

A Reputation Search Engine That Collects People's Opinions Using Information Extraction Technology

KENJI TATEISHI,[†] YOSHIHIDE ISHIGURO[†]
and TOSHIKAZU FUKUSHIMA[†]

In this paper, we propose a new search engine, “Reputation Search Engine (RSE),” that collects many kinds of opinions by using information extraction techniques. Although the Internet is one of the best mechanisms for acquiring and storing people’s opinions, there is no appropriate tool to gather them. The RSE can collect opinions quickly and at a lower cost than conventional methods. To develop the RSE we first define the vague word “opinion” as a fact composed of three entities: “object name,” “attribute expression” and “evaluative expression.” This definition allows us to use an information extraction approach. We describe the RSE architecture, which divides the task of opinion extraction into two steps by exploiting a model that applies the above definition. The prototype RSE system utilizes two of the three entities, “object name” and “evaluative expression,” to process opinions, and it implements the first step of the opinion extraction task. The system is based on a dictionary and pattern-matching rules. The dictionary contains evaluative expressions, each of which has a label of positive or negative. This label information enables the system to classify extracted opinions into positive or negative types. The experimental results show the advantage of our opinion extraction method over an SVM text classification method in two representative domains on the Internet: computers and alcoholic beverages.

1. Introduction

People’s opinions are important for corporate activities. Companies often collect opinions on their products by conducting surveys through questionnaires.

The Internet is one of the best mechanisms for acquiring and storing people’s opinions, since everyone has the opportunity to express his or her opinions on web pages. If there were a way to collect them more promptly and at a lower cost than conventional methods, we believe such an approach would become a substitute for the traditional questionnaire survey.

Currently, general-purpose search engines and Internet monitoring services can be used to find opinions. However, the former tools usually produce many irrelevant results. According to our research, when a certain product name is entered as a search keyword into a general-purpose search engine, the proportion of web pages that include opinions in the search results is, on average, only 16%¹⁾.

Internet monitoring services such as eWatch and gala are well known but have proved to be problematic. These services inform their customers of injurious information on their products written on the Internet. However,

these are costly and take a long time to produce results since these services are usually performed by human staff rather than by software.

In this paper, we propose a search engine, “Reputation Search Engine (RSE),” that collects many kinds of opinions by using information extraction technology. The goal of the RSE is to show users the reputation of a given object. This reputation can be expressed intuitively in the form of a radar chart, as shown in **Fig. 1**. Each of the axes represents an attribute of the evaluation: performance, advertising, support, cost, or design. The evaluation is based on the proportion of positive opinions among all opinions on an attribute category. Using this radar chart, users can determine which attributes of the product are good or bad, compared with other products.

To develop the RSE we first need to define the term “opinion,” since it is vague and has many different meanings. We propose an opinion model that defines “opinion” as a fact composed of three entities: “object name,” “attribute expression” and “evaluative expression.” This definition allows us to take an information extraction (IE) approach. In other words, we can consider the task of collecting

<http://www.ewatch.com>
<http://www.gala.jp>

[†] Internet Systems Research Laboratories, NEC Corp.

Table 1 Opinion model.

Entity	Explanation	Examples
Object Name	A product name, a brand name, a personal name, etc.	Mobile777, NEC, Albert Einstein, etc.
Attribute Expression	An aspect of the object, or a part of the object	performance, cost, design, support, advertising, etc.
Evaluative Expression	An evaluation of the object that is either positive or negative	good, bad, like, hate, handy, expensive, etc.

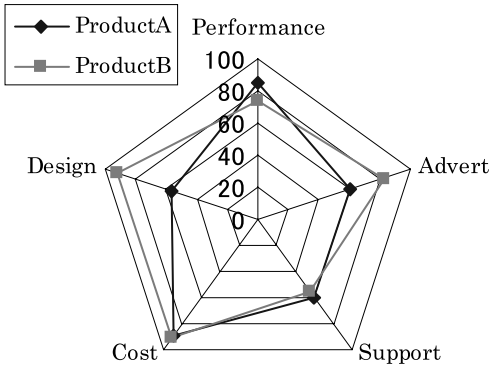


Fig. 1 Reputations of product A and product B.

opinions as that of extracting the tuples of these entities.

We have developed a prototype RSE based on two of the three entities, “object name” and “evaluative expression,”¹⁾ as our first step. The prototype system extracts opinions by using a dictionary and pattern-matching rules. This approach is often used in conventional IE systems, especially Named Entity extraction systems^{2),3)}. The dictionary contains evaluative expressions, and each of them has a label of positive or negative. Using this label, the system classifies the extracted opinions into positive or negative types.

In an experiment, we compared our opinion extraction method, based on the dictionary and the pattern-matching rules, with a text classification method using SVM. The result showed the advantage of our method in two representative domains on the Internet: computers and alcoholic beverages. Another experiment showed that the performance of our method largely depended on the size of the dictionary.

2. Reputation Search Engine

2.1 Requirements

As mentioned above, the goal of the RSE is to depict reputation in a radar chart, like that shown in Fig. 1. To achieve this goal, the following functions are required:

- (1) A function to extract opinions on a given

- object from many web sources;
- (2) A function to classify the extracted opinions as either positive or negative;
- (3) A function to classify the extracted opinions into one of the attribute categories.

After extracting opinions by (1), the RSE classifies them by (2) and (3). Therefore, it is not adequate for function (1) to only find web documents that include opinions: It is also necessary to locate the opinions in web documents and extract sentences that describe them.

2.2 An Opinion Model

To fulfill the above requirements, we first clarify our definition of the term “opinion.” The definition of the word in this paper is quite strict. We only deal with opinions that express an “evaluation,” either positive or negative. This definition makes it possible for us to use an IE approach in the development of the RSE.

We define opinion as a fact composed of the following three entities: an object name, an attribute expression, and an evaluative expression. An object name includes a product name, a personal name and a brand name. An attribute expression means an aspect of the object, or a part of the object. An evaluative expression contains an evaluation of the object that is either positive or negative. **Table 1** shows examples of these entities.

Not all opinions contain all three entities. Here, we deal with two types of opinions, (1) and (2):

- (1) Mobile777 design has become stale. — (Mobile777, design, stale)
- (2) Mobile777 is expensive. — (Mobile777, -, expensive)

Sentence (2) has no attribute expression while (1) has all three entities. We regard (2) as omitting an attribute expression, such as “price.”

2.3 Architecture

The RSE is made of three components. **Figure 2** shows the architecture of the RSE. The first component (a), corresponding to function (1), finds tuples of three entities as opinions

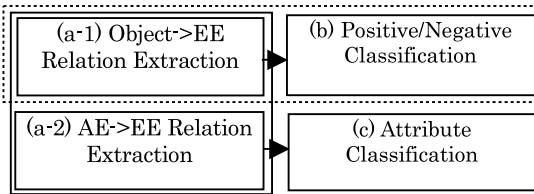


Fig. 2 Architecture of RSE (EE: Evaluative Expression, AE: Attribute Expression).

from given documents. Component (a) can be divided into two sub-components, (a-1) and (a-2). (a-1) extracts the relation between an object name and an evaluative expression. Similarly, (a-2) extracts the relation between an attribute expression and an evaluative expression. The second component (b), corresponding to function (2), classifies the extracted opinions into positive or negative by using the evaluative expression as a clue for the classification. The third component (c), corresponding to function (3), classifies the extracted opinions into one of the attribute categories by using the attribute expression as a clue for the classification.

2.4 Our Approach

We plan to develop the RSE in two steps, and we achieve the first step in this paper. The objective of the first step is to evaluate the effectiveness of our IE-based opinion extraction approach. Since (a-1) and (a-2), the main components of the RSE, involve the task of predicate-argument extraction, we expect them to share the same algorithm and just differ in the dictionary and pattern-matching rules, which are described in Sections 3.2 and 3.4. We first develop a prototype system dealing with (a-1) and (b) by using a simple two-entity model incorporating “object name” and “evaluative expression.” We will then expand the prototype system by using the complete model in future work.

3. A Prototype System

3.1 Overview

We have developed a prototype system that achieves (a-1) and (b), as represented by the dotted area of the architecture in Fig. 2. The method we have employed is to use a dictionary and pattern-matching rules. This method is well known in the field of information extraction, especially Named Entity extraction^{2),3)}. Since an object name is given to the RSE from a user in the same way as in other search engines, the dictionary that the prototype RSE needs is an evaluative expression dictionary.

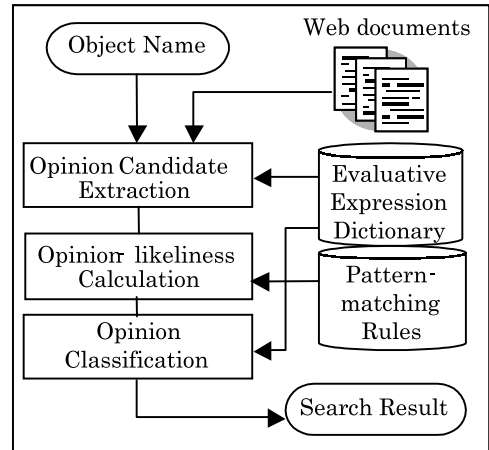


Fig. 3 Process of prototype RSE.

The pattern-matching rules mainly describe the relation of an object name and an evaluative expression and are used to obtain opinion-likelihood, which means how similar to an opinion an opinion candidate is. Each of the evaluative expressions in the dictionary has a label of positive or negative. Using this labeling, the system classifies the extracted opinions into positive or negative. **Figure 3** illustrates the process of the prototype RSE. When the RSE receives an object name, it first finds opinion candidates from the crawled web documents by using the evaluative expression dictionary (see Sections 3.2 and 3.3). An opinion candidate is defined as a passage that contains both an object name and an evaluative expression. The RSE then calculates opinion-likelihood scores using the pattern-matching rules (see Section 3.4). It finally classifies the extracted opinions into positive or negative (see Section 3.5) and outputs the results.

3.2 Evaluative Expression Dictionary

The evaluative expression dictionary is created by extracting expressions by hand from Web sites such as ‘Yahoo! Message Boards,’ where opinions on objects are often discussed. **Table 2** shows the top 20 most frequent evaluative expressions in 1000 collected expressions.

The dictionary was prepared on a domain basis, in this case, the computer and the alcoholic beverage domains, because there are many domain-specific expressions, as shown in Table 2. When we develop an RSE that covers more than one domain, users will need to select a domain-specific dictionary when they input

Table 2 Top 20 most frequent evaluative expressions (left: computer domain; right: alcoholic beverage domain).

Evaluative Expression	Lbl	Evaluative Expression	Lbl
良い (good)	Pos	おいしい (tasty)	Pos
使いやすい (handy)	Pos	良い (good)	Pos
満足 (satisfactory)	Pos	好き (like)	Pos
高い (high)	Pos	飲みやすい (soft)	Pos
最高 (greatest)	Pos	すっきり (clear)	Pos
便利 (useful)	Pos	すばらしい (fantastic)	Pos
お気に入り (favorite)	Pos	お薦め (recommendable)	Pos
速い (fast)	Pos	うまい (delicious)	Pos
すばらしい (fantastic)	Pos	しつかり (steady)	Pos
快適 (comfortable)	Pos	コクがある (rich)	Pos
不満 (dissatisfied)	Neg	さわやか (fresh)	Pos
きれいな (clear)	Pos	お気に入り (favorite)	Pos
優れる (superior)	Pos	面白い (interesting)	Pos
十分 (sufficient)	Pos	まろやか (mellow)	Pos
可愛い (pretty)	Pos	さっぱり (light)	Pos
魅力 (attractive)	Pos	いまいち (unsatisfied)	Neg
遅い (slow)	Neg	悪い (bad)	Neg
楽しむ (enjoy)	Pos	まずまず (passable)	Pos
悪い (bad)	Neg	上品 (elegant)	Pos
安い (reasonable)	Pos	強すぎる (too strong)	Neg

an object name.

There are two kinds of evaluative expressions. One kind expresses (a) people's emotion or feeling, such as “良い (good)” or “好き (like)”. The other expresses (b) the property or the nature of an object, such as “速い (fast)” or “小さい (small)”.

Each of the evaluative expressions in the dictionary has a label of positive or negative. The expressions of type (b) can sometimes have both a positive and a negative meaning. In this case, based on the assumption that this ambiguity can be reduced by creating the dictionary on a domain basis, we selected the label whose frequency was larger. For example, ‘小さい (small)’ appeared eight times on the computer domain, and in seven of these instances the word was used in a positive sense.

3.3 Opinion Candidates Extraction

An opinion candidate is a passage where both an object name and an evaluative expression appear. Specifically, if a passage satisfies the following search condition, the passage is selected as an opinion candidate:

Table 3 Examples of opinion candidates.

(a)	Mobile777、これは良い!! (Mobile777 is a good product!)	
(b)	Mobile777 は悪いマシンではない。(Mobile777 is not a bad product.)	
(c)	Mobile777 を持っております。ICQ を使いたのですがどうすれば良いでしょうか? (I have a Mobile777, and I want to use ICQ on it. What is a good way to do that?)	×
(d)	Mobile777 が良いという人もいます。が...(Some say that Mobile777 is a good product, but...)	×
(e)	PC の調子が悪いので、Mobile777 を使っていますが...(Since there is something wrong with my PC, I've used a Mobile777....)	×

W_O AND (W_{E1} OR ... OR W_{EN})

W_O : Object name

{ W_E } : Evaluative expression

We define a passage as sentences composed of one sentence that includes an object name and n sentences before and after that sentence.

3.4 Opinion-likelihood Calculation

Not all opinion candidates are correct. **Table 3** shows examples of opinion candidates for “Mobile777”. Here, (a) and (b) are correct examples, but (c) and (d) are incorrect ones because these sentences do not convey opinions. Furthermore, (e) is also incorrect because it does not contain an opinion on “Mobile777” but an opinion about another product. These examples reveal that opinion candidate extraction alone cannot provide high precision.

Opinion-likelihood calculation gives opinions like (a) and (b) high scores and opinions like (c) to (e) low scores by using the context of an opinion.

Pattern-matching rules mainly describe the relation between an object name and an evaluative expression and are used for the opinion-likelihood calculation. The rules are written by using regular expressions. **Table 4** shows six representative examples. There are six other rules for comparison or conjecture, etc., making 12 rules in all. In this table, $_OBJ_$ is replaced with the object name, and $_EXP_$ is replaced with the evaluative expression.

The system applies these rules to each opinion candidate and creates an array with n dimensions. n is the number of rules; in this case, $n = 12$. Each element of the array has a 1/0 value: 1 means that the opinion

Table 4 Pattern-matching rules (`_OBJ_`: an object name; `_EXP_`: an evaluative expression).

ID	Pattern-matching rules
1	<code>_OBJ_.*([。 。 ? !]).*_EXP_ </code> <code>_EXP_.*([。 。 ? !]).*_OBJ_</code> (Two expressions exist in different sentences.)
2	<code>_OBJ_{0,12}_EXP_ EXP_{0,12}</code> (Two expressions are located in close proximity.)
3	<code>_OBJ_.*_EXP_</code> (<code>_OBJ_</code> is located before <code>_EXP_</code> .)
4	<code>_OBJ_{0,12}{(は が も)}.*_EXP_</code> (<code>_OBJ_</code> is the subject of the sentence.)
5	<code>_EXP_{0,12}([。 。 ? !])</code> (<code>_EXP_</code> is close to the last of the sentence.)
6	<code>_EXP_.*か?</code> (<code>_EXP_</code> is in the question sentence.)

candidate satisfies the rule, and 0 means it does not. For example, (a) is converted into (0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0) and (c) is converted into (1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0).

Opinion-likeness scores are assigned by each array and learned from a learning data set. If there are more than one learning data on an array, the score is their average. As a result, (a) gets a high score of 0.76 and (c) gets a low score of 0.19.

3.5 Opinion Classification

The system classifies opinions by using labels attached to evaluative expressions (see Section 3.2) and negative expressions.

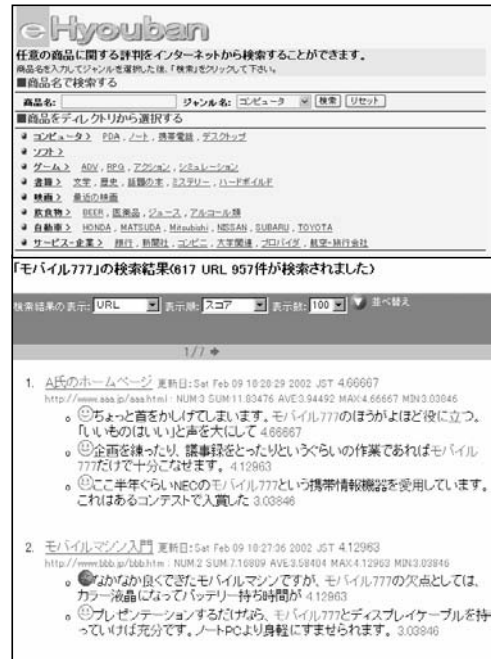
The system first looks up the label of the evaluative expression and then counts the frequency of negative expressions located closely before and after the evaluative expression. If the frequency is an even number, the system classifies the opinion as the label. Alternatively, if the frequency is an odd number, the system classifies the opinion as the reverse of the label.

Here, there are two groups of negative expressions. One contains grammatically negative expressions, “ない (not or no)”. The other contains lexically negative expressions, negating the sentence or the predicate with a lexicon such as “いがかたい (hard to say)”, “ほど遠い (far from)”, “非 (un-, in-, non-)” and “疑わしい (doubt)”. The system employs both groups.

4. Examples

Figure 4 shows screenshots of the prototype RSE. The extracted opinions are grouped by URL and displayed in the order of the opinion-likeness score. An icon that represents positive (a smiling icon) or negative (a crying icon) is shown at the side of each opinion.

The system collects Web documents in two

**Fig. 4** Screenshots of prototype RSE.

ways. One is by utilizing the results of general-purpose search engines, in which the system first gathers URLs by throwing an object name at them and then collects their Web documents. The other way is by using crawlers for message boards. Since each of the crawlers is tuned to the format of a message board, their posted dates can be added to extracted opinions.

These posted dates enable temporal analysis. **Figure 5** shows an example where opinions on a certain product have been counted for a five-day period. From this chart, it can be observed that the proportion of negative opinions after the product release date increases drastically, whereas it was low before. This observation means that this product did not meet consumers' expectations.

5. Experiment

5.1 Evaluation of Opinion Extraction

We compare the proposed method with a text classification method in which term frequencies are used as the feature space. The result proves the advantage of our IE approach in two representative domains on the Internet.

We selected two product domains: computers and alcoholic beverages, then chose four product names for each domain, i.e., four PDA products and four beers.

We selected these two domains by taking into

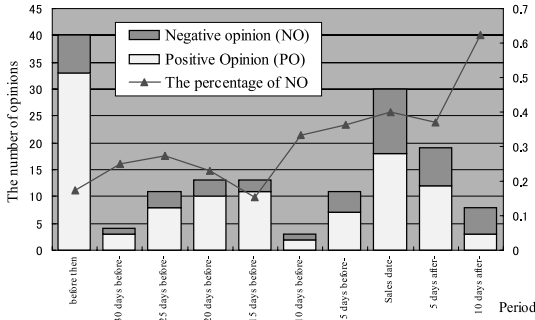


Fig. 5 Example of temporal analysis.

account the tendency of opinions on the Internet. There are two groups of product's domains. One contains (a) products whose functionalities tend to cause the basis of opinions. The other contains (b) products for which personal tastes tend to cause the basis of opinions. We selected one representative domain from each (a) and (b). Initially, we chose several domains as candidates on which the total number of opinions expressed on the Internet was large, i.e. computer and car domains for (a), and book, movie and alcoholic beverage domains for (b). We then considered the number of opinions expressed on each product. Book and movie domains of type (b) were inadequate for use in this experiment because there were so many products in these domains that the number of opinions on each product tended to be small. On the other hand, the computer domain of type (a) was adequate because there were a vast number of opinions on this domain, even though there were many products in it. Consequently, we selected the computer and alcohol beverage domains in terms of the type of product domains and the number of opinions on these domains.

The experimental target consists of passages from 2,400 web documents that were retrieved when we gave each product name to a general-purpose search engine and picked the top 300 web documents for each product name. 1200 documents in all (150 documents for each product) were used for the learning data and the remaining 1,200 documents were used for the test data.

We defined a passage, in this experiment, as three sentences: one sentence including a product name and one sentence before and one after that sentence. We gave each passage a label "1" when it was an opinion relevant to the product name and gave a label "0" when it was not.

Table 5 The number of learning and test data.

Domain	Passages	Learning Passages	Test Passages
Computer	2092	1106 (201/905)	986 (199/787)
Alcohol	2016	1083 (212/871)	933 (194/739)

Table 5 shows the number of passages used for the learning and the test data.

We then collected evaluative expressions from the learning data whose labels were "1". Consequently, 132 expressions for the computer domain and 63 expressions for the alcoholic beverage domain were found. In addition, opinion-likeness scores were also obtained from this learning data. These were calculated from 311 opinion candidates in the computer domain and 411 in the alcoholic beverage domain.

We compared the performance with that of Support Vector Machine (SVM) classification. SVM is known for being able to input a high dimension vector without the curse of dimensionality⁴⁾, and it has been reported that it can achieve higher accuracy than other machine learning methods such as the decision tree or naive Bayes⁵⁾. We used the same learning and test passages as the proposed method. We constructed a model from the learning data by using TinySVM, an SVM package developed by the Nara Institute of Science and Technology. We used noun, verb and adjective word frequencies for the SVM features.

Figure 6 shows the results of the proposed method. Since the rate of precision rises in proportion to the opinion-likeness score, it is clear that the rules we employed in Section 3.4 worked well. In addition, the average precision rate for the top ten results was 72% for the computer domain, 84% for the alcoholic beverage domain, and 78% in total.

Table 6 shows a comparison of the SVM classification and the proposed method. When we regard opinion candidates whose opinion-likeness scores were over 0.43 as opinions, the recall of the proposed method is closest to that of the SVM classification. Under this condition, the proposed method is superior to SVM classification in precision.

5.2 Size of the Evaluative Expression Dictionary

To extract all opinions in the test data (to

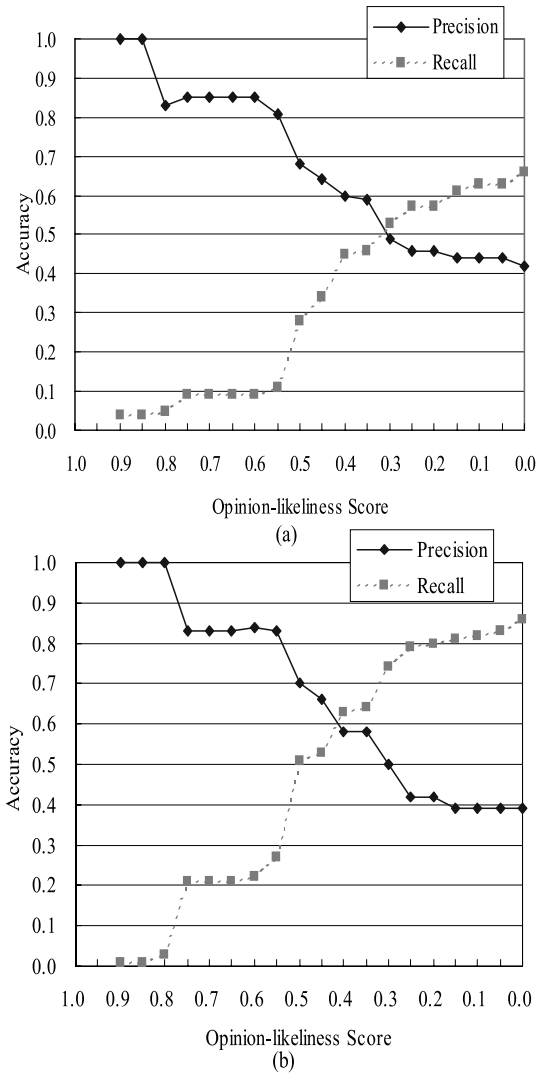


Fig. 6 Accuracy of proposed method for (a) computer domain and (b) alcoholic beverage domain.

Table 6 Comparison between SVM classification and the proposed method (above: computer domain, below: alcohol beverage domain).

Accuracy	SVM Classification	Our method
Precision	47%(82/172)	62%(86/139)
Recall	41%(82/199)	43%(86/199)
Accuracy	SVM Classification	Our method
Precision	51%(112/219)	59%(118/199)
Recall	57%(112/194)	61%(118/194)

achieve 100% recall), 89 expressions for the computer domain and 56 expressions for the alcoholic beverage domain were required. Among these expressions, about half of the expressions

(44/89) for the computer domain and about 40% of the expressions (23/56) for the alcoholic beverage domain were not common to the dictionary created from the learning data. This result means that the recall of our method largely depends on the size of the dictionary.

5.3 Evaluation of Opinion Classification

Next, we evaluated the ability to classify opinion into positive or negative types.

We used 137 correct opinions for the computer domain and 166 correct opinions for the alcoholic beverage domain. These were retrieved by the RSE using the test data. The evaluative expression dictionary was the same as in Section 5.1. We used the grammatically negative expressions “ない (not or no)” and fixed the scope for finding negative expressions to 12 bytes after the evaluative expression.

The precision was 87% (119/137) for the computer domain, 93% (154/166) for the alcoholic beverage domain, and 90% (273/303) in all. The overall accuracy was found to be high.

The most noticeable area of failure was when the label of evaluative expressions in the dictionary was wrong, which was in 18/30 instances. For example, we gave “高い (high)” a positive label, but it was used in a negative meaning like “(価格が高い) high price” in some extracted opinions. This means that ambiguity was not completely eliminated by creating the dictionary on a domain basis. The RSE will have to take account of the attribute expressions to achieve better opinion classification.

6. Related Work

The RSE can be categorized as a specialized search engine that uses an approach of information extraction technology. The objective of our research is to create a radar chart as reputation by extracting three entities of opinions: object names, attribute expressions and evaluative expressions.

Current specialized search engines are based on a text classification approach^{6)~9)}. This approach works well when the goal is just to find opinion-related documents, but our IE approach is more suitable when the goal is to create a radar chart because the text classification approach cannot discover entities of opinions. In addition, our IE approach is superior to the text classification approach with respect to accuracy (see Section 5.1).

No previous work in the IE field has given

attention to the opinion extraction. Our opinion extraction was inspired by previous work on named entity extraction. There are two approaches to building extraction systems²⁾: the knowledge engineering approach and the automatically trainable approach. We selected the former because the method of using pattern-matching rules and a word dictionary is more effective in MUC and IREX evaluation^{2),3)}.

Two research efforts have been undertaken in parallel with ours. One is research on classifying opinions into positive or negative ones^{10)~12)}, while the other is research on collecting opinion-related expressions^{13),14)}. The former research learns features of opinions using opinions on Internet review sites and incorporating their ratings as learning data, and classifies new articles as positive or negative. The latter study collects subjective expressions¹³⁾ or expressions on people's emotions¹⁴⁾.

However, since both groups of research do not take into account entities of opinions (object names, attribute expressions and evaluative expressions), they cannot determine to which object and attribute the opinion refers. They are therefore unsuitable for realizing our goal.

In addition, our evaluative expressions deal with a broader range of expressions than the latter research in that we collect not only subjective expressions that indicate people's emotion or feeling, but also objective expressions that indicate the properties or the nature of an object. It is necessary for products whose functionalities tend to cause the basis of opinions, such as those in car or computer domains, to cover objective expressions like “重い (heavy),” “大きい (large),” “速い (fast),” and “安い (inexpensive)” as well as subjective expressions like “楽しい (happy),” “好き (like),” “快適 (comfortable),” and “魅力 (attractive).”

7. Conclusions

In this paper, we proposed a new Internet search engine, the RSE. It collects people's opinions by using IE technology. After introducing the opinion model, we described how we developed the prototype RSE, which extracted people's opinions by using a dictionary and pattern-matching rules. The experimental result showed the advantage of our method in two representative domains on the Internet: computers and alcohol beverages.

Planned future work is as follows:

- (1) Expand the opinion model to that of three entities. Add (a-2) and (c) in Fig. 2 by creating an attribute expression dictionary and another set of pattern-matching rules.
- (2) Automatically construct an evaluative expression dictionary. The result in Section 5.2 revealed that a large dictionary is necessary to enable a high recall rate. There has been research on automatic collection of opinion-related expressions^{13),14)}. We plan to utilize this work to reduce the burden of manually creating the dictionary.
- (3) Prove the effectiveness on other products' domains. In the experiment, we evaluated the proposed method in two domains, computer and alcoholic beverage domains. However, this is not adequate for proving the generality. Since the RSE uses an evaluative expression dictionary on a domain basis and pattern-matching rules common to all domains, we plan to conduct additional experiments on many other domains by changing the dictionary and confirm the generality of our framework on a broad range of domains.
- (4) Expand the proposed method to other languages, as we expect that the same framework can be applied to languages other than Japanese. In this case, however, we need to change pattern-matching rules according to the grammar of the language as well as create an evaluative expression dictionary. We need to examine features of other languages, and develop and evaluate the RSE for them.

References

- 1) Tateishi, K., Ishiguro, Y. and Fukushima, T.: A Reputation Search Engine that Gathers People's Opinions from the Internet (in Japanese), Technical Report NL-144-11, Information Processing Society of Japan, pp.75-82 (2001).
- 2) Appelt, D. and Israel, D.: Introduction to Information Extraction Technology, Tutorial for IJCAI-99 (1999).
- 3) Takemoto, Y., Fukushima, T. and Yamada, H.: A Japanese Named Entity Extraction System Based on Building a Large-Scale and High-quality Dictionary and Pattern-matching Rules (in Japanese), *IPSJ Journal*, Vol.42, No.6, pp.1580-1591 (2001).
- 4) Taira, H. and Haruno, M.: Feature Selection in SVM Text Categorization, *Proc. Sixteenth*

National Conference on Artificial Intelligence (AAAI-99), pp.480–486 (1999).

- 5) Dumais, S., Platt, J., Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization, *Proc. Seventh International Conference on Information and Knowledge Management (CIKM-98)*, pp.148–155 (1998).
- 6) Steele, R.: Techniques for Specialized Search Engines, *Proc. Internet Computing 2001* (2001).
- 7) Chakrabarti, S., van den Berg, M. and Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery, *Proc. 8th International World Wide Web Conference (WWW8)* (1999).
- 8) Shakes, J., Langheinrich, M. and Etzioni, O.: Dynamic reference sifting: A case study in the homepage domain, *Proc. 6th International World Wide Web Conference (WWW6)* (1997).
- 9) McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: Building Domain-Specific Search Engines with Machine Learning Techniques, *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace* (1999).
- 10) Pang, B., Lee, L. and Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proc. Conference on Empirical Methods in Natural Language processing (EMNLP)*, pp.77–86 (2002).
- 11) Turney, P.: Thumbs Up? Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.417–424 (2002).
- 12) Dave, K., Lawrence, S. and Pennock, D.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, *Proc. 12th International World Wide Web Conference (WWW12)* (2003).
- 13) Riloff, E. and Wiebe, J.: Learning Extraction Patterns for Subjective Expressions, *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.105–112 (2003).
- 14) Liu, H., Lieberman, H. and Selker, T: A Model of Textual Affect Sensing using Real-World Knowledge, *Proc. 2003 International Conference on Intelligent User Interfaces (IUI2003)* (2003).

(Received September 20, 2003)

(Accepted February 3, 2004)

(Editor in Charge: *Masatoshi Arikawa*)



Kenji Tateishi received his B.S. degree in physics from Science Univ. of Tokyo in 1997 and M.S. degree in information engineering from Univ. of Kyushu in 1999. He joined Internet Systems Research Laboratories, NEC Corp. in 1999. He has been engaged in the field of Information Extraction and Information Retrieval on the Internet. He received Best Paper Award for the 64th IPSJ National Convention in 2002. He is a member of IPSJ.



Yoshihide Ishiguro received both B.E. and M.E. degrees in electrical engineering from Kyoto University in 1988 and 1990. He is a principal researcher of Internet Systems Research Laboratories, NEC Corporation and has been engaged in the field of groupware, software agent and information retrieval since his joining NEC in 1990. From 1998–1999, he was a visiting researcher at Georgia Institute of Technology. He is a member of the Association for Computing Machinery, the Information Processing Society of Japan, and Japanese Society for Artificial Intelligence.



Toshikazu Fukushima received his B.S. degree in physics from the Univ. of Tokyo and joined NEC Corp. in 1982. He is currently a senior manager of Ubiquitous Intelligence Technology Group in Internet Systems Research Laboratories, NEC Corp. He received his Ph.D. from Kyushu Univ. in 1998, Best Paper Award for Young Researcher of the 45th IPSJ National Convention in 1992, Best Paper Award of the 53rd IPSJ National Convention in 1996, the 23rd IPSJ Best Paper Award in 1992, the 6th IPSJ Sakai Special Researcher Award in 1997, and the 51st OHM Technology Award in 2003. He is a member of IPSJ, JSAI, ANLP, JSIK, INFOSTA and ACM.