Evaluation of the Document Categorization in “Fixed-point Observatory”

Yoshihiro Ueda Mamiko Oka Katsunori Houchi
Service Technology Development Department
Fuji Xerox Co., Ltd.
3-1 Minatomirai 3-chome, Nishi-ku, Yokohama-shi, Kanagawa 220-8401, Japan
Ueda.Yoshihiro@fujixerox.co.jp oka.mamiko@fujixerox.co.jp houchi.katsunori@fujixerox.co.jp

Akio Yamashita
FXPAL Japan
Fuji Xerox Co., Ltd.
1-1 Roppongi 3-chome, Minato-ku, Tokyo 106-0032, Japan
akio.yamashita@fujixerox.co.jp

Abstract

“Fixed-point observatory” is a prototype to support users to grasp recent trends in the fields of their interest from large-scale information. It consists of content-based categorizer, named-entity-based categorizer and multiple-document summarizer. We have evaluated the content-based categorizer, which adopts the simple “bag-of-words” model. Though the quality seems be sufficient for rough classification, it might be improved to use the categorizer in other applications.

Keywords: content-based categorizer, vector space model, bag-of-words.

1 Introduction

Many sort of office workers including executives, marketing researchers, planners and R&D engineers have to be aware of the recent trends in the field of their interest. For example, they would like to know which products sell well and what are the required functionalities and the signs of troubles.

“Fixed-point Observatory” is the concept of a system to support such office workers to filter and view the information from their points of view. The system consists of several modules; named-entity-based categorizer, content-based categorizer and multiple-document summarizer.

Named-entity-based categorizer provides a hierarchical view of the interested field by categorizing the documents according to the class hierarchy of the named entities contained in the documents. For example, a text that contains “Sony” is classified as a document in the “company” class, “manufacturer” class and “electronic appliance manufacturer” class. For the purpose, we give at most 4 category levels for each named entity.

Content-based categorizer routes the incoming text into one or more predefined categories. This functionality reinforces the named-entity-based categorizer. Named entities require “context” to identify their categories. For example, “ Rakuten” is the name of an e-commerce company and is also the name of a baseball team. The person who is surveying current activities of the e-commerce companies might want to avoid baseball news articles. Content-based categorizer provides the right context to a document by assigning the appropriate categories to it.

Multiple-document summarizer provides the shared topics in the classified documents by showing them in the form of short phrases. The summary is created by extracting the shared semantic fragments from the documents [1]. Unlike other systems that use keyword enumeration to overview the classified documents (e.g. Scatter/Gather [2]), users can grasp the relationship among the keywords. Especially how the selected named entities appear in the topics gives a valuable clue.

These functionalities are provided in the forms of software modules and these modules are provided for system integration.

In this paper, first we will show how these modules are combined to be effective in “Fixed-point Observatory.” Then the content-based categorization method will be briefly described in Section 3 and the evaluation result and analysis will be shown in Section 4.

2 System Overview

We have constructed a sample "Fixed-point Observatory" integration whose target is the websites that supply RSS such as news sites and blogs.

The system structure is shown in Fig. 1. At first, RSS is obtained from each site specified to the system. Then each HTML page listed in the RSS is collected and the body text is extracted from each page. Classified articles are represented by HTML hyperlink structure and can be accessed by ordinary Web browsers.

Fig.2 is a snapshot of the browser screen. Category names used in the content-based categorization are shown in the top area (1). The categories currently used are based on the classification of ordinary newspapers.

Named-entity-based hierarchical view of documents in the selected category is shown in the bottom left area (2). The hierarchy is a part of the whole hierarchy defined for each content-base category. For example, country names are essential for the "international" category and "economical" category and "technological" category require company names.

The titles in the selected named-entity category are listed in the bottom-right area (3). Each title is accompanied by short phrase summaries [3] with the specified named entity occurrences emphasized, which support the user to determine whether or not to read the document. Each title has a hyperlink to the original HTML page.

Multiple-document phrase summary is located at the top of the article list area (3). The specified named entity occurrences in the phrases are also emphasized.

3 Content-based categorizer

Content-based categorizer routes documents to predefined categories. The categorization method adopts a simple "bag-of-words" model in which the word vector for a document is the one created by the relevant document search [4]. No thesauri are applied to any words. The vector for each predefined category is composed from the word vectors of sample documents associated with the categories.

A word vector is calculated for each target document and is compared with all predefined categories' word vectors to identify the closest category for the document.

Each element of the vector for a predefined category is a score for a word and is calculated as follows.

$$ElemScore_i = \sqrt{\frac{ldf_i^{\,2} * ng}{gdf_i * nl^2}} *(1+ntf_i)$$

where ***ng*** is the number of documents in all categories and ***nl*** is the number of documents in the specified category. ***gdf*** (global document frequency) is the number of documents in all categories that contains the word$_i$ and ***ldf*** (local document frequency) is the number of documents in all categories that contains the word$_i$. ***ntf*** (normalized term frequency) is the number of the word occurrences normalized by the total word occurrences.

Score between a category and a document *Score(doc, cat)* is the sum of the element scores of

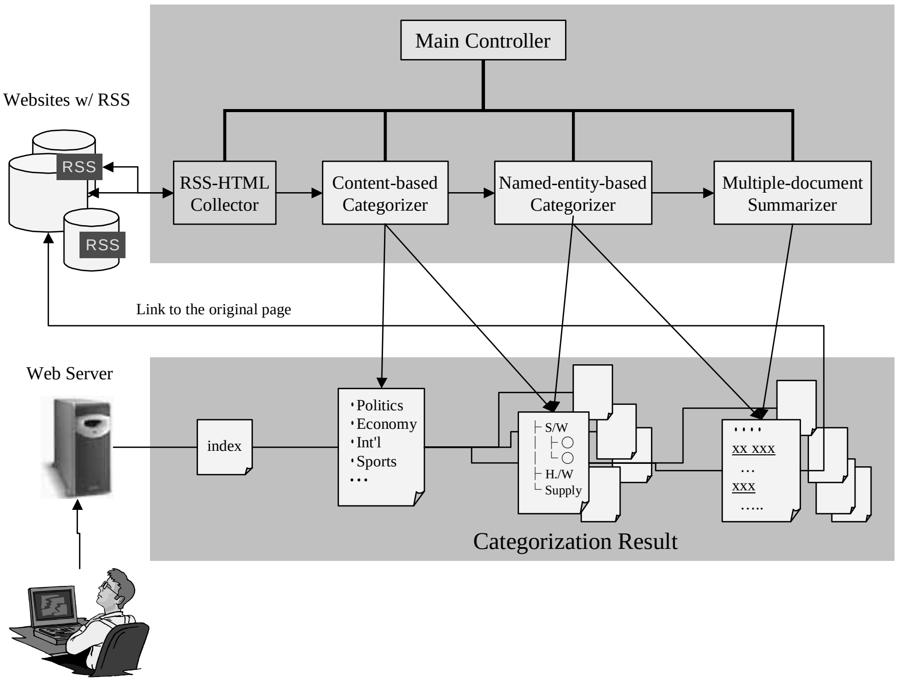

Fig 1. System Overview

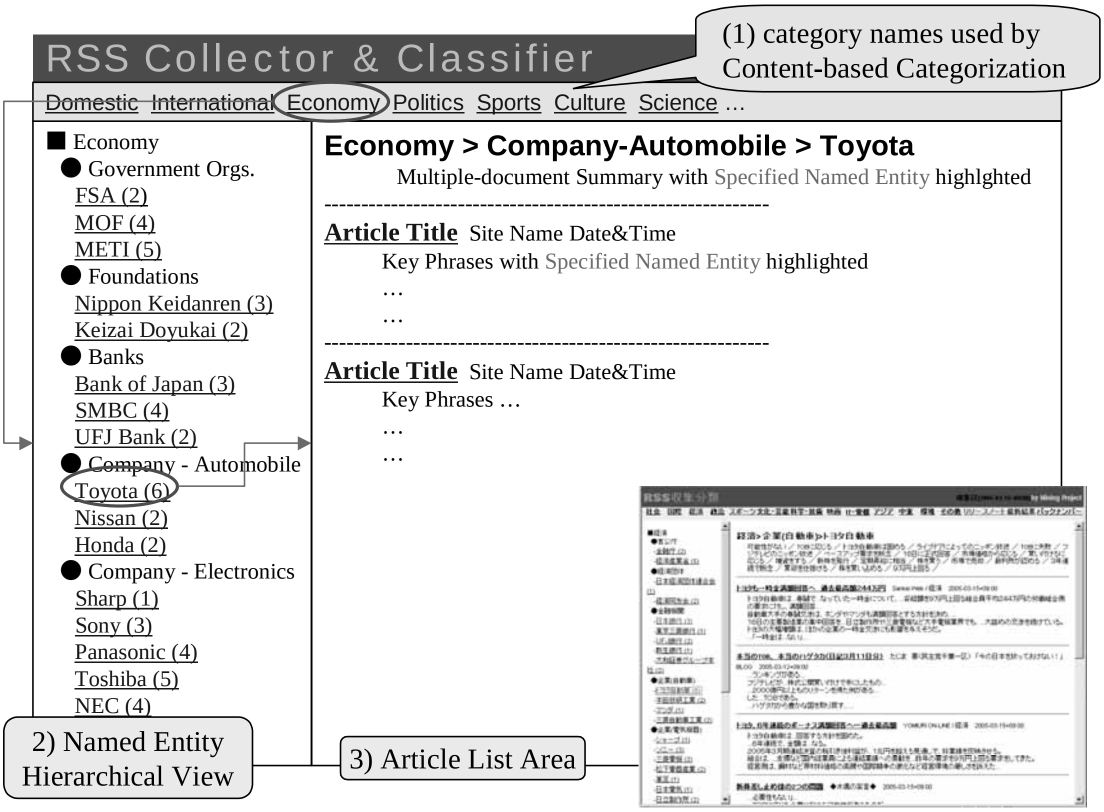
RSS Collector & Classifier
(1) category names used by Content-based Categorization
Domestic International Economy Politics Sports Culture Science …
■ Economy
● Government Orgs.
FSA (2)
MOF (4)
METI (5)
● Foundations
Nippon Keidanren (3)
Keizai Doyukai (2)
● Banks
Bank of Japan (3)
SMBC (4)
UFJ Bank (2)
● Company - Automobile
Toyota (6)
Nissan (2)
Honda (2)
● Company - Electronics
Sharp (1)
Sony (3)
Panasonic (4)
Toshiba (5)
NEC (4)
Economy > Company-Automobile > Toyota
Multiple-document Summary with Specified Named Entity highlghted
Article Title Site Name Date&Time
Key Phrases with Specified Named Entity highlighted
…
…
Article Title Site Name Date&Time
Key Phrases …
…
…
2) Named Entity Hierarchical View
3) Article List Area

Fig. 2 System View

the vector created by inserting the document to the category's document set and extracting only the shred words from the combined vector.

This formula is only an experimental one and has not been compared with any other candidates.

The words used in the vector can be selected according to their part of speech because words of some parts of speech appear in most documents. The categorizer also provides the option of whether to use compound words as a single word or a set of words.

A document can be assigned with more than one category by setting the relaxing ratio R, which tells the category (cat_i) satisfying the following condition is also assigned to the document.

$$\frac{Score(doc, cat_i)}{Score(doc, cat_{\max})} > R$$

"Others" category is also provided for the documents that are not close to any predefined categories. The "minimum relevance score" is introduced to control this.

This NTCIR-5 evaluation experiment taught us that the categorizer must be tuned to endure large-scale categorizations. The current implementation of the categorizer has sufficient performance for the presumed application shown in Section 2.

4 Evaluation and Analysis

4.1 Application Method

Here we describe how we applied the content-based categorizer to the given subtasks.

For theme categorization subtask, each theme is treated as a predefined category. Thus, the subtask can be considered to route 1119 patent documents into 2520 categories. However, to rank 100 themes for each patent document (to follow the instruction), we use the internal score matrix instead of the categorizer output itself. The categorizer output is used only to mark confidences; i. e. the confidence for a category is set to 1 if the categorizer assigned it to a patent document, otherwise to 0.

We have performed several runs with different settings. Differences among runs are as follows.

- **Selection of the part of the text**: The whole patent document is larger than the target of our presumed application. Thus, we use some part of the patent text for training and categorization target. The selections are (A) abstract, (B) claims, and (C) the combination of technical field, prior art and subject to be solved.

- **Selection of part of speech**: Basic selection is nouns including verb-stem noun (“Sa-hen meishi”), compound words and “Katakana” unknown words. Some runs omit compound words or “Katakana” unknown words and another run uses all independent words.

The number of the sample documents associated with each theme is 30 ~ 50. Though more samples are desirable to obtain a better result, the number of the theme is beyond our presumed application (the system described in Section 2 uses only 12 predefined categories). The documents are extracted equally from the whole training set.

For the f-term categorization subtask, a single categorization is not sufficient.

For each viewpoint, we have prepared a category set whose categories correspond to the elements of the viewpoint and executed the categorization with each category set with a specific viewpoint. We have extracted equal number of ranked patent documents in each category set to produce a single list.

4.2 Evaluation Result and Analysis

Table 1 is the comparison of the f-measure from selected results of theme categorization.

Comparison of text used - #11, #15 and #7: (C) The combination of technical field, prior art and subject to be solved is the best. The run using (B) claims follows this and the run using (C) abstract produced the poorest result.

This difference mainly is supposed to be caused by the length of the description. Another reason might be their uniformity. Descriptions in one abstract is the purpose and another the means. Claims tend to use less common words.

Comparison of POS used - #11, #19, #17 and #21: More parts of speech seems to produce better result. However, this consideration is not so reliable because they use the abstract and the comparison among the poor result.

We cannot get any analysis from f-term categorization because no run produced a good result.

This might be caused from the application method. Extracting equal number of ranked documents in each category set may drop many correct elements because many f-terms that share a single viewpoint are assigned to some patent documents.

To select the appropriate part of the document for each viewpoint might produce a better result. For example, “subject to be solved” part might be appropriate if the viewpoint for the purpose, and “claims” part for “means.”

4.3 Knowledge from the practice

Though we could not obtain any explicit evidence from the experiments, we have been considering that our categorization is hard to be applied to detailed categorizations from several trials not described here. On the other hand, we are confident of the quality of the content-based categorizer is sufficient for rough categorization from the experience of daily execution of the sample integration. However, the result of the theme categorization might tell that the categorizer needs to be enhanced to apply other than RSS collection and classification.

Conclusion

Here we introduced the concept and a sample implementation of “Fixed-point Observatory” that incorporate content-based categorization and other NLP-based features to support users to find tendency from large-scale information source.

From this evaluation experiment, we have come to know that;

the quality of categorizer still needs improvement in other application area, and that

the system must be tuned up or re-designed to be applied to larger-scale document set.

References

[1] Yoshihiro Ueda and Takahiro Koyama. Multiple Document Summarization by Extracting Shared Semantic Fragments. Proc. The 6th Annual Meeting of The Association for Natural Language Processing: 360-363. (in Japanese) 2000.

[2] Douglass Cutting, David Karger, Jan Pedersen, and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, Proc.

Table 1. Summary of the result of theme categorization

Our Run ID (Priority order)	# of samples (max)	Text Part used	POS used	f-measure
# 11	50	abstract	Noun, Compound, Unknown	31.2
# 15	50	claims	Noun, Compound, Unknown	32.2
# 7	30	technical field, prior art and subject to be solved.	Noun, Compound, Unknown	37.8
#19	50	abstract	Compound, Unknown	20.0
#17	50	abstract	All independent words	32.0
#21	50	abstract	Noun, Unknown	30.7
Best of all participant				52.7
Average of all participant				31.1

the 15th Annual International ACM/SIGIR Conference: 318-329. 1992.

[3] Yoshihiro Ueda, Mamiko Oka, Takahiro Koyama and Tadanobu Miyauchi. Toward the "At-a-glance" Summary: Phrase-representation Summarization Method. Proc. COLING-2000: 878-884. 2000.

[4] Hiroshi Umemoto, Tadanobu Miyauchi and Yoshihiro Ueda. Document Retrieval in Consideration of the Amount of Term Frequencies. Proc. the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization: 5-163-165. 2001.