Chapter I

Information Extraction: Methodologies and Applications

Jie Tang
Tsinghua University, China

Mingcai Hong
Tsinghua University, China

Duo Zhang
Tsinghua University, China

Bangyong Liang
NEC Labs, China

Juanzi Li
Tsinghua University, China

ABSTRACT

This chapter is concerned with the methodologies and applications of information extraction. Information is hidden in the large volume of Web pages and thus it is necessary to extract useful information from the Web content, called information extraction. In information extraction, given a sequence of instances, we identify and pull out a subsequence of the input that represents information we are interested in. In the past years, there was a rapid expansion of activities in the information extraction area. Many methods have been proposed for automating the process of extraction. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still presents many challenging research problems. In this chapter, we will investigate the problems of information extraction and survey existing methodologies for solving these problems. Several real-world applications of information extraction will be introduced. Emerging challenges will be discussed.
INTRODUCTION

Information extraction (IE), identifying and pulling out a subsequence from a given sequence of instances that represents information we are interested in, is an important task with many practical applications. Information extraction benefits many text/Web applications, for example, integration of product information from various Web sites, question answering, contact information search, finding the proteins mentioned in a biomedical journal article, and removal of the noisy data.

We will focus on methodologies of automatic information extraction from various types of documents (including plain texts, Web pages, e-mails, etc.). Specifically, we will discuss three of the most popular methods: rule learning based method, classification model based method, and sequential labeling based method. All these methods can be viewed as supervised machine learning approaches. They all consist of two stages: extraction and training.

In extraction, the subsequences that we are interested in are identified and extracted from given data using learned model(s) by different methods. The extracted data are then annotated as specific information on the basis of the predefined metadata.

In training, the model(s) are constructed to detect the subsequence. In the models, the input data is viewed as a sequence of instances. For example, a document can be viewed as either a sequence of words or a sequence of text lines (it depends on the specific application).

All these methodologies have immediate real-life applications. Information extraction has been applied, for instance, to part-of-speech tagging (Ratnaparkhi, 1998), named entity recognition (Zhang, Pan & Zhang, 2004), shallow parsing (Sha & Pereira, 2003), table extraction (Ng, Lim, & Koo, 1999; Pinto, McCallum, Wang, & Hu, 2002; Wei & Croft, 2003), and contact information extraction (Kristjansson, Culotta, Viola, & McCallum, 2004).

In the rest of the chapter, we will describe the three types of the state-of-the-art methods for information extraction. This is followed by presenting several applications to better understand how the methods can be utilized to help businesses. The chapter will have a mix of research and industry flavor, addressing research concepts and looking at the technologies from an industry perspective. After that, we will discuss the challenges the information extraction community has faced. Finally, we will give concluding remarks.

METHODOLOGIES

Information extraction is an important research area, and many research efforts have been made so far. Among these research works, rule learning based method, classification based method, and sequential labeling based method are the three state-of-the-art methods.

Rule Learning-Based Extraction Methods

In this section, we review the rule based algorithms for information extraction. Numerous information systems have been developed based on the method, including: AutoSlog (Riloff, 1993), Crystal (Soderland, Fisher, Aseltine, & Lehnert, 1995), (LP)2 (Ciravegna, 2001), iASA (Tang, Li, & Lu, 2005), Whisk (Soderland, 1999), Rapier (Califf & Mooney, 1998), SRV (Freitag, 1998), WIEN (Kushmerick, Weld, & Doorenbos, 1997), Stalker (Muslea, Minton, & Knoblock, 1998, 1999), BWI (Freitag & Kushmerick, 2000), and so forth. See Muslea (1999), Siefkes and Sinha (2005), and Peng (2001) for an overview. In general, the methods can be grouped into three categories: dictionary based method, rule based method, and wrapper induction.
Related Content

Productivity Analysis of Public Services: An Application of Data Mining
[www.igi-global.com/chapter/productivity-analysis-public-services/44284?camid=4v1a](www.igi-global.com/chapter/productivity-analysis-public-services/44284?camid=4v1a)

Visual Mobility Analysis using T-Warehouse
[www.igi-global.com/article/visual-mobility-analysis-using-warehouse/49638?camid=4v1a](www.igi-global.com/article/visual-mobility-analysis-using-warehouse/49638?camid=4v1a)

A Hybrid Method for High-Utility Itemsets Mining in Large High-Dimensional Data
[www.igi-global.com/article/hybrid-method-high-utility-itemsets/1823?camid=4v1a](www.igi-global.com/article/hybrid-method-high-utility-itemsets/1823?camid=4v1a)

On Data Mining and Knowledge: Questions of Validity
[www.igi-global.com/chapter/data-mining-knowledge/44288?camid=4v1a](www.igi-global.com/chapter/data-mining-knowledge/44288?camid=4v1a)