Extracting and Classifying Text about Business in Website of Company

Kenichiro Ando (s1610224)

School of Information Science, JAIST, s1610224@jaist.ac.jp

Extended Abstract

In recent years, information on the Web is exploded, since people can easily publish their comments or opinions through a blog, Social Networking Service (SNS), curation site and so on. Therefore, there exists a lot of not only useful but also incorrect information on the Web. When users search through the Web, they need to judge whether obtained information is correct or not. One of the ways to verify reliability of information is an author of the web page. Based on the hypothesis that information provided by experts would be more reliable than one by non-experts, information of an author can be helpful to judge reliability of information. For example, when searching for something about law, information in a website of a law office or in an article written by a lawyer can be highly reliable. In this way, useful clues for guessing reliability of a website becomes more important in the future as information on the Web is rapidly increased.

This thesis focuses on estimating reliability of websites of companies, since they are often hit by a search engine. Our goal is to automatically extract information about business of a company, then classify a company's website into several categories of business types. Here, "information about business" is defined as description or text of business that a company involves. Since information about business is a kind of a profile of a company, it is regarded as an author's information of a company's website in this study. Our final goal is to develop a search engine that can show search results with authors' information of web pages in order to help users to distinguish reliable information. However, it is inappropriate to show authors' information (i.e. information about business in this study) as is, since it is usually a long text. Instead, information about business is classified into predefined business categories, then the identified categories of business are shown in search results.

The proposed method consists of two steps: (1) extraction of information about business, and (2) classification of a website into business categories. In the first step, four kinds information about business are extracted: "Description", "Keywords", "Description of company's business", and "Description of company's activity". "Description" and "Keywords" are texts marked up by a meta tag whose name attributes are "description" and "keywords", respectively. "Description of company's business" is a text about a type of business of a company, which is supposed to be written in a table in a separate web page that explains an outline of a company. "Description of company's activity" is a text

about activities and enterprises of a company, which is supposed to be written in a separate web page.

First, an HTML source file of a company's website is analyzed and represented by Document Object Model (DOM). A DOM tree representing hierarchical structure of HTML tags is obtained. Next, a DOM node including information about business are identified by heuristic rules that check whether specific keywords exist in a text or an URL of a link. "Description" and "Keywords" are marked up by HTML tags and can be easily extracted. "Description of company's business" is supposed to be in a table. We first identify a DOM node corresponding to a heading of a table, then identify a DOM node that includes "Description of company's business" in the neighborhood of the heading. To extract "Description of company's activity", we first detect a link to a separate web page that summarizes activity of a company from a top page of a website. In the detected web page, not only desired texts but also unnecessary texts such as advertisement and navigation. Therefore, only main contents of the web page are extracted as "Description of company's activity". The algorithm proposed by Kato et al. is used to identify main contents.

In the second step, supervised machine learning is applied for classification of business category. Features used for machine learning are content words in a text of information about business, which are automatically extracted in the first step. In addition, content words in a top page of a company's website is also used as features. Features (words) appearing in information about business are more weighted in the feature vector. A classifier of business categories is trained using Naive Bayes (NB) and Random Forest (RF). The number of business categories is 28. They are defined by modifying the categories in the hierarchy of Japanese in the Open Directory Project (ODP), which is one of the web directory services.

Several experiments are carried out to evaluate our proposed method. First, our methods to extract information about business are evaluated by measuring precision, recall and F-measure. One hundred company web side obtained from ODP is used as a test data. The gold information about business is manually annotated in the test data. It is compared to the information extracted by our proposed methods. As for extraction of "Keywords" and "Description", precision, recall, and F-measure are 1. As for extraction of "Description of company's business", precision, recall, and F-measure are 1, 0.95, and 0.97, respectively. As for extraction of "Description of company's activity", precision, recall, and F-measure are 1, 0.91, and 0.95, respectively. These results indicate that all kinds of information about business can be accurately extracted. Next, proportion of the number of websites that contain each type of information about business is measured to investigate how frequently companies publish their business in their website. They are 0.8 for "Keywords", 0.85 for "Description", 0.7 for "Description of company's business", and 0.36 for "Description of company's activity". It is found that not many companies release their activity, but other types of information about business are released by most of companies.

Next, our methods of classification of business category are evaluated. A set of 29,364 websites in ODP is used as a dataset. It is divided into the training

data (90%) and the test data (10%). The classifiers are trained from the training data, and accuracy of classification on the test data is measured. A baseline system is a classifier with features extracted from a top page only. In addition, in order to reveal a ceiling of this task, a human judges categories of business for 300 websites randomly chosen from the test data. The accuracy of the proposed method using NB and RF are 0.270 and 0.508, respectively. On the other hand, the accuracy of the baseline with NB and RF are 0.252 and 0.493. The accuracy of human judgment is 0.717. The proposed methods slightly outperform the baselines. Comparing machine learning algorithms, the accuracy of RF is much better than that of NB. These results show the effectiveness of our proposed methods. However, it is far from human judgment; there is much room for improvement. When the human classifies the website, he tries to find specific keywords that strongly indicate a category of business. If a method to identify such specific keywords is implemented, performance of classification of business category would be much improved.

In the future, to improve the accuracy of classification of business category, several parameters of Naive Bayes and Random Forest should be optimized on a development data. Another important future work is to develop a system that can show the identified category of business in results of a search engine.