Chapter 1

Company Acquisition Relations Extraction From Web Pages

Jie Zhao
Anhui University, China

Jianfei Wang
Anhui University, China

Jia Yang
University of Science and Technology of China, China

Peiquan Jin
University of Science and Technology of China, China

ABSTRACT

In this chapter, we study the problem of extracting company acquisition relation from huge amounts of webpages, and propose a novel algorithm for a company acquisition relation extraction. Our algorithm considers the tense feature of Web content and classification technology of semantic strength when extracting company acquisition relation from webpages. It first determines the tense of each sentence in a webpage, where a CRF model is employed. Then, the tense of sentences is applied to sentences classification so as to evaluate the semantic strength of the candidate sentences in describing company acquisition relation. After that, we rank the candidate acquisition relations and return the top-k company acquisition relation. We run experiments on 6144 pages crawled through Google, and measure the performance of our algorithm under different metrics. The experimental results show that our algorithm is effective in determining the tense of sentences as well as the company acquisition relation.

INTRODUCTION

Web has been the most important information source in the world. It has been widely accepted that webpages include useful competitive intelligence for enterprises (Zhao and Jin, 2009). Enterprise competitive intelligence refers to the intelligence related with competitors, competitive environment, and competitive strategies, among which the competitive strategies is the most difficult type of competitive intelligence to obtain in real applications. The competitive strategies about a certain company usually
hide in a lot of phenomenon. For example, the future developing strategy of a company, e.g., IBM, may be reflected by its acquirement actions. There is a real case regarding IBM. Recently, this company is advancing its development on big data analysis (IBM, 2015). But if we check the acquirement history of IBM in recent years, we can find that since 2009 it keeps buying companies on data analysis, such as SPSS, Emptoris, and Netzza. Thus, if we can extract the company acquirement relation about concerned competitors from the Web and detect their early intends in strategical development, it will be helpful to conduct deeply analysis on enterprise competitive intelligence (Zhao and Jin, 2010a).

Company acquirement relation can be represented as a triple \(<\text{ORG1}, \text{verb\_tense}, \text{ORG2}>\), where \(\text{ORG1}\) is the acquirer while \(\text{ORG2}\) is the acquiree and \(\text{verb\_tense}\) refers to the tense of the acquiring action. The \(\text{verb\_tense}\) usually indicates the state of the acquirement relation. For instance, in a sentence saying “IBM has acquired Algorithmics”, we can detect that there exists an acquirement relation between IBM and Algorithmics and this action has been actually completed due to the verb tense “has acquired”.

Compared with traditional techniques in relation extraction, the extraction of company acquirement relation in the Web has two key issues:

1. How to detect the tense of a sentence?

   In English sentences, the sentence tense has an important impact on the effectiveness of company acquirement relations extraction. For example, the sentence “IBM has acquired Algorithmics” is with the tense of present perfect, which indicates the validness of the acquirement relation between IBM and Algorithmics. However, in a sentence with future tense such as “IBM will acquire Algorithmics”, the acquirement relation may not be valid. Another challenging issue is that some sentences imply negative semantics, e.g. “IBM rumored to acquire Algorithmics”.

2. How to determine the semantic strength of sentences?

   Web information usually has multiple styles, as different Web information sources have differing format when describing events or news. Some typical sources of Web information are blog, BBS (Bulletin Board System), Wikipedia, and news pages. Besides, webpages may differ in the time factor, e.g., the publication dates of webpages are usually different. As a result, there are various forms of company acquirement relations in the Web, so we cannot simply detect them by analyzing one Web page. The more reasonable way is to check a set of related webpages and detect their semantic strength in asserting the validness of a possible company acquirement relation. This problem is essentially an issue of classification, as we have to find the webpages showing positive semantics and others showing negative semantics, and then make further decision on the validness of company acquirement relations.

   Traditional ways in relation extraction only considered the limited characters of sentences, such as morphology and syntax (Zhou et al., 2005; Zhang et al., 2005; García and Gamallo, 2011), and lacked the analysis on the tense of sentences as well as the semantic strength of sentences in determining the validness of relations. Some recent works showed that the tense of sentences has a very important impact on relation extraction (Zhao and Jin, 2010b). Therefore, we introduce the tense of sentences into the extraction of company acquirement relations.

   In this paper, aiming at solving the challenging problems in extracting company acquirement relations from the Web, we present a new algorithm. The idea of our algorithm is to introduce the tense of sentences into the extraction process. In particular, we first label the tense of each sentence in each Web