Cluster Labeling based on Concepts in a Machine-Readable Dictionary

Fumiyo Fukumoto
Interdisciplinary Graduate School of
Medicine and Engineering
Univ. of Yamanashi
fukumoto@yamanashi.ac.jp

Yoshimi Suzuki
Interdisciplinary Graduate School of
Medicine and Engineering
Univ. of Yamanashi
ysuzuki@yamanashi.ac.jp

Abstract

This paper addresses the issue of cluster labeling and presents a method for assigning labels by using concepts in a machine-readable dictionary. We assume that salient terms in the cluster content have the same hypernym because hypernymic semantic relation represents a generalization that goes from specific to generic. Our experimental results reveal that hypernymic semantic relations can be exploited to increase labeling accuracy, as the results of 0.441 F-score improves over the two baselines.

1 Introduction

With the exponential growth of information on the Internet, finding and organizing relevant materials on the Internet is becoming increasingly difficult. Internet directories such as Yahoo! and Google, which classify Web pages into pre-defined hierarchical categories, provide one solution to the problem. Categories in the hierarchical structures are carefully defined by human experts and documents are well-organized. However, manual category-tagging is extremely costly. Moreover, categories on some Internet is often insufficient in finding relevant documents for users. Because these categories tend to have some bias in both defining and classifying documents. Cluster labeling is one of the techniques to attack the problem.

Most of the work on cluster labeling identifies salient terms in the cluster content that characterize the cluster in contrast to other clusters. Salient terms are extracted by using statistical feature selection, *e.g.*, maximum sum of the individual term frequencies of documents assigned to a cluster (Cutting et al., 1992), an adapted versions of Information Gain (Geraci et al., 2007), χ^2 method (Popescul and Ungar, 2000), and the Jensen-Shannon Divergence (Camel et al., 2009). Other works based on salient terms are frequent phrases by (Osinski and Weiss, 2005), and integration of hierarchical information by (Muhr et al., 2010). However, the suggested terms, even when related to each other, tend to represent different aspects of the topic underlying the cluster, and it is often the case that a good label does not occur directly in the document. Carmel *et. al* addressed the issue and presented a method to use Wikipedia as an external knowledge. They showed the effectiveness of the method. However, Wikipedia is the free online encyclopedia, and everyone can access and edit the information. Therefore, it is often included noise information such as categories which do not characterize the cluster in the pages. Chin *et. al* presented a method to use WordNet (Chin et al., 2006). They used machine learning through extending the given term set with synonyms, hypernyms, hyponyms and so on. However, their method needs training data to determine the actual weights. Through supervised training in the labeling process the actual influence of synonyms, hypernyms, hyponyms information remains unclear.

This paper focuses on cluster labeling, and presents a method for assigning labels automatically by using concepts in a machine-readable dictionary. Similar to Chin *et. al* work, we focused on semantic relation in a dictionary, namely hypernymic semantic relation that represents a generalization, *i.e.*, goes from specific to generic (Fellbaum, 1998), and used it in the cluster labeling process. We assume that salient terms in the cluster content have the same hypernym in a hierarchical structure of a dictionary. The hypernym represents generic concepts of a set of documents, thus can be a label of a cluster.

2 Cluster Labeling

The procedure for cluster labeling consists of four steps: documents clustering, term weighting, hy-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1371–1375,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

pernym extraction, and ranking labels.

2.1 Documents clustering

The first step is to classify documents into a set with semantically similar documents. In the document clustering, we do not know how many clusters there are in a given input documents. Moreover, the algorithm should allow each data point to belong to more than one cluster because of multi-label classification. We used a graph-based unsupervised clustering technique developed by (Reichardt and Bomholdt, 2006); we call this the RB algorithm. This algorithm detects the node configuration that minimizes the energy of the material. The energy function, called the Hamiltonian, for assignment of nodes into communities clusters together those that are linked, and keeps separate those that are not by rewarding internal edges between different clusters. Here, "community" or "cluster" have in common that they are groups of densely interconnected nodes that are only sparsely connected with the rest of the network. Only local information is used to update the nodes which makes parallelization of the algorithm straightforward and allows the application to very large networks. Moreover, comparing global and local minima of the energy function allows the detection of overlapping nodes. Reichardt et al. evaluated their method by applying several data including a large protein folding network, and reported that the algorithm successfully detected overlapping nodes (Reichardt and Bornholdt, 2004). We thus used the algorithm to cluster documents. Let d_i ($1 \leq i \leq n$) be a document in the input, and σ_i be a label assigned to the cluster in which d_i is placed. The Hamiltonian H is defined as:

$$H(\{\sigma_i\}) \;=\; -\sum_{i<j}(A_{ij}(\theta) - \gamma p_{ij})\delta_{\sigma_i \sigma_j}. \quad (1)$$

δ denotes the Kronecker delta. The function $A_{ij}(\theta)$ refers to the adjacency matrix of the graph, which is defined as:

$$A_{ij}(\theta) = \begin{cases} 1 & \text{if } sim(d_i, d_j) > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$sim(d_i, d_j)$ in Eq (2) refers to cosine similarity between d_i and d_j. The matrix p_{ij} in Eq. (1) denotes the probability that a link exists between d_i and d_j, and is defined as:

$$p_{ij} \;=\; \sum_{i<j} \frac{A_{ij}(\theta)}{N(N-1)/2}, \quad (3)$$

where N refers to the number of documents and $\frac{N(N-1)}{2}$ is the total number of document pairs. As the parameter γ in Eq. (1) increases, each document is distributed into larger number of clusters. Eq. (1) thus shows comparison of the actual values of internal or external edges with its respective expectation value under the assumption of equally probable links and given data sizes. The minima of the Hamiltonian H are obtained by simulated annealing (Kirkpatrick et al., 1983). We applied simulated annealing for T runs [1].

2.2 Term weighting

For the results of clustering, we extracted salient terms from each clusters obtained by the RB algorithm. We tested four metrics which are commonly used as feature selection, *i.e.*, TF$*$IDF, mutual information, χ^2 statistics, and information gain. The terms we used are noun words in the documents. Each term is scored according to its contribution to the metrics between the cluster and other clusters. The top k scored terms are then selected as a candidate of the cluster salient terms.

2.3 Hypernym extraction

The third step is to extract hypernym for each term selected by term weighting method. We used Japanese word and concept dictionaries of EDR [2]. The word dictionary consists of 270,000 words. Each word has concept identifier as well as lexical and grammatical information. The concept identifier is to identify words and their concepts. The concept dictionary consists of 410,000 concepts. Each concept is linked to other concepts, and the link is a relation between concepts, namely super-sub relation. We used this super-sub relation as hypernymic semantic relation. Let W = $\{w_1$, w_2, $\cdots$, $w_n\}$ be a set of words in a cluster selected by feature selection. For each pair of words, w_i and w_j, we identified its hypernym c_k by using Eq. (4).

$$c_k \;=\; hy(w_i) \cap hy(w_j) \quad (4)$$

where $hy(w_i)$ and $hy(w_j)$ satisfy $min(dis(hy(w_i), w_i))$ and $min(dis(hy(w_j), w_j))$, respectively. In Eq. (4), $hy(x)$ refers to the hypernym of a word x. $min(dis(hy(w_i), w_i))$ shows the minimum distance between $hy(w_i)$ and w_i. We extracted hypernym by using Eq. (4), and regarded these as label candidates.

[1] We set T to 1,000 in the experiments

[2] http://www2.nict.go.jp/r/r312/EDR/index.html

Second level	Third level	Fourth level
sports	gymnastics	winter ski
music	opera	song
medicine	pharmacy	pharmaceuticals
education	school	teacher
architecture	house	flat
nature	environment	lebensraum
plants	botany	dicots
religion	religion in India	Buddhism
military	national defense	army
earth	geology	geomorphology
organism	anthropology	anthropologist
economy	labour	labour market
management	post	mail service
agriculture	animal care	poultry
animals	zoology	animal physiology
international law	UN	UNSC
finance	stock	bond

Table 1: Categories

2.4 Ranking Labels

The final step for cluster labeling is to rank label candidates according to their scores. The score of candidate c is obtained by using Eq. (5).

$$Score(c) \quad = \quad -\log\frac{freq\_p(c)}{N} \qquad (5)$$

where $N = \frac{1}{freq\_p(w_i)+freq\_p(w_j)}$. w_i and w_j are words selected by feature selection. $freq\_p(x)$ is the number of senses that the word x has.

3 Experiments

3.1 Data

We used two types of test data: one is a collection that correct labels occur directly in the documents. Another is that a label does not appear in the documents. The data we used is RWCP corpus labeled with UDC codes selected from 1994 Mainichi newspaper (RWC, 1998). It consists of 27,755 documents organized into fine-grained categories, 9,951 categories with a seven-level hierarchy. We used categories/labels assigned to the second, third and fourth level of a hierarchy, each of which has more than five documents [3]. Each level consists of 17 categories shown in Table 1. For each category, we randomly selected five documents, and created each type of test data. We extracted the top 20 scored words by term weighting as a candidate of the cluster salient terms.

	RB						EM		
Level	γ	θ	C	Prec	Rec	F	Prec	Rec	F
2nd	.1	.8	12	.601	.673	.635	.583	.673	.625
3rd	.9	.9	9	.620	.703	.659	.500	.742	.598
4th	1.0	.2	9	.398	.647	.493	.333	.633	.437
Avg	–	–	10	.539	.674	.595	.472	.682	.553

Table 2: Clustering Results

3.2 Clustering accuracy

For each category, we randomly selected five documents and created a training data to estimate two parameters, γ and θ. We represented each document as a vector of noun word frequencies, and applied RB algorithm. For evaluation of document clustering, we used F-score, especially to capture how many documents does the algorithm actually detect more than just one category. Precision was defined by the percentage of documents appearing in the correct clusters compared to the number of documents appearing in any cluster, and recall was defined by the percentage of documents within the correct clusters compared to the total number of documents to be clustered. For comparison of clustering algorithm, we used the EM algorithm that is widely used as a soft clustering technique(Nock et al., 2009). We set the initial probabilities by using the result of k-means clustering, where k is set to the number of correct clusters,1 7. We used up to 30 iterations to learn the model probabilities. The results are shown in Table 2.

Table 2 shows average performance between two types of test data. γ and θ in Table 2 denote the values that maximized the F-score obtained by using the training data. "C" refers to the number of clusters obtained by RB. The overall results obtained by the RB algorithm were better to those obtained by the EM algorithm regardless of the level of a hierarchy.

3.3 Labeling accuracy

We tested two types of document collection. For evaluation of cluster labeling, we used 11-point average precision. For comparison of the method, we used two baselines: (i) a feature selection by TF$*$IDF, and (ii) the use of Wikipedia for labeling. The method using Wikipedia is based on (Camel et al., 2009) [4]. The difference is that we used RB for clustering, and TF$*$IDF to extract salient

[3] We did not use categories assigned to the top level, as it was defined by only one label.

[4] We used Wikipedia downloaded from http://download.wikimedia.org.jawiki.

Level	Labels are included						Labels are not included				
	EDR				TF∗IDF	Wiki	EDR				Wiki
	TF∗IDF	MI	χ^2	IG			TF∗IDF	MI	χ^2	IG	
Second	0.460	0.318	0.281	0.288	0.150	0.236	0.500	0.304	0.288	0.272	0.153
Third	0.533	0.276	0.281	0.396	0.220	0.187	0.523	0.340	0.334	0.343	0.140
Fourth	0.310	0.254	0.214	0.256	0.183	0.194	0.299	0.262	0.193	0.220	0.142
Average	**0.434**	0.283	0.259	0.284	**0.184**	**0.206**	**0.441**	0.302	0.272	0.278	**0.145**

Table 3: The results of cluster labeling

Second level	Third level	Fourth level
vertebrate, life-form	**contest, sport**	**ski, athlete**
music, **opera**	**music**, **opera**	**song**, **music**
sick, hypofunction	sick, food	sick, antibiotic
book, building	rule, human	guide, rule
cook, **building**	**building**, activity	**building**, activity
nature, **natural phenomenon**	think, information	study, phenomenon
plants,**botany**	**botany**, **tree**	animals and plants, animal
religion, **human**	**belief**, **statue**	life, plants
reader, staff	**military**, **military affairs**	**military**, **army**
earth, **planet**	**geology**, message	**geology**, loss
organism, **life**	**life anthropology**	human, animal
economy social economy	**labour**, **worker**	**labour market**, **market**
management, **organization**	money, market	information, service
agriculture, vegetables	food, cook	care vegetables
animals, **mammals**	zoo, plant	care, food
international law, law	**UN**, USA	**UNSC**, society
bank, money	**stock**, **share**	**bond**, market

Table 4: Lists of top 2 terms (The top 20 term weighting scored terms)

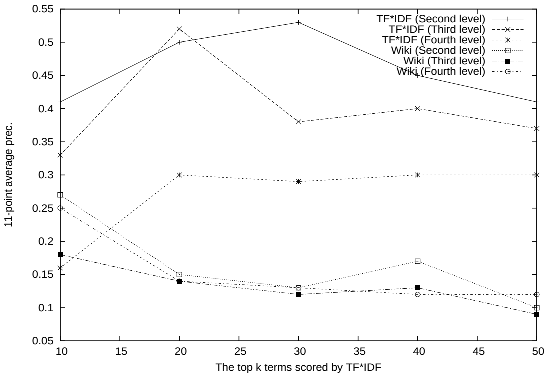

Figure 1: Performance against the top k terms

terms. The results are shown in Table 3. As can be seen clearly from Table 3, the results obtained by concepts based method were better than TF∗IDF and Wikipedia in both types of data. The results obtained by concepts based method show that there is no significant difference between two types of data, while the results by Wikipedia go down when we tested data that correct labels do not occur in the documents. This shows that the use of concepts in a dictionary improves overall performance. Table 4 shows a list of top 2 terms identified by concepts based method. Bold font terms are correctly identified by the method. Table 4 shows that more than half of the terms are correctly identified in the second and third level of a hierarchy.

We note that we set the number of scored terms to 20. To examine how the number of scored terms affects the overall performance, we performed an experiment by varying the values. Figure 1 shows performance plots against the top k terms scored by TF∗IDF. The best performance by both methods was around the top 20 terms scored by TF∗IDF term weighting method. The larger the number of scored terms becomes low precision. This is reasonable because a good label for a cluster generally consists of a few words.

4 Conclusions

We focused on cluster labeling, and presented a method for assigning labels by using concepts in a machine readable dictionary. Comparison with baselines showed improvements regardless of the level of a hierarchy. Future work will include: (i) incorporating hierarchical structure of documents, and (ii) applying the method to other data and thesaurus such as ODP dataset and WordNet.

Acknowledgement

The authors would like to thank the referees for their comments on the earlier version of this paper. This work was partially supported by the Telecommunications Advancement Foundation.

References

D. Camel, E. Yom-Tov, A. Darlow, and d. Pelleg. 2009. What Makes a Query Difficult? In *Proc. of the 32rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 139–146.

O. S. Chin, N. Kulathuramaiyer, and A. W. Yeo. 2006. Automatic Discovery of Concepts from Text. In *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 1046–1049.

D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. Scatter/ gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

C. Fellbaum. 1998. *An WordNet Electronic Lexical Database*. The MIT Press.

E. Gabrilovich and S. Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

F. Geraci, M. Pellegrini, m. Maggini, and F. Sebastiani. 2007. Cluster Generation and Labeling for Web Snippets: A Fast, Accurate Hierarchical Solution. *Internet Mathematics*, 3(4):413–443.

S. Kirkpatrick, C. G. Jr., and M. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

M. Muhr, R. Kern, and M. Granitzer. 2010. Analysis of Structural Relationships for Hierarchical Cluster Labeling. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–23.

R. Nock, P. Vaillant, C. Henry, and F. Nielsen. 2009. Soft Membershipfs for Spectral Clustering with Application to Premeable Language Distinction. *Pattern Recognition*, 42:43–53.

S. Osinski and D. Weiss. 2005. A Concept-Driven Algorithm for Clustering Search Results. *IEEE Intelligent Systems*, 20(3):48–54.

A. Popescul and L. H. Ungar. 2000. Automatic Labeling of Document Clusters.

J. Reichardt and S. Bomholdt. 2006. Statistical Mechanics of Community Detection. *PHYSICAL REVIEW E*, 74.

J. Reichardt and S. Bornholdt. 2004. Detecting Fuzzy Community Structure in Complex Networks with a Potts Model. *PHYSICAL REVIEW LETTERS*, 93(21).

RWC. 1998. *RWC Text Database*. In Real World Computing Partnership.

Z. S. Syed, T. Finin, and A. Joshi. 2008. Wikipedia as an Ontology for Describing Documents. In *Proc. of the International Conference on Weblogs and Social Media 2008*, pages 136–144.