Extraction of Topical Consumer Products from Weblogs

Shinichi Nagano and **Masumi Inaba** and **Yumiko Mizoguchi** and **Takahiro Kawamura**
Corporate R&D Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan
shinichi3.nagano@toshiba.co.jp

Abstract

This paper proposes a new algorithm of associated topic extraction, which detects related topics in a collection of blog entries commenting on a specified topic. The main feature of the algorithm is to evaluate how important a topic is to the collection, according to the popularity of blog entries through Trackbacks and comments. Another feature is to utilize product ontology for excluding unrelated topics. Evaluation results show that the proposed algorithm can capture users' impressions of associated topics more accurately than TF-IDF.

Introduction

Consumer Generated Media (CGM) has been one of the major "word-of-mouth" media, and has a significant impact not just on customers' purchase decision processes but also on companies' product marketing strategies. Analysis of CGM posed a challenge in product marketing(Facca & Lanzi 2005; Mishne & de Rijke 2006). Sentiment analysis is a typical CGM analysis, which extracts evaluation expressions from CGM contents commenting on a specified product to summarize its reputation. Its technology is still in the early phase of application to practical product marketing.

We have developed a reputation analysis system(Kawamura *et al.* 2007), which retrieves blog entries commenting on a specified product, and then extracts reputation expressions and related products from the blog entries. The system indicates the overall rating of the product reputation (positive vs. negative), and suggests related products that are much discussed in the blog entries.

This paper proposes a new algorithm of associated topic extraction for suggestion of related products. The main feature is the evaluation of how important a topic is to a collection of blog entries, according to the popularity of the blog entries, through Trackbacks and comments. Another feature is the utilization of product ontology for topic filtering to extract products relevant to or similar to the given product. The paper also briefly evaluates the proposed algorithm in comparison with TF-IDF, a widely-used topic detection algorithm.

Associated Topic Extraction

The purpose of associated topic extraction is to suggest similar or comparable products that are much discussed in retrieved blog entries commenting on a specified product. The straightforward approach to the extraction is to count frequency of occurrence of proper nouns in blog entries through morphological analysis. Morphological analysis segments and tokenizes text strings into a series of meaningful terms, and then adds the terms to complemental information such as part of speech, pronunciation, semantic information. In the straightforward approach, any blog entry and any noun string appearing in blog entries are uniformly enumerated. Although it may appear to be a plausible approach such a uniform approach does not accord with blog readers' impression.

This paper proposes a new algorithm for associated topic extraction, abbreviated to AT. The main feature is that AT finds a blog thread, which is a set of blog entries spanned by Trackbacks, and then biases the frequency occurrence of proper nouns according to the thread size. This is derived from the following belief. Suppose that blog entries conversing about a target product receive many Trackback pings. Then, the entries may be especially interesting to many bloggers, and products mentioned in the entries may be similar to the target product or competitive with it. The degree of association $AT(t)$ for each extracted topic t is formulated as follows. Let $TF(t)$ and $IF(t)$ be the topic occurrence frequency t and the inverse document frequency for topic t. Given a set of blog entries conversing about a certain product, then $AT(t)$ is defined as follows:

$$AT(t) = (TF_R(t) + TF_{IR}(t)) \times IF(t) \qquad (1)$$

Note that $TF_R(t)$ and $TF_{IR}(t)$ weight a topicality of topic t by correlation between blog entries.

Another feature of AT is to utilize product ontology. Product ontology is composed of an is-a relation representing hierarchical categorization of products and an instance-of relation representing mappings of individual products with product categories. Referring to product ontology, AT finds only similar or competitive products mentioned in the blog threads conversing about a target product to exclude out proper nouns other than product names. We have developed product ontology including more than 400 thousand consumer products including cars, DVDs, and home appli-

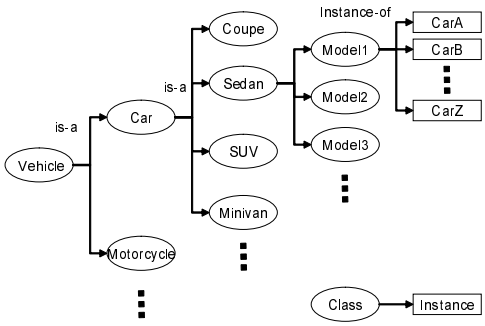

Figure 1: Product ontology

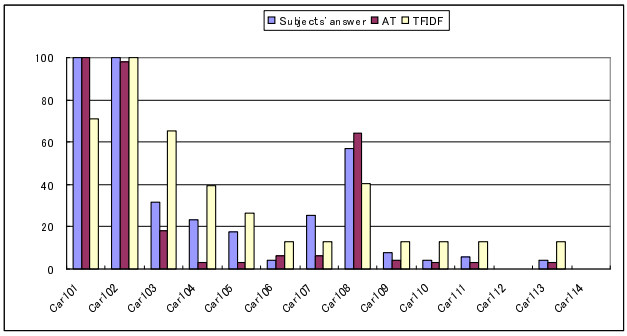

Figure 2: Impression evaluation for blog thread 1

ances. A fragment of the product ontology is illustrated in Fig.1.

Evaluation

Experiment is done as a part of feasibility studies for practical product marketing. We first prepare three blog thread as collections, which comment on the same car in the sedan category and include totally 171 blog entries. We then apply AT to each collection and then evaluate the results of AT with the two performance measures, recall and precision, which are widely used in IR and NLP. In addition, we ask 18 trial subjects to extract topical cars from the collections and to make their ratings in 4 levels as subjective impression. We then compare the averages of the subjects' answers with the AT and TF-IDF scores.

The results of precision and recall evaluations are shown in Table 1, where all precision scores are close to 100%. One of the reasons why these extremely high scores are achieved is that product ontology filtering functions effectively enough to exclude any noun other than product names. Another reason is that very few blog entries in the collections comment on subjects other than car and thus the collections are easy to analyze. However, several terms are incorrectly extracted as product names, and thus reduce precision scores. On the other hand, all recall scores are completely over 80% as shown in Table 1, even though absence of some car models from product ontology reduce recall scores. The results of precision and recall measures depend on the maintenance status both of a morphological analysis dictionary

Table 1: Evaluation of precision and recall measures

Thread no.	(a)	(b)	Recall	Precision
1	21	14	94.3%	100.0%
2	101	19	86.0%	97.6%
3	49	19	84.3%	100.0%

(a) Number of blog entries, (b) number of car models.

Table 2: Differences with the correct answer

Thread no.	D1	D2	D1-D2
1	6.3	11.2	5.0
2	8.3	11.5	3.2
3	6.5	7.9	1.4

D1: Difference between AT and subjects' answers
D2: Difference between TF-IDF and subjects' answers

and a product ontology. Thus, it is expected that experiments in categories other than cars will produce different results of performance evaluation.

The results of impression evaluation are shown in Figure 2. Scores of each method are normalized between 0 and 100. The results show that AT is more well-controlled and closer to subjects' impression than TF-IDF. For example, AT gives considerably higher scores than TF-IDF to the cars that subjects feel highly topical, as Car101 and Car108 shown in Figure 2. The cars appear frequently in the blog entries which receive a lot of Trackbacks and comments, and thus AT raises a topicality of each car. Next, Table 2 shows the difference between AT scores and subjects' answers (D1) and the difference between TF-IDF scores and subjects' answers (D2). Note that both D1 and D2 are averaged per car. Table 2 shows that AT is extremely closer to subjects' impression than TF-IDF in all blog thread. As a result, we can say that the proposed weighting formula accurately express the strength of blog readers' impression.

Conclusions

Refinement of the proposed algorithm is a future work to gain higher qualities of extraction results. We are also planning to develop novel technologies for ontology building and maintenance.

References

Facca, F., and Lanzi, P. 2005. Mining interesting knowledge from weblogs: a survey. *Data and Knowledge Engineering* 53(3):225–241.

Kawamura, T.; Nagano, S.; Inaba, M.; and Mizoguchi, Y. 2007. Mobile service for reputation extraction from weblogs - public experiment and evaluation. In *Proceedings of Twenty-Second Conference on Artificial Intelligence (AAAI-07)*.

Mishne, G., and de Rijke, M. 2006. A study of blog search. In *Proceedings of 28th European Conference on Information Retrieval (ECIR)*.