Chapter 7.12
Search Engine–Based Web Information Extraction

Gijs Geleijnse
Philips Research, The Netherlands

Jan Korst
Philips Research, The Netherlands

ABSTRACT

In this chapter we discuss approaches to find, extract, and structure information from natural language texts on the Web. Such structured information can be expressed and shared using the standard Semantic Web languages and hence be machine interpreted. In this chapter we focus on two tasks in Web information extraction. The first part focuses on mining facts from the Web, while in the second part, we present an approach to collect community-based meta-data. A search engine is used to retrieve potentially relevant texts. From these texts, instances and relations are extracted. The proposed approaches are illustrated using various case-studies, showing that we can reliably extract information from the Web using simple techniques.

INTRODUCTION

Suppose we are interested in *'the countries where Burger King can be found'*, *'the Dutch cities with a university of technology'* or perhaps *'the genre of the music of Miles Davis'*. For such diverse factual information needs, the World Wide Web in general and a search engine in particular can provide a solution. Experienced users of search engines are able to construct queries that are likely to access documents containing the desired information. However, current search engines retrieve Web pages, not the information itself[1]. We have to search within the search results in order to acquire the information. Moreover, we make implicit use of our knowledge (e.g. of the language and the domain), to interpret the Web pages.

DOI: 10.4018/978-1-60566-112-4.ch009

Table 1. Comparison between the Web as a corpus and 'traditional' corpora

Web Corpus	Newspaper Corpus
Redundancy. Because of the size of the Web, we can expect information to be duplicated, or formulated in various ways. If we are interested in a fact, we have to be able to identify just one of the formulations to extract it.	**No or fewer redundancy.** Especially for smaller corpora, we cannot expect that information is redundantly present.
Temporal and unreliable. The content of the Web is created over several years by numerous contributors. The data is thus unreliable and may be out-dated.	**Constant and reliable**. In corpus-based IE, it is assumed that the information in the corpus is correct and up-to-date.
Multilingual and heterogeneous. The Web is not restricted to a single language and the texts are produced by numerous authors for diverse audiences.	**Often monolingual and homogeneous.** If the author or nature (e.g. articles from the Wall Street Journal) of the corpus is known beforehand, it is easier to develop heuristics or to train named entity recognizers.
No representative annotated corpora. As no representative annotated texts are available, the Web as a corpus is currently less suited for supervised machine learning approaches.	**Annotated test corpora available.** In order to train supervised learning based named entity recognizers (NERs), test corpora are available where instances of a limited number of classes are marked within the text.
Dynamic. The contents of the Web changes continuously, results of experiments may thus also change over time.	**Static.** Experimental results are independent of time and place as the corpora are static.
Facts and opinions. As a multitude of users contribute to the Web, its contents are also suited for opinion mining.	**Facts only.** Information Extraction tasks on Newspaper corpora mainly focus on the identification of facts.

Apart from factual information, the Web is the de-facto source to gather community-based data as people with numerous backgrounds, interests and ideas contribute to the content of the Web. Hence the Web is a valuable source to extract opinions, characterizations and perceived relatedness between items.

In this chapter, the focus is on gathering and structuring information from the 'traditional' Web. This structured information can be represented (and shared) using the standard Semantic Web (SW) languages. Hence, this chapter focuses on the automatic creation of content for the SW. For simplicity, we abstract from the SW standards RDF(S)/OWL.

The Web-as-a-Corpus vs. Traditional Text Corpora

Information extraction (IE) is the task of identifying instances (or *named entities*) and relations between those instances in a collection of texts, called a text corpus.

In the nineties, the Message Understanding Conferences (MUC) focused on the recognition of named entities (such as names of persons and organizations) in a collection of texts (Chinchor, 1998). Initially, this work was mostly based on rules on the syntax and context of such named entities. For example, two capitalized words preceded by *mr.* will denote the name of a male person. As the creation of such rules is a laborious task, approaches became popular where named entities were recognized using machine learning (Mitchell, 1997), for example in (Zhou & Su, 2002; Brothwick, 1999; Finkel, Grenager, & Manning, 2005). However, such approaches typically make use of annotated training sets where instances (e.g. *'Microsoft'*) are labeled with their class *('Organization'*).

Traditional information extraction tasks focus on the identification of named entities in large text corpora such as collections newspaper articles or biomedical texts. In this chapter however, we focus on the Web as a corpus. In Table 1 the most important differences between the two can be found.