Mining the Internet for Concepts

Ramon F. Brena
Tecnológico de Monterrey, Mexico

Ana Maguitman
Universidad Nacional del Sur, Argentina

Eduardo H. Ramirez
Tecnológico de Monterrey, Mexico

INTRODUCTION

The Internet has made available a big number of information services, such as file sharing, electronic mail, online chat, telephony and file transfer. However, services that provide effective access to Web pages, such as Google, are the ones that most contributed to the popularization and success of the World Wide Web and the Internet. Pages published at the World Wide Web belong to many different topic areas, such as music, fishing, travel, etc. Some organizations have tried to organize pages in a predefined classification, and have manually built large directories of topics (e.g. *Dmoz* or the *Yahoo!* directory). But given the huge size and the dynamic nature of the Web, keeping track of pages and their topic manually is a daunting task. There is also the problem of agreeing on a standard classification, and this has proved to be a formidable problem, as different individuals and organizations tend to classify things differently. Another option is to rely on automatic tools that mine the Web for "topics" or "concepts" related to online documents. This approach is indeed more scalable than the manual one. However, automatically classifying documents in topics is a major research challenge. This is because the document keywords alone seem to be insufficient to directly convey the meaning of the document to an autonomous system. In some cases, the main difficulty is due to the ambiguity of the terms encountered in the document. Even if the ambiguity problems were solved there is still no guarantee that the vocabulary used to describe the document will match that used by the autonomous system to guide its search.

Central to automatic approaches is the notion of "semantic context", which loosely means the subject or topic where a task like searching is embedded. Of course, we need a way to computationally represent this notion of context, and one possibility is to see context as a collection of interrelated terms in the sense that they appear together in a number of related pages (Ramirez & Brena, 2006). For instance, the word "Java" appears together with "roasted" when talking about coffee, but appears more frequently with "code" when talking about a programming language. Semantic contexts allow performing searches on the Web at the concept level, rather than at the more basic keyword level. In this chapter we present recent advances in automated approaches in web concept mining, emphasizing our own work about mining the Web for semantic contexts.

BACKGROUND

The notion of semantic contexts is closely linked to that of *semantic similarity*. Two documents are semantically similar if they belong to the same topic or to similar topics. Likewise, two words are semantically similar if they represent similar concepts.

The study of semantic similarity between documents and terms has long been an integral part of information retrieval and machine learning, and there is extensive literature on measuring the semantic similarity between entities (Resnik, 1995; Lin, 1998; Hatzivassiloglou et al. 1999; Landauer et al. 1997; Turney, 2001; Maguitman et al. 2005; Ramirez & Brena, 2006; Sahami et al. 2006; Bollegala, 2007). These methods can be roughly classified into two major categories: knowledge-based and corpus-based approaches.

Knowledge-based approaches rely on the use of predefined directories or ontologies to identify semantic relations. Measures of semantic similarity between entities take as a starting point the structure of a directory or ontology where the entities have been previously classified. Examples of such ontologies include

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

WordNet for the case of terms, and *Dmoz* or the *Yahoo! Directory* for the case of document. Ontologies are a special kind of network and early proposals to estimate semantic similarity have used path distances between the nodes in the network representation (Rada et al. 1989). Some successors have looked into the notion of information content (Resnik, 1995; Lin, 1998) or have combined both distance and information content (Jiang & Conrath, 1997) to assess semantic similarity. In an information theoretic approach, the semantic similarity between two entities is related to their commonality and to their differences. Given a set of entities in a hierarchical taxonomy, the commonality of two entities can be estimated by the extent to which they share information, indicated by the most specific class in the hierarchy that subsumes both. Once this common classification is identified, the meaning shared by two entities can be measured by the amount of information needed to state the commonality of the two objects. Generalizations of the information theoretic notion of semantic similarity for the case of general ontologies (i.e., taxonomies that include both hierarchical and non-hierarchical components) have been proposed by Maguitman et al. (2005).

Keeping these ontologies up-to-date is expensive. For example, the semantic similarity between terms changes across different contexts. Take for instance the term java, which is frequently associated with the java programming language among computer scientist. However, this sense of java is not the only one possible as the term may be referring to the java coffee, the java island or the java Russian cigarettes, among other possibilities. New words are constantly being created as well as new senses are assigned to existing words. As a consequence, the use of knowledge-based approaches has disadvantages because they require that the ontologies be manually maintained.

Corpus-based approaches, on the other hand, can help to automate the process of keeping the ontology up-to-date. They are based on information exclusively derived from large corpora (such as the World Wide Web). A well-known approach of corpora-based method is latent semantic analysis (Landauer et al. 1997), which applies singular value decomposition to reduce the dimensions of the term-document space, harvesting the latent relations existing between documents and between terms in large text corpora. Less computationally expensive techniques are based on mapping documents to a kernel space where documents that do not share any term can still be close to each other (Cristianini et. al 2001; Liu et al. 2004). Another corpus-based technique that has been applied to estimate semantic similarity is PMI-IR (Turney, 2001). This information retrieval method is based on pointwise mutual information, which measures the strength of association between two elements (e.g., terms) by contrasting their observed frequency against their expected frequency.

In general, automatic methods to compute similarity between texts have applications in many information retrieval related areas, including natural language processing and image retrieval from the Web, where the text surrounding the image can be automatically augmented to convey a better sense of the topic of the image. Automatic methods to identify the topic of a piece of text have also been used in text summarization (Erkan and Radev 2004), text categorization (Ko et al. 2004), word sense disambiguation (Schutze 1998), evaluation of text coherence (Lapata and Barzilay 2005) and automatic translation (Liu and Zong 2004), but most of them use human-defined categories for topics, becoming thus prone to the problems we mentioned before, like disagreement, difficulty to update, etc.

MAIN FOCUS

We discuss two completely corpus-based methods proposed by the authors that can be used to identify semantic contexts and applied to mine the Web for concepts. The first is based on the notion of k-core and the second is based on the use of incremental methods.

K-Core Method

In the k-core method, the main idea is to find groups of *k* keywords (for instance, 4 keywords) that appear together in a big number of pages. The basic assumption is that words with related meanings appear together more often than unrelated words. The number of co-occurrences is readily obtained using the current indexing-based search methods; for instance, *Google* search includes this number in every single search results page (upper-right corner). So, co-occurrences based methods could be extremely efficient and well integrated with internet search technology.

A "Semantic Context" is defined in this method as the relative weights of keywords in a given topic. For instance, when talking about coffee, the word sugar is

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/mining-internet-concepts/10991

Related Content

Learning Exceptions to Refine a Domain Expertise
Rallou Thomopoulos (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1129-1136).*
www.irma-international.org/chapter/learning-exceptions-refine-domain-expertise/10963

View Selection in DW and OLAP: A Theoretical Review
Alfredo Cuzzocrea (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 2048-2055).*
www.irma-international.org/chapter/view-selection-olap/11101

Soft Computing for XML Data Mining
K. G. Srinivasa, K. R. Venugopaland L. M. Patnaik (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1806-1809).*
www.irma-international.org/chapter/soft-computing-xml-data-mining/11063

Variable Length Markov Chains for Web Usage Mining
José Borgesand Mark Levene (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 2031-2035).*
www.irma-international.org/chapter/variable-length-markov-chains-web/11098

Transferable Belief Model
Philippe Smets (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1985-1989).*
www.irma-international.org/chapter/transferable-belief-model/11091