Clustering Blogs with Collective Wisdom*

Nitin Agarwal, Magdiel Galan, Huan Liu, Shankar Subramanya
Computer Science and Engineering
Arizona State University
Tempe, AZ 85287
{Nitin.Agarwal.2, Magdiel.Galan, Huan.Liu, Shankara.Subramanya}@asu.edu

Abstract

Blogosphere is expanding in an unprecedented speed. A better understanding of the blogosphere can greatly facilitate the development of the Social Web to serve the needs of users, service providers and advertisers. One important task in this process is clustering blog sites. Clustering blog sites presents new challenges. We propose to tap into collective wisdom in clustering blog sites, present statistical and visual results, report findings, and suggest future work extending to many real-world applications.

1 Introduction

With an explosive growth of the blogosphere there is a need for automatic and dynamic organization of the blog sites in order to enhance the structured search and access to blog sites. Clustering of these blog sites is a promising way to achieve the automatic organization of the content. Blog site clustering not only helps better organize the information but also aids convenient accessibility to the content. Clustering blog sites helps in optimizing the search engine by reducing the search space. We only need to search the relevant cluster and not the entire blogosphere. Though there exist a good number of traditional clustering methods, they are not designed to consider the unique characteristics of the blogosphere. A prominent feature of the Social Web is that many bloggers voluntarily write, tag, and catalog their posts in order to reach the widest possible audience who will share their thoughts and appreciate their ideas. In the process a new kind of collective wisdom is generated. We propose to leverage the collective wisdom in clustering blogs.

The problem of clustering blog sites could be formally defined as, given m blog sites, $S_1, S_2, ..., S_m$, we construct k disjoint clusters of the m blog sites, such that $k \leq m$. We exploit the collective wisdom while forming clusters of these blog sites. The collective wisdom is available in the form of predefined labels for each blog site. A single blog site could be tagged under multiple labels.

With the proposed new framework for clustering we intend to explore new ways for clustering. We also show that the clusters thus obtained are more meaningful as compared to traditional ways for clustering. Moreover, conventional approaches for clustering have inherent shortcomings like the curse of dimensionality and sparsity [2], semantic similarity is not captured by the similarity measure very well [4], obtained clusters are sometimes not very meaningful [4], and number of clusters needs to be known *a priori*.

2 Proposed Approach - WisClus

We perform our study in a controlled environment by collecting data from a blog site directory available at BlogCatalog. We first briefly describe the blog site data available at BlogCatalog before discussing the proposed approach.

BlogCatalog (http://www.blogcatalog.com) is a blog directory allowing bloggers to label the blogs under a given hierarchy. The structure of this hierarchy keeps changing as more blogs are submitted to BlogCatalog, although in a controlled fashion. At the time of writing, BlogCatalog had in total 56 level-1 nodes (or labels). The maximum depth of the hierarchy is 3. Later we experiment with varying granularity of structural information.

We leverage the label information to cluster the blog sites. A naïve way could be to treat all the blog sites that have the same label as one cluster resulting in too many clusters and moreover some of the clusters thus obtained might be related and would be better if they are merged into one cluster. Also many blog sites are tagged under more than one labels, which makes it difficult to form clusters in the naïve way. To achieve this, we cluster similar labels.

Clustering the similar labels can be formulated as an optimization problem. Assume we have t labels, $l_1, l_2, ..., l_t$ and are clustered into k clusters, $C_1, C_2, ..., C_k$, then opti-

*This work is in part supported by AFOSR and ONR grants to the third author.

978-0-7695-3261-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ICWE.2008.9

mal clustering is obtained if, for any two labels l_i and l_j,

$$\min \sum d(l_i, l_j), \forall (l_i, l_j) \in C_m, 1 \leq m \leq k, i \neq j \quad (1)$$

$$\max \sum d(l_i, l_j), \forall l_i \in C_m, \forall l_j \in C_n, 1 \leq (m, n) \leq k, m \neq n \quad (2)$$

Here $d(l_i, l_j)$ refers to a distance metric between the labels l_i and l_j. (1) minimizes the within-cluster distance between the cluster members and (2) maximizes the between-cluster distance. The optimal solution for the above min-max conditions is NP-complete because of the combinatoric nature of the problem [3].

Other approaches like K-means requires the value of k *a priori*, which is difficult to specify. Another approach to cluster the blog sites is based on the tags assigned to the blog posts and the blog site. Each blog site can be profiled based on these accumulated tags. A simple cosine similarity distance metric could be used to find similarity between different blog sites. However, the vector space model of the blog sites based on the tags is high-dimensional and sparse. We use a SVD based clustering algorithm as the baseline to avoid the curse of dimensionality. We chose top 25 eigenvectors to transform the blogs to reduced concept space. Pairwise similarity between blogs was computed using cosine similarity between reduced concept space vectors of the blogs. More details could be found in [1].

Based on the above discussion and limitations with the vector space model, we propose an approach to achieve blog site clustering leveraging the "collective wisdom" of the bloggers. Often bloggers specify more than one predefined labels for a particular blog site. Such blog sites help in establishing links between these labels. This results in a *label relation graph*. For example, labels like `Computers` and `Technology`; `Computers` and `Internet`; `Computers` and `Blogging` were linked by the bloggers. The category labels are the vertices of the label relation graph. The number of blog sites that create the links between various labels is termed as *link strength*, which could be treated as the edge weights of the label relation graph. Note that here collective wisdom is used to construct the label relation graph. Using this label relation graph, different labels can be clustered or merged. We call this link-based clustering, *WisClus*. We experiment with different thresholds for the link strength in Section 3.

WisClus clustering approach is highly time sensitive and adaptive to the current interests, since the labels of a blog site could change depending on what the blogger is blogging about. This results in dynamic as well as adaptive clustering, every time new blog posts appear, there will be new edges appearing in label relation graph or the link strength changes as blogger specifies different labels, the clustering results would change. Since the blogosphere provides more emphasis on the freshness of the content, the proposed clustering approach would also reflect similar dynamics. Also,

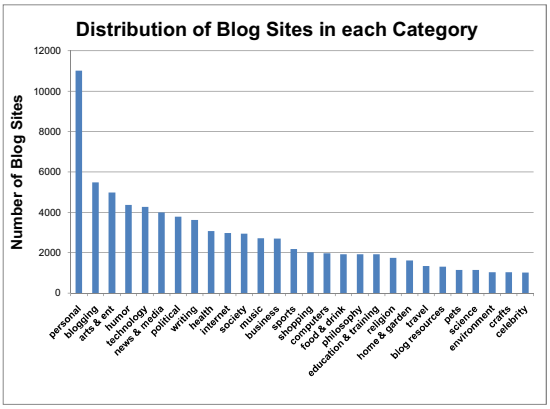

Figure 1. Distribution of blog sites with respect to the labels.

WisClus is resistant to noise since false category labeling is rare and would have minimum impact on the collective wisdom.

3 Experiments and Analysis

Data Collection: We started with 4 bloggers from different categories as the starting points and crawl their social networks, recursively in a breadth-first fashion. For each blogger thus crawled, we collect their blog site URL, blog site labels, blog post tags, and blog post snippets.

Since the blog site labels have a hierarchical structure, to decide what level gives the best clustering results we construct three different datasets:

1. *Top-level*: The labels of all the blog sites are abstracted to their top most parent level labels.

2. *All-category*: The full hierarchical structure of the labels is considered.

3. *One node-split*: According to the distribution of blog sites in various top level labels, illustrated in Figure 1, `Personal` has the largest number of blog sites[1]. Hence, we split `Personal` into its child labels, to reduce the skewed distribution of blog sites.

Results and Analysis: The experiments are designed to evaluate two issues:

1. What granularity of label hierarchical structural information generates best clustering? For this we study the clustering results for the three variants of the dataset.

2. Which one of the two clustering approaches, WisClus or the baseline approach, performs the best?

Link Strength in WisClus: We experiment with different threshold values for link strength varying from 3, 4, 5, 7, and 10. Due to space constraints we just present the cluster

[1]For the sake of space constraint and the analysis presented here, we limit the labels in this chart that have at least 1000 blog sites.

	Number of clusters	Highest degree	Lowest degree	Largest cluster size	Smallest cluster size	Coverage Total %	Coverage 1st cluster %	Coverage 2nd cluster %
All-categories, >= 3	3	17	1	48	2	79.78	76.98	1.76
All-categories, >= 4	5	11	1	27	2	67.57	58.05	4.31
All-categories, >= 5	6	8	1	15	2	54.76	42.3	6.375
All-categories, >= 7	4	6	1	10	2	44.63	33.64	4.61
All-categories, >= 10	1	3	1	5		21.67	21.67	

Table 1. Various statistics to compare clustering results for different threshold values for WisClus.

Figure 2. WisClus results for link strength $\geq$ 5 for All-category dataset.

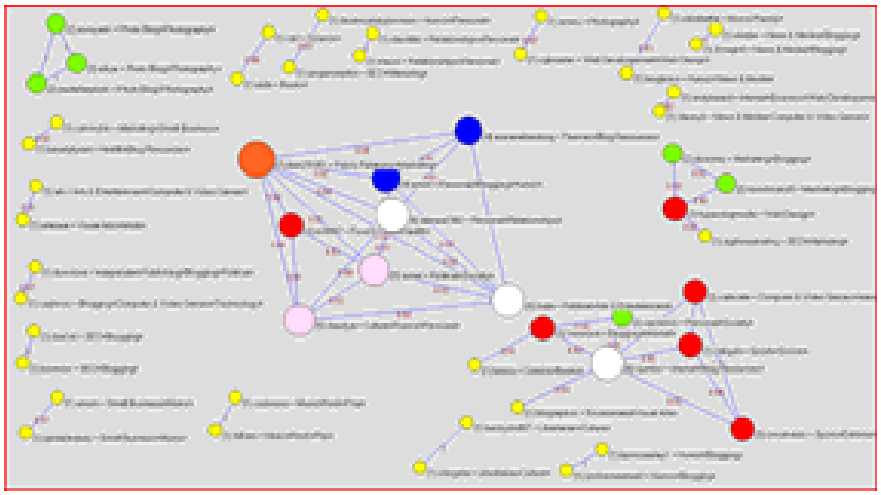

Figure 3. Clusters obtained using baseline clustering approach for All-category dataset.

Category	Number of clusters	Highest degree	Lowest degree	Largest cluster size	Smallest cluster size	Coverage Total %	Coverage 1st cluster %	Coverage 2nd cluster %
All-categories, >= 5	6	8	1	15	2	54.76	42.3	6.375
Top-level	1	16	1	22		100	100	
One node-split	3	9	1	21	2	82.87	76.44	3.88

Table 2. Various statistics to compare clustering results for different label structure for WisClus.

visualization results for threshold values of 5 in Figure 2[2]. Link Strength is denoted by the values on the edges. Names of the nodes depict the labels assigned by the bloggers to the blog sites. Here a node represents all the blog sites that are labeled as the label of the node. A cluster of labels would represent a cluster of all the blog sites that are labeled with one of these labels. Some nodes like `Internet>Web Design` depict the hierarchical structure of labels. Here the blog sites are labeled `Web Design` which is a child of `Internet`. We present detailed statistics for clustering results for all the threshold values in Table 1 for comparison. For threshold $>= 3$, total coverage is highest but we have a single large cluster and 2 very small clusters depicted by the cluster coverages. Similar is the case for threshold $>= 4$, 7, and 10. This indicates highly unbalanced clusters are achieved at other thresholds as compared with threshold $>= 5$. Hence we set threshold=5 for rest of the experiments. More results and analysis are reported in [1].

Label Hierarchy: We consider all the three variants of the dataset, i.e., Top-level, All-category, and One node-split for this experiment. Due to space limitations we present clustering results for All-category in Figure 2. Statistics of clusters obtained from WisClus for different datasets are reported in Table 2. Although the total coverage is maximum for Top-level label structure, it creates only one cluster hence treating all the labels as semantically related. Similarly, results for One node-split show that the cluster size is again highly unbalanced. There are only 3 clusters with the 1^{st} cluster having majority of coverage and the difference between 1^{st} and 2^{nd} cluster is very small. Results for All-categories has the lowest coverage but the clusters are not as unbalanced. This shows that leveraging the complete structure of collective wisdom gives best results as compared to exploiting a part of it. This proves that the more collective wisdom is available the better it is.

WisClus vs. Baseline Clustering: Here we compare WisClus algorithm and baseline algorithm to study the advantages of collective wisdom. Results obtained using baseline clustering algorithm are presented in Figure 3. Here nodes represent the blog sites or bloggers. For easier comparison we also display the labels of their blog sites besides their name. For example, a node label like, `emom=Small Business:Moms`, tells us that the blogger `emom` has a blog site with labels `Small Business` and `Moms`. However, we do not use the label information while clustering in baseline approach. We report the differences between the two approaches based on the results as follows:

1. Clusters obtained from baseline approach are too fragmented (lot of 2-member clusters) as compared to WisClus.

2. As a result, clusters are too focussed. This affects the insertions of new blog site later on. Cluster configurations are highly unstable in such a focussed clustering.

[2]Pajek was used to create the visualizations.

3. Several clusters from baseline clustering, have members whose blog site labels are semantically unrelated. For example, `bluemonkey jammies = Humor:Personal` and `emperoranton = SEO: Marketing` are clustered together. The reason for semantically incoherent clusters is the susceptibility of vector space clustering to text noise, predominantly found in blogs. Moreover, blogs are dynamic in nature with the blogger occasionally posting about different topics. However WisClus gives high-quality, semantically coherent clusters.

4. Several clusters obtained from baseline approach have members that have exactly the same labels. For example, the cluster with bloggers `emom` and `geraelindsey` have the same labels, i.e., `Small Business` and `Moms`. Clustering blog sites that have different yet related theme/topics are more helpful. WisClus generates clusters of blog sites with labels like, `Technology`, `Computers`, `Internet`, and `Technology>Gadgets`.

4 Conclusions

Clustering blogs is a challenging task with many real-world applications. Classic clustering methods do not take advantage of some characteristics of the blogosphere. We proposed WisClus a blog clustering algorithm that leverages collective wisdom. We evaluate different values of link strengths and various levels of label hierarchies, compare WisClus with a classic SVD-based clustering algorithm, present results statistically and visually, and summarize findings that deepen our understanding. Since WisClus mainly relies on the label information by the bloggers, it is adaptive to the blog dynamics. WisClus is a proof-of-concept project using forms of collective wisdom in the Social Web. Our future work includes the integrative use of multiple sources of information such as labels, tags, and posts in clustering, and the intelligent use of the clustering results to help focused search in the blogosphere.

References

[1] N. Agarwal et al. Clustering with Collective Wisdom - A Comparative Study. Technical Report TR-08-004, Arizona State University, 2008.

[2] M. Devaney and A. Ram. Efficient feature selection in conceptual clustering. In *ICML*, 1997.

[3] C. H. Q. Ding et al. A Min-max Cut Algorithm for Graph Partitioning and Data Clustering. In *ICDM*, 2001.

[4] J. Z. Huang, M. Ng, and L. Jing. Text Clustering: Algorithms, Semantics and Systems. PAKDD Tutorial, 2006.