From Web Content Mining to Natural Language Progressing

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street, Chicago, IL 60607-7053
liub@cs.uic.edu

Abstract

This tutorial introduces some important tasks of Web content mining and their connections with natural language processing (NLP), and encourages NLP researchers to join the Web content mining research.

1 Introduction

Web mining consists of usage mining, structure mining, and content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from hyperlinks. Web content mining aims to extract/mine useful knowledge from Web page contents. This tutorial focuses on Web content mining and its extensive connections with natural language processing (NLP).

In the past few years, there was a rapid expansion of activities in Web content mining because of the huge amount of valuable information of almost any imaginable type on the Web and significant economic benefits of such mining. However, due to the heterogeneity and the lack of structure of the Web data, automated discovery of useful knowledge still presents challenging problems. This tutorial introduces several such problems. These problems all have strong connections with NLP. In the tutorial, special attentions will be paid to such connections. Many real-life examples are also given to help participants understand research concepts and see how the technologies may be deployed to real-life applications. The tutorial thus has a mix of research and industry flavor, addressing seminal research ideas and looking at the technology from an industry angle.

2 Tutorial Outline

1. **Structured data extraction**
   The problem
   Wrapper induction
   Automated extraction
   NLP connection
2. **Information integration**
   The problem
   Some integration techniques
   Web query interface integration
   NLP connection
3. **Information synthesis**
   The problem
   Exploiting information redundancy
   Using syntactic patterns
   NLP connection
4. **Opinion mining**
   The abstraction
   Document level sentiment classification
   Sentence level sentiment analysis
   Feature-based opinion mining & summarization
   Comparative sentence and relation mining

3 Presenter Biography

Bing Liu is an associate professor of Computer Science at the University of Illinois at Chicago. His research interests include data, Web and text mining. He has published extensively in these areas in leading conferences, e.g., KDD, WWW, AAAI, IJCAI, ICML and SIGIR. He has served or serves as program chairs, vice chairs or senior PC members of many data mining related conferences, including KDD, WWW, ICDM, SDM, CIKM and PAKDD, and also as an associate editor of IEEE Transactions on Knowledge and Data Engineering, and an associate editor of SIGKDD Explorations.