Chapter 5), and the various semantic representation formats for the data created in social networks and on social media sites.

Provenance and provenance mining is an important aspect here, especially when data is combined from multiple services. We will expand on the subject of provenance and especially its importance in relation to social data (Chapter 6).

We will describe extensions to social semantic vocabularies specifically designed for community mining purposes (SIOCM).

In the last three chapters, we describe how the combination of web intelligence and social semantic data can be used to derive knowledge from the Social Web, starting at the community level (macro), and then moving through group mining (meso) to user profile mining (micro).

1.2 THE WORLD WIDE WEB

"The World Wide Web ("WWW" or simply the "Web") is a global, read-write information space. Text documents, images, multimedia and many other items of information, referred to as resources, are identified by short, unique, global identifiers called Uniform Resource Identifiers (URIs) so that each can be found, accessed and cross-referenced in the simplest possible way."[1]

As the reader will note, the Web is not a synonym for the Internet. The Internet refers to the physical network and protocols, while the Web refers to a framework running on top of it. The public debut of the Web was on August 6, 1991, when Tim Berners-Lee (one of its creators) posted a short summary of the World Wide Web project on the alt.hypertext newsgroup.[2] Berners-Lee's breakthrough was to combine hypertext with the Internet in an easy-to-deploy framework, that went well beyond previous applications like Gopher.

1.2.1 HISTORY AND EVOLUTION OF THE WEB

In 1980, Tim Berners-Lee wrote some software (ENQUIRE) which allowed one to browse through "cards" about resources at CERN (the European Organization for Nuclear Research) via bidirectional hyperlinks. By 1990, Berners-Lee was working on a GUI (graphical user interface) to perform browsing of hypertext, naming it the "World Wide Web" [Chakrabarti, 2002], and the Web's operation began in the following year, as detailed above.

In 1992, other programs were developed to browse hypertext on the Web such as Viola, Erwise, Midas, and Cello. By early 1993, the first version of Mosaic was completed by Mark Andreessen at the NCSA (National Center for Supercomputer Applications). Simultaneously, CERN developed a new improved HTML protocol (HTTP) to send HTML (Hypertext Markup Language) documents and other data over the growing Internet. To do so, they developed a server called CERN httpd (Hyper Text Transfer Protocol Daemon).

When it was announced that Gopher was no longer free to use, CERN announced that the World Wide Web would be free to anyone, with no fees, this producing a rapid shift away from Gopher and toward the Web. This resulted in an exponential growth in traffic on the Web and also in website development (detailed information can be found in Chakrabarti [2002]).

By October 1994, "*Sir Tim Berners-Lee, inventor of the World Wide Web, left the European Organization for Nuclear Research (CERN) and founded the World Wide Web Consortium (W3C) at the Massachusetts Institute of Technology Laboratory for Computer Science (MIT/LCS) with support from the Defense Advanced Research Projects Agency (DARPA), which had pioneered the Internet, and the European Commission.*"[3] In the same year, the Mosaic Communications Corporation was also created. Later, Mosaic changed its name to the more well known Netscape Communications Corporation.

[1] http://en.wikipedia.org/wiki/World_Wide_Web.
[2] https://groups.google.com/forum/#!msg/alt.hypertext/eCTkkOoWTAY/bJGhZyooXzkJ.
[3] http://en.wikipedia.org/wiki/W3C.

To summarize, from a technical point of view the three key features that allow the existence of the Web are:

- Uniform Resource Identifier (URI), which is a universal system for referencing resources on the Web, such as webpages, images, videos, etc.
- Hypertext Transfer Protocol (HTTP), which specifies how the client (browser) and server communicate with each other.
- Hypertext Markup Language (HTML), used to define the structure and content of hypertext documents (webpages).

A term often conflated with URI is URL (Uniform Resource Locator). A URL is a type of URI that not only identifies a resource but allows one to locate that resource at the address that the URI points to.

1.2.2 A SHORT EXPLANATION OF HOW THE WEB WORKS

The Web is based on the simple but powerful idea of browsing a hyperlinked collections of hypertext documents [Morisseau-Leroy et al., 2001]. Usually, a blade/rack/tower server running some server software for the Web (Apache, Caucho Resin, PWS, etc.) is called the "web server." However, the computer can be any machine, even a desktop computer or a low-cost Raspberry Pi microcomputer.

It is important to differentiate the hardware server from the server software. The same holds for the client computer of the person who is trying to browse the Web (see Fig. 1.1). The software for browsing hypertext documents on the Web (Chrome, Firefox, Safari, Opera, Internet Explorer, etc.) is called the client software, whereas the actual computer is the hardware client. This is why in Fig. 1.1, we show client and server software which is running on different computers.

Let us imagine a person who is at home and is trying to access the website for the University of Tokyo. To do so, they will have some browser software installed on their computer, for example, Firefox. To begin with, they could type in the URL (if known) for the University of Tokyo's main page in English[4] and then press "enter." Alternatively they could use a search engine to find and click on the web address. At this point, the browser generates an "HTTP request" (back arrow on Fig. 1.1). This request is then routed over many different networks (composed of routers, switches, firewalls, satellite connections, etc.) until it reaches the server whose address corresponds to the desired one.[5] Once that happens, the server (software), for example, Apache, resolves the request and sends the requested page back to the client. The webpage (in this case "index_e.html") is divided into small pieces of data called datagrams (or data packages) and these are sent back to the client over the Internet. At the client, these packages are reassembled to get the original data.

[4] `http://www.u-tokyo.ac.jp/index_e.html`.

[5] To reduce complexity and provide the reader with just an overview of how the Web works, we have avoided an explanation of datagram routing (IP tables, IP numbers, DNS, Internet registries, etc.).

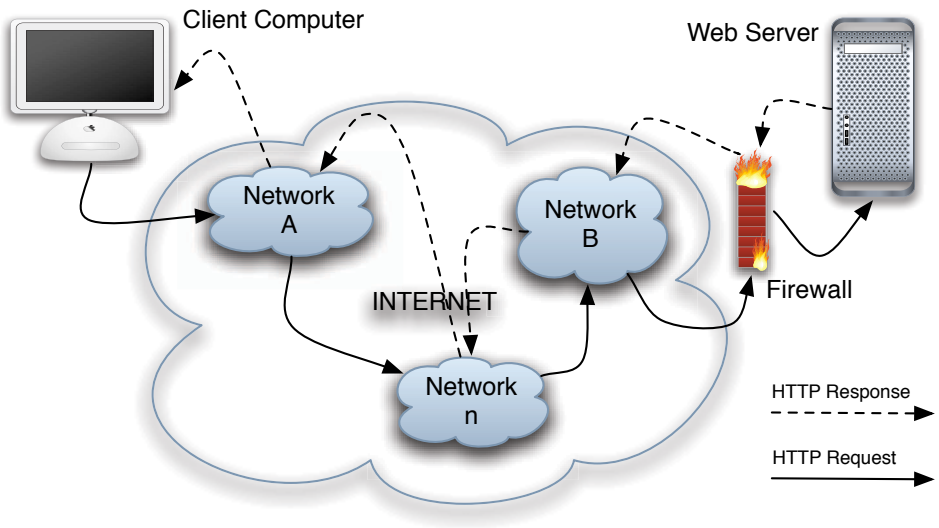

Figure 1.1: World Wide Web requests over the Internet.

In this way, the person who is using the client software is able to read a copy of the webpage "index_e.html" which was originally stored on the web server "www.u-tokyo.ac.jp."

1.3 GLOBAL IMPACT OF THE INTERNET AND THE WEB

The Internet and the World Wide Web have enabled a global information and communications revolution. Firstly, the Internet has triggered the pervasiveness of instant communications. Today it is possible to instantly locate and talk to people, via textual chats or online multicast videoconferencing systems. We are able to perform telephone calls from software to landlines using different technologies such as SIP phones, Skype, or WebRTC for web-based videoconferences. Prices are now much cheaper than the traditional high prices of calls from fixed lines to fixed lines. Email technology has become a vital tool for any organization or person. We can send messages and data over the Internet to any person in the world using a cell phone or laptop when connected to the Internet with just one touchscreen tap or mouse click.

Similarly, the Web provides many valuable services which have made our lives easier, including allowing anyone to publish and discuss his or her interests through services such as blogs, discussion forums, podcasts (personalized on-demand audio and video shows), etc. While the Web began life as a research application, it has rapidly become mature and has been adopted at a faster rate than any other communications technology thus known.

In addition to organizations, anyone is now able to have a personal website (usually for free). Blogs have also become quite popular over the past 5–10 years: these are systems whereby a person can self-publish any content he or she likes. In the same manner, other types of blogs and channels enriched with audio or video are starting to gain in popularity (e.g., SoundCloud, Vine, etc.). On Vine, the content is distributed as a short video clip produced by any user of the service (often armed with just a mobile phone camera). More recently, we have been witness to the YouTube phenomenon, one of the most visited sites on the Web and a major source of all online video traffic.

Over the past two decades, the Web has become one of the most popular transmission platforms for various forms of entertainment, and in parallel it has developed into a recognized global ecommerce framework. It is a new sales channel, a new learning space, a new communications metaphor, and so much more. The Web has reached almost every known organization and today it is possible to find online counterparts to many established institutions: universities, schools, banks, travel agencies, cinemas, etc.

Many different ways and paradigms of carrying out ecommerce on the Web have also been established, for example:

- Business to Consumer (B2C) where an organization's site is focused on selling products or services to the end consumer, e.g., Amazon.com, barnesandnoble.com, etc.
- Business to Business (B2B) where an organization's site is oriented to sell to another organization. A good example of a B2B site is Alibaba.
- Consumer to Consumer (C2C) where any person can sell or buy services from other persons or companies. One of the most popular sites of this type is eBay, and other prominent examples include Japanese site Kakaku.com and Etsy.
- Mobile to Consumer (M2C) where an organization has developed a mobile or responsive website that has been enhanced for mobile devices smartphones or tablets.

These new paradigms have brought about many advantages because—via the Web—it is now possible to buy almost any product from a variety of countries at a lower cost than was previously possible, creating new opportunities for both companies and individuals.

1.4 WEB MINING, THE SOCIAL WEB, AND THE SEMANTIC WEB

As the Web grows larger (in terms of documents and servers serving those documents) and ever more diverse (in terms of content types, topics, languages, etc.), it becomes more difficult for users to find relevant information or to carry out tasks as efficiently. Therefore some mining of relevant knowledge and patterns can assist users in navigating through this ever-expanding information resource. We will discuss various "web mining" techniques in Chapter 2.

The Web was originally envisaged to be not just an information resource, but a Web where users could collaborate and communicate with others around shared topics of interest, and also one where machines could interpret the data that the Web contained to help users with their daily online activities.

As mentioned, the Web has evolved in the first direction to have a more social aspect whereby users can interact with each other, often around content items like videos or status updates. We refer to this in the book as the Social Web (previously referred to as Web 2.0; O'Reilly [2005]), and incorporates aspects such as social software and social media. We discuss the Social Web in more detail in Chapter 3.

In parallel, various efforts have been ongoing in the second direction to make portions of the Web more machine readable for various reasons: data reuse, interoperability and interpretability, improved search around knowledge objects, etc. This direction is called the Semantic Web, and we will give an introduction to its underlying framework in Chapter 4.

1.5 CONCLUSION

In this chapter, we have introduced the book and given a short history of the Web. In the next chapter, we will look at web mining processes and frameworks that can be used to cluster and derive semantic information from the Web (and the Social Web).