Utilizing Wikipedia as a Knowledge Source in Categorizing Topic related Korean Blogs into Facets

Dongkwon LIM†　Daisuke YOKOMOTO††　Kensaku MAKITA†
Takehito UTSURO††　Tomohiro FUKUHARA†††

†College of Engineering Systems, School of Science and Engineering, University of Tsukuba
††Graduate School of Systems and Information Engineering, University of Tsukuba
†††Center for Service Research,
National Institute of Advanced Industrial Science and Technology

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 not usually 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-topics. In such a situation, the framework of *faceted search* [8], which has been well studied in the information retrieval community, can be a solution.

In this paper, we propose a framework of categorizing Korean blog posts according to their sub-topics, where, given a search query, those blog posts are collected from the Korean 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, Figure 1 illustrates a result of faceted search for an initial topic keyword *"global warming"* within the Korean 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 the evaluation, we can achieve about 50∼70 % accuracy.

2 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. We use the search engine "Naver Open API"[1] for Korean. With this API, we restrict the target as "blog", indicating that we only collect blog posts. For each query, this search engine API returns a ranked list of at most 1,000 blog posts.

Then, out of all the returned blog posts, we focus on major 4 Korean blog hosts[2], and manually generate patterns for extracting the body text from each blog post. For each initial topic keyword t_0, we construct the set $P(t_0)$ of blog posts, each of which is with its body text successfully extracted from the blog post. Here, for each t_0 of the 6 topics we list in section 4 for evaluation, we have about $700 \sim 850$ blog posts included in the set $P(t_0)$.

3 Categorizing Blog Posts into Facets

3.1 Set of Facets

First, for each initial topic keyword t_0, we construct the set $F(t_0)$ of facets from the whole entries in the Korean 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 $\mathrm{df}(P(t_0), t(f_0))$ of the title $t(f_0)$ of f_0 over the set of collected blog posts $P(t_0)$ is more than or equal to 5 into the set $F(t_0)$ of facets:

$$F(t_0) \;=\; \Big\{\, f \;\Big|\; \mathrm{df}(\; P(t_0),\; t(f)\;) \;\geq 5 \Big\}$$

[1] http://dev.naver.com/openapi/ (in Korean)
[2] blog.naver.com, blog.daum.net, blog.cyworld.com, blog.paran.com

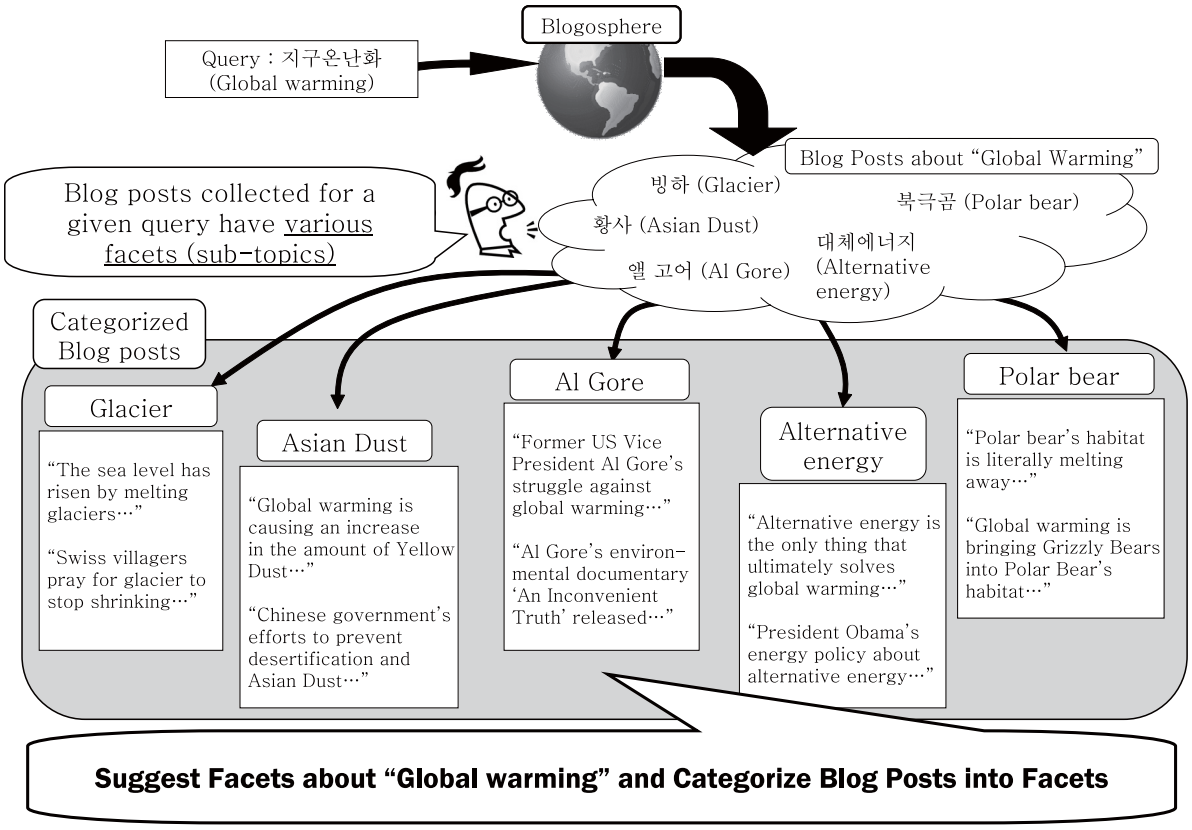

Figure 1: Framework of Categorizing Blog Posts into Facets

3.2 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 $\vec{I}(e)$ is generated as below:

$$\vec{I}(e) \quad = \quad \Big(\; w(r_1), \; \ldots, \; w(r_n) \; \Big)$$

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, $n = |\vec{I}(e)| = |R(e)|$ holds.

The weight $w(r)$ of each dimension is given as the product below:

$$w(r) = c(type(r)) \times \mathrm{idf}(W, r) \times \mathrm{idf}(F(t_0), r)$$

Here, $\mathrm{idf}(X, r)$ is the idf (inverse document frequency) of a related term r over the set X of Wikipedia entries. Note that whether a related term r is included in a Wikipedia entry e $(\in X)$ or not is not measured against the body text of the entry e, but against the set $R(e)$ of related terms extracted from e.

$$\mathrm{idf}(X, r) \quad = \quad \log \frac{\big|X\big|}{\Big|\big\{e \in X \big| r \in R(e)\big\}\Big|}$$

Also, $\mathrm{idf}(W, r)$ is intended to measure the idf of a related term r over the set W of the whole entries in the Korean version of Wikipedia, while $\mathrm{idf}(F(t_0), r)$ is intended to measure the idf of a related term r over the set $F(t_0)$ of facets constructed in the previous section from t_0. The coefficient $c(type(r))$ is defined as 3 when $type(r)$ is the entry title or the title of a *redirect*, as 2 when $type(r)$ is a bold-faced term, and as 0.5 when $type(r)$ is an anchor text of a hyperlink to another entry in Wikipedia.

3.3 Term Frequency Vector of a Blog Post

From each blog post p $(\in P(t_0))$ retrieved in section 2 given an initial topic keyword t_0, a term frequency vector $\vec{G}(p)$ is generated as below:

$$\vec{G}(p) = \big(\; freq(p, r_1), \; \ldots, \; freq(p, r_n) \; \big)$$

where, for each dimension of the vector, $r_i \in R(e)$ $(i = 1, \ldots, n)$ holds and for the length n of

Table 1: Accuracy of Pairs of ⟨Blog Post, Facet⟩ and Examples of Facets

Topics	Accuracy of Pairs of <Blog post, Facet> (%)	# of Facets	Examples of Facets
흡연 (Smoking)	60.7 (51 / 84)	22	간접 흡연 (Passive smoking), 전자담배 (Electronic cigarette), 폐암 (Lung cancer), 재떨이 (Ashtray), 폐기종 (Emphysema), 조산 (Preterm birth), 유산 (Miscarriage), 골다공증 (Osteoporosis), 공기 청정기 (Air purifier), 말보로 (Marlboro), 니코틴 (Nicotine), 세계 금연의 날 (World No Tobacco Day), etc.
스마트폰 (Smartphone)	64.3 (63 / 98)	28	아이폰 (iPhone), Samsung Galaxy S, 애플 (Apple Inc.), 아이폰 4 (iPhone 4), 스냅드래곤 (Snapdragon), CYON Optimus Q, Wi-Fi, 블랙베리 (BlackBerry), iOS, 테더링 (Tethering), 넷북 (Netbook), 피처 폰 (Feature phone), 넥서스원 (Nexus One), 모바일 장치 (Mobile device), 윈도 폰 7 (Windows Phone 7), Samsung Omnia II, 증강현실 (Augmented reality), 윈도 모바일 (Windows Mobile), etc.
구조조정 (Restructuring)	56.3 (27 / 48)	14	재무개선작업 (Debt restructuring), 고용 (Employment), 이명박 정부 (Lee Myung-bak government), 김대중 (Kim Dae-jung), 대한민국 금융위원회 (Financial Services Commission), 재단법인 (Foundation), 전국민주노동조합총연맹 (Korean Confederation of Trade Unions), IMF 구제 금융 요청 (IMF crisis), 대한민국 금융감독원 (Financial Supervisory Service), etc.
지구온난화 (Global warming)	69.8 (88 /126)	37	빙하 (Glacier), 이산화탄소 (Carbon dioxide), 남극 (Antarctica), 지구 (Earth), 황사 (Asian Dust), 탄소세 (Carbon tax), 기상이변 (Meteorological disasters), 폭염 (Heat wave), 대기 오염 (Air pollution), 온실효과 (Greenhouse effect), 빙하기 (Ice age), 교토의정서 (Kyoto Protocol), 온실 기체 (Greenhouse gas), 탄소 (Carbon), 버락 오바마 (Barack Obama), 북극 (Arctic), 북극곰 (Polar bear), 뎅기열 (Dengue fever), 경제학 (Economics), 기후학 (Climatology), 대체 에너지 (Alternative energy), 앨 고어 (Al Gore), 불편한 진실 (An Inconvenient Truth), 종말론 (Eschatology), 4대강 정비 사업 (The Four Major Rivers Project), 가뭄 (Drought), 세계 자연보호 기금 (World Wide Fund for Nature), 기후 (Climate), 투발루 (Tuvalu), 키리바시 (Kiribati), etc.
독도 (Liancourt Rocks)	47.4 (74 / 156)	43	독도 분쟁 (Liancourt Rocks dispute), 우산도 (Usan-do), 일본강치 (Japanese Sea Lion), 반크 (VANK), 가상공간 (Cyberspace), 천암함 침몰 사건 (ROKS Cheonan sinking), 괭이갈매기 (Black-tailed Gull), 독도레이서 (Dokdo Racer), 독도의 날 (Dokdo Day), 다케시마의 날 (Takeshima Day), 쓰시마 섬 (Tsushima Island), 일본 (Japan), 이명박 정부 (Lee Myung-bak government), 요미우리 신문 (Yomiuri Shimbun), 울릉도 (Ulleungdo), 우체통 (Post box), 배타적 경제 수역 (Exclusive Economic Zone), 안용복 (An Yong-bok), 시마네 현 (Shimane Prefecture), etc.
저작권 (Copyright)	59.2 (77 / 130)	36	한국음악저작권협회 (Korea Music Copyright Association), 파일 공유 (File sharing), 유튜브 (YouTube), 소리바다 (Soribada), 표절 (Plagiarism), 저작권 침해 (Copyright infringement), 영화관 (Movie theater), 전자책 (e-book), 공정이용 (Fair use), Mnet, 지적재산권 (Intellectual property), 형법 (Criminal law), 크리에이티브 커먼즈 (Creative Commons), 싸이월드 (Cyworld), 스타벅스 (Starbucks), etc.

the vector, $n = |\vec{I}(e)| = |R(e)|$ holds. The term frequency $freq(p, r)$ of each dimension is given as the frequency of a related term r in the blog post p.

3.4 Similarity of a Wikipedia Entry and a Blog Post

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

$$Sim(e,p) \;=\; \vec{I}(e) \cdot \vec{G}(p) \;=\; \sum_{r \in R(e)} w(r) \times freq(p,r)$$

3.5 Assigning a Facet to a Blog Post

In this section, we describe how to assign a facet to each blog post p ($\in P(t_0)$). In principle, to each blog post p, we simply assign a facet f ($\in F(t_0)$) which maximizes the similarity $Sim(f, p)$ of the facet f and the blog post p:

$$f \;=\; \operatorname*{argmax}_{f' \in F(t_0)} Sim(f', p)$$

Then, we generate a pair $\langle p, f \rangle$ of a blog post p and a facet f assigned to p. Finally, in the evaluation of the next section, for each facet f ($\in F(t_0)$), we collect five pairs of $\langle p, f \rangle$ with the highest similarities $Sim(f, p)$ into the set PF_{eval} for evaluation:

$$\langle p_1, f \rangle,\ \langle p_2, f \rangle,\ \langle p_3, f \rangle,\ \langle p_4, f \rangle,\ \langle p_5, f \rangle$$
$$(\text{ s. t. } \quad i < j, \quad Sim(f, p_i) \geq Sim(f, p_j)\ )$$

4 Evaluation

For evaluation, we pick up the 6 topics listed in Table 1 as the initial topic keywords. In Table 1, we also show the number of facets extracted according to the procedure in section 3.1 as well as the examples of extracted facets[3].

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

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

[3]From the results of automatically extracting facets according to the procedure in section 3.1, we manually remove useless facet candidates such as those corresponding to hypernym of the initial topic keyword (e.g., a hypernym *"mobile phone"* of *"smartphone"*) and closely related common words (e.g., a facet candidate *"earth"* for *"global warming"*). The number of removed facet candidates is 25 in total for the 6 topics.

Then, we measure the following *accuracy*:

$$\text{accuracy} = \frac{\begin{array}{c}\#\text{ of pairs } \langle p, f\rangle\ (\in PF_{eval}) \text{ for which} \\ p \text{ is closely related to both } t_0 \text{ and } f\end{array}}{\left|PF_{eval}\right|}$$

Table 1 shows the accuracy for each of the 6 topics for evaluation. For each of the 6 topics for evaluation, we achieved about 50∼70 % accuracy.

Examples of assigning an *appropriate* facet to blog posts include the case for an initial topic keyword *"smoking"* and a facet *"miscarriage"*, where those blog posts and the Wikipedia entry with the title *"miscarriage"* share related terms such as *"pregnancy"*, *"fertilization"*, *"ovum"* and *"fetus"*. On the other hand, examples of assigning an *inappropriate* facet to blog posts include the case for the initial topic keyword *'smoking"* and a facet *"electronic cigarette"*, where those blog posts should be assigned a much more *appropriate* facet such as *"passive smoking"*. However, in this case, the Wikipedia entry with the title *"passive smoking"* has just a small number of related terms such as *"cigarette"*, *"smoking ban"* and *"smoker"*, and most of those related terms are also shared with the Wikipedia entry with the title *"electronic cigarette"*. In such a case, it is not easy to assign the appropriate facet *"passive smoking"* to those blog posts, but the inappropriate facet *"electronic cigarette"* is assigned.

Furthermore, most cases of assigning an *inappropriate* facet to blog posts are due to the fact that the set $F(t_0)$ of facets for the initial topic keyword t_0 does not include any *appropriate* facets for certain number of blog posts. For example, in the case of an initial topic keyword *"smartphone"*, bloggers introduce a lot of names of smartphone devices, where most of them are not listed as titles of entries in the Korean version of Wikipedia. In such cases, we could not include names of those smartphone devices in the set $F(t_0)$ of facets, and thus certain number of blog posts are assigned *inappropriate* facets included in $F(t_0)$.

5 Related Works

In TREC 2009 blog track [4], 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. [1] invented a multi-faceted blog search framework, where various facets are introduced in terms of topics, bloggers, links, and sentiments. [3] also proposed a framework of generating a faceted search interface for Wikipedia.

Another related works include techniques of clustering and summarizing search results [2], or those of clustering search results and assigning cluster labels [6,7]. 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 employed in this paper is originally invented in our related works [5,9], where it is applied to the task of categorizing Japanese blog posts into facets.

6 Conclusion

In this paper, we proposed a framework of categorizing Korean blog posts according to their sub-topics, where, given a search query, those blog posts are collected from the Korean 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 utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a facet label. In the evaluation, we achieved about 50∼70 % accuracy.

References

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

[2] J. Harashima and S. Kurohashi. Summarizing search results using PLSI. In *Proc. 2nd Workshop on NLPIX*, pages 12–20, 2010.

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

[4] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2009 blog track. In *Proc. TREC-2009*, 2009.

[5] K. Makita, D. Yokomoto, T. Utsuro, and T. Fukuhara. Analyzing temporal distribution of sub-topics in blogs related to a topic. In *Proc. 3rd DEIM*, 2011. (in Japanese).

[6] T. Shibata, Y. Bamba, K. Shinzato, and S. Kurohashi. Web information organization using keyword distillation based clustering. In *Proc. WI-IAT*, pages 325–330, 2009.

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

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

[9] D. Yokomoto, D. Lim, K. Makita, T. Utsuro, Y. Kawada, T. Fukuhara, N. Kando, M. Yoshioka, H. Nakagawa, and Y. Kiyota. Utilizing Wikipedia as a knowledge source in categorizing topic related blogs into facets. In *Proc. 3rd DEIM*, 2011. (in Japanese).