Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 169 – 177

Procedia
Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

Utilizing Wikipedia in categorizing Topic related blogs into Facets

Daisuke Yokomoto[a], Kensaku Makita[a], Takehito Utsuro[a]*, Yasuhide Kawada[b] and Tomohiro Fukuhara[c]

Daisuke Yokomoto[a], Kensaku Makita[a], Takehito Utsuro[a]*, Yasuhide Kawada[b], and Tomohiro Fukuhara[c]

[a]*University of Tsukuba, 1-1-1, Tennodai, Tsukuba, 305-8573, JAPAN*
[b] *Navix Co., Ltd., Tokyo, 141-0031, JAPAN*
[c] *National Institute of Advanced Industrial Science and Technology,Tokyo 135-0064, JAPAN*

Abstract

Given a search query, most existing search engines simply return a ranked list of search results. However, it is often the case that those search result documents consist of a mixture of documents that are closely related to various sub-topics. This paper proposes a framework of categorizing blog posts according to their sub-topics. In our framework, the sub-topic of each blog post is identified by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a sub-topic label. We achieve to quickly overview the distribution of sub-topics over the whole collected blog posts.

Keywords: Blog, Faceted Search, Wikipedia, Topic Analysis , Search Engine

1. Introduction

As blog services and blog tools are becoming more and more popular, people have been able to express one's own interests as well as opinions on the Web. Search engines are then used for accessing various information that can be found in the blogosphere, where, given a search query, a ranked list of blog posts is provided as a search result. However, such a search result in the form of a ranked list is usually not helpful for a user to quickly identify blog posts that satisfy his/her information need. This is especially true when, given a search query, the search result is a mixture of blog posts that focus on various sub-

* Corresponding author

Email address: utsuro@iit.tsukuba.ac.jp

doi:10.1016/j.sbspro.2011.10.595

topics. In such a situation, the framework of faceted search[1], which has been well studied in the information retrieval community, can be a solution.

In this paper, we propose a framework of categorizing Japanese blog posts according to their sub-topics, where, given a search query, those blog posts are collected from the Japanese blogosphere. In our framework, the sub-topic of each blog post is regarded as a facet of an initial topic keyword, and a facet is automatically assigned to each blog post. For example, Fig. 1 illustrates a result of faceted search for an initial topic keyword "global warming" within the Japanese blogosphere. In this result, a number of collected blog posts regarding "global warming" are categorized into facets by identifying each blogger's interest in a blog post. This procedure of assigning a facet to a blog post is realized by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a facet label. In its Japanese version, about 760,000 entries are included (checked at August, 2011). Given an initial topic keyword, candidates of its facets are collected from Wikipedia. Then, for each facet candidate, its Wikipedia entry body text is analyzed as fundamental knowledge source for the facet itself, and terms strongly related to the facet are extracted. Those terms are then used for assigning this facet to a blog post. With this framework, it becomes much easier to quickly overview the distribution of sub-topics over the whole blog posts collected with a certain search query. In the evaluation, we show that the proposed method of assigning a facet to a blog post is effective and promising.

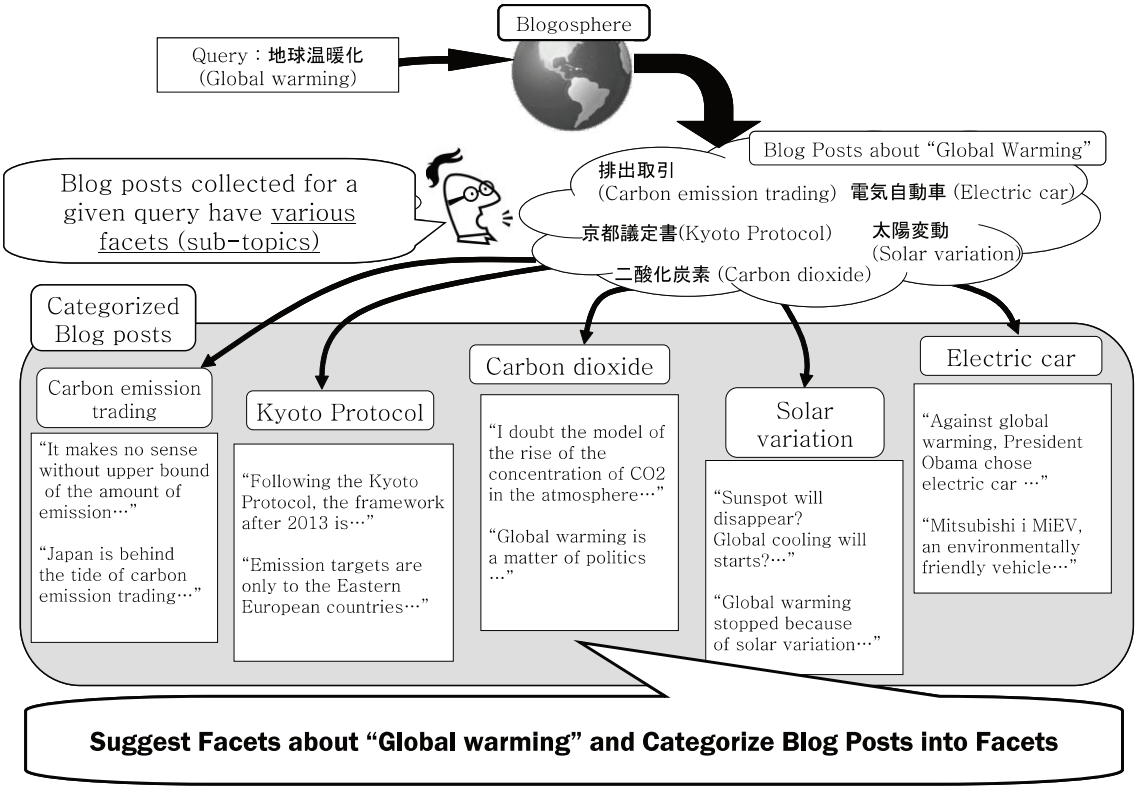

Fig. 1 Framework of Categorizing Blog Posts into Facets

2. Related Works

In TREC 2009 blog track[2], faceted blog distillation task was studied, where three facets, namely, *opinionated/personal/in-depth* are introduced and participants are required to assign facets to blog feeds.[3] invented a multi-faceted blog search framework, where various facets are introduced in terms of topics, bloggers, links, and sentiments. [4] also proposed a framework of generating a faceted search interface for Wikipedia. Compared to those previous works[2][3][4], the proposed method is innovative in that it realizes a novel technique of automatically generating sub-topic oriented facets for blog posts collected from the blogosphere. Our work is related to [4] in that both techniques collect facet candidates from Wikipedia. In [4], it is also presented how to rank facet hierarchies, where the cost of navigation through Wikiepedia facet hierarchies is modeled and is utilized in facet hierarchy ranking. However, compared to our technique, that of [4] modeled the cost of navigation only within the Wikipedia facet hierarchy, where the target of navigation is also Wikipedia articles. Our technique is different from that of [4] in that, in our technique, given the set of blog posts collected with an initial query as the target of navigation, facet candidates that are not frequently observed in the collected blog posts are removed. As a future work, it is also possible to introduce the formalization of the navigation cost of [4] into our task.

Another related works include techniques of clustering and summarizing search results[5], or those of clustering search results and assigning cluster labels [8][6][7][8]. Compared with those techniques on search results clustering, the proposed technique is advantageous in that it is capable of assigning facets to even quite a small number of blog posts, simply because it utilizes Wikipedia as a knowledge source for assigning facets to blog posts.

The technique presented in this paper is also related to previous works on assigning Wikipedia concepts to document clusters (e.g., [9]) and those combining Wikipedia concepts as well as important terms extracted from the cluster content in cluster labeling (e.g., [10]). However, those previous works mostly concentrate on clustering standard document sets such as those of newspaper articles with broad range of topics. In this paper, on the other hand, we focus on extracting facets from Wikipedia, given the set of blog posts collected with an initial query, where the collected blog posts cover much narrower range of sub-topics. Compared with the tasks studied in those previous works, the task of facet categorization of topic related blog posts studied in this paper is relatively difficult to tackle.

3. Retrieving Blog Posts with an Initial Topic Keyword

Given an initial topic keyword t_0, this section describes how to retrieve blog posts with t_0 as a search query. With this procedure, we intend to collect candidates of blog posts that are closely related to t_0.

First, we use an existing Web search engine API, which returns a ranked list of blog posts, given a topic keyword. For the evaluation in section 6, during the period from July to September, 2010, we used the Japanese search engine "Yahoo! Japan" API (http://www.yahoo.co.jp/ (in Japanese)) for Japanese. Blog hosts are limited to major 8 hosts (fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp) for Japanese. For each query, this search engine API returns a ranked list of at most 1,000 blog posts. A list of blog feeds is then generated from the returned ranked list of blog posts by simply removing duplicated feeds. From the retrieved blog feeds, we next collect blog posts that include the initial topic keyword t_0 in the body text into the set $D(t_0)$ of blog posts as candidates for those closely related to t_0.

4. Similarity of a Wikipedia Entry and a Blog Post

4.1. Idf vector of related terms extracted from a Wikipedia entry

Given a Wikipedia entry e, we automatically extract terms that are closely related to e. From the body text of each Wikipedia entry e, we extract bold-faced terms, anchor texts of hyperlinks, and the title of a

redirect, which is a synonymous term of the title of the target page. Then, we construct the set $R(e)$ of extracted related terms from the entry e.

Next, from each entry e, an idf vector $\boldsymbol{I}(e)$ is generated as below:

$$\boldsymbol{I}(e) = (w(r_1), \ldots, w(r_n)) \quad (1)$$

where, for each dimension of the vector, $r_i \in R(e)$ $(i = 1, \ldots, n)$ holds and for the length n of the vector, $n = |\boldsymbol{I}(e)| = |R(e)|$ holds. The weight $w(r)$ of each dimension is the idf (inverse document frequency) of a related term r over the set W of all entries in the Japanese version of Wikipedia:

$$w(r) = \mathrm{idf}(W, r) \quad (2)$$

$$\mathrm{idf}(W, r) = \log(|W| / |\{ e \in W \mid r \in R(e) \}|) \quad (3)$$

4.2. Term Frequency Vector of a Blog Post

Given a Wikipedia entry e and a blog post d $(\in D(t_0))$ retrieved with an initial topic keyword t_0, a term frequency vector $\boldsymbol{G}(d,e)$ is generated as below:

$$\boldsymbol{G}(d,e) = (freq(d,r_1), \ldots, freq(d,r_n)) \quad (4)$$

where, for each dimension of the vector, $r_i \in R(e)$ $(i=1,\ldots,n)$ holds and for the length n of the vector, $n = |\boldsymbol{G}(d,e)| = |R(e)|$ holds. The term frequency $freq(d,r)$ of each dimension is given as the frequency of a related term r in the blog post d.

4.3. The Similarity Measure

The similarity $Sim(e,d)$ of a Wikipedia entry e and a blog post d is defined as the inner product of the idf vector $\boldsymbol{I}(e)$ of e and the term frequency vector $\boldsymbol{G}(d,e)$ of d.

$$Sim(e,d) = \boldsymbol{I}(e) \bullet \boldsymbol{G}(d,e) = \Sigma_{r \in R(e)} w(r) \times freq(d,r) \quad (5)$$

5. Categorizing Blog Posts into Facets

5.1. Set of Facets

First, for each initial topic keyword t_0, we construct the set $F(t_0)$ of facets from all the entries in the Japanese version of Wikipedia. Let f_0 be a Wikipedia entry, where the initial topic keyword t_0 is included in the body text of f_0. We consider f_0 as a candidate of a facet for t_0. Then, we collect such f_0 that the document frequency $\mathrm{df}(D(t_0), t(f_0))$ of the title $t(f_0)$ of f_0 over the set of collected blog posts $D(t_0)$ is more than or equal to 30 into the set $F(t_0)$ of facets:

$$F(t_0) = \{ f \mid \mathrm{df}(D(t_0), t(f)) \geq 30 \} \quad (6)$$

5.2. Assigning a Facet to a Blog Post

In this section, we describe how to assign a facet to each blog post d $(\in D(t_0))$. In the evaluation of the next section, for each facet f $(\in F(t_0))$, as candidates of blog posts for evaluation, we collect 20 blog posts with the highest similarities $Sim(f, d)$:

$$d_1, d_2, \ldots, d_{19}, d_{20} \quad (\text{s. t.} \quad i < j, \quad Sim(f, d_i) \geq Sim(f, d_j)) \quad (7)$$

Next, to each candidate blog post d, we simply assign a facet f^{mx} $(\in F(t_0))$ which maximizes the similarity $Sim(f, d)$ of the facet f and the blog post d:

$$f^{mx} = \mathrm{argmax}_f \, Sim(f, d) \quad (\text{s. t. } f \in F(t_0)) \tag{8}$$

Then, we generate a pair $<d,f>$ of a blog post d and a facet f assigned to d and collect them into the set DF_{eval} for evaluation.

6. Evaluation

6.1. Evaluation Results

For evaluation, we pick up the 9 topics listed in Table 1 and Table 2 as the initial topic keywords. In Table 1 and Table 2, we also show the number of facets extracted according to the procedure in section 5.1 as well as the examples of extracted facets. Here, from the results of automatically extracting facets, we manually remove useless facet candidates such as those corresponding to hypernym of the initial topic keyword (e.g., a hypernym *"addiction"* of *"alcoholism"*) and closely related common words (e.g., a facet candidate *"alcoholic beverage"* for *"alcoholism"*). The number of removed facet candidates is 6 in total for the 9 topics.

For each pair $<d,f>$ $(\in DF_{eval})$ of a blog post d and a facet f assigned to d, we manually judge whether the following two criteria are satisfied:

- The blog post d is closely related to the initial topic keyword t_0.
- The blog post d is closely related to the facet f.

Manual evaluation was done by an evaluator other than the authors, while one of the authors fully checked the evaluation result by the evaluator, and corrected errors in the evaluation result by the evaluator.

In this evaluation procedure, we first manually judge whether the facet automatically assigned to the blog post is *appropriate* or *inappropriate*. Then, in the case where the automatically assigned facet is *inappropriate*, we manually assign a *reference* facet to each blog post.

Then, we measure the following *accuracy*:

$$\text{accuracy} = (\#\text{ of pairs } <d,f> (\in DF_{eval}) \text{ for which } d \text{ is closely related to both } t_0 \text{ and } f) \,/\, |DF_{eval}| \tag{9}$$

Table 1 and Table 2 show the accuracy for each of the 6 topics for evaluation. For most of the 9 topics for evaluation, we achieved about 30-70 % accuracy.

Examples of assigning an *appropriate* facet to blog posts include the case for an initial topic keyword *"smoking"* and a facet *"passive smoking"*, where those blog posts and the Wikipedia entry with the title *"passive smoking"* share related terms such as *"symptom of passive smoking"* and *"(Japanese) health-promotion law"*. Another example is the case for an initial topic keyword *"organ transplantation"* and a facet *"Japanese legislation for organ transplantation"*, where those blog posts and the Wikipedia entry with the title *"(Japanese) health-promotion law"* share related terms such as *"Liberal Democratic Party (Japan)"*, *"A plan"*, and *"D plan"*.

6.2. Error Analysis

On the other hand, cases of assigning an *inappropriate* facet to blog posts can be roughly categorized into the following 6 types:

1. The blog post is closely related to the initial topic.
   - (1a) The reference facet is included in the set $F(t_0)$ of facets for the initial topic keyword.
   - (1b) Although the reference facet is *not* included in the set $F(t_0)$ of facets for the initial topic keyword, it is listed in Wikipedia as an entry.
   - (1c) The reference facet is *not* listed in Wikipedia as an entry.
   - (1d) Although the blog post is closely related to the initial topic, it loosely covers certain sub-topics of the initial topic keyword, and it is difficult to assign a specific facet to it.
2. The blog post is *not* closely related to the initial topic.
   - (2a) The facet assigned by the proposed method is closely related to the blog post.
   - (2b) The facet assigned by the proposed method is *not* closely related to the blog post.

Most cases of the type (1a) are due to the fact that the reference facet and the *inappropriately assigned* facet share many related terms in their Wikipedia entries. For example, for an initial topic keyword *"smoking"*, the reference facet *"smoking ban"* and the *inappropriately assigned* facet *"nicotine addiction"* share many related terms such as *"nicotine"*, *"smoking"*, and *"addiction"*.

The type (1b) includes the case for an initial topic keyword *"organ transplantation"*, where the reference facet is *"diseased kidney transplant"* and the *inappropriately assigned* facet is *"immunosuppressive drug"*. In this case, in the set $D(t_0)$ of the blog posts collected for the initial topic keyword *"organ transplantation"*, the title of the facet *"diseased kidney transplant"* has relatively low document frequency compared with other facet titles such as *"immunosuppressive drug"*. With this reason, the reference facet *"diseased kidney transplant"* is not included in the set $F(t0)$ of facets for the initial topic keyword *"organ transplantation"*. Moreover, in the description of the Wikipedia entry with the title *"immunosuppressive drug"*, it is stated that *"immunosuppressive drug"* is often used in *"organ transplantation"* including *"kidney transplant"*. Thus, those blog posts and theWikipedia entry with the title*"immunosuppressive drug"* share many related terms.

The type (1c) includes the case for an initial topic keyword *"smartphone"*, where the reference facet is *"moTweets"* (name of an application) and the *inappropriately assigned* facet is *"Pocket PC"*. In this case, the reference facet is somehow known in the blogosphere, while it is before the reference facet is listed in Wikipedia as an entry.

The type (1d) includes the case for an initial topic keyword *"smoking"*, where the blog post is somehow related to *"smoking manner"*. However, according to the nature of Wikipedia, it is not usual that an entry with a title such as *"smoking manner"* is found in Wikipedia. In this case, the blog post is *inappropriately assigned* a facet *"smoking ban fascism"*. In such a case, an alternative strategy should be listing up several related facets, each of which actually exists in Wikipedia as an entry.

The type (2a) includes the case for an initial topic keyword *"global warming"*, where the facet assigned to the blog post by the proposed method is *"wind power"*. In this case, although the blog post is

closely related to the assigned facet *"wind power"*, it is not related to the initial topic keyword *"global warming"*.

Finally, most cases of the type (2b) are due to errors of assigning facets to the blog post, where the blog post is related to neither the initial topic keyword nor the assigned facet.

Table 1 Accuracy of Pairs of <Blog Post, Facet> for 5 Initial Topics

Initial Topic	Facet	Accuracy of Pairs of <Blog Post, Facet> (%)	(# of pairs $<d, f>$ ($\in DF_{eval}$) for which d is closely related to both t_0 and f) / $\lvert DF_{eval} \rvert$
Smoking	Passive smoking	87.5	7 / 8
	Smoking ban	100	5 / 5
	Smoking ban fascism*	50.0	3 / 6
	others (17 facets)	63.6	21 / 33
	total (20 facets)	69.2	36 / 52
Organ Transplantation	Japanese legislation for organ transplantation*	100	9 / 9
	Immunosuppressive drug	46.2	6 / 13
	Uwajima Tokushukai Hospital*	100	4 / 4
	others (15 facets)	65.0	26 / 40
	total (18 facets)	68.2	45 / 66
Global warming	Kyoto protocol	75.0	9 / 12
	Renewable energy	62.5	5 / 8
	Ecotax	100	4 / 4
	others (23 facets)	64.8	35 / 54
	total (26 facets)	67.9	53 / 78
Medical error	Physician	56.2	9 / 16
	Medical error lawsuit	54.5	6 / 11
	Japan Council for Quality Health Care*	100	4 /4
	others (11 facets)	48.6	18 / 37
	total (14 facets)	54.4	37 / 68
Population ageing	Birth dearth	80.0	8 / 10
	Social security	77.8	7 / 9
	Pension	71.4	5 / 7
	Others (9 facets)	17.1	6 / 35
	total (12 facets)	47.5	28 / 59

Facets marked with * are without English entries in Wikipedia, most of which are found only in Japanese society.

Table 2 Accuracy of Pairs of <Blog Post, Facet> for 4 Initial Topics

Initial Topic	Facet	Accuracy of Pairs of <Blog Post, Facet> (%)	(# of pairs $<d, f>$ ($\in DF_{eval}$) for which d is closely related to both t_0 and f) / $\mid DF_{eval} \mid$
Toyota Prius	Electric car	22.2	4 / 18
	Toyota	35.3	6 / 17
	Hybrid vehicle	75.0	9 / 12
	others (14 facets)	21.7	5 / 23
	total (17 facets)	34.3	24 / 70
Smartphone	Android (operating system)	100	4 / 4
	W-ZERO3*	66.7	2 / 3
	Willcom	44.4	4 / 9
	others (34 facets)	26.6	17 / 64
	total (37 facets)	33.8	27 / 80
Alcoholism	Driving under the influence	76.9	10 / 13
	Mental disorder	16.7	2 / 12
	Narcotic	8.3	1 / 12
	others (12 facets)	25.0	8 / 32
	total (15 facets)	30.4	21 / 69
Restructuring	Job hunting	20.0	3 / 15
	Employment	50.0	4 / 8
	Termination of employment	100	1 / 1
	Others (6 facets)	3.8	1 / 26
	total (10 facets)	18.0	9 / 50

Facets marked with * are without English entries in Wikipedia, most of which are found only in Japanese society.

7. Conclusion

In this paper, we proposed a framework of categorizing Japanese blog posts according to their sub-topics, where, given a search query, those blog posts are collected from the Japanese blogosphere. In our framework, the sub-topic of each blog post is regarded as a facet of an initial topic keyword, and a facet is automatically assigned to each blog post. This procedure of assigning a facet to a blog post is realized by utilizingWikipedia entries as a knowledge source and eachWikipedia entry title is considered as a facet label. In the evaluation, we showed that the proposed method of assigning a facet to a blog post is effective and promising.

References

[1] Tunkelang D. *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan \& Claypool Publishers, 2009.

[2] Macdonald C., Ounis I., and Soboroff I., Overview of the TREC-2009 Blog Track. In *Proceedings of TREC-2009*, 2009.

[3] Fujimura K., Toda H., Inoue T., Hiroshima N., Kataoka R, and Sugizaki M., BLOGRANGER - a multi-faceted blog search engine., In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, Aggregation, Analysis and Dynamics,* 2006.

[4] Li C., Yan N. Roy S. B., Lisham L. and Das G., Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In *Proceedings of the 19th WWW*, 2010, p. 651-660.

[5] Harashima J.and Kurohashi. S. Summarizing search results using {PLSI}. In *Proceedings of the 2nd Workshop on NLPIX*, 2010, p. 12-20.

[6] Toda H., Kataoka R., and Oku M., Search result clustering using informatively named entities. In *International Journal of Human-Computer Interaction*, 2007, p. 3-23,.

[7] de Winter W. and de Rijke M. Identifying facets in query-biased sets of blog posts. In *Proceedings of ICWSM*, 2007, p. 251-254.

[8] Shibata T., Bamba Y., Shinzato K., and Kurohashi S. Web information organization using keyword distillation based clustering. In *Proceedings of WI-IAT*, 2009, p. 325-330.

[9] Hu J., Fang L., Cao Y., Zeng H.-J, Li H., Yang Q., and Chen Z. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st SIGIR*, 2008, p. 179-186.

[10] Carmel D., Roitman H., and Zwerdling N. Enhancing cluster labeling using Wikipedia, In *Proceedings of. The 32nd SIGIR*, 2009, p. 139-146.