Clustering blog entries based on the hybrid document model enhanced by the extended anchor texts and co-referencing links[1]

Hiroshi Ishikawa, Masashi Tsuchida, Hajime Takekawa
Graduate School of Science and Technology, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Japan

Abstract. In this paper, we propose a document vector space model where weights of noun terms vary depending on positions within the texts of blog entries as search results. We extend "extended anchor texts" (i.e., extra texts surrounding anchor texts) with the exponential potential such that the weight of a noun term decreases exponentially as the distance between the term and link increases. In order to cluster blog entries as search results, we use the hybrid vector space model which takes into account both texts including extended anchor texts and co-references of Web pages through links described in blog entries. We evaluate the effects of our scheme on clustering blog search results.

Keywords: Clustering, blogspace, extended anchor text, vector space model, link analysis

1 Introduction

As a phenomenon related to Web 2.0, a tremendous number of blogs have emerged each day. One of the salient features is the explosive growth of recent blog population. As blog entries are more frequently updated and more individual opinions can be read in real-time than general Web pages, blogs are increasingly important sources of information. The purposes of reading blogs range from individual interests to mining and "word of mouth" marketing.

Due to these situations, there is an urgent demand for technologies for gaining accurate information from a large scale of information sources rapidly. Most blog searches such as Technorati [Technorati 2008], display the search result set linearly by sorting the elements based on certain kinds of ranks such as relevance. As we live in the age of information explosion, such methods alone are insufficient to rapidly access necessary information.

As a solution to this problem, clustering a search result set based on similarities of the elements is considered as promising. Assuming that an appropriate label is automatically assigned to each cluster, the user is expected to more rapidly reach the necessary information by accessing only the clusters relevant to the purpose of the search.

Some Web (meta) search engines, such as Clusty [Clusty 2008] not only display the result set linearly but also display the folders of elements dynamically generated from the result of preceding clustering. Clusty allows the user to cluster the blog entries returned by dedicated blog search although the deployed technologies are not deeply described.

[1] This work was partially supported by KAKENHI (B) (19300026).

Generally, as the author of Web pages deliberately describes links to other Web pages, those pages are expected to be strongly related to each other. It is possible to increase the efficiency of clustering Web pages by analyzing link structures. It is possible to cluster elements based on their similarities, which inside use either texts or links, or both of them. Further, words in particular positions are often thought of more highly than those in other positions. Anchor texts associated with hyperlinks and title texts are such typical examples.

In this paper, in order to effectively cluster blogspace (i.e., a set of blog entries relevant to topics), we propose a document vector space model where the weights of terms vary depending on positions within the texts. We define “extended anchor texts” by extra texts surrounding anchor texts and extend them with the exponential potential such that the weight of a term decreases exponentially as the distance between the term and link increases. In reality, we use the hybrid vector space model which takes both texts including extended anchor texts and links within blog entries into account. In order to validate our approach as a preliminary work, we evaluate the effectiveness of our essential scheme.

We compare our work with related works. Shen et al. [Shen et al. 2006] proposed and utilized the extended anchor text model, which by their definitions consists of just a fixed number of words containing anchor texts, equally weighted. They used the models to classify Web pages, but not blog entries. Zhu et al. [Zhu et al. 2004] used a similar extended anchor model to automatically generate cluster labels in clustering generic Web pages.

2 Hybrid vector space model

We describe our hybrid vector space model deployed in clustering blog entries, which consists of texts and links. We have to take the following salient features of blog entries into account. Links contained by blog entries include hyperlinks, trackbacks, and comments. As the most common ways, hyperlinks link a page to another page by using the anchor tags as “ *texts* ”. Texts surrounded by the corresponding anchor tags are called anchor texts. The anchor texts and the contents of hyperlink destinations are assumed to be highly similar. However, the origins of hyperlink jumps can be disclosed by analyzing Web access logs. For the reasons such as the purpose to conceal the existence of the hyperlink from the link destinations and the preference of the blog writers, urls are sometimes written directly in blog entries as a part of comments like “HP is *url*". We call such pseudo links direct links, where the sentences expressing the contents of the link destinations equivalent to anchor texts are not directly specified. We also consider texts surrounding such direct links as one kind of extended anchor texts deployed in clustering.

We use the hybrid vector space model consisting of link and document (text) vector space models in order to cluster blog entries. We describe the general flow of processes realizing our approach. First we make the link vector space model by focusing on co-references of Web pages through links. Next we make the document vector space model of blog entries with respect to the texts and weight it with our extended anchor text scheme. Then we hierarchically cluster the blog entries based on the similarities. When we merge two clusters, we use the furthest entries belonging to either of the clusters in order to measure the similarity between the two clusters. We use the hierarchical agglomerative clustering method with the prescribed cutoff threshold to avoid only one big cluster.

We collectively call hyperlinks and direct links explicit links. We describe the link

vector space model based on such explicit links except image links because the images are rarely co-referenced and inappropriate for measuring similarities based on them.

First we extract all explicit links from each blog entry. The total number of distinct links throughout all blog entries is denoted as n. The counts of the url L_p appearing in the blog entry D_i is the component W_{ip} of the matrix corresponding to the link vector space. We cluster blog entries hierarchically based on the cosine similarity measure as follows:

$$sim(D_i, D_j) = \frac{W_{i1}W_{j1} + ... + W_{in}W_{jn}}{\sqrt{W_{i1}^2 + ... + W_{in}^2} \times \sqrt{W_{j1}^2 + ... + W_{jn}^2}} \quad (n \geq 1) ...(1)$$

This similarity measure also can apply to the document vector space model as described later. In order to construct the document vector space model, we extract only nouns except numerals and pronouns from blog entries and titles by using Japanese morphological analyzer. The varieties of verbs and adjectives, which we exclude, are not so rich; they are not so helpful to distinguish each entry. The total number of distinct nouns throughout all blog entries is denoted as n. The component W_{ik} of the document vector space model is calculated by combining tf_{ik} for frequencies of the noun term t_k appearing in the blog entry D_i and df_k for the counts of documents containing the noun term t_k as follows:

$$W_{ik} = tf_{ik} \times \left(\log \frac{n}{df_k} + 1 \right) \quad (1 \leq k \leq n) ...(2)$$

Blog titles are considered as the most concise abstracts of the blog entries. In general, the nouns contained by the blog titles are more highly weighted than those contained by the blog entries (bodies). The similarity can be measured by using the formula (1). In order to avoid the weighting of the insignificant words (i.e., stop words) such as “today”, “previously”, we don’t weight noun terms having tf*idf equal to or less than the average value of tf*idf even if they are included in extended anchor texts.

3 Extended anchor texts

We describe the general concept of our extended anchor text. We pay our attention to both texts surrounding so-called anchor texts and texts surrounding direct links. We don’t exclude image links for this time. For the blog writers, the information contained by the images is assumed to be important as well. Images are very important as sources of information as well. That is, terms surrounding image links are also more highly weighted.

Some blog writers describe explicit links around the beginning of the blog entries while others describe them in other positions such as the middle or end. We will explain how to determine the positions at the end of this section. We assume that the extended anchor texts follow the links at the beginning and precede the links at the other positions. We measure the distance forward and backward from the origin according to the positions, respectively.

We weight noun terms in the extended anchor texts. Prior to determining the weighting scheme, we measured how many noun terms averagely constitute one sentence. We randomly selected 250 sentences from a set of blog entries. The number of noun terms varies from 0 to 15. One sentence on the average contains 3.5 noun terms.

We assume that the origin is the position where the url is described. The unit of measuring the distance of the noun term is one noun term. The *ith* noun term from the link

(the origin) has the distance *i-1(i>0)*. From the preparatory experiments, we assume that one sentence consists of three noun terms although a triplet of nouns does not always correspond to the same sentence. We weight noun terms sentence by sentence. Three noun terms within d_i corresponding to *ith* statement has the distances (3i-3,3i-2,3i-1). Weights g_i for noun terms within d_i are defined using the weight g for noun terms within d_1 as follows:

$$G = \{g_1, g_2, ..., g_M\} = \{g, \frac{g}{2}, ..., \frac{g}{2^{M-1}}\} \quad (M \geq 1)$$

How to determine g will be explained later. We use the exponentially decreasing potential in order to more strongly emphasize the effects of the distance than the linearly decreasing potential. On the other hand, traditional anchor texts and other extended anchor texts are regarded as constantly weighted independent of the distance from the origin.

We have analyzed the blog entries in order to determine the directions of weighting such as forward and backward from the origin (i.e., explicit link). We classify the origins of url into the beginning part, middle part, and ending part. After that, the anchor texts or explicit links are described. On the other hand, the blog writers describe sources of information and related Web pages collectively in the ending part. In this case, the titles and the concise descriptions of the Web pages are described in the anchor texts or followed by direct links. In the middle part, related noun terms precede anchor texts or explicit links. We assume the directions as follows: The origins precede extended anchor texts in the beginning part. The origins follow extended anchor texts in the middle or ending part.

We have done some experiments in order to determine the beginning part. We have counted 379913 noun terms in 9117 blog entries. One entry contains 41.67 noun terms on the average. As one sentence contains 3.5 noun terms on the average, one entry contains 11.9 statements on the average. By the prior analysis, urls appear around the second sentence on the average for the beginning part. Considering the length of individual blog entries, we decide the noun terms appearing until $C_i/5$ as the beginning part, C_i being the noun counts in the entry D_i. If urls are detected in the beginning part, the weighting direction is forward, otherwise backward.

4 Evaluation framework and hybrid vector space model

Here we describe the general framework for evaluating our approach through experiments. We have prepared a dataset of 37279 Japanese blog entries for the whole experiments, collected through crawling “Livedoor Blog” [Livedoor Blog 2008] and randomly selected blog entries co-referencing specific Web news articles. We use this dataset in order to construct the hybrid vector space model. Then we compare three clustering approaches including our proposed approach in order to evaluate the effectiveness of our approach.

As a preliminary work, our experiments focus on the following objectives: (1) Analysis of explicit links and construction of link vector space model, (2) Construction of the extended anchor text model with the exponential potential, (3) Validation of the relevance of the extended anchor texts to the feature terms of the entries, and (4) Validation of the effectiveness of the extended anchor texts and links in clustering

We have analyzed the dataset and have found that 4775 blog entries, about 13 % of the

whole dataset, have explicit links (i.e., urls) and distinct 10790 urls appear for 23693 times in all. Around 80 % of distinct urls, 8035 urls are referenced only once. Only one entry references more than one Web page. The number of urls co-referenced by 2 to 10 entries is 2714 in all. That by 11 to 30 entries is 28; that by 31 to 147 (largest) entries is 13.

According to the analysis, most of urls are not co-referenced. Almost all urls co-referenced by more than and equal to 8 entries are either of shopping sites, auction sites, or application services embedded in the blog entries. The blog entries referencing “8 or more times co-referenced” urls can be categorized as advertisement- or affiliate-oriented blog entries. Those referencing “2 to 7 times co-referenced” urls are expected as the most promising as valuable information sources. This criterion is just the first approximation requiring elaboration; the threshold 7 or 8 here will change depending on the dataset.

We construct the link vector space model based on the urls extracted through the above analysis and cluster 4775 blog entries containing urls by using the link vector space model. If two clusters have the positive value as the similarity with respect to the urls, those two clusters are merged into one cluster. As the result, 3346 clusters are made. Among them, 390 clusters have co-references of urls and the remaining 2956 clusters have no co-references. Only one cluster has 147 blog entries, which co-reference advertizing sites and can be categorized as blog entries with commercial objects.

In order to construct the document vector space model, we also used the same dataset of consisting of 37279 blog entries as we used when we constructed the link vector space model. By focusing on entries containing a particular word, we evaluate the effectiveness of clustering based on our approach. As for the extended anchor texts, we consider the two nearest statements from the origin as the range of the extended anchor texts and we merge two clusters only if the similarity between them is equal to or larger than 0.4. We use the value 4.0, calculated by the following formula (3), as the weight g for the extended anchor texts. N is the total number of the entries and D_m denotes the entry m $(1<=m<=N)$.

$$g = \frac{\sum_{m}^{N}\{Max_{D_m}(tf \times idf) - Avg_{D_m}(tf \times idf)\}}{N} \quad ...(3)$$

5 Evaluation of relevance of extended anchor texts

We will validate the relevance of our extended anchor texts with respect to the topics by checking whether noun terms related to the topics are successfully included by the extended anchor texts. We actually used three topics for the experiments, but we describe only the topic ”natto” (fermented soybeans) due to the limited space. The dataset contains references to Web pages related to these topics. As a background story, the broadcaster#2 aired the TV program, the production of the broadcaster#1, insisting that just taking natto every morning is an effective diet, which influenced consumers but was later found to be a frame-up.

We extract only blog entries dealing with “natto”, “diet”, “frame-up” and then use 16 entries referencing information sources related to this topic. We expect to extract the noun terms deeply associated with natto and frame-up from the extended anchor texts in the entries. Half of the ten entries contain explicit links and half contain normal anchor texts.

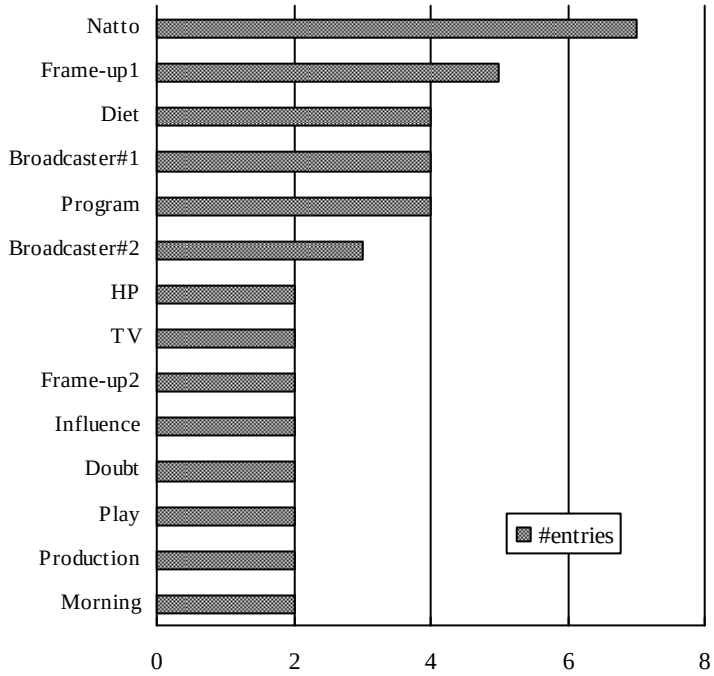

Figure 1. The number of entries with a specific term.

Among them, the noun terms related to this topic and those less related the topic appeared 63 and 103 times, respectively, in the extended anchor texts. Figure 1 illustrates the noun terms weighted in the extended anchor texts of at least two entries. We found that 63 (total, 166) noun terms could be extracted as related to the topic according to the definition of our extended anchor texts, especially from the beginning and ending parts.

In case of the explicit links described in the middle part where we only weight the extended anchor texts backward from the origin (explicit links), we sometimes failed to extract related noun terms although there are a lot of successful cases. So we have to refine or change our current rule of weighting the extended anchor texts. We think that it is possible to remedy this problem if we divide the whole blog entry into fragments by using some information such as line breaks and apply to such fragments our current rule of extracting noun terms from the extended anchor texts.

6 Comparison of clustering methods

In order to validate the effectiveness of our approach to clustering based on the hybrid vector space model enhanced by the extended anchor texts and co-referencing links, we compare the following three clustering methods: (Method1) Clustering based on document (text) vector space model, (Method2) Clustering based on both document- and link-vector space model, and (Method3, our approach) Clustering based on both document vector space model enhanced by our extended anchor text model and link vector space model.

We made experiments on clustering a set of blog entries containing specific noun terms

Table 1. The feature terms of clusters by the clustering method.

Method	Cluster	Feature terms
1	I	Natto Diet Broadcast Frame-up Dictionary
	II	Frame-up Cholesterol Natto Amino Broadcater1
2	III	Natto Frame-up Diet Broadcast Cholesterol
3	IV	Natto Broadcast Morning/Evening Dictionary Diet
	V	Natto Frame-up Cholesterol Diet Doubt

(i.e., search term) selected from the 37279 entry dataset. We actually did three sets of experiments by using three topics but we explain only about the ”natto” case as the others showed coherent results. We used “natto” as the search term for this clustering experiment deploying Method 1, Method 2, and Method 3. An article in a major news site reported that a frame-up of effective diet with natto by a certain broadcaster came to light. In advanced we found 5 entries co-referencing and discussing this topic. We name these “Co-A”, “Co-B”, “Co-C”, “Co-D”, and “Co-E”. There were 172 entries containing “natto”. We obtained 143, 137, and 138 clusters by Method 1, Method 2, and Method 3, respectively.

First, we focus on the cluster with 13 entries containing noun terms such as “diet” and “frame-up”, constructed by Method 1. We name this cluster “Cluster I”. We summed the tf*idf for each noun term and sorted them in descending order. We take the top 5 noun terms as feature terms for each cluster such as Cluster I (See Table 1). The entries Co-A, Co-B, Co-C belong to Cluster II as another cluster by Method 1. The entry Co-D belongs to Cluster I. Co-E belongs to another cluster. As a result of analysis of Cluster I, among 13 entries, only three entries were found related to the term “frame-up”. Two entries describe “natto diet”. Six entries describe “shortage of natto”. The rest is on the other topics.

As a result of clustering by Method 2, Cluster I and Cluster II were merged in Cluster III with 17 entries. “Frame-up” appeared again in the feature terms for the cluster. However, the topics are a mixture of “frame-up” and “shortage” (See Figure 2). As a result of clustering by Method 3, entries describing “shortage of natto” and ”natto diet” moved from Cluster I to Cluster IV. Entries about “shortage of natto” moved from other clusters than Cluster I to Cluster IV. The size of Cluster IV was 11. In addition, Co-A, Co-B, Co-C, and Co-D, entries about “frame-up” belong to Cluster V. The size of Cluster V is 6. Cluster V contains another entry about “frame-up”, originally belonging to Cluster I and Co-E, describing opinions about the news plainly. In summary, Method 3 more sharply separated topics and produced fewer mixtures of topics than Method 1 and Method 2.

In this experiment, our proposed method (Method 3) performed the best clustering in comparison to the other methods. We think that even if similarities based on co-references are not so large, the method by the extended anchor texts can compensate them.

The same author sometimes links the specific pages not related to the topic in question,

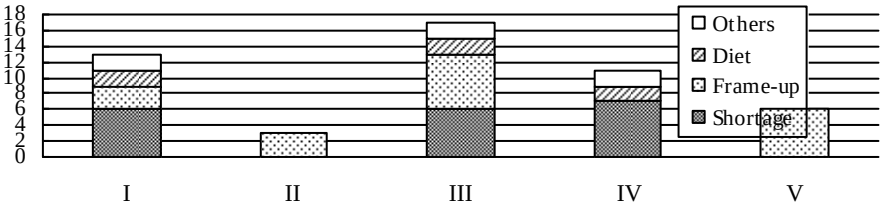

Figure 2. The number of entries with topics costituting clusters.

such as blog ranking and the author's own pages. In order to avoid the extracting of such "spam links" as co-references, we have to take special measures in dealing with explicit links in the blog entries written by the same author. While there may be errors, if we delete such spam blogs as commercial blogs and news blogs, the clustering method enhanced with the link vector space model is expected to allow clustering of entries describing more of the authors' opinions of the topics than of the description itself of the topics because such entries tend to omit the description by just linking Web pages as references.

In summary, the major advantage of deployment of co-referencing links in clustering is the possibility of clustering entries with explicit links, which describe fewer details about the topics and express more opinions about them. However, this approach may lead to mixtures of topics as a result of wrong clustering caused by spam links to non-relevant pages such as ranking pages and the authors' own pages. Especially, our extended anchor texts model is effective in labeling clusters with the extracted feature terms after clustering.

7 Conclusions

In this paper, we proposed the extended anchor texts model enhanced by exponential weighting. We also proposed blogspace clustering based on hybrid of document vector space model with extended anchor texts and link vector space model. Through the experiments, we have validated that it is possible to extract feature terms more relevant to the topic by using our extended anchor texts. We have found that use of co-referencing links is effective to blogspace clustering. The experiments have verified that the proposed scheme is at least effective for the dataset. Although some specific constants, such as three as the average number of nouns for one sentence, may depend on the dataset or the Japanese language, our essential schemes are neutral with respect to the dataset and the language.

One of the remaining issues as a preliminary work is to improve the rule of weighting the extended anchor texts from explicit links described in the middle of the blog entries. As the size of the dataset we used in the experiments was not so large and the sizes of constructed clusters were relatively small, the scalability of our approach with respect to large-scale blogspace is another big issue. As a third issue, the explicit links include lot of affiliate links to advertizing and commercial sites. We have to explore how to delete the affiliate links in order to more effectively cluster blog entries. Cleansing noun terms in constructing the document vector space model is another issue.

References

[Clusty 2008] Clusty http://clusty.com/ Accessed 2008.
[Livedoor Blog 2008] Livedoor Blog http://blog.livedoor.com/ Accessed 2008.
[Shen et al. 2006] D. Shen, et al.: A Comparison of Implicit and Explicit Links for Web Page Classification, Proc. Intl. Conf. WWW, 643-650, 2006.
[Technorati 2008] Technorati http://www.technorati.com/ Accessed 2008.
[Zhu et al. 2004] J. Zhu, et al.: PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation, ACM Trans. Internet Technology,185-208, 2004.