Towards Conceptual Indexing of the Blogosphere through Wikipedia Topic Hierarchy

Mariko Kawaba[a], Daisuke Yokomoto[b], Hiroyuki Nakasaki[b],
Takehito Utsuro[b], and Tomohiro Fukuhara[c]

[a] NTT Cyber Space Laboratories, NTT Corporation, Yokosuka, Kanagawa, 239-0847, JAPAN
[b] University of Tsukuba, Tsukuba, 305-8573, JAPAN
[c] University of Tokyo, Kashiwa, 277-8568, JAPAN

Abstract. This paper studies the issue of conceptually indexing the blogosphere through the whole hierarchy of Wikipedia entries. About 300,000 Wikipedia entries are used for representing a hierarchy of topics. Based on the results of judging whether each blog feed is relevant to a given Wikipedia entry, this paper proposes how to judge whether there exist blog feeds to be linked from the given entry. In our experimental evaluation, we achieved over 90% precision in this task.

Keywords: blogosphere, Wikipedia, blog feed retrieval, topics

1 Introduction

Weblogs or blogs are considered to be one of personal journals, market or product commentaries. While traditional search engines continue to discover and index blogs, the blogosphere has produced custom blog search and analysis engines, systems that employ specialized information retrieval techniques. With respect to blog analysis services on the Internet, there are several commercial and non-commercial services such as *Technorati*, *BlogPulse* (Glance et al., 2004), and *kizasi.jp* (in Japanese). With respect to multilingual blog services, *Globe of Blogs* provides a retrieval function of blog articles across languages. *Best Blogs in Asia Directory* also provides a retrieval function for Asian language blogs. *Blogwise* also analyzes multilingual blog articles.

In terms of conceptual indexing of the blogosphere, existing services for blog retrieval can be roughly divided into two types. The first type is that of keyword based blog search function of search engines such as *Yahoo! blog search* (in Japanese) and *Google blog search* (in Japanese). In this type of indexing, not only keywords, but also subjective expressions as well as time series changes are used for indexing. This type of indexing is too fine-grained compared to actual needs for indexing the blogosphere. Since the number of indices is extremely huge, it is definitely impossible for users to grasp the whole structure of the index hierarchy. Therefore, unless each user comes up with queries appropriate for their search needs in the blogosphere, it is difficult for them to easily access the blogosphere with certain information needs. The second type is that of manually indexing the blogosphere through a directory of manually created categories such as *Technorati*. This type of indexing is, on the other hand, too coarse-grained compared to actual needs for indexing the blogosphere, and such indexing lacks coverage in the whole blogosphere. It is also quite difficult to manually update such a directory of categories when blog feeds of new topics are created in the blogosphere.

Based on this observation, this paper takes an approach of conceptually indexing the blogosphere through the whole hierarchy of Wikipedia entries. In our approach, we regard Wikipedia as a large scale ontological knowledge base for conceptually indexing the blogosphere. We regard Wikipedia also as a large scale encyclopedic knowledge base which includes well known facts and relatively neutral opinions. In its Japanese version, about 623,000 entries are included (checked in October, 2009). For the purpose of conceptually indexing the blogosphere, Wikipedia has an advantage over any other ontological knowledge resource. Although many blog feeds with new topics keep being created rapidly, in Wikipedia, new entries for describing those new topics are also rapidly created, and existing entries also keeps being updated rapidly.

More specifically, this paper proposes how to link Wikipedia entries to blog feeds in the Japanese blogosphere, where about 300,000 Wikipedia entries are used for representing a hierarchy of topics. Furthermore, based on the results of judging whether each blog feed is relevant to a given Wikipedia entry, this paper also examines how to judge whether there exist blog feeds to be linked from the given entry. In the following sections, first, we examine correlation between the number of hits of a Wikipedia entry title and existence of blog feeds to be linked from the entry. We empirically examine the range of the number of hits and conclude that the entries with the range over 10,000 tend to have relevant blog feeds. Actually, according to our manual evaluation of this range, about 80% of Wikipedia entries have at least one relevant blog feed. Second, we apply SVMs to the task of judging whether a blog feed is relevant to a given Wikipedia entry. Based on the learned SVMs model, we automatically judge whether there exists at least one blog feed which is relevant to the given Wikipedia entry. In our experimental evaluation, we achieved over 90% precision in this task.

2 Wikipedia

In the evaluation of this paper, we collected about 400,000 Japanese entries in November, 2007, and removed entries such as data logs and historical eras as noise. The resulting 305,986 entries are used in the evaluation. The hierarchical structure of Wikipedia can be represented as a undirected graph of categories, where at each category, one or more Wikipedia entries are listed. The version we used in this paper has 29,970 categories and from the root category, topmost 8 categories, namely, *“academia”*, *“technology”*, *“nature”*, *“society”*, *“geography”*, *“humans”*, *“culture”*, and *“history”* are directly connected. From those topmost 8 categories, about 700 categories are directly connected.

3 Criterion on Judging Relevance between a Wikipedia Entry and a Blog Feed

In this paper, in order to judge whether a blog feed is relevant to the description in a Wikipedia entry, we roughly follow the criterion studied in the blog distillation task (Macdonald et al., 2007) in TREC 2007 blog track. The blog distillation task can be summarized as *Find me a blog with a principle, recurring interest in X*. For a given target X, systems should suggest feeds that are principally devoted to X over the time span of the feed, and would be recommended to subscribe to as an interesting feed about X. Here, systems analyze the multiple posts of a given feed.

4 Analyzing Correlation between Hits of the Wikipedia Entry Title and Existence of Blog Feeds

In this section, before applying machine learning techniques to the task of judging relevance between a Wikipedia entry and a blog feed, we examine the issue of the number of hits of each Wikipedia entry title in the blogosphere. More specifically, in the analysis on existence of blog feeds to be linked from a Wikipedia entry, we examine correlation between the number of hits of a Wikipedia entry title and existence of blog feeds to be linked from the entry.

4.1 Preliminary Analysis

We first identified the range of the numbers of the hits of a Wikipedia entry title in the Japanese blogosphere, where blog feeds to be linked from each entry most exist compared to the rest range. The resulting range is over 10,000. We further examined the tendency of blog feeds linked from Wikipedia entries and observed certain differences between the range of 10,000 $\sim$ 500,000 and that over 500,000. In the range over 500,000, most Wikipedia entry titles can be regarded as general terms and some of the blog feeds linked from those entries should be linked rather from descendant concepts of those entries. With this result, we classify the numbers of the hits in the Japanese blogosphere into the three ranges, i.e., 1,000 $\sim$ 10,000, 10,000 $\sim$ 500,000, and over 500,000 [1].

4.2 Sample Wikipedia Entries for Evaluation

In the strategy of sampling Wikipedia entries for evaluation, we take an approach of selecting Wikipedia entries through Wikipedia categories. Here, we prefer sample entries to be distributed over various categories, simply because future application of this work is to estimate the topic distribution in the Japanese blogosphere. The detailed procedure is given below: We first heuristically allocate each Wikipedia entry to three of the topmost 8 categories or the second topmost 700 categories. For each Wikipedia entry, we ranked categories according to ascending order of the distance (here, we use the number of edges) from the entry, and selected the topmost three categories. We then restrict the candidate categories as those with more than 50 entries allocated in this procedure. We randomly select 35 categories from those candidates. Finally, from each of the 35 categories, we selected 75 entries from the range of 1,000 $\sim$ 10,000 of the hits in the Japanese blogosphere, 168 from that of 10,000 $\sim$ 500,000, and 149 from that over 500,000, which amount to 392 in total.

4.3 Collecting Blog Feeds for Evaluation

For each of the sample Wikipedia entries for evaluation, we consider the title of the entry as the query for collecting blog feeds for evaluation. The details of how to retrieve Japanese blog feeds for evaluation given a query are in Kawaba et al. (2008) and Nakasaki et al. (2008). In order to collect candidates of blog feeds for a given query, we use existing Web search engine APIs, which return a ranked list of blog posts, given a topic keyword. We use the Japanese search engine “Yahoo! Japan” API[2], where blog hosts are limited to major 11 ones[3]. We re-rank the list of blog feeds according to the number of hits of the query keyword in each blog feed. In terms of the criterion presented in section 3, Kawaba et al. (2008) and Nakasaki et al. (2008) reported that the procedure above outperformed the original ranking returned by “Yahoo! Japan” API, although the evaluation is small scale.

It is important to note here that future application of this work is not to generally detect the topic of a given blog feed, but to estimate the topic distribution in the Japanese blogosphere. In this application, it is more important to judge whether there exist blog feeds to be linked from a given Wikipedia entry. According to our preliminary analysis, whether there exist blog feeds to be linked from a given Wikipedia entry can be mostly judged considering the top-ranked 20 blog feeds returned by the procedure above.

4.4 Results of the Analysis

For each of the 392 entries selected in section 4.2, 20 blog feeds are collected according to the procedure of the previous section, and their relevance are manually judged based on the criterion

[1] Out of the titles of the whole 305,986 Wikipedia entries, about 8% (24,075 entries) are with the number of hits as zero, about 56% (172,471 entries) are with 1 $\sim$ 1,000, about 21% (63,835 entries) are with 1,000 $\sim$ 10,000, about 14% (40,852 entries) are with 10,000 $\sim$ 500,000, and about 1% (4,753 entries) are with those over 500,000.

[2] `http://www.yahoo.co.jp/` (in Japanese)

[3] `FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, jugem.jp, yaplog.jp, webry.info.jp, hatena.ne.jp`

Table 1: Manual Analysis on Existence of Blog Feeds to be Linked from a Wikipedia Entry (with top ranked 20 feeds): Criterion

Label	Description
C1	20∼10 feeds relevant to the given entry.
C2	9∼5 feeds relevant to the given entry.
C3	4∼1 feed(s) relevant to the given entry.
HU	At least one feed is relevant to an immediate ascendant concept of the given entry.
HL	At least one feed is relevant to an immediate descendant concept of the given entry.
HR	At least one feed is relevant to a related concept of the given entry.
E	None of above.

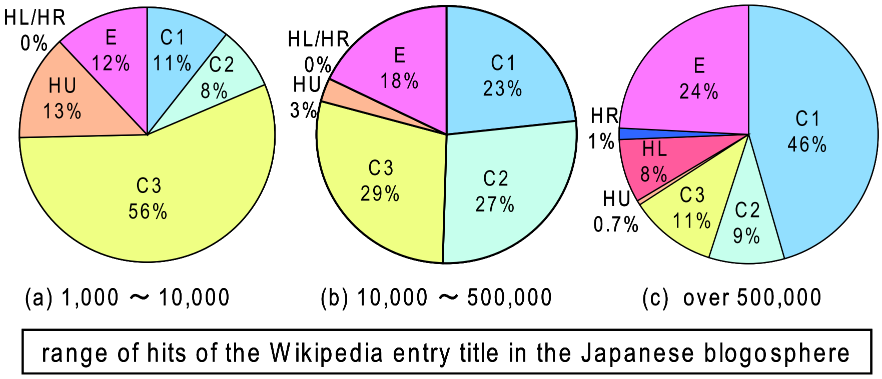

Figure 1: Manual Analysis on Existence of Blog Feeds to be Linked from a Wikipedia Entry: Distribution per Range of Hits of the Wikipedia Entry Titles

presented in section 3. Then, based on the judgments of 20 blog feeds, each entry is assigned one of 7 labels listed in Table 1. Final statistics is shown in Figure 1, where the distribution of those 7 labels is given for each of the three ranges.

As can be seen from this result, it is clear that entries with the number of hits over 10,000 tend to have many relevant blog feeds in the Japanese blogosphere. Furthermore, for the entries with the number of hits around over 500,000, some of the blog feeds linked from them are judged to be more relevant to an immediate descendant concept of those entries. Examples of the entries with the number of hits around 10,000 ∼ 500,000 which have relevant blog feeds are *"kitchen garbage"*, *"adoption"*, *"department store's basement food floor"*, and *"guide dog"*. On the other hand, entries with those around 1,000 ∼ 10,000 tend to have relatively small number of relevant blog feeds. Examples of those entries are usually very specific ones such as *"Xenopus"*, which is a genus of highly aquatic frogs native to Sub-Saharan Africa. Finally, as for examples of entries with the number of hits over 500,000, entries such as *"piano"* have many relevant blog feeds.

Table 2: Features of Linking Wikipedia Entries to Blog Feeds

Feature Label	Description
t-hits	hits of the given Wikipedia entry title, where the number of hits is counted in the blog feed
H-hits	sum of the hits of all the related terms in the range of hits over 500,000, where the numbers of hits are counted in the blog feed
M-hits	sum of the hits of all the related terms in the range of hits 10,000 $\sim$ 500,000, where the numbers of hits are counted in the blog feed
L-hits	sum of the hits of all the related terms in the range of hits below 10,000, where the numbers of hits are counted in the blog feed
H-num-b	the number of the related terms in the range of hits over 500,000, which are observed in the blog feed
M-num-b	the number of the related terms in the range of hits 10,000 $\sim$ 500,000, which are observed in the blog feed
L-num-b	the number of the related terms in the range of hits below 10,000, which are observed in the blog feed
all-num-b	the number of all the related terms, which are observed in the blog feed
H-num-w	the number of the related terms in the range of hits over 500,000, which are collected from the given Wikipedia entry
M-num-w	the number of the related terms in the range of hits 10,000 $\sim$ 500,000, which are collected from the given Wikipedia entry
L-num-w	the number of the related terms in the range of hits below 10,000, which are collected from the given Wikipedia entry
all-num-w	the number of all the related terms, which are collected from the given Wikipedia entry
char-len	character length of the given Wikipedia entry title
1-char	true if the given Wikipedia entry title is one kanji (Chinese character) word

5 Linking Wikipedia Entries to Blog Feeds by SVM

We apply Support Vector Machines (SVMs) (Vapnik, 1998) to the task of judging whether a blog feed is relevant to a given Wikipedia entry[4]. In this task, we train three classifiers, where the first one is applicable to all the Wikipedia entries with the hits 1,000 $\sim$ 10,000, the second is applicable to all the Wikipedia entries with the hits 10,000 $\sim$ 500,000, and the third is applicable to all the Wikipedia entries with the hits over 500,000.

Features employed in this paper are summarized in Table 2, where all of them are independent of specific Wikipedia entries. Thus, the classifiers trained with those features are applicable to arbitrary Wikipedia entries.

Most features are based on terms that are closely related to the given Wikipedia entry, and are automatically extracted from the body text of the given entry. The related terms we extract from the body text of the given entry can be categorized as follows: 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. From each entry, we extracted 15 related terms on the average. Then, we further classify those related terms into three ranges according to the number of hits in the Japanese blogosphere. This is simply from the observation in section 4 on the correlation between the number of hits of a Wikipedia entry title and existence of blog feeds to be linked from the entry.

For each of the features ‘t-hits’, ‘H-hits’, ‘M-hits’, and ‘L-hits’, sum of the hits are classified into 5 ranges, and each of the 5 ranges is represented as a binary feature, whose value indicates whether the sum of the hits is within the corresponding range.

[4] Details of the procedure and evaluation results are in Kawaba et al. (2009).

6 Binary Judgment on Existence of Blog Feeds to be Linked from an Wikipedia Entry

Based on the relevance judgment results given by the SVMs classifier, this section reports analysis on the task of judging whether, for each Wikipedia entry, there exist blog feeds to be linked from the entry.

6.1 The Procedure

In the evaluation of this section, for each entry, we use all of the 20 collected blog feeds. This means that, after the 10-fold cross-validation evaluation of the previous section, we apply each of the 10 trained SVMs classifiers to those ignored blog feeds with the class c as "$c = -$". We regard the distance from the separating hyperplane to each test blog feed as a confidence measure. We then introduce a lower bound of the distance from the separating hyperplane to each test blog feed, where blog feeds with this distance smaller than the lower bound are rejected.

Based on the relevance judgment results given by the SVMs classifier, each entry is assigned one of the labels below according to the following procedure. Suppose that a Wikipedia entry as well as the 20 blog feeds for evaluation are given, then, i) judge the entry as 'C1' if more than or equal to 10 blog feeds are judged as relevant, ii) judge the entry as 'C2' if from five to nine blog feeds are judged as relevant, iii) judge the entry as 'C3' if from one to four blog feed(s) are judged as relevant, iv) otherwise judge the entry as having no relevant blog feed.

Next, as for the judgment on whether, for each Wikipedia entry, there exist blog feeds to be linked from the entry, we consider the following case:

> "Blog Feeds Exist = C1,C2":
> for each entry, if both the manual evaluation and the system judgment above are 'C1' or 'C2', then the judgment by the system is correct for the entry. This criterion is especially intended to prefer entries to which sufficient number of blog feeds can be linked.

Finally, we employ the following evaluation measures:

$$\text{precision} = \frac{\text{the number of the entries for which the judgment by the system is correct}}{\text{the number of the entries for which the system judges as ``Blog Feeds Exist''}}$$

$$\text{recall} = \frac{\text{the number of the entries for which the judgment by the system is correct}}{\text{the number of the entries for which manual evaluation is ``Blog Feeds Exist''}}$$

6.2 Results of the Analysis

Tables 3 summarizes the evaluation results for each of the three ranges. We evaluate all the combination of features for learning SVMs classifier described in the previous section as well as several values for lower bounds of confidence, and then pick up results with maximum F-measure and maximum precision with F-measure at least around 40%. Here, the performance when regarding entries with the number of hits over 10,000 as "Blog Feeds Exist" are shown as baseline. Compared to the baseline, our proposed method achieves much improvement in precision than in F-measure. We achieved precisions over 90% and F-measures around 60% or more with lower bound of confidence as certain values. We are now working on applying machine learning framework to the task of binary judgment on existence of blog feeds to be linked from an Wikipedia entry.

Table 3: Binary Judgment on Existence of Blog Feeds to be Linked from an Wikipedia Entry: Evaluation Results (%)

(a) range of hits of the Wikipedia entry title: 1,000 ∼ 10,000

condition	feature set	precision / recall / F-measure of judgment as “Blog Feeds Exist”
baseline (“Blog Feeds Exist” = hits of the Wikipedia entry title over 10,000)	—	0/0/0
maximum F-measure (lower bound of confidence as 0.5)	t/H/M/L-hits + H/M/L-num-b + H/M/L-num-w + char-len	44.4/85.7/**58.5**
maximum precision (lower bound of confidence as 1.9)	t/H/M/L-hits + H/M/L-num-b + all-num-w	**100.0**/28.6/44.5
(maximum F-measure/maximum precision, without lower bound of confidence)	H-hits or t-hits + M-hits + H/M-num-b	35.5/78.6/48.9

(b) range of hits of the Wikipedia entry title: 10,000 ∼ 500,000

condition	feature set	precision / recall / F-measure of judgment as “Blog Feeds Exist”
baseline (“Blog Feeds Exist” = hits of the Wikipedia entry title over 10,000)	—	50.6/100.0/67.2
maximum F-measure (lower bound of confidence as 0.1)	t/H/M/L-hits + H/M/L-num-b + H/M/L-num-w + char-len	89.6/65.2/**75.5**
(maximum F-measure, without lower bound of confidence)		88.2/65.2/75.0
maximum precision (lower bound of confidence as 1.1)	t/H/M/L-hits + H/M/L-num-b	**100.0**/31.5/47.9
(maximum precision, without lower bound of confidence)	t-hits + L-num-b	94.1/34.8/50.8

(c) range of hits of the Wikipedia entry title: over 500,000

condition	feature set	precision / recall / F-measure of judgment as “Blog Feeds Exist”
baseline (“Blog Feeds Exist” = hits of the Wikipedia entry title over 10,000)	—	55.0/100.0/71.0
maximum F-measure/maximum precision (without lower bound of confidence)	t/H/M/L-hits + H/M/L-num-b + H/M/L-num-w + 1-char	79.0/77.1/**78.0**
maximum precision (lower bound of confidence as 1.4)		**93.3**/33.7/49.5

7 Related Works

The blog distillation task (Macdonald et al., 2007) in TREC 2007 blog track is related to the task we examine in this paper. Among the participants of the task, the best performing system was Elsas et al. (2007), which employs query expansion using hyperlinks in Wikipedia. Compared to this approach, we integrate more information other than hyperlinks, such as bold-faced terms and the title of a *redirect*, where those terms are carefully distinguished in terms of their numbers of hits in the Japanese blogosphere. Previous works on text classification (e.g., Wang and Domeniconi (2008)) as well as text clustering (e.g., Hu et al. (2009)) using Wikipedia knowledge are also based on techniques which extract related terms such as hyponyms, synonyms, and associated terms. Among them, text classification using Wikipedia knowledge by Wang and Domeniconi (2008) apply machine learning techniques to the task of classifying documents into certain number of classes, where Wikipedia knowledge are used as features. Major differences between our work and those works are: i) first of all, the underlying purpose of our work is to estimate topic distribution in the blogosphere, where Wikipedia is used as the topic hierarchy, ii) the overall frameworks of evaluation differ. In our evaluation, given a Wikipedia entry as a sample topic, candidate blog feeds are collected and their relevance to the given topic are examined through the proposed technique. Other related works include Mihalcea and Csomai (2007) which studied how to link important keywords in a document to an appropriate Wikipedia entry.

8 Concluding Remarks

This paper studied the issue of conceptually indexing the blogosphere through the whole hierarchy of Wikipedia entries. Based on the results of judging whether each blog feed is relevant to a given Wikipedia entry, this paper proposed how to judge whether there exist blog feeds to be linked from the given entry. In our experimental evaluation, we achieved over 90% precision in this task. Future work includes scaling up empirical evaluation throughout the whole Wikipedia entries whose titles are with the number of hits over 1,000. It is also important to invent techniques for discovering blog feeds which are not relevant to any of existing Wikipedia entries and thus are not properly indexed by the proposed method.

References

Elsas, J., J. Arguello, J. Callan and J. Carbonell. 2007. Retrieval and feedback models for blog distillation. In *Proceedings of the TREC (Text REtrieval Conference) 2007 (Notebook)*, 170–175.

Glance, N., M. Hurst and T. Tomokiyo. 2004. BlogPulse: Automated Trend Discovery for Weblogs. In *Proceedings of the WWW (International World Wide Web Conference) 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Hu, X., X. Zhang, C. Lu, E. K. Park and X. Zhou. 2009. Exploiting Wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 389–396.

Kawaba, M., H. Nakasaki, T. Utsuro and T. Fukuhara. 2008. Cross-lingual blog analysis based on multilingual blog distillation from multilingual Wikipedia entries. In *Proceedings of International Conference on Weblogs and Social Media*, 200–201.

Kawaba, M., D. Yokomoto, H. Nakasaki, T. Utsuro and T. Fukuhara. 2009. Linking Wikipedia entries to blog feeds by machine learning. In *Proceedings of the 3rd International Universal Communication Symposium*.

Macdonald, C., I. Ounis and I. Soboroff. 2007. Overview of the TREC-2007 blog track. In *Proceedings of the TREC (Text REtrieval Conference) 2007 (Notebook)*, 31–43.

Mihalcea, R. and A. Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 233–242.

Nakasaki, H., M. Kawaba, T. Utsuro, T. Fukuhara, H. Nakagawa and N. Kando. 2008. Cross-lingual blog analysis by cross-lingual comparison of characteristic terms and blog posts. In *Proceedings of the 2nd International Symposium on Universal Communication*, 105–112.

Vapnik, V. N. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Wang, P. and C. Domeniconi. 2008. Building semantic kernels for text classification using Wikipedia. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 713–721.