

SK-TOMS: Opinion Mining System Using Reusable Knowledge Representations

Kyungkoo Min, June Sup Lee

Personalization Technologies Development Team, SK Telecom
11, Euljiro 2-ga, Jung-gu, Seoul, Korea
mingk24@sktelecom.com, junelee@sktelecom.com

Abstract. SK-TOMS¹ is an opinion mining system being developed by SKTelecom, which is specialized for extracting sentiments from plain texts. It is crucial to keep detailed knowledge base of a specific domain to extract sentiments from plain texts of the domain. We formed an independent dictionary called ‘sentiment sheets’ to represent domain knowledge, which can be reused in multiple domains. In this paper, we explain system structures, GUIs, and the implementation of ‘sentiment sheets’.

Keywords: opinion mining system, sentiment analysis, sentiment sheets, SK-TOMS

1 Introduction

With the influence of Internet, the production and consumption of information do not go through the traditional way any longer. Information spreads in real time and an individual can express one’s opinions about any specific issue easily on the web. User generated contents increase exponentially, which affects the behaviors of other people. For example, consumers refer to other users’ reviews when they buy something on Internet shopping malls like Amazon.com. Therefore, it is meaningful to analyze a lot of opinions from unspecified individuals on Internet. An opinion mining system is a kind of text mining system, which is specialized for extracting sentiments from large amount of plain texts from unspecified individuals. It can be used as a decision making tool for consumers while it can be used as a marketing tool for enterprises.

In this paper, we introduce an opinion mining system called SK-TOMS which is in the middle of development process at SKTelecom. SK-TOMS divides a subject into items and topics according to whether the analysis target can be materialized or not. Items and topics have different characteristics, for example, items tend to have more common sets of keywords than topics. We apply different sets of ‘sentiment sheets’ to items and topics. SK-TOMS then extracts sentiment expressions from web texts with the help of ‘sentiment sheets’ and analyzes if those sentiment expressions are positive or negative opinions with respect to a given subject. The analyzed data can be

¹ SK Telecom Opinion Mining System

provided to other applications by Open APIs and it can also be viewed to users by a GUI.

2 System Structure

SK-TOMS collects documents from the Internet and databases like Oracle, MySQL, MS-SQL and assigns five levels of scores (Very positive, Positive, Normal, Negative, Very negative, ranging from 2 to -2 respectively) to each document according to the degree of preferences about a given subject (item or topic) represented in the document. Each subject may have sub-features and these sub-features can be scored in the same way. These scoring processes can also be executed multiple times during a various length of time to keep track of the history of opinions.

SK-TOMS is composed of three layers: collection layer, analysis layer, and presentation layer. In this paper, we explain each layer in more details.

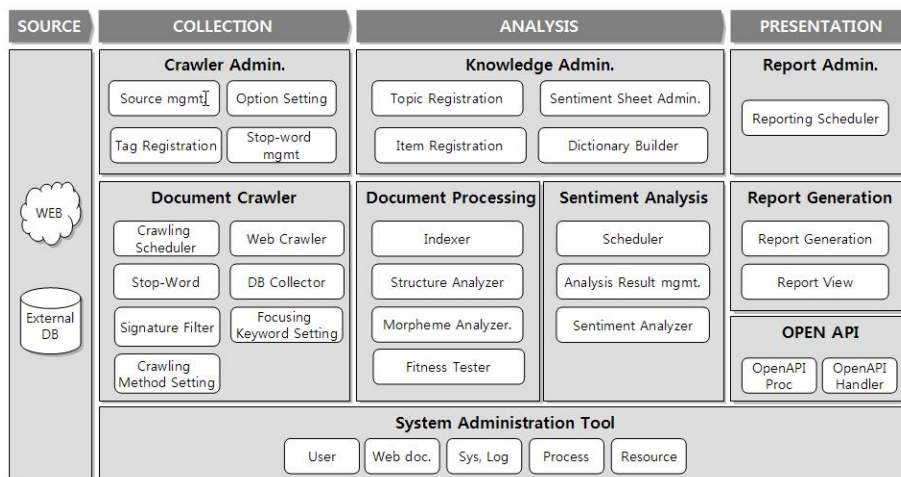


Fig. 1. Structure of SK-TOMS

2.1 Collection Layer

The collection layer is composed of a document crawler and a crawling administration tool. The core part of the collection layer is the document crawler which is made up of a web crawler and a DB collector. The web crawler analyzes the structure of collected texts and saves title, body, reply and the date of publication separately. We use focused crawling method to set bounds to gathering unrelated documents. The crawler uses meta-search method. It use search engine to find documents related to given subject as well as collects documents from seed URL.

After the collection, it filters out spams and eliminates duplicate texts by signature filtering and removes html tags.

2.2 Analysis Layer

The analysis layer is in charge of analyzing texts from collection layer into sentiment expression and the result is stored as structured form. The core function block of this layer is fitness filter and sentiment analyzer. The fitness filter filtrate unrelated text with a given subject. The sentiment analyzer extracts pros/cons opinion from sentiment expression about given subject. We divide subject into ‘topic’ and ‘item’ according to whether analysis target is materiality or not.

Sentiment analyzer extracts opinion from sentiment expression based on using pattern matching method and sentiment dictionary which is formed eight categories (commerce, travel, politics, economy, society, culture, sports, entertainment)

Sentiment Sheet. We use concept of sentiment sheet. Sentiment sheet is an idea for convenience and effectiveness of the system. When we extract opinion from some subject, we have to apply proper knowledge dictionary to system. Because, an expression with many meanings can be translated differently from the existing state of things. SK-TOMS manage knowledge dictionary as a set of sentiment sheets. The sentiment sheets are divided into multiple ‘sentiment prototype’ which is the smallest set of sentiment expression. For example, if we want to know consumer’s review about ‘iPod Touch’, we make ‘mp3’ sentiment sheet and then import sentiment prototype, like as design, color, price, sound quality and so on, into sentiment sheet. We can use sentiment sheet in case of analyzing other mp3 player. Also sentiment prototype is reused as a unit of other sentiment sheet.



Fig. 2 User interface of SK-TOMS. It shows the example of user movie review analysis. Circle graph shows the Pros/Cons of total user review (top left) and bar graph means sub-feature of movie review like as director, actor, sound and etc (bottom left). Left side line graphs are the number of analyzed documents (top right) and the result of time series analysis (bottom right).

2.3 Presentation Layer

The presentation layer takes a role of reporting by GUI interface. Users take a view of analyzed results as a various forms like as five-level Pros/Cons analysis, Pros/Cons analysis as sub-feature, volume of analysis, and etc, and monitor the results for various length of time. Besides, system administrator can observe crawling/ analysis results and processor status. It includes also Open API for providing results to external web sites. We issue authentication key to user who want to use the result of SK-TOMS. If the user submits parameter about analysis report with the key, the results are sent off as a XML form. Figure 2 show the screenshot of SK-TOMS.

3 Future Works

SK-TOMS has been developed for the purpose of autonomous customer review analysis and brand monitoring for marketing and making sure the monitoring system which deals with unfavorable business environment. As a result of this research, we built demonstration framework and set an example of 37 topics and 25 items and contents with sentiment expression more than 58,000.

Table 1. The performance of sentiment extraction on the sentence

Subject Type	Recognition Rightly	Absence of Sentiment Dictionary	Language Analysis Error
Item	82.2%	15.7%	2.1%
Contents	72.9%	26.0%	1.1%
Topic	65.6%	32.1%	2.4%
Total	76.3%	21.6%	2.1%

The accuracy of sentiment extraction on the sentence is 76.3% on the average and most of the recognition errors are due to lack of dictionary (table 1). If we endeavor to make dictionaries, then the performance will be improved. But, we have another problem. Although we can raise accuracy by dictionary expansion, there is little guarantee that the sentiment expression is related to the given subject. Of course we use 'fitness tester' to filter out unsuitable documents, but it is not sufficient method because, document can contain sentence that is no relation with the subject. We are in the process of developing module to overcome this problem by using mutual information and we are going to develop that SK-TOMS can handle multilingual documents besides Korean.

References

1. Wang, G., Arakai, K.: OMS-J: An Opinion Mining System for Japanese Weblog Reviews Using a Combination of Supervised and Unsupervised Approaches. NAACL HLT Demonstration Program, 19--20 (2007)
2. Chaovalit, P., Zhou, L.: Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, 38th Hawaii International Conference on System Sciences, 43--50 (2005)
3. Jindal, N., Liu, B.: Review Spam Detection, In Proceedings of WWW-2007, 1189--1190, (2007)
4. Popescu, A., Etzioni, o.: Extracting Product Features and Opinions from Reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 339 -- 346 (2005)