A Novel Approach for Crawling the Opinions from World Wide Web

Surbhi Bhatia, Department of Computer Engineering, Banasthali Vidyapith, Rajasthan, India
Manisha Sharma, Department of Computer Science, Banasthali University, Jaipur, India
Komal Kumar Bhatia, Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, India

ABSTRACT

Due to the sudden and explosive increase in web technologies, huge quantity of user generated content is available online. The experiences of people and their opinions play an important role in the decision making process. Although facts provide the ease of searching information on a topic but retrieving opinions is still a crucial task. Many studies on opinion mining have to be undertaken efficiently in order to extract constructive opinionated information from these reviews. The present work focuses on the design and implementation of an Opinion Crawler which downloads the opinions from various sites thereby, ignoring rest of the web. Besides, it also detects web pages which frequently undergo updation by calculating the timestamp for its revisit in order to extract relevant opinions. The performance of the Opinion Crawler is justified by taking real data sets that prove to be much more accurate in terms of precision and recall quality attributes.

KEYWORDS


INTRODUCTION

With the exponential increase in the trend of taking decisions based on the reviews and feedbacks collected on the World Wide Web, there has been consistent accumulation of large unstructured user generated contents over the internet. The products which are bought to the leaders they choose and policies they make, everything is carried forward by the opinions that are shared and communicated on the social platforms. These feedbacks may vary to a large extent as whether or not to buy a certain product and it influences the policy making process and many more. But, to look at the reviews in a concise and appropriate manner is a cumbersome task which needs to be achieved with mining. Recent studies (Pang, Lee, & Vaithyanathan, 2002; Hu & Liu, 2004; Hu & Liu, 2004) have been focused consistently and actively carried on to mine opinions. The primary objective of opinion mining is to categorically present these opinions to users illustrating the preference, featuring at the document level and then more precisely at the sentence level. Proper web crawler system needs to be developed and implemented for various social networking sites, mobile sites, E commerce sites and review sites.

Previously, many methods have been developed to analyze the reviews (Liu, 2011). Different researchers have developed various methodologies but still the familiar concept to retrieve the reviews is under consideration. The common aspect which has discussed here is about the crawler. Traversing the web documents and gathering information through text is done via crawler (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001; Brin & Page, 1998; Burner, 1997). Some parts of the web pages...
are constantly updated like reviews posted by the autonomous users. These dynamically generated web pages like database driven web page (Sharma & Dixit, 2008) make the crawler to revisit that web page so as to frequently collect the new and the updated opinions. Here the effort has been to calculate the frequency of revisiting the web page in context of collecting opinions.

Web crawlers are programs or bots that traverse web pages and create an index of the pages (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001). The primary objective of web crawler is to serve for the later processing of search engine by storing a copy of the visited URL in a file or a repository which will further index the downloaded HTML pages for fast searching.

The general architecture of the Web crawler is given in Figure 1. The Web crawler contains several components explained below:

- **URL Database**: It stores the collection of URL’s;
- **URL queue**: The seed URLs are given in the URL queue which follows the FIFO approach;
- **Data Fetcher**: It works to collect URL from the queue one by one and fed the data from the World Wide Web;
- **Web Pages**: It is a repository of the stored web pages;
- **DNS Repository**: The HTTP socket to threads is created after resolving DNS by applying threads;

Figure 1. General architecture of the web crawler
DE-ForABSA: A Novel Approach to Forecast Automobiles Sales Using Aspect Based Sentiment Analysis and Differential Evolution
www.igi-global.com/article/de-forabsa/217482?camid=4v1a