Chapter XXIII
Information Extraction from Blogs
Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium

ABSTRACT

This chapter introduces information extraction from blog texts. It argues that the classical techniques for information extraction that are commonly used for mining well-formed texts lose some of their validity in the context of blogs. This finding is demonstrated by considering each step in the information extraction process and by illustrating this problem in different applications. In order to tackle the problem of mining content from blogs, algorithms are developed that combine different sources of evidence in the most flexible way. The chapter concludes with ideas for future research.

INTRODUCTION

A blog (short for Web log) is a Web based publication consisting primarily of periodic content. The content is usually displayed in a reverse chronological order. Blogs are typically social media and provide commentary on a large variety of topics on a particular subject, such as products (e.g., cars, food), people (e.g., politicians, celebrities), politics, news or health. The communication medium is primarily text, although we see an increasing focus on photographs ( photoblog), sketchblogs, videos (vlog) or audio (podcasting), or on combinations of these media. A descriptive textual component is usually present, because text is an important component in human communication. Many blogs are built in an interactive dialogue setting, but a blog can also have the form of a personal diary. Other people engage themselves to complement, freely tag or comment the content, and authors of blogs prefer to link to other content. The people who write the blogs are usually not professionals.

Blogs are very creative forms of human expression and have in our society an influence on our convictions, political opinions and societal relationships that is often underestimated. Blogs are a mirror of a society, and many different parties have an interest in monitoring their content. Businesses, lawyers, sociologists and politicians
want to know the topics that are of most concern to citizens. Police and intelligence services might find valuable links or cues to crime tracking. Citizens are interested in finding soul mates with common interests. We humans have no trouble aggregating the different media and inferring messages and interpretations from them. If we design machines that help people to search blogs, to monitor blogs, mine or summarize them, we expect from these machines a certain degree of understanding of the blog contents. Assigning a semantic meaning to blogs brings us to the domain of artificial intelligence. This chapter will treat the topic of information extraction from blogs.

In previous work we have defined information extraction as:

Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources providing additional aids to access and interpret the unstructured data by information systems. (Moens, 2006, p. 225)

Information extraction is used to get some information out of unstructured data. Written and spoken text, pictures, video and audio are all forms of unstructured data. Unstructured does not imply that the data is structurally incoherent (in that case it would simply be nonsense), but rather that its information is encoded in such a way that makes it difficult for computers to immediately interpret it. Information extraction is the process that adds meaning to unstructured, raw data, whether that is text, images, video or audio. Consequently, the data becomes structured or semi-structured and can be more easily processed by the computer.

In other words, information extraction presupposes that although the semantic information in a text and its linguistic organization is not immediately computationally transparent, it can nevertheless be retrieved by taking into account surface regularities that reflect its computationally opaque internal organization. An information extraction system will use a set of extraction patterns, which are either manually constructed or automatically learned, to take information out of a source and put it in a more structured format. When structuring this information, it is not the purpose to replace the unstructured data by the extracted information, which would be equal to imposing a certain view on the data. The goal is to complement the unstructured low level data with semantic labels so that their automated retrieval, linking, mining and visualization become more effective (Moens, 2006).

The unstructured data sources we are mainly concerned with in this chapter are written texts, possibly enriched with free tags, comments and hypermedia links. Information extraction aims here at identifying certain information for use in subsequent information systems. State-of-the-art information extraction techniques are applied to well-formed texts, i.e., consistent with the standards of an official language. However, blog data is notorious for being incoherent and full of grammatical and spelling errors. Sometimes a community or jargon language is used. The focus of this chapter is on the problems encountered by using a state of the art information extraction system when dealing with blogs.

This chapter is organized as follows. We continue with some background (next section) on information extraction in general and information extraction from blogs in particular. We outline the history of information extraction. In a subsequent section we consider the different steps in an information extraction task and focus on particular issues when dealing with blog data. We discuss tokenization and lexical analysis, natural language processing and finally information extraction. In the latter part of the chapter we go deeper into a few specific applications: topic and thread detection, opinion mining, and argumentation detection. Wherever possible, we illustrate our findings with our own research experiences. We
17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage: www.igi-global.com/chapter/information-extraction-blogs/22015?camid=4v1


Related Content

Challenges, Opportunities, and Trends in Quality K-12 Online Environments
[www.igi-global.com/chapter/challenges-opportunities-and-trends-in-quality-k-12-online-environments/137417?camid=4v1a](www.igi-global.com/chapter/challenges-opportunities-and-trends-in-quality-k-12-online-environments/137417?camid=4v1a)

Dynamic Backfilling Algorithm to Increase Resource Utilization in Cloud Computing
[www.igi-global.com/article/dynamic-backfilling-algorithm-to-increase-resource-utilization-in-cloud-computing/217692?camid=4v1a](www.igi-global.com/article/dynamic-backfilling-algorithm-to-increase-resource-utilization-in-cloud-computing/217692?camid=4v1a)

Design of Vision Measurement Device for Seeding Robot based on Ant Colony Algorithm and Nonlinear Circuit System
[www.igi-global.com/article/design-of-vision-measurement-device-for-seeding-robot-based-on-ant-colony-algorithm-and-nonlinear-circuit-system/182263?camid=4v1a](www.igi-global.com/article/design-of-vision-measurement-device-for-seeding-robot-based-on-ant-colony-algorithm-and-nonlinear-circuit-system/182263?camid=4v1a)

Quantitative Evaluation of Web2.0 Application
[www.igi-global.com/chapter/quantitative-evaluation-of-web20-application/143384?camid=4v1a](www.igi-global.com/chapter/quantitative-evaluation-of-web20-application/143384?camid=4v1a)