Preliminary Experiment of Ontology Maintenance for Ontology Metrics Formalization

Takayuki Iida, Masumi Inaba, Yumiko Mizoguchi,
Shinichi Nagano, and Masanori Hattori

Corporate R&D Center, Toshiba Corporation, Japan
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582, Japan
takayuki1.iida@toshiba.co.jp

Abstract. This paper addresses issues concerning metrics that measure quality of the ontology or cost of work, and then presents requirements for ontology metrics by referring to an experiment on ontology maintenance. The metric makes sense if it enables estimation of manpower costs and schedule for ontology maintenance, but there is no metric appropriate for this purpose. In this paper, we perform a preliminary experiment in which evaluation expression ontology is maintained for a reputation analysis system from weblogs. We measure six naive metrics for the ontology, namely, the number of instances, error rate of instance, precision and recall, and slope and variance of the precision/recall, and then discuss metrics from an engineering viewpoint.

Keywords: Ontology, Ontology Maintenance, Metrics, Reputation analysis

1 Introduction

Consumer Generated Media (CGM) have become one of the major "word-of-mouth" media both for consumers and companies. A blog is an important example of CGM. Some consumers write reviews of products or describe their impressions or experiences of them, and others compare a specific product with similar products or with products competitive with it. Blogs often furnish their readers with information that influences their purchase decisions. Analysis of CGM has become an important issue in marketing[1][2]. Therefore, we have developed a reputation analysis system [3].

A feature of our reputation analysis system is accurate analysis to understand the meaning of context by using evaluation expression ontology and product ontology. To preserve the quality of the analysis result, it is necessary to maintain the ontologies periodically. For the cost issue concerning ontology, we have developed a maintenance tool that helps engineers manually build and modify the ontologies.

However, the lack of engineering metrics to measure the quality of ontology and the maintenance cost is an issue. To ensure the continual operation of our ontology-based system, it is necessary to estimate the quality and the cost, and then to make a detailed maintenance schedule. Therefore, we discuss requirements for ontology engineering in order to formalize some clear metrics.

In the paper, we report on the performance of maintenance of three genres of evaluation expression ontology. It took 50 hours per genre; a total of 150 hours. We

measured six naive metrics, such as the number of instances, precision and recall, slope and variance of the precision/recall. Then, based on the experimental result, we discuss the requirements for metrics.

The remainder of this paper is organized as follows. Section 2 presents an overview of our reputation analysis system. Section 3 addresses the issues concerning ontology maintenance. Section 4 presents an experiment on ontology maintenance, and then Section 5 discusses the result of the experiment. Section 6 refers to related work. Section 7 concludes the paper.

2 Reputation Analysis System

2.1 System Overview

We have developed a reputation analysis system, which retrieves blog entries commenting on a specified product, and then extracts reputation expressions and related products from the blog entries. The main feature of the system is that it analyzes the contents of retrieved blog entries using ontology. This makes it possible (1) to indicate the overall rating of the product reputation (positive vs. negative), (2) to extract associated products that are much discussed in the blog entries, and (3) to sort the blog entries by reputation relevance and blog popularity.

Since this paper refers to evaluation expression ontology, it describes only positive/negative determination. A part of our evaluation expression ontology is shown in Fig. 1. Attributes, such as functionality, design and speed, are defined upper-node. Expressions, such as best, good and bad, are defined lower-node. Evaluation expression ontology is constructed for every genre, considering the difference in meaning according to target genre.

The positive/negative determination is one of the text summarization technologies. It performs the morphologic analysis and the syntactic analysis of blog contents retrieved from the Internet, and then evaluates a positive or negative rating of the blog contents.

2.2 Maintenance Tool

Since the system quality using ontology is heavily dependent on the quality of ontology, constant maintenance is needed. Examples of ontology maintenance are dealing with words that differ in meaning with the passage of time or depending on age group, adding slang and colloquial expressions, and adding attributes of new products.

A screen shot of the maintenance tool is shown in Fig.2. On the left side, an engineer can browse a result of positive/negative determination, while on the right side an engineer can edit an ontology. The main features are as follows:

・Application analysis checking support

This feature enables confirmation of the result of analysis and editing of an ontology interactively. It can perform all ontology maintenance from an edit operation to an analysis operation. It can show the result before maintenance and after maintenance simultaneously so that change of an analysis result can be grasped easily.

・Instance adding support

This feature enables addition of multiple instances simultaneously that are composed of different types of writing, such as hiragana, katakana, or kanji characters.

・Consistency checking

An inconsistent ontology means one part of the ontology does not agree with another. In order to keep consistency, this feature detects multiple entries or problematic dependent relationships.

With the developed tool, the engineer checks whether positive/negative determination and extraction is right or wrong, edits an inappropriate instance or adds a new instance or class, and then confirms the result of analysis by edited ontology.

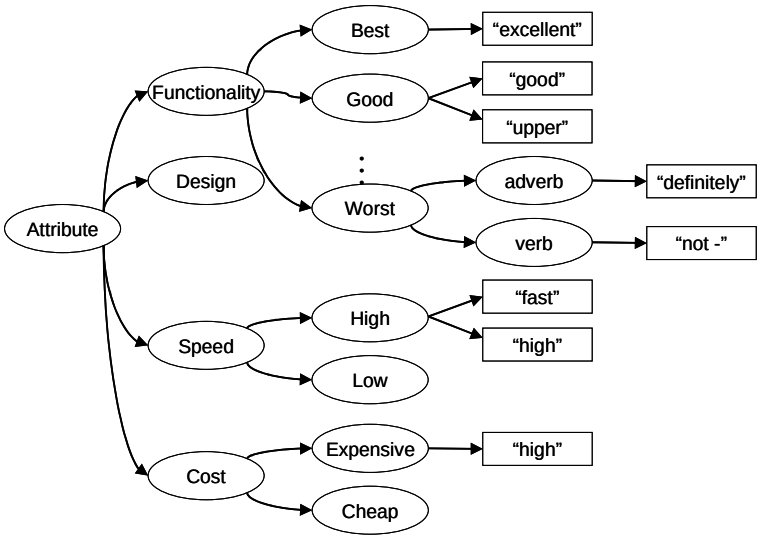

Fig. 1. Part of evaluation expression ontology

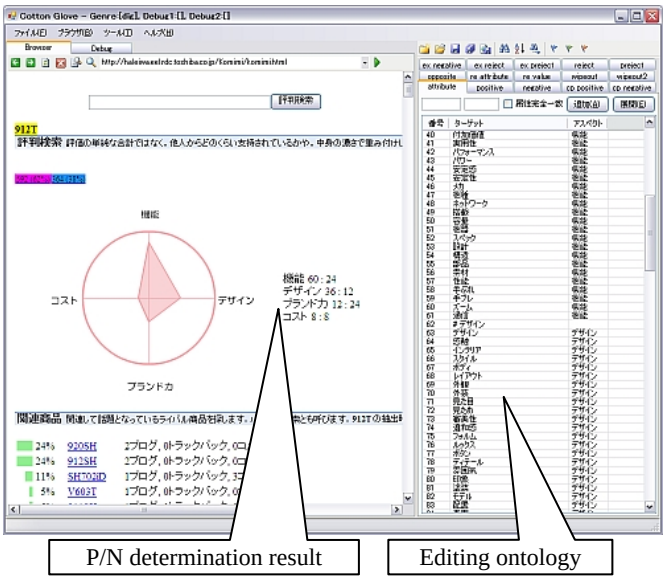

Fig. 2. Ontology maintenance tool

3 Maintenance Issue

With regard to ontology maintenance at present, a pressing issue is that there are no clear metrics that measure quality of ontology and maintenance cost.

Metrics are needed because lack of engineering measures for quality and cost causes the following problems. There are two phases of maintenance: the initial construction phase and the operation phase. In the initial construction phase, we intend to construct new genre ontology, whereas it is unclear when it should be decided that ontology maintenance is finished. Additionally, because the relation of maintenance cost and quality is unclear, the detailed schedule cannot be estimated.

On the other hand, the problem in the operation phase is the decrease of ontology quality with time elapsed, as described in Section 2. Because the meaning of new words differs depending on the age group, for example, between young people and elder people, system quality is affected. Although, it is desirable to detect the decrease automatically, the lack of metrics makes it impossible to detect it automatically at present. So we deal with it by scheduled maintenance. Therefore, we perform maintenance when it is not required, or perform maintenance after lowering the targeted accuracy. The ideal is to perform appropriate maintenance when it is needed.

Next, we discuss why formulating a metric is difficult. It is easy to calculate the number of classes or instances as the ontology size. But, it is not directly related to the accuracy, because there are redundant entries that are not entirely used by the system and inappropriate entries.

The metrics to evaluate system quality at present are precision and recall. Precision and recall are degree of accuracy when ontology is adapted to the external system,

and we estimate quality of ontology by that. But precision and recall need to be given the correct data manually each time, and therefore calculation cost is high.

An inappropriate metric is one that cannot measure value automatically in the operation phase. As described above, we need to formalize a metric that measures quality and cost. Therefore, in the next section, we discuss the requirements and issues concerning a metric by referring to an evaluation experiment.

4 Maintenance Experiment

4.1 Experiment Overview

For evaluating the ontology maintenance, an engineer performs ontology maintenance using our developed maintenance tool, and measures six naive metrics such as the number of instances per maintenance time and precision/recall. Although it is desirable that multiple engineers measure the metrics for ontology maintenance to ensure a more detailed evaluation, this is performed by only one engineer for the preliminary experiment.

The outline of the experiment is as follows. An engineer performs maintenance of three genres of evaluation expression ontology using the tool. It takes 50 hours per genre; a total of 150 hours. Target genres are digital appliance (DIG), movie DVD (DVD), and facilities and restaurant (POI). We assume that initial construction has been finished for the ontology for each genre. The quality of each genre ontology is different. The order of better quality genre is POI, DVD, DIG. This is utilized for evaluating the relation of between the initial quality and maintenance work. The maintenance data for each genre are shown in Table 1.

Table 1. Maintenance data

Genre	Time (hour)	# of entries	# of sentences
DIG	50	687	6,111
DVD	50	556	7,643
POI	50	651	7,236

4.2 Naive Metrics

We use the following six naive metrics in the experiment.

- The number of increased instances.
- Entry error rate of instance
- Precision
- Recall
- Slope of Precision/Recall
- Variance of Precision/Recall

The metrics are described below.

First, the number of increased instance is a measure of the relation between ontology size and accuracy. It involves measuring the number of instances before maintenance and after maintenance, and the number of instances added per maintenance time.

Second, entry error rate of instance is ratio of inappropriate instances judged by an engineer with seasoned knowledge of ontology, in all instances added by maintenance. It evaluates the instance that cannot be used by the system, because the system does not use all instances.

Third, precision shows the ratio of sentences that actually include positive/negative expression to sentences that include positive/negative expression judged by system analysis. A low precision value indicates the existence of extraction errors.

Then, recall shows the ratio of sentences that include positive/negative expression judged by system analysis to sentences that include positive/negative expression judged manually. A low recall value means some sentences that include evaluation expression are not extracted.

Finally, slope of precision/recall measures a percentage of increase and variance of precision/recall measures variability of index values of multiple products. In this experiment, we measured data for 10 products.

4.3 Experiment Results

We collect blog entries of 10 target products per genre and analyze the positive/negative determination to evaluate the ontology maintenance. Transition of precision and recall is calculated by comparing the human check and system result. The number of increased instances is measured every two hours, and precision/recall is measured every ten hours. The data of the evaluation target is shown in Table 2.

Table 2. Evaluation data

Genre	# of products	# of entries (avg)	# of sentences (avg)
DIG	10	114.3	1191.6
DVD	10	44.1	632.0
POI	10	77.1	880.3

We evaluate each of the six naive metrics shown in Section 4.2. First, the number of instances for each genre is shown in Table 3, and transition of the number of instances is shown in Fig.3. The result shows that the number of instances is proportional to the maintenance time. Finally, about 400-800 instances are added per genre by 50-hours maintenance. The ratio of newly added instances to the total is about 6.5%(DIG), 6.6%(DVD), 3.0(POI). The values in Table 3 indicate that POI genre can not add many expressions, because POI genre has many instances compared to other genres and has already added specific expressions.

Table 3. Instance data

Genre	Before maintenance	After Maintenance	# of increased instance	Increased instance percentage of total
DIG	9,907	10,595	+688	6.5%
DVD	11,179	11,970	+791	6.6%
POI	13,071	13,482	+411	3.0%

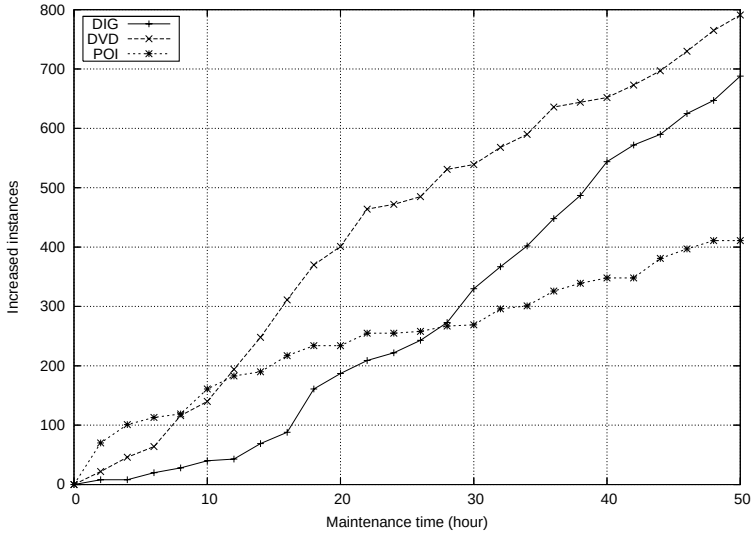

Fig. 3. Number of Increased instances

Second, Table 4 shows that the result in the case that an engineer with seasoned knowledge of ontology judged whether newly added instances were appropriate or not. The result indicates about 10 percent of instances are inappropriate for each genre. For example, the word is not suitable for evaluation expression or added in the wrong class. About 60 percent of inappropriate instances can be detected by the consistency checking feature of the tool. Since 95 percent of instances remaining after editing or deleting are suitable, this result is reasonable for operation.

Table 4. Error rate of instance

		Inappropriate		Detected	
Genre	# of added instances	# of instances	Percentage of total	# of instances	Percentage of total
DIG	688	83	12.1%	50	60.2%
DVD	791	115	14.5%	79	68.7%
POI	411	49	11.9%	28	57.1%

Third, the transition of precision/recall is shown in Fig.4 and Fig.5. Slope and variance of precision/recall is shown in Table 5. Variance is an average value

calculated for every 10 hours. Precision/recall value of DIG and DVD is proportional to maintenance time, and the appreciation rate of DIG, for which added instance is lower than for other genres, is high. Although the rate of added instance is 6.5%, it contributes to improving the accuracy because added instances are genre-specific expressions. Examples of instances for DIG are "high-image-quality" and "high-musical-quality".

Conversely in genre POI, precision and recall do not rise. The number of added instances is constant and the result does not change. The reason that the values do not rise is newly added instances are not used frequently by the system. Judging by the result, system quality cannot be measured by only the number of instances.

The values of variance vary widely for DIG genre that does not contain many words. In contrast, the value of variance for POI genre is about the same, because many instances that are used by the system frequently are added. It is thought that if the value of variance is low, ontology is well maintained.

As seen in the case of POI genre, when the instances are added to a certain level, it does not seem worthwhile to perform maintenance due to poor work performance. When considering the estimation of maintenance time, the metric that measures a limit of the accuracy is very important. But, from the result of this experiment, the situation cannot be estimated.

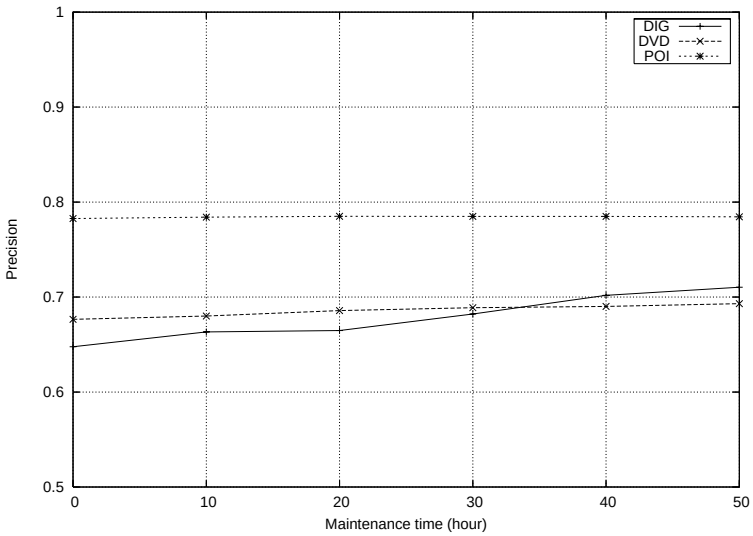

Fig. 4. Transition of precision (average)

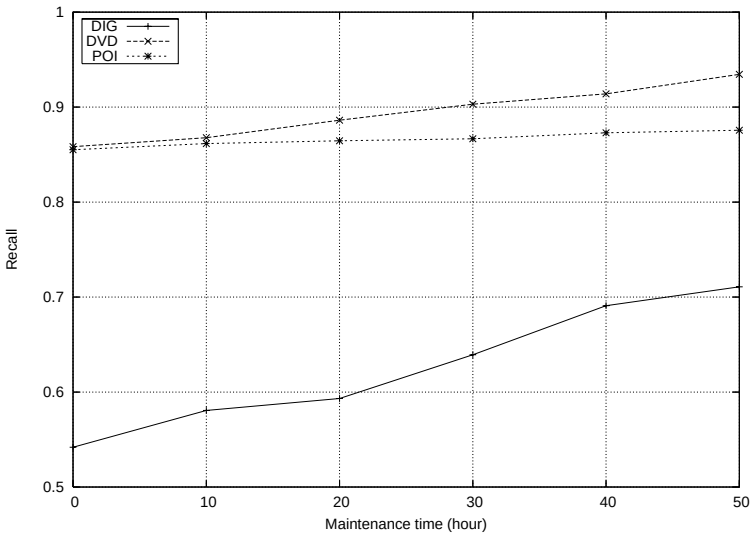

Fig. 5. Transition of recall (average)

Table 5. Slope and variance of precision/recall

	Slope		Variance	
Genre	Precision	Recall	Precision	Recall
DIG	0.127	0.349	0.996	0.523
DVD	0.033	0.153	1.129	0.282
POI	0.003	0.040	0.185	0.185

5 Discussion

In this section, we point out the inadequacy and problematic points of six naive metrics based on the experiment results, and then describe the requirements for the metrics.

・The number of instances

As you can see from the result of this experiment, condition of ontology cannot be estimated by only the number of instances. This is because added words are uncommon and not used frequently by the system. However, it is useful to estimate the size of ontology due to the low calculation cost.

In addition, we do not measure the balance of positive and negative expressions, because is the numbers are almost the same in this experiment. But we do measure the balance of each class. For example in the reputation analysis service, if the expression of positive is extremely large, the result has more positive terms than negative ones even if it is a correct analysis. In order to perform an impartial analysis, it is

preferable that the number of positive instances and negative instances are almost the same.

In the case of product ontology, the situation can be estimated if only a particular category is well maintained but other categories are not maintained.

・Precision/Recall

Precision and recall are useful metrics for measuring quality, because they are applied to the external system. However, it is a problem to give the correct data manually each time, because it requires considerable time. A time-consuming metric is impractical for automatic detection of deterioration.

Therefore, it is necessary to formulate a new metric similar to precision and recall, and that does not require correct data or requires minimal correct data; for example, a method of estimating the condition of ontology to compare the added words with a part-of-speech extracted from a set of documents automatically. If we can formalize a metric that does not require correct data, deterioration can be detected when the accuracy decreases and the most suitable maintenance can be performed.

・Slope and variance of precision/recall

From the experimental result, it is observed that slope and variance of precision/recall varied according to condition of ontology. By using this metric, we can estimate that a condition reaches almost its limit if slope value is low or variance value is low, and finish the ontology maintenance. But, we have to consider a metric that does not need to label correct data.

・Maintenance target data

From the experimental result, we can estimate the approximate number of registered instances from the size of maintenance target data. However, condition of ontology cannot be estimated from only the number of instances. It is necessary to clarify the relation of maintenance time and increase of accuracy combined with other metrics.

・Operator variation

In this experiment, ontology maintenance and evaluation were performed by another operator. It is necessary to deal with the bias of manual procedure for each operator. In particular, the concept of evaluation expression has no criterion contrary to a product name, and therefore operator variation exists. We will also need to consider a metric that measures the bias of manual procedure when it is performed by one operator or multiple operators.

The results described above show that precision/recall is a useful metric for measuring quality of ontology. However, precision/recall has problems in terms of the cost of adding correct data, biased evaluation data, and operator variation. It is necessary to formalize metrics that are low in cost and impartial.

6 Related Work

This section refers to related work of ontology evaluation, metrics, construction, and maintenance.

A growing number of ontology metrics and measures have been suggested and defined[4][5]. But many of them are based on structural notions without taking into account application of external system. Therefore they are not appropriate for scheduling of detailed maintenance and automatic detection of decrease.

Tools to support construction and maintenance include "HOZO"[6], "Protégé"[7], and "DODDLE-OWL"[8]. "HOZO" handles the role concept explicitly and has a distributed development environment. "Protégé" is the most widely used ontology editor and can be customized to provide domain-friendly support for creating knowledge models and entering data. "DODDLE-OWL" makes reuse of existing ontologies and supports the semi-automatic construction of taxonomic and other relationships in domain ontologies from documents.

The ontology wiki has been proposed for general users in order to ameliorate the difficulties of ontology construction and enable use of collaborative knowledge.

7 Conclusion

This paper presents issues concerning metrics that measure quality of ontology or cost of work, and then discusses important requirements by referring to an evaluation experiment. From the results of this experiment, we confirmed that the condition of ontology cannot be inferred by only the number of instances and we found there is a tendency for the slope and variance values to differ according to the maintenance condition. In future work, we are planning to define metrics that are low in cost and impartial based on the experiment results.

References

1. F. Facca and P. Lanzi: “Mining interesting knowledge from weblogs: a survey”, Data and Knowledge Engineering, 53, 3, pp. 225–241 (2005).
2. G. Mishne and M. de Rijke: “A study of blog search”, Proceedings of 28th European Conference, 2006
3. T. Kawamura, S. Nagano, M. Inaba, Y. Mizoguchi: “Mobile Service for Reputation Extraction from Weblogs - Public Experiment and Evaluation”, Proceedings of Twenty-Second Conference on Artificial Intelligence (AAAI-07), 2007.
4. D. Vrandecic and Y. Sure: “How to design better ontology metrics”, Proceedings of 4th European Semantic Web Conference (ESWC), 2007.
5. J. Brank, M. Grobelnik, and D. Mladenic: “A survey of ontology evaluation techniques”, Proceedings of the conference on Data Mining and Data Warehouses (SiKDD), 2005.
6. HOZO, http://www.hozo.jp/
7. Protégé, http://protege.stanford.edu/
8. DODDLE-OWL, http://doddle-owl.sourceforge.net/en/