Effects of Document Clustering in Modeling Wikipedia-style Term Descriptions

Atsushi Fujii, Yuya Fujii, Takenobu Tokunaga

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. Existing search engines provide an enormous volume of information, but retrieved information is not organized. Hand-compiled encyclopedias provide organized information, but the quantity of information is limited. In this paper, aiming to integrate the advantages of both tools, we propose a method to organize a search result based on multiple viewpoints as in Wikipedia. Because viewpoints required for explanation are different depending on the type of a term, such as animal and disease, we model articles in Wikipedia to extract a viewpoint structure for each term type. To identify a set of term types, we independently use manual annotation and automatic document clustering for Wikipedia articles. We also propose an effective feature for clustering of Wikipedia articles. We experimentally show that the document clustering reduces the cost for the manual annotation while maintaining the accuracy for modeling Wikipedia articles.

Keywords: Term description, Wikipedia, Clustering

1. Introduction

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. Existing search engines provide an enormous volume of information, but retrieved information is not organized. Hand-compiled encyclopedias provide organized information, but the quantity of information is limited. Even Wikipedia, which is a large free encyclopedia on the Web, does not include all the terms. In addition, descriptions in Wikipedia articles are restricted by authors' viewpoints.

To solve the quantity and quality problems above, we have been developing a method to produce encyclopedic dictionaries automatically from Web pages and patent documents (Fujii, 2008). Our method extracts paragraph-style term descriptions from a source text collection and classifies those descriptions into domains. Our method also extracts related terms for each headword and summarizes multiple descriptions into a single text. We have produced an encyclopedic dictionary including approximately 1 900 000 Japanese terms as headwords, which is available at an encyclopedic search site called "CYCLONE"[1]. CYCLONE has been used for various research purposes, including looking up a definition for a term and searching for patent documents.

Aiming to enhance the encyclopedic search, Fujii (2010) proposed a method to model Wikipedia-style term descriptions, i.e., how each term is described and structured in Wikipedia. In encyclopedias including Wikipedia, a single term is usually described from one or more viewpoints. In addition, a set of viewpoints required to describe a term is different depending on the type of the term in question. For example, while viewpoints for an animal include "ecology", "species", and "appearance", viewpoints for a disease include "symptom", "diagnosis", and "treatment". Thus, it is time-consuming to manually model term descriptions for various term types. To address this problem, a model of a viewpoint structure for each term type was extracted from Wikipedia. The resultant model was used to classify texts retrieved by a search engine in response to a keyword, such as "influenza", and produce a summary for that keyword from multiple viewpoints.

However, because categories annotated to Wikipedia articles are not well organized, in Fujii (2010) Wikipedia articles must be manually annotated according to the term type, prior to modeling term descriptions. In this paper, we investigate effects of document clustering in reducing the cost for the manual categorization. Although we target Japanese in this paper, our method is language-independent.

2. Related work

Fujii and Ishikawa (2004) summarized information on the Web based on viewpoints. However, because they targeted only the computer domain, a set of viewpoints was always the same. In addition, they used hand-crafted rules to extract sentences for each viewpoint, which is not scalable.

Biadsy et al. (2008) used Wikipedia to produce a biographic summary for a person. They used articles in Wikipedia to determine whether a sentence is a description for a person or not. However, they did not use viewpoints for summarization purposes.

Blair-Goldensohn et al. (2008) used aspects, which correspond to viewpoints in this paper, for summarizing customer reviews. For example, reviews for a restaurant are summarized based on aspects, such as "location" and "price". Although aspects are extracted from the texts to be summarized, we use Wikipedia as external knowledge to model viewpoints.

Sauper and Barzilay (2009) used Wikipedia articles for a term type, such as "disease", to model descriptions for that term type. Given a specific term, such as "influenza", they searched the Web for texts associated with "influenza" and

[1] http://cyclone.cl.cs.titech.ac.jp/

selected representative texts to generate a description for "influenza". However, in their method a type of an input term must be identified manually and only two term types ("American film actors" and "disease") were used for evaluation purposes. In addition, they did not address problems associated with lexical ambiguities (i.e., homonymy and polysemy) for a target term. For example, if a target term is "kiwi (bird/fruit)", their method cannot distinguish different meanings for "kiwi".

Fujii (2010) addressed the above problems related to Sauper and Barzilay (2009) and modeled term types and viewpoints in a single framework. However, in this method, Wikipedia articles must be manually annotated according to the term type. To reduce the manual cost, we investigate effects of document clustering for Wikipedia articles.

3. Methodology

3.1. Overview

Figure 1 depicts an overview of our method, in which the left and right regions correspond to processes for modeling term descriptions and classifying texts, respectively. Although Figure 1 is mostly based on Fujii (2010), the contribution of this paper is document clustering for articles in the modeling process. While in this section we explain the entire process in Figure 1, we elaborate on our clustering method in Section 3.2.

In Figure 1, the numbers of term types and viewpoints for each term type are two and three, respectively, without loss of generality. For each term type, articles describing terms associated with that term type are collected. Although categories annotated to Wikipedia articles can be useful for this purpose, Wikipedia categories are not well organized as expected. For example, articles for individual animals and articles for movies associated with animals are classified into the "animal" category together. Unlike Fujii (2010), in which articles were manually categorized according to the term type, in this paper we perform document clustering to automatically collect articles for each term type.

Articles in Wikipedia are usually structured based on "sections". For example, sections for "influenza" include "Etymology", "History", "Microbiology", and "Symptoms and diagnosis". We use each section heading as a single viewpoint. However, section headings for the same term type can vary depending on the author. Thus, we divide the collected articles into sections, and extract high-frequent section headings as viewpoints for the term type in question.

We use Support Vector Machines (SVM) to learn two types of classifiers. We use the One-Vs-Rest method to target more than two categories. Given a set of article fragments for each viewpoint, we learn a classifier for viewpoints ("viewpoint classifier"). We use words in article fragments as features.

We also learn a classifier for term types ("term classifier"), for which we combine all the article fragments in a term type as a single text. In Figure 1, we use texts for "animal" and "disease" to produce a term classifier. We use words in article fragments as features. However, unlike the viewpoint classifier, we also use words in the title of each article as features. For example, both "stomach cancer" and "lung cancer" include the word "cancer" that can be a strong clue associated with diseases.

In summary, our term description model consists of the term classifier and the viewpoint classifier for each term type modeled in the term classifier. This feature contrasts with Sauper and Barzilay (2009), which did not model term types.

The classification process is driven by a set of texts retrieved in response to a keyword, such as "influenza". For each of the texts, which can be either the entire Web page or a snippet extracted by a search engine, we first use the term classifier to determine the term type of "influenza". In Figure 1, because "disease" is selected as the term type for "influenza", we use the viewpoint classifier for "disease" to determine the viewpoint from which "influenza" is described in the input text. Finally, for each viewpoint, we select the top N texts with greater SVM scores, so that a user can obtain information about "influenza" from different viewpoints with a minimal cost. In addition, texts not describing "influenza", which are usually assigned a small score, can be discarded.

As a result, users can obtain encyclopedic information for terms not included in Wikipedia. Even if the target term is included in Wikipedia, the description is often restricted by author's viewpoints. However, our method collects various information about that term from the Web. In addition, individual articles in Wikipedia often lack some viewpoints. However, by collecting high-frequent viewpoints across articles, viewpoints for a term can be augmented.

3.2. Clustering for Wikipedia Articles

For clustering of Wikipedia articles, a method to represent each article by a feature vector and a clustering algorithm are important. While we use the repeated bisection clustering algorithm (Zhao and Karypis, 2005), we use the following feature types.

- Category in article (CAT)

  Each article in Wikipedia is annotated with one or more categories. Although Wikipedia categories are not well organized, category information is somewhat useful to determine the term type of an article title.

- Bag of words in article (BOW)

  This feature is usually used for representing target documents in clustering. Because Japanese sentences lack lexical segmentation, we use MeCab to perform morphological analysis and extract nouns, verbs, adjectives, out-of-dictionary words, and symbols as index terms.

- Character bigram in article title (CBG)

  Article titles for the same term type often look similar orthographically and share specific suffixes.

- Section heading in article (SEC)

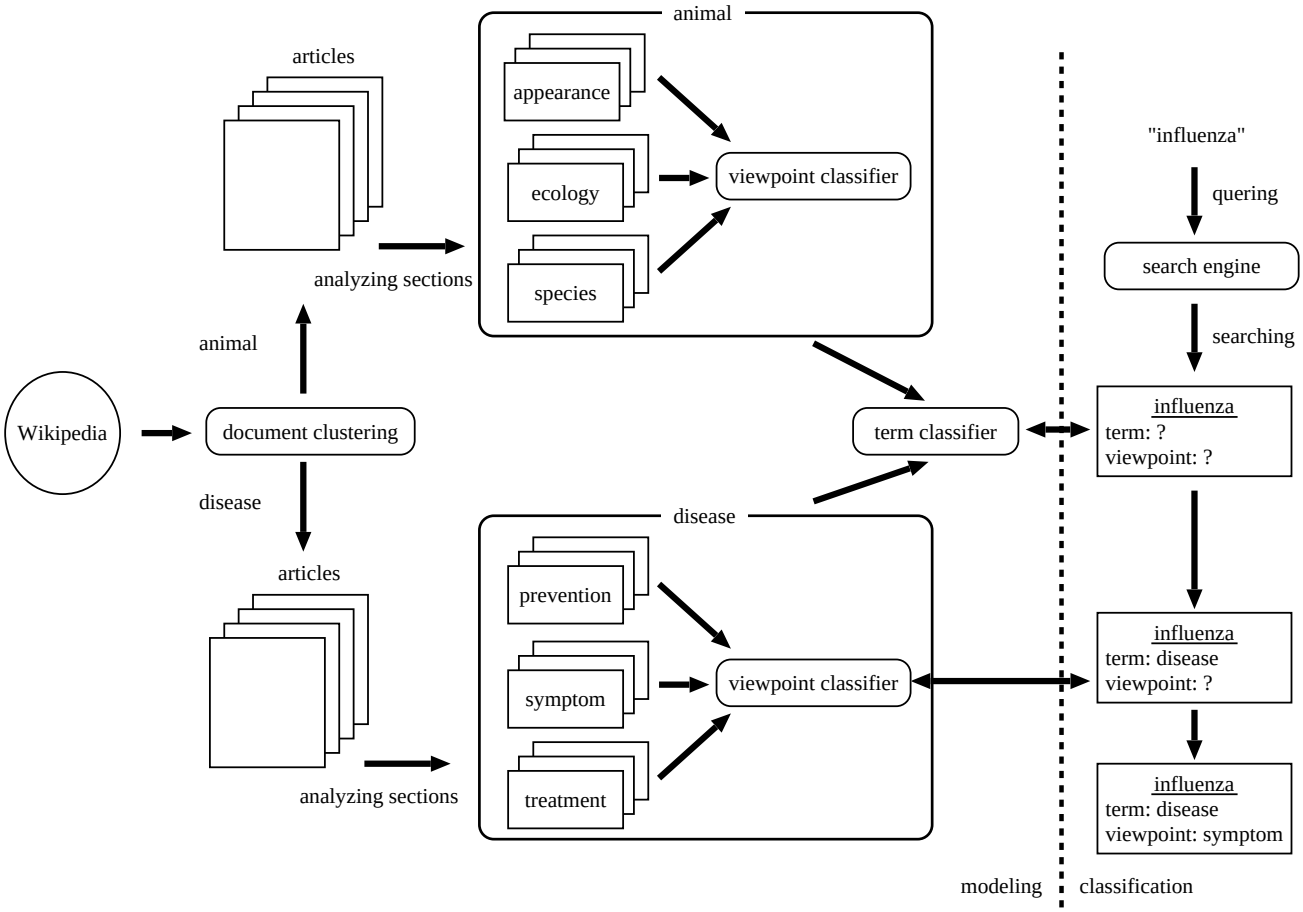

Figure 1: Overview of modeling Wikipedia articles and classifying online texts.

Articles for the same term type often share section headings.

- Extended section heading in article (EXS)

  Because the section headings in an article can be different depending on the author, the vocabulary mismatch problem is crucial if we rely only on SEC. To resolve this problem, we propose the notion of "extended section heading". For a target article, we search Wikipedia for such articles that share at least one category with the target article and use all section headings in the collected articles as features for the target article.

Among the above features, the extended section heading is proposed in this paper. In Section 4., we investigate the effectiveness of the individual features and any combinations of features.

4. Evaluation

We used the same data set as in Fujii (2010), which consists of 6 144 articles manually annotated with one of the following 20 term types.

- general: animal, company, cooking, disease, fish, insect, movie, person, plant, sports
- technical: astronomy, chemistry, construction, electricity, geology, informatics, law, mathematics, physics, veterinary

We performed experiments for the general and technical term types independently. Because the repeated bisection clustering requires a predefined number of clusters, we set this value to 10 in each experiment.

First, we evaluated the effectiveness of feature types in clustering for articles. As evaluation measures for clustering, we used purity, inverse purity, and F-measure (Amigó et al., 2009). Each of these measures compares a set of clusters to be evaluated with the set of reference categories. While purity focuses on the maximum precision for each cluster, inverse purity focuses on the maximum recall for each category. F-measure is a harmonic mean of purity and inverse purity. Purity and inverse purity are calculated by Equation (1).

$$\begin{aligned} purity(C,A) &= \frac{1}{N}\sum_j \max_k |c_j \cap a_k| \\ inv\text{-}purity(C,A) &= \frac{1}{N}\sum_k \max_j |c_j \cap a_k| \end{aligned} \quad (1)$$

$C = \{c_1, c_2, \ldots, c_J\}$ and $A = \{a_1, a_2, \ldots, a_K\}$ denote the set of clusters and the set of reference categories, respectively, and N is the total number of documents.

Table 1 shows purity, inverse purity, and F-measure of clustering results for different feature types. In Table 1, for

Table 1: Evaluation for clustering of Wikipedia articles.

	General						Technical					
	CAT	BOW	CBG	SEC	EXS	CBG+CAT+EXS	CAT	BOW	CBG	SEC	EXS	CBG+CAT+EXS
Purity	.797	.841	.829	.671	.897	.905	.691	.692	.658	.384	.794	.782
Inv-Purity	.793	.833	.810	.593	.908	.917	.599	.576	.618	.303	.724	.757
F-measure	.795	.837	.819	.630	.903	.911	.642	.629	.637	.339	.757	.769

the sake of simplicity we show the results for the individual feature types and CBG+CAT+EXS, which achieved the best F-measure among different combinations of feature types. Looking at Table 1, CBG+CAT+EXS generally outperformed the other feature types in clustering for articles. Comparing the results for the individual feature types, EXS, which is proposed in this paper, was most effective. Although CAT itself was not effective as predicted, BOW, which is usually effective for document clustering, was also not effective in our experiments.

Second, we evaluated the effectiveness of CBG+CAT+EXS in the term classification (see Figure 1). In real world usage, target texts for the classification are any texts. However, as in Fujii (2010), we used Wikipedia articles divided based on sections as input texts. We performed 5-fold cross-validation. Table 2 shows the accuracy of the term classification for Fujii (2010) and our method (CBG+CAT+EXS). Although Fujii (2010) outperformed our method in the term classification, unlike Fujii (2010), our method does not require manual categorization for Wikipedia articles.

Table 2: Evaluation for term classification.

	General	Technical
Fujii (2010)	87.3%	77.2%
CBG+CAT+EXS	80.4%	69.2%

5. Conclusion

Aiming to integrate the advantages of a search engine and an encyclopedia, we proposed a method to organize a search result based on multiple viewpoints as in Wikipedia. Because viewpoints required for explanation are different depending on the type of a term, such as animal and disease, we modeled articles in Wikipedia to extract a viewpoint structure for each term type. To identify a set of term types, we independently used manual annotation and automatic document clustering for Wikipedia articles. We proposed an effective feature for clustering of Wikipedia articles and showed its effectiveness experimentally. We also showed that the document clustering reduces the cost for the manual annotation while maintaining the accuracy for modeling Wikipedia articles.

6. References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 807–815.

Sasha Blair-Goldensohn, Ryan McDonald, George Reis, Tyler Neylon, Kerry Hannan, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW2008 Workshop on NLP Challenges in the Information Explosion Era*.

Atsushi Fujii and Tetsuya Ishikawa. 2004. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 645–651.

Atsushi Fujii. 2008. Producing an encyclopedic dictionary using patent documents. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Atsushi Fujii. 2010. Modeling Wikipedia articles to enhance encyclopedic search. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2591–2595.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 208–216.

Ying Zhao and George Karypis. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.