Using WebCrawling and DataMining Technique to Analyse Business Domain

Arathy R P
Master of Computer Applications
Mohandas College of Engineering And Technology
arathy096@gmail.com

Mr. Rajesh D
Asst. professor, Master of Computer Application
Mohandas College of Engineering And Technology

***Abstract*—It has become an era where everything is on the web with ever more chances of data utilization on the web. Still, there are obstacles to make the use of the web efficiently. With too much information, Internet users have often come across information that are not relevant for their use. On top of that, until recently, most of web content have not contained semantic information, posing difficulties for mechanical analysis. The Semantic Web emerged as a way to tackle those poor qualities of the web. In this study, what we aimed at is building a small business knowledge base to provide useful information for small business owners for their marketing strategies systems for their services. The knowledge base was built according to the concept of the Semantic Web. To build the knowledge base, first, it is needed to conduct web crawling from different web sources including social media. However, the crawled data typically come in informal and do not have any semantic information. So we devised text mining techniques to catch useful information from them and generate formal knowledge for the knowledge base.**

Keywords—semantic web; knowledge base; knowledge discovery, small business, social data analysis, web crawling, data mining

I. INTRODUCTION

As more and more web content is produced on the web, the need to sort out the fast growing data is increasingly in demand. On top of that, most available web content had been written in markup languages such as HTML [1], which is designed for web browsers to parse so that they can draw graphical layouts mainly containing texts, links, pictures, and sometimes other types of multimedia content

Against the background where knowledge on the web was fast-growing and still mostly did not contain semantic information in it, a new paradigm was proposed: Semantic Web. Those languages are designed to describe information semantically by adopting formal language structures and schemas. In semantic web languages, schemas play a role in defining the structure of information and possible relations among different types of information. The content must be written according to the schemas, and it amounts to instances of the schemas. This semantic web content is highly available for various tasks including queries[5], reasoning[6] and visualization.

In this background, our work is about building a knowledge base of small business domain and publishing it on the web to provide the knowledge for online users who want to get information about various types of businesses including restaurants, stores, etc. Among businesses, a lot of small businesses are not well-known to the public. Even though some static features of the small businesses such as their names, locations, contacts are available on some websites, more detailed features such as their reputations are hard to catch since they have shortage of means of marketing themselves by hiring or encouraging online users to write reviews for their services.

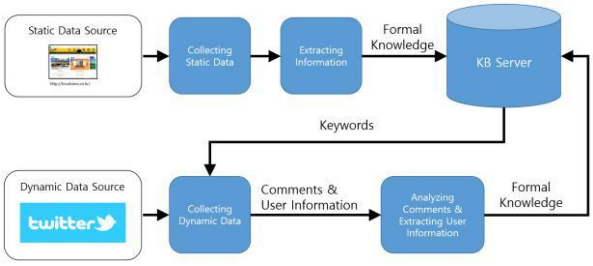
To obtain these hard-to-catch types of information of small businesses, our work is focused on utilizing social networking sites as data sources. In these sites, users post comments about their everyday life, and in some cases, the comments involve where they went, spent time, and dined out with their opinions whether they are positive or not. These user opinions are useful to understand their preferences, and reputations of places or services they used. However, users' comments on social networking sites usually are written in natural languages not in formal formats. Therefore, proper text mining techniques are needed to catch required information and also guarantee the reliability of the information.

In this study, we aimed to implement an automated knowledge discovery on small businesses by using local business websites and Twitter as data sources and analyzing the data with data mining techniques. We defined websites which contain information about local small businesses as static data sources and Twitter as a dynamic data source. We first tried to obtain keywords about small businesses and use the keywords for search queries on Twitter. To analyse users' comments obtained from the queries, we devise text mining techniques to figure out how much positive or negative the comments mean. To get user information who wrote the obtained comments, we defined a list of keywords that can be clues for users' identities such as age, gender, and occupation. Finally, these data were processed to form semantic information about user preferences on small businesses and aggregated to provide useful information for relevant requests from small business owners.

II. SYSTEM ARCHITECTURE

The overall processes for the knowledge discovery are overviewed in the Figure 1. It involves web crawling processes from static and dynamic data sources on the Internet, information extraction and analysis processes from each data source, and the knowledge base server, which contains the knowledge database.

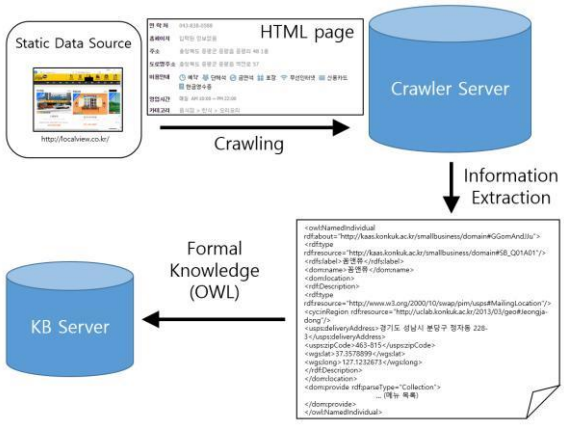
Fig. 1 Overview of the Knowledge Discovery Process



A. *Collecting Static Data*

Among various types of data on small businesses, we defined data that are not produced by users and likely to remain unchanged for a long time as static data. These data include their names, locations, contacts, types of businesses, etc. To obtain static data of small businesses, we used localview.com as the source[3]. The overall process of crawling the static data is shown in the Figure 2.

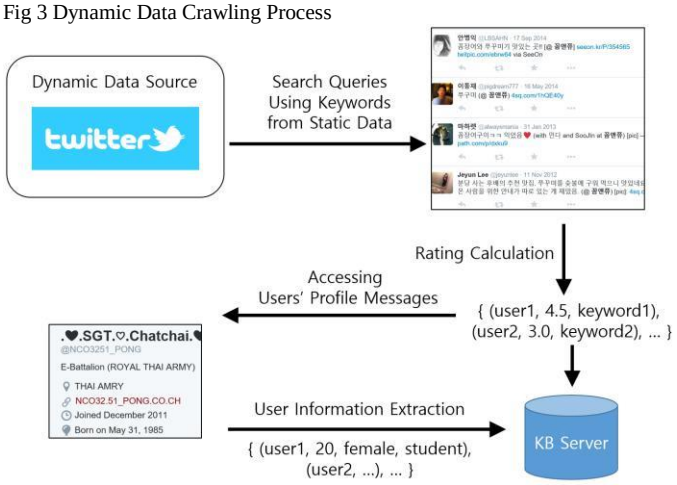
Fig 2 Static Data Crawling Process



B. *Collecting Dynamic Data*

We also defined another type of data: dynamic data, which are produced by users and more likely to change over time than static ones. Those data include businesses' reputations for their services. We chose Twitter as the dynamic data source. We used businesses' names obtained from the static data source as keywords to perform search queries on twitter, and collected tweets containing the keywords and profile messages of users who wrote the tweets. To produce meaningful results, there is the need to perform additional processes including text mining techniques. These steps are seen in the Figure 3, and the additional process to analyze comments will be explained in the next section.

Fig 3 Dynamic Data Crawling Process



C. *Analyazing Comments*

Each of tweets containing businesses' names can reflect the user's opinion about the business. However, as for Tweeter, users' comments just come in words and do not have any ratings. Therefore, some kind of analysis is needed to figure out how much positive or negative each of the tweets are about the businesses.

To tackle this challenge, we devised a method to measure how much positive a user's comment on social media where the comments usually do not come with ratings. The first step is to collect users' comments on websites where the comments come with ratings. And the next step is to split each comment by word and count the number of each word occurring in the comments. At the same time, for each word, it is needed to add up the ratings of comments where the word occurred. After figuring out the number of occurrences and the sum of the ratings of each word, dividing the latter by the former produces the average rating for each word. Naturally, words that often appear in positive or high-rating reviews show high average ratings, and words that often appear in negative or low-rating reviews show low average ratings.

With the table consisting of words and their average ratings, it is possible to measure ratings of any comments. The way goes as follows: a comment is split by word, then dividing the sum of the average ratings of all the words in the comment by the number of words in the comment produces the final figure, which can be considered as the rating for the comment.. Those words are considered to have the median value of the maximum rating. For instances, if the maximum rating is 5.0, then the median value is to be 2.5. As a result, the more positive words a comment has, the more rating it gets

.

Fig 4 Comment Rating Example

Word	Rating
love	4.0
tasty	5.0
but	2.0
pricey	1.0
...	...

"I **love** this restaurant!
Foods here are very **tasty but** a bit **pricey**."

$$\text{Rating} = \frac{4.0 + 5.0 + 2.0 + 1.0 + \ldots}{13 \text{ (number of words)}}$$

D. Extracting User Information

Besides users' opinions, information about the users are also a great resource to support small businesses' decision making in their marketing or operating strategies. Especially, users' age, gender, occupation can be considered as crucial factors. Users on Tweeter usually show these pieces of user information on their profile messages. But the messages are just written in natural language often with their own unique self-expressions. By observing a lot of the profile messages of Tweeter users, we could identified several keywords that can be a dead give-away showing their identities. With the keywords, we could figure out users' gender, age, and occupations from their profile messages by exact keyword matching while excluding profile messages which do not contain any of the keywords. Each user's information forms a pair with his or her tweet's rating calculated by the previous step, and is stored in the knowledge base server.

III. EXPERIMENT

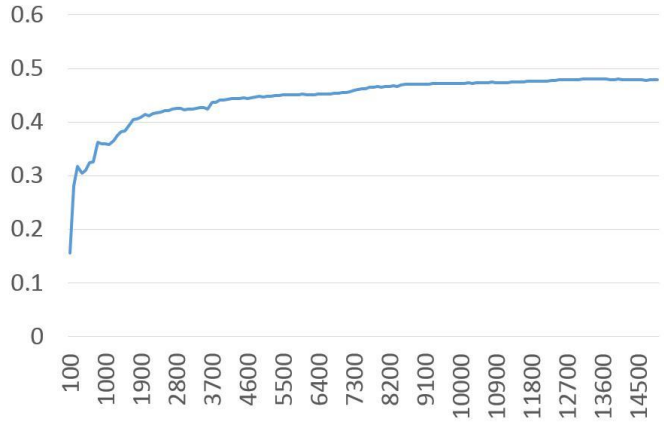
We conducted an experiment for measuring the performance of the process of analyzing comments, which is designed to produce ratings of users' comments.

First of all, to obtain users' comments with ratings, we used menupan.com where users post comments on restaurants with ratings ranging from 0.0 to 5.0., which was developed by Team Intelligent Data Systems Laboratory from Seoul National University [7].

We tried to measure the correlation coefficient between the actual ratings and the ratings the suggested method produces while adjusting the number of comments to create the table of words' ratings from 100 to 15,000. For the actual ratings, we set aside 8,329 comments and tried to measure ratings of the comments by applying our method.

The Figure 4 shows correlation coefficients of ratings the suggested method produced against the actual ratings. The y-axis represents correlation coefficient, and the x-axis represents the number of comments used to create the table of words' ratings. As shown in the graph, the more data are given for the table, the stronger correlation is seen. The maximum coefficient is around 0.48, which reflects a clear linear correlation between the two datasets.

Fig 5 Correlation Coefficients of the Suggested Method's Ratings against the Actual Ratings of the Users on menupan.com



IV. CONCLUSION

Our study was about building a small business knowledge base in an automated way by exploiting data on the web. We first defined two types of data: static data and dynamic data. For the static data, we collected data from a local website containing information about small businesses. And for the dynamic data, we used Twitter and collect tweets and users' information through open API it provides.

What we aimed to collect from the dynamic data is businesses' reputations among online users. However, users' comments on social media usually do not have ratings to indicate how much positive or negative their reactions are. So we devised a method to measure ratings of comments.. By using words' ratings obtained from the method, we could calculate ratings of comments from Tweeter. We also showed correlation coefficients of the method's results against the users' actual ratings on the website.

For future works, we will try to obtain more various information such as popular services of businesses from social media. And, in addition to Tweeter, we will also look to use other web sources. There is also the need to make sure the data are reliable since there always can be fake reviews from users who write reviews out of malicious purposes or are hired by businesses to write glowing reviews on social media. Sometimes information from users or websites can be out of date or not consistent with each other. These factors should be considered in the future as well to improve the reliability of our knowledge discovery system.

Another challenge will be filtering out data that contain target keywords but are not relevant. For instance, a business's name can be such a general term that people can use the term in different contexts in most cases. Therefore, to obtain relevant information for a search keyword, recognizing the context related to the keyword is essential when dealing with such keywords.

V. ACKNOWLEDGMENT

This project was supported by many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them. I thank my colleagues from who provided insight and expertise that greatly assisted the project.

I owe my deep gratitude to our project guide Prof. Rajesh D who took keen interest on my project work and guided me all along, till the completion of our project work by providing all the necessary information for developing a good system.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of MCA which helped me in successfully completing the project work. Also, I would like to extend our sincere esteems to all staff in laboratory for their timely support.

REFERENCES

[1] Edd Dumbill, Forbes, Volume, Velocity, Variety: What You Need to Know About Big Data, http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/

[2] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific american 284.5 (2001): 28-37.

[3] W3C, RDF - Semantic Web Standards, http://www.w3.org/RDF/

[4] W3C, OWL - Semantic Web Standards, http://www.w3.org/2001/sw/wiki/OWL

[5] W3C, Query - W3C, http://www.w3.org/standards/semanticweb/query

[6] W3C, Inference - W3C, http://www.w3.org/standards/semanticweb/inference

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.