Representative Entry Selection for Profiling Blogs

Jinfeng Zhuang, Steven C.H. Hoi, Aixin Sun
School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
{zhua0016,chhoi,axsun}@ntu.edu.sg

Rong Jin
Computer Sci. & Eng. Dept.
Michigan State University
East Lansing, MI 48824
rongjin@cse.msu.edu

ABSTRACT

Many applications on blog search and mining often meet the challenge of handling huge volume of blog data, in which one single blog could contain hundreds or even thousands of entries. We investigate novel techniques for profiling blogs by selecting a subset of representative entries for each blog. We propose two principles for guiding the entry selection task: *representativeness* and *diversity*. Further, we formulate the entry selection task into a combinatorial optimization problem and propose a greedy yet effective algorithm for finding a good approximate solution by exploiting the theory of submodular functions. We suggest blog classification for judging the performance of the proposed entry selection techniques and evaluate their performance on a real blog dataset, in which encouraging results were obtained.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting methods

General Terms

Algorithm, Experimentation

Keywords

Blog profiling, entry selection, blog classification

1. INTRODUCTION

In general, a blog contains a number of elements, including blog entries (or posts), tags, comments, links, and others. Among various elements, entries are the most important element for content analysis of a blog. To simplify the problem, in our study, each blog is treated as a set of entries or a sequence of entries. Given this entry set representation, some conventional web search and mining techniques can be applied for the blog search and ming applications by treating each entry as a web page. However, they often encounter the *efficiency* challenge since one single blog can contain hundreds or even thousands of entries. In addition, blogs often contain *noisy* entries that may degrade the performance of blog search and mining tasks. For example, some blog posts may be only a few sentimental words or just a single URL, which are irrelevant to the main theme of the blog. These motivate us to find some concise representation for blogs to enhance the efficiency and effectiveness of blog applications.

To this end, we introduce the problem of **blog entry selection** for profiling blogs with representative entries. For the rest of the paper, we first formally define the problem and then present our technique. Finally, we conduct an empirical study for assessing the performance of the proposed technique.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

2. REPRESENTATIVE ENTRY SELECTION

Formally, let B_i denote a blog, which is represented by a set of N_i entries, i.e., $B_i = \{B_{i1}, \ldots, B_{iN_i}\}$, where N_i is the total number of entries in B_i; $B_{ij} \in \mathbb{R}^n$ is the j-th entry of B_i and n is the number of dimensions of the blog entry vectors. The problem of blog entry selection is defined as:

DEFINITION 1 (ENTRY SELECTION). *Given a blog B_i of N_i entries and a predefined entry selection size m, an entry selection problem is to select a subset of entries $S_i \subseteq B_i$, where $|S_i| = \min\{m, N_i\}$, such that the selected entries S_i best represent blog B_i in a blog data mining task.*

To solve the entry selection task, we propose two principles as follows:

- **Representativeness**. An entry selection task should select the entries that are *representative* to the blog.
- **Diversity**. It should choose *diverse* entries to maximize the information of the selected entries

The first principle is to ensure that the selected entries are the most important and closest to the theme of the blog and the second principle is critical to avoiding the selection of *redundant* entries. Next we show our approach to formulate these two principles formally.

Given a blog B_i, we calculate the centroid of its associates entries as:

$$B_i^C = \frac{1}{N_i} \sum_{j=1}^{N_i} B_{ij} \tag{1}$$

Then we define the **representativeness** measure of a blog entry B_{ij} below:

$$r(B_{ij}; B_i) = sim(B_{ij}, B_i^C) \tag{2}$$

where $sim(\cdot, \cdot)$ is a similarity function, e.g., cosine similarity between two document vectors.

There are other possible methods for measuring representativeness. We employ clustering techniques to remove the

noisy entries and define the representativeness measure on the major clusters. Specifically, we first cluster the entries in the blog into k clusters and then treat the "small" clusters as outliers and drop them before the entry selection phase. In the next phase, for an entry $B_{ij} \in B_i$, we measure its representativeness below:

$$r(B_{ij}; B_i) = sim(B_{ij}, c_j^*) \times \frac{l(B_{ij})}{max_j\ l(B_{ij})} \quad (3)$$

where $l(B_{ij})$ is the number of distinct words in the entry B_{ij}.

The **diversity** of a given subset of selected entries S_i is measured by:

$$d(S_i) = \sum_{B_{ij}, B_{ik} \in S_i} d(B_{ij}, B_{ik}). \quad (4)$$

Similar measures are often adopted in other applications [1, 3]. Based on the above measures of representativeness and diversity, for a candidate subset of entries S_i, we define the quality evaluation function below:

$$f(S_i; B_i) = r(S_i; B_i) + \lambda d(S_i) \quad (5)$$

where the function $r(S_i; B_i)$ measures the *representativeness* of S_i with respect to the all entries in B_i, the function $d(S_i)$ measures the *diversity* among the entries in S_i, and λ is a parameter to balance the tradeoff between them.

The task of blog entry selection is then reduced to the issue of finding the optimal subset S_i for the following optimization:

$$\max_{|S_i|=k} \quad f(S_i; B_i), \quad (6)$$

where k is the number of entries to be selected. The above problem is a combinatorial optimization problem, which is often NP-hard to find the global optima.

To solve the optimization problem above, we explore the theory of submodular functions for seeking the sub-optimal solution. In particular, we show the following theorem.

THEOREM 1. *The objective function in (5) is a submodular function.*

Given a submodular function $f(\mathcal{S})$ and the related combinatorial optimization problem above, a straightforward solution is to employ a greedy approach (denoted by GES): we start an empty set for $\mathcal{S}$; in each iteration, we expand the set $\mathcal{S}$ with the element e that maximizes the difference $f(\mathcal{S} \cup e) - f(\mathcal{S})$. We keep on expanding $\mathcal{S}$ till the number of elements in $\mathcal{S}$ is k. The submodular theorem in [2] provides the performance guarantee for the greedy algorithm above for finding the sub-optimal solution.

3. EXPERIMENTAL RESULTS

In our experiment, we use blog classification to examine the performance of the proposed entry selection techniques. In our experiments, we crawled a dataset from the BlogFlux[1], which consists of 5,000 blogs with 840,150 entries in English. These blogs come from 10 popular categories and each blog belongs to one or more categories. For experimental evaluations, we partition the dataset into two parts: one half for training and the other half for test. Table 1 shows the statistics of the dataset.

[1]http://dir.blogflux.com

Table 1: The statistics of our experimental dataset

	Training Set	Test Set
# blogs	2,500	2,500
# entries	424,948	415,202

For performance measure, we use F_1 metric, which is widely adopted in text categorization tasks. For comparison, we employ *Macro*-F_1 and *Micro*-F_1 measures over 10 categories to evaluate different entry selection algorithms. To examine the performance of the proposed entry selection technique comprehensively, we compare several different schemes in our experiments:

(1) **random**: the baseline method by random sampling.
(2) **newest**: a method by sampling newly posted entries.
(3) **GES**$_c$: the proposed greedy ES algorithm with the *centroid* based representativeness.
(4) **GES**$_\ell$: the proposed greedy ES algorithm with the *cluster* based representativeness.
(5) **all**: simply use all entries without entry selection.

Table 2 shows the results of the entry selection technique when selecting 30 entries for each blog.

Table 2: Results for entry selection techniques

Method	random	newest	GES$_c$	GES$_\ell$	all
Micro-F_1	0.691	0.667	0.702	0.701	0.698
Macro-F_1	0.705	0.678	0.713	0.717	0.710

Several observations can be made from the results. First, the proposed algorithms GES$_c$ and GES$_\ell$ produce better results than the *random* approach. This result verifies the effectiveness of the proposed technique. Second, we observe that the classification performance of the selected entries by the proposed technique is better than the one using all entries when m is 30. One possible reason is that the subset of entries selected by the proposed technique contains less noise than the original blog and thus can result in better classification results. This result again validates the effectiveness and importance of the proposed entry selection technique.

4. CONCLUSION

This paper investigated a new research problem of profiling a blog by selecting a subset of representative blog entries. We formulated it into a combinational optimization problem based on two principles and proposed an algorithm to solve it by exploiting the theory of submodular functions. We have conducted experiments to examine the effectiveness of the proposed entry selection techniques by evaluating the blog classification performance. In future work, we will study more effective entry selection techniques and apply our techniques on other web applications.

5. REFERENCES

[1] X. Geng, T.-Y. Liu, T. Qin, and H. Li. Feature selection for ranking. In *SIGIR*, pages 407–414, 2007.

[2] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, pages 265–294, 1978.

[3] D. Xin, H. Cheng, X. Yan, and J. Han. Extracting redundancy-aware top-k patterns. In *SIGKDD*, pages 444–453, 2006.