Discovering Dynamic Logical Blog Communities Based on Their Distinct Interest Profiles

Guozhu Dong
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435, USA
guozhu.dong@wright.edu

Neil Fore
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435, USA
neilfore@gmail.com

***Abstract* — This paper addresses the problem of identifying dynamic logical blog communities based on the distinct interests shared by blogs in the communities. This paper is motivated by the facts that the blog space is highly dynamic both in the participating bloggers and in the interests/issues of concern to them in their blogs, and that many organizations are interested in identifying the evolution/emergence of various blog communities and the current interests/issues of concern to the blogs of those communities. Using dynamic shared distinct interests to define logical blog communities allows us to better identify and track the issues of concern to the bloggers than using statically chosen keywords or statically defined blog communities. The paper discusses algorithms to solve the above problem, which attempt to maximize discriminativeness and diversity of the distinct interests of the blog communities. Experiments are reported to evaluate the performance of the algorithms and to demonstrate their utility.**

Keywords-distinct interest; clustering with description; weblog community discovery; community description; contrast pattern.

I. INTRODUCTION

Weblogs are a popular and easily accessible avenue for individuals and social communities to express their opinions and to interact with each other on matters of interest. The ability to quickly understand the evolution of, including the formation of new, blog communities, and the ability to quickly identify the current interests and issues of concern to various communities of bloggers, are important to various social organizations. The purpose of this paper is to study the problems associated with providing such abilities. It presents algorithms for solving those problems and reports an experimental evaluation on those algorithms.

The blog space is both large (containing a large number of blogs written by a large number of bloggers) and fast-changing. Moreover, the issues/interests of concern to bloggers are dynamic and fast-changing, reflecting what is happening in the real world. As a result, static blog communities and static blog community descriptions are not adequate to best provide the two abilities discussed in the previous paragraph. We need to identify logical blog communities, based on the distinct interests/issues shared by blogs in those communities, in a dynamic manner.

Distinct interests of blogs play a significant role in solving those problems discussed above. A distinct interest of a blog is a combination of a (small) number of words that occur in the given blog but rarely occur in other blogs. Distinct interests of a blog capture highly distinguishing focuses of the blog when compared against the other blogs. A logical blog community is one in which the blogs have significant shared distinct interests. (We will provide precise criteria used to determine logical blog communities below.)

In general, a distinct interest corresponds to a *contrast pattern* (CP). For this, it will be convenient to use the term “cluster” (of blogs) as a synonym of blog community. A pattern is a small set of words. A CP (a.k.a. emerging pattern) [3] for a given clustering (corresponding to number of blog communities) is a pattern that occurs much more frequently in blogs of one (its home) cluster than in the other clusters. This difference in frequency makes a CP a highly discriminative pattern to describe its home cluster and distinguish that cluster against the other clusters.

In this paper, we will use distinct interests as the basis to form logical blog clusters, and to identify the shared distinct interests of the formed blog clusters. To that end, we use a recently proposed Contrast Pattern based Clustering algorithm, called CPC [6]. Intuitively speaking, in the blog community discovery context, the CPC algorithm aims to form blog clusters whose associated distinct interests have high quality and high diversity (as CPs of the clusters). A CP's quality is defined in terms of its support in its home cluster and the length ratio of its closed pattern over its minimal generator patterns; the diversity of CPs is defined in terms of their shared items and shared matching data (tuple sets). CPCQ is a clustering quality index introduced in [7], based on the quality/diversity of CPs; that paper also demonstrated that CPCQ consistently prefers expert-given clusterings to other generated clusterings. CPC aims to form clusters to maximize the CPCQ score.

ISBN: 978-1-61208-163-2

Since CPCQ and CPC automatically discover the most important distinct interests of the dynamically formed blog clusters of the dynamic collections of blogs posted recently (e.g., in the last hour), they are well suited to the dynamic nature of blog communities. The results they produce are more desirable than those blog communities (and their descriptions) that correspond to static communities of bloggers who write blogs of interest or correspond to a static set of keywords. Those two static approaches are often based on what were known about the past, and may not be able to capture what is happening now.

Table I illustrates some of these ideas using 7 short weblogs. The best clustering (indicated in column 1) is mostly suggested by the distinct interests (CPs) present in the blogs. The patterns {disease}, {diet}, {suffer}, {suffer, treatment}, etc. are distinct interests (CPs) of blogs in cluster 1, and the patterns {music}, {song}, and {music, song} are distinct interests (CPs) of blogs in cluster 2. (A minimum support threshold of 25%, i.e., 2 weblogs, is used here.)

TABLE I. CPC CLUSTERING OF SYNTHETIC WEBLOGS

Cluster	Weblogs
1	disease, diet, exercise
1	disease, suffer, treatment
1	best, diet, help, weight
1	help, suffer, treatment
2	album, artist, music, song
2	best, music, popular, song
2	band, music, release, song

The contrast patterns {music} and {song} are high-quality distinct interests of blogs in cluster 2 because each occurs in all 3 weblogs of cluster 2. These two patterns do not share common words and hence are very diversified. The pattern set {{music}} (of one pattern) makes a distinct interest set/profile of cluster 2 (since its matching data covers the entire cluster 2). The patterns {treatment} and {diet} do not occur together in any given weblog, but together they cover the entire cluster 1. Each of these two patterns is a distinct interest of cluster 1. Moreover, the pattern set {{treatment}, {diet}} is a diversified set (profile) of the distinct interests of cluster 1.

Since no other clustering is better than the given clustering in terms of quality, diversity, and total coverage of CPs, the given clustering maximizes the CPCQ score. Moreover, the clustering is also considered high-quality since it has two diversified distinct interest profiles for each cluster: {{disease}, {help}} and {{treatment}, {diet}} for cluster 1, and {{music}} and {{song}} for cluster 2.

TABLE II. DISTINCT INTEREST PROFILES OF BLOG CLUSTERS

DIPs (2) for Cluster 1	{{diet},{treatment}}}, {{disease, help}}
DIPs (2) for Cluster 2	{{music}}, {{song}}

In practice, the blogs collected are often not clustered, and/or they are clustered but the clusters do not have known blog collection descriptions. The blog cluster profiles can help users get a sense of the main issues of interest in the blog cluster, and get a feel on what the blogs are roughly about.

Related Work: (A) Our work is related to weblog analysis and tracking (e.g., [1]). Reference [1] presented a tool for tracking and analyzing blogs, which can identify frequent terms used in blogs, and the influential bloggers and relationship among bloggers in given blogs. However, it did not consider forming blog clusters together with their distinct interest profiles, as is done in this paper.

(B) This work is also related to [4], which presented methods to discover descriptions of weblog clusters/communities using sets of single-word patterns, but it did not consider how to form blog clusters and did not consider using multi-word patterns. Regarding distinct interest profile formation, this paper generalizes [4] by using multi-word contrast pattern descriptions, which are more general and are capable of capturing more subtle distinct interests for blog clusters, and also gives methods to discover high-quality clusters (or communities) in given weblog collections. Experiments confirm that the approach of this paper is more effective.

(C) This work is also related to document collection summarization (e.g., [8,9]), and multiple document collection summarization (e.g., [2,4]). Document summarization can be divided into three types, namely sentences-based (where a summary consists of sentences extracted from the original documents), templates-based (where predefined templates are filled as summaries), and term-based (where a set of carefully selected terms is used to form the summaries). This work is novel in its use of the quality and diversity of the contrast patterns, which can contain multiple terms, to form the clusters and to describe the clusters.

(D) To some extent, this paper is also related to frequent pattern based clustering [5]. While [5] used frequent patterns in the formation of clusters, it did not consider using the quality and diversity of the contrast patterns of the clusters in the clustering process.

Organization of the Paper: Section II formally defines the problems to be studied in the paper. Section III presents the CPC clustering algorithm and the CPCQ clustering quality index. (Their details can be found in [6] and [7].) Section IV reports an experimental evaluation of the algorithms on real blog data. Section V concludes the paper.

II. PROBLEM DEFINITION

In this paper, we consider two technical problems for dynamic weblog-based community analysis. The first problem is for producing distinct interest profiles for weblog data with a known community/cluster structure. The second is for clustering and describing the found communities (clusters) using their distinct interest profiles, for weblog data without a known community structure.

Below, we use "distinct interest" (DI) and "contrast pattern" (CP) as synonyms. (However, only carefully selected CPs are included in the distinct interest profiles (DIPs).) Preliminaries on CPs are given in Section III.

Definition: Given k (≥2) (blog) collections (communities) D_1,…, D_k, a *distinct interest profile* (DIP) consists of k sets S_1,…, S_k of distinct interests for the collections.

Problem 1 (Distinct Interest Profile): Given k (≥2) (blog) collections (communities) D_1,…, D_k, find a succinct and informative DIP to describe the collections.

The “succinct” requirement on the descriptions ensures that low processing load is required when users examine a DIP. The “informative” requirement ensures that users can quickly estimate the main themes/concerns/interests/issues of the blogs in given collections. Technically, we can measure the informativeness by the F-score of clusterings induced by the DIP on data with known classes. (The F-score measures the agreement between the DIP-induced clustering and the original collections viewed as classes.)

Problem 2 (Distinct Interest Based Community Discovery and Description): Given a (blog) collection D and natural number $k \geq 2$, form a high-quality clustering of D having k clusters C_1,…, C_k together with a succinct and informative DIP to describe the clusters.

The “high quality of formed clustering” requirement will be measured by the F-score and the CPCQ value (discussed in the next section). Earlier studies [6] show that clusterings maximizing the CPCQ value often are highly similar to expert-given classes (on datasets with known classes).

III. CONEPTS AND TECHNIQUES OF CPCQ AND CPC

In this section, we provide high-level descriptions of our technical approaches to solve the two problems listed in Section II. Specifically, after first providing some technical preliminaries, we describe the CPCQ clustering quality measure and the CPC clustering algorithm. Roughly speaking, CPC attempts to discover clusters that maximize the quality and diversity of the CPs within each cluster. Similarly, CPCQ evaluates the quality of a given clustering by finding optimal groups of high-quality, diversified CPs in each cluster.

A. Preliminaries

As discussed in Section I, a contrast pattern (CP) for a given clustering (a set of collections) is a pattern that occurs much more frequently in blogs of one (its home) cluster than in the other clusters.

For a pattern P, we will use |P| to denote its item length (cardinality) and mt(P) to denote its matching tuple set -- mt(P) is the set of blogs in a dataset (or cluster, which can be clear from the context) that contain the pattern P.

Each pattern P is associated with an equivalence class (EC) of patterns defined as EC(P)={P' | mt(P') = mt(P)}. In a sense, all patterns in a common EC have the same practical “meaning”, since they match the same set of blogs/objects. Each EC can be concisely described by a closed pattern Pmax and a set of minimal generator (MG) patterns. The closed pattern is the unique longest pattern in the EC and the MG patterns are those minimal (under the set containment relation) in the EC. An EC contains precisely those patterns X satisfying “X is a superset of at least one MG pattern” (of the EC) and “X is a subset of the closed pattern”. The MGs of an EC can be viewed as different minimal descriptions of the mt(P) dataset.

For example, for the data given in Table I, the EC of CPs containing {music} consists of the following CPs: {music}, {song}, and {music, song}. {music} and {song} are the MGs, and {music, song} is the closed pattern. Moreover, mt(P) is equal to the entire cluster 2 for each CP P in the EC.

Both the MGs and the closed patterns of such ECs will be important for quality and diversity evaluation of clusters and for cluster descriptions. Below, when we say "pattern" P, we refer to any MG pattern in the EC of P.

Thinking of patterns in terms such equivalence classes (determined by the patterns’ matching tuple sets) is an important aspect of both CPC and CPCQ.

B. The CPCQ Clustering Quality Measure

For the CPCQ cluster quality measure [7], a high-quality clustering is one having a large number of high-quality, diversified CPs in each of its clusters.

A CP P is considered to have high quality if it is short, its closed pattern is long, and its support in its home cluster is high.

- If P is short, its home cluster is more easily distinguishable from the other clusters by using P. For blog data, a short P can be viewed as a short distinct interest of the blogs in the home cluster of P.
- If its closed pattern is long, its matching blogs (i.e., mt(P)) are more coherent. All of the words in the closed pattern of P are coincident distinct interests of the home cluster of P.
- If P's support in its home cluster is high, it will account for a large number of blogs in that cluster.

Technically, given a MG pattern P, we use the term *length ratio* to denote the ratio of P's closed pattern length to P's length, or |Pmax|/|P|. We prefer higher length ratios. For example, for the CP (of the EC of) {music}, the length ratio is 2.

The diversity requirement of CPCQ is motivated by the fact that natural concepts (captured by clusterings) (e.g., the gender, male/female, concepts) often can be easily distinguished/characterized in many highly different ways. The diversity of CPs is measured by the average of diversity between CP pairs. Two CPs are considered diversified if they two CPs share few items/blogs (i.e., their item/data overlap is low, and their item/data diversity is high). To measure the abundance/diversity of CPs in each cluster, CPCQ builds a number of diversified CP groups for each cluster. Ideally, the average pairwise data- and item-overlap among CPs should be low within each CP group, and each CP group should cover its entire cluster. Among CP groups, the average pairwise item overlap among CPs from different CP groups should be low; however, data overlap among CP groups is not considered since each CP group can cover its home cluster. The high-quality CP groups of high-quality, diversified CPs found for the clusters can be used to describe/represent the clusters.

A DIP can be formed by taking one CP group from each cluster in a clustering.

More details on CPCQ are given in [7], including a greedy algorithm to search for the multiple high-quality, diversified CP groups to assess the CPCQ value.

In this paper, we use the CP groups constructed by the CPCQ group-building algorithm [7] as the DIPs to describe/characterize the clusters.

C. The CPC Clustering Algorithm

The CPC algorithm constructs clusters on the basis of patterns to maximize the CPCQ score of the resulting clustering. A main challenge for CPC is that it only has access to the frequent patterns, since CPs are only determined after the clusters are known. Hence the CPC algorithm must guess and evaluate which frequent patterns should become CPs and which of such CPs should be put into the same cluster.

To address the challenge, a relationship is defined between CPs to measure their suitability of belonging to the same cluster. This relationship, termed *Mutual Pattern Quality* (MPQ), measures the number and quality of *other* CPs that can be gained by assigning two diversified CPs to the same cluster. Specifically, given two patterns P_1 and P_2 sharing few blogs/tuples, MPQ(P_1,P_2) is high if a relatively large number of (mutual) patterns share many matching blogs with both P_1 and P_2. If MPQ(P_1,P_2) is high, then patterns P_1 and P_2 are likely to belong to be CPs of the same cluster; if MPQ(P_1,P_2) is low, P_1 and P_2 are likely to be CPs of separate clusters. The MPQ formula will be given below.

Using MPQ, the CPC algorithm constructs clusters bottom-up by first finding a set of weakly-related seed patterns (having low MPQ values among the CPs in the set) to initially define the clusters, and then repeatedly adding diversified patterns that have high MPQ values with CPs of a certain cluster to that cluster. Once clusters are completely defined in terms of CPs, blogs (and other CPs) can be assigned to clusters based on their matching CPs. The details are given in [6].

MPQ Formulae: Given two patterns P_1 and P_2 sharing very few blogs/tuples, MPQ(P_1,P_2) is defined as

$$MPQ(P_1,P_2) = \frac{PQ2(P_1,P_2)}{PQ1(P_1) * PQ1(P_2)}. \qquad (1)$$

In (1), PQ2(P_1,P_2) is the *Joint Overlap-Weighted Pattern Quality* of P_1 and P_2, given by

$$PQ2(P_1,P_2) = \qquad (2)$$

$$\sum_X \left(\frac{|mt(P_1) \cap mt(X)| * |mt(P_2) \cap mt(X)|}{|mt(X)|} * \left(\frac{|X_{max}|}{|X|} \right)^2 \right).$$

In (2), X is any pattern except P_1 or P_2. PQ2(P_1,P_2) is high if X shares many blogs/tuples with P_1, it shares many blogs/tuples with P_2, and it matches few blogs/tuples elsewhere in the dataset. These properties indicate that X is a CP alongside P_1 and P_2 if, and only if, P_1 and P_2 are CPs of the same cluster. Additionally, PQ2(P_1,P_2) is high if X has a high length ratio.

To favor the most exclusive connections between P_1 and P_2, PQ2(P_1,P_2) is normalized by the *Overlap-Weighted Pattern Quality* (PQ1) of each argument. PQ1 for a pattern Q is given below:

$$PQ1(Q) = \qquad (3)$$

$$\sum_P \left\{ |mt(P) \cap mt(Q)| * \left(\frac{|P_{max}|}{|P|}\right)^2 \middle| \, P \neq Q \right\}.$$

A high PQ1(P) value indicates that many high-quality patterns share many blogs/tuples with P. Normalizing PQ2(P_1,P_2) by PQ1(P_1)*PQ1(P_2) is necessary since more mutual patterns (i.e., patterns contributing non-zero values to PQ2/MPQ) are likely to be found between two patterns P_1 and P_2 if their PQ1 values are high. Conceptually, this normalization is analogous to calculating the correlation between two events A and B, which is strong if Prob(AB)/(Prob(A)*Prob(B)) >> 1.

Unlike most clustering algorithms, CPC constructs clusters on the basis of patterns before assigning blogs/tuples. That is, CPC-produced clusters are completely determined/described by DIs/CPs when applied to weblogs. The accuracy of these clusters, then, reflects of the utility of DIs and DIPs for discovering blog communities as well as describing them.

IV. Experimental Evaluation on Weblog Data

This section reports our experimental evaluation of CPC for the discovery of blog communities in real weblog data, and of CPCQ for the discovery of the associated DIPs.

Specifically, we report CPCQ scores, which are used as an internal quality measure by CPC, and F-scores, for various settings. (F-score measures agreement between CPC-generated clusters and the given categories; it has a maximum value of 1.0.) We also show the DIPs found by CPCQ for given communities and for CPC-produced clusterings.

These experiments were performed on weblog collections extracted from four categories of the BlogCatalog dataset [10]: health, music, sports, and business. Each dataset was preprocessed by stemming, removing duplicate weblogs, and tokenizing (i.e., words were treated as items).

A. Succinctness and Informativeness of DIPs

As stated earlier, the DIP-based community descriptions are generated by the CPCQ algorithm. For consistency, all CPCQ results (scores and DIPs) use a minimum support threshold (minS) of 3% and two CP groups (and hence two DIPs). This leads to brief descriptions while allowing informative CP groups (DIPs) to be found.

Table III shows the DIP based descriptions and CPCQ scores for the health and music collections treated as given clusters C_1 and C_2, and for the clusters created by CPC (k=2, minS=10%) from the union of health and music. One can clearly estimate the themes of the clusters from the DIPs. Moreover, given the similarity of the DIPs in the two cases, the high F-score for the CPC clustering suggests a high degree of informativeness for the DIPs of the health/music collections (as a given clustering). In each case, the CPCQ score is relatively low, mostly because the DIPs do not cover the majority of their home clusters.

TABLE III. CLUSTER DESCRIPTIONS FOR HEALTH AND MUSIC CATEGORIES

	Categories, as given CPCQ: 0.53		**CPC clusters (minS=10%)** F-score: 0.901, CPCQ: 0.71	
	DIP 1	**DIP 2**	**DIP 1**	**DIP 2**
C_1	{diseas}, {help, hair}	{diet}, {treatment, suffer}, {peopl, insur}	{diseas}, {dai, drink}	{diet}, {treatment, suffer}
C_2	{music}	{song}	{song}	{album}

To evaluate DIP based descriptions for a larger number of weblog collections, we repeated the above using four collections. Table IV shows the DIPs generated from the original categories, and Table V shows those generated for the clusters created from the union of the four collections by CPC (k=4, minS=3%).

TABLE IV. CLUSTER DESCRIOTIONS FOR HEATH, MUSIC, SPORTS AND BUSINESS CATEGORIES

Category	**CPCQ-generated CP groups (minS=3%)**, CPCQ: 0.247	
	DIP (CP group) 1	**DIP (CP group) 2**
Health	{health, diet}, {treatment, suffer}	{disord}, {health, heart}
Music	{releas, song}	{guitar}, {releas, post, music}
Sports	{season, team}	{final, win}, {am, sport}
Business	{busi, monei, market}	{busi, monei, internet}

TABLE V. CLUSTER DESCRIPTIONS FOR CPC-GENERATED CLUSTERS, k=4, minS=3%

Cluster	**CPCQ-generated CP groups (minS=3%)** F-score: 0.77, CPCQ score: 0.325	
	DIP (CP group) 1	**DIP (CP group) 2**
1	{bodi, food}, {suffer, medic}	{symptom}, {health, fit}
2	{band, song}, {youtub, music}	{releas, song}
3	{team, game}	{season, team}
4	{busi, market}	{busi, monei}

ISBN: 978-1-61208-163-2

Again, it is clear that one can estimate the cluster themes from the DIPs. Moreover, the DIPs are similar for the two cases, and the F-score for the CPC clustering is reasonably high. The DIPs shown in Table IV for health and music are not identical to those in Table III, due to the presence of two new clusters.

B. CPC Clustering Quality

To measure the quality/accuracy of blog clusters created by CPC, we use the CPCQ measure as well as the F-score of the CPC clusterings. Tables VI and VII show these values for various CPC clusterings for k=2, 3, and 4. For brevity, not all results are shown, but the cases with the highest and lowest F-scores are included.

TABLE VI. CPCQ AND F-SCORE FOR CPC CLUSTERINGS, k=2

	heath, music		health, business		sports, business	
minS	CPCQ	F-score	CPCQ	F-score	CPCQ	F-score
10%	0.71	0.901	0.36	0.853	0.407	0.858
5%	0.535	0.893	0.33	0.74	0.38	0.882
3%	0.664	0.89	0.355	0.77	0.352	0.826
2%	0.664	0.899	0.338	0.704	0.347	0.794

TABLE VII (a). CPCQ AND F-SCORE FOR CPC CLUSTERINGS, k=3

	health, music, sports		health, music, business	
min S	CPCQ	F-score	CPCQ	F-score
5%	0.393	0.827	0.33	0.76
3%	0.552	0.849	0.334	0.806
2%	0.404	0.824	0.35	0.723
1%	0.452	0.843	0.397	0.743

TABLE VII (b). CPCQ AND F-SCORE FOR CPC CLUSTERINGS, k=4

	health, music, sports, business	
minS	CPCQ	F-score
5%	no clusters at minS=5%	
3%	0.325	0.77
2%	0.229	0.767
1%	0.247	0.742

The F-score of CPC varies significantly based on the dataset, which may indicate more similarity between certain weblog collections. CPC's F-score also decreases with increasing k. This is partly due to the same reason, but is also expected since fewer CPs can be expected to exist among a larger number of clusters. Nonetheless, the F-scores achieved here demonstrate consistently high accuracy, confirming the usefulness of DIPs (i.e., CPs) in discovering weblog communities.

Notice that in three of the five cases, the highest F-score coincides with the highest CPCQ score, indicating that CPCQ can be used to estimate the optimal minS value for CPC when categories (and hence F-scores) are not known. However, the results show no clear trend in F-scores or CPCQ scores as minS is varied for any case; therefore, a range of minS values should be tried to estimate an optimal minS value.

C. Comparison against DCR-induced Clusterings

Below, we compare the F-score of CPC-produced clusterings to those induced by discriminative collection representatives (DCR) [4]. Note that DCRs are created from known blog collections (that paper builds a very simple DCR-based classifier to recover the blog categories) while CPC attempts to discover the blog clusters/collections together with the DIP-based descriptions. In other words, F-score of the approach of [4] is supervised class recovery (classification) accuracy, whereas F-score given by our CPC approach here gives unsupervised class recovery accuracy.

To make the comparison, we constructed three datasets as described in [4]; each set contains the first 1000 weblogs from a weblog collection C_1 (shown in column 1 of Table VIII) and 100 weblogs from each of 10 other categories (of BlogCatalog [10]). For each dataset, we selected the CPC clustering (k=2) having the highest CPCQ score among minS in {10%, 5%, 3%, 2%}.

TABLE VIII. F-SCORE COMPARISON: DCR AND CPC

C_1	F-score (C_1)		F-score (total)	
	DCR	CPC	DCR	CPC
Sports	0.772	0.734	0.713	0.78
Music	0.785	0.846	0.727	0.83
Health	0.756	0.733	0.668	0.765

Interestingly, overall (total) F-scores are higher for CPC clusterings in all 3 cases, while F-scores for C_1 are higher for DCR-induced clusterings in 2 of the 3 cases. This suggests that CPC's ability to dynamically discover/utilize the DIPs for the remaining 10 categories gives it an advantage over DCR-induced clustering, despite the fact that CPC is unsupervised.

We note that F-scores for these CPC clusterings are lower, on average, than for the k=2 clusterings reported in the previous section. A possible explanation is that more patterns may be shared between C_1 and the union of 10 categories (and hence fewer high quality CPs exist) than between C_1 and a single other category.

V. CONCLUDING REMARKS

This paper addressed the problem of identifying dynamic logical blog communities based on the distinct interests shared by blogs in the communities. Using dynamic shared distinct interests to define logical blog communities allows us to better identify and track the issues of concern to the bloggers than using statically chosen keywords or statically defined blog communities. The paper discusses a Contrast Pattern based Clustering (CPC) algorithm to solve the above problem, which attempts to maximize discriminativeness and diversity of the distinct interests of the blog communities. Experiments over real weblog data indicate that the contrast pattern based clustering quality measure and the contrast pattern based clustering methods can help discover natural weblog communities/clusters, and can discover succinct and informative distinct interest profiles to describe the communities/clusters. Potential future research topics include improving the CPC algorithm by considering the history of blog communities, their distinct interest profiles, and the blogger themselves, in the process of forming new blog communities and their distinct interest profiles.

REFERENCES

[1] N. Agarwal, S. Kumar, H. Liu, and M. Woodward. BlogTrackers: A Tool for Sociologists to Track and Analyze Blogosphere. AAAI Conf. on Weblogs and Social Media, 2009.

[2] L. Chen and G. Dong. Succinct and informative cluster descriptions for document repositories. pp. 109-121, Int'l Conf. on Web-Age Information Management, 2006.

[3] G. Dong and J. Li: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. pp. 43—52, KDD 1999.

[4] G. Dong and T. Sa: Analyzing and Tracking Weblog Communities Using Discriminative Collection Representatives. pp. 256-264, SBP, 2010.

[5] B. C. M. Fung, K. Wang, and M. Ester: Hierarchical Document Clustering using Frequent Itemsets. SDM 2003.

[6] N. Fore and G. Dong. CPC: A Contrast Pattern based Clustering Algorithm. To appear.

[7] Q. Liu and G. Dong: A Contrast Pattern based Clustering Quality Index for Categorical Data. pp. 860-865, IEEE ICDM, 2009.

[8] C. Y. Lin and E. Hovy. Automated multi-document summarization in neats. Proceedings of the Human Language Technology Conference, 2002.

[9] D. Shen, J. T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. pp. 2862-2867, IJCAI, 2007.

[10] R. Zafarani and H. Liu. Social Computing Data Repository at ASU, http://socialcomputing.asu.edu/datasets/BlogCatalog, July 24, 2011.