Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums

Jiang-Ming Yang[†], Rui Cai[†], Yida Wang[‡], Jun Zhu[§], Lei Zhang[†], and Wei-Ying Ma[†]

[†]Microsoft Research, Asia. {jmyang, ruicai, leizhang, wyma}@microsoft.com

[‡]CSSAR, Chinese Academy of Sciences. wangyida@cssar.ac.cn

[§]Dept. Computer Science and Technology, Tsinghua University. jun-zhu@mails.tsinghua.edu.cn

ABSTRACT

Web forums have become an important data resource for many web applications, but extracting structured data from unstructured web forum pages is still a challenging task due to both complex page layout designs and unrestricted user created posts. In this paper, we study the problem of structured data extraction from various web forum sites. Our target is to find a solution as general as possible to extract structured data, such as *post title*, *post author*, *post time*, and *post content* from any forum site. In contrast to most existing information extraction methods, which only leverage the knowledge inside an individual page, we incorporate both *page-level* and *site-level* knowledge and employ Markov logic networks (MLNs) to effectively integrate all useful evidence by learning their importance automatically. *Site-level* knowledge includes (1) the linkages among different object pages, such as *list pages* and *post pages*, and (2) the interrelationships of pages belonging to the same object. The experimental results on 20 forums show a very encouraging information extraction performance, and demonstrate the ability of the proposed approach on various forums. We also show that the performance is limited if only page-level knowledge is used, while when incorporating the site-level knowledge both precision and recall can be significantly improved.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrival**]: Miscellaneous - Data Extraction; Web; H.3.3 [**Information Storage and Retrival**]: Information Search and Retrieval - Information filtering; I.5.1 [**Pattern Recognition**]: Models - Statistical

General Terms

Algorithms, Performance, Experimentation

Keywords

Web forums, Structured data, Information extraction, Site-level knowledge, Markov logic networks (MLNs)

This work was performed when the 3^{rd} and the 4^{th} authors were visiting Microsoft Research, Asia.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

1. INTRODUCTION

The rapid growth of Web 2.0 is making web forums (also named bulletin or discussion board) an important data resource on the Web. The strong driving force behind web forums is the power of users and communities. With millions of users' contribution, plenty of highly valuable knowledge and information have been accumulated on various topics including recreation, sports, games, computers, art, society, science, home, health, etc. [1]. As a result, recent years have witnessed more and more research efforts trying to leverage information extracted from forum data to build various web applications. Some exemplary applications include extracting question-answer pairs for QnA service [5]; collecting review comments for business intelligence [6], and discovering expertise networks in online communities [18].

To use forum data, the fundamental step in most applicants is to extract structured data from unstructured forum pages represented in HTML format by removing useless HTML tags and noisy content like advertisements. Only after extracting such structured data can we further exploit forum data to discover communities, find emerging topics, model user interests, etc. However, automatically extracting structured data is not a trivial task due to both complex page layout designs and unrestricted user created posts. This problem has become a major hindrance for efficiently using web forum data.

In this paper, we study the problem of structured data extraction from web forum sites. Our target is to find a solution as general as possible to extract structured data such as *post title*, *post author*, *post time*, and *post content* from any forum site. Because the number of forums is very large, at least tens of thousands, the task is very challenging. However, no matter how complex the forum page layouts are, forum sites have some intrinsic characteristics which make it possible to find a general solution. The most relevant work for structuring web forum data is web data extraction, which has been an active research topic in recent years [2,7, 10,11,17,19–21]. In general, web data extraction approaches can be classified into two categories: *template-dependent* and *template-independent*.

Template-dependent methods, as the name implies, utilize a wrapper as an extractor for a set of pages which were generated based on the same layout template. A wrapper is usually represented in the form of a regular expression or a tree structure. Such a wrapper can be manually constructed, semi-automatically generated by interactive learning [19], or even discovered fully automatically [17]. However, for web forums, different forum sites usually employ