Using Tags and Clustering to Identify Topic-Relevant Blogs

Conor Hayes
Digital Enterprise Research Institute
National University of Ireland, Galway
Ireland
conor.hayes@deri.org

Paolo Avesani
ITC-IRST
Via Sommarive 18
38050 Povo (Trento)
Italy
avesani@itc.it

Abstract

The Web has experienced an exponential growth in the use of weblogs or blogs. Blog entries are generally organised using tags, informally defined labels which are increasingly being proposed as a 'grassroots' answer to Semantic Web standards. Despite this, tags have been shown to be weak at partitioning blog data. In this paper, we demonstrate how tags provide useful, discriminating information where the blog corpus is initially partitioned using a conventional clustering technique. Using extensive empirical evaluation we demonstrate how tag cloud information *within* each cluster allows us to identify the most topic-relevant blogs in the cluster. We conclude that tags have a key auxiliary role in refining and confirming the information produced using typical knowledge discovery techniques.

Keywords

Tag, Blog, Tag Cloud, Clustering, Relevance

1. Introduction

A weblog (blog) is a website containing journal-style entries presented in reverse chronological order and generally written by a single user. Over the past few years, there has been an exponential growth in the number of blogs [17] due to the ease with which blog software enables users to publish to the web, free of technical or editorial constraints.

However, the decentralised and independent nature of blogging has meant that tools for organising and categorising the blog space are lacking. Advocates of the so-called web 2.0 school of thought have proposed emergent organisational structures such as 'tag clouds' to tackle this problem. Tags are short informal descriptions, often one or two words long, used to describe blog entries (or any web resource). There is no globally agreed list of tags the user can choose from, nor is there an agreed best practice for tagging. Tag clouds refer to aggregated tag information, in which a taxonomy or 'tagsonomy' emerges through repeated collective usage of the same tags.

In previous work we presented an empirical evaluation of the role for tags in providing organisational support for blogs [8]. In comparison to a simple clustering approach, tags performed poorly in partitioning the global document space. However, we discovered that, *within* the partitions produced by content clustering, tags were extremely useful for the detection of cluster topics that appear coherent but are in fact weak and meaningless. The key observation was that semantically meaningful clusters were more likely to contain higher proportions of high-frequency tags than weak clusters. This allowed us to construct a score for each cluster called the $\mathcal{T}_r$ score, which allowed the detection of semantically weak clusters that could not be detected automatically by standard techniques based on intra- and intercluster distance.

Our overall conclusion is that using a single global tag cloud as a primary means of partition is imprecise and has low recall. On the other hand, partitioning the blog document space using a conventional technique such as clustering produces multiple topic-related or *local* tag clouds, which can provide discriminating secondary information to further refine and confirm the knowledge produced by the clustering. Furthermore, local tag clouds establish topic-based relationships between tags that are not observable when considering the global tag cloud alone. The work described in this paper builds upon this supporting role for tags.

Our previous work was motivated by the need to build a blog recommender system in which a registered blogger would be regularly recommended posts by other bloggers with similar interests. A key issue is to recommend posts about a topic by relevant bloggers, that is, bloggers who write regularly in a non-trivial way about a particular topic. The Google search engine has successfully used link analysis to identify the most relevant pages for a particular query. However, recent research on the blog domain would suggest that the majority of blogs are unconnected [9]. Our own blog dataset consisting of over 7000 blogs monitored over a 6-week period had almost no internal links. Instead, we propose using tags to automatically gauge blogger relevance in a cluster-based recommender system.

Local tag clouds exhibit the same power law tag frequency distribution as global tag clouds. However, the ratio between high-frequency and low-frequency tags varies according to cluster strength. We define an *a-tag* as a tag in a local tag cloud that has a frequency greater than 1. An *a-blog* is a blog belonging to a cluster that contributes an a-tag i.e. an a-tag that is shared by at least one other blog in the cluster. A *c-blog* is a blog from the same cluster that contributes a unique single tag not shared by any other blog in the cluster, a c-tag. The key observation of this paper is that a-blogs form subclusters that are consistently the most relevant blogs to the cluster concept and that are more likely to stay together as blog data is clustered over time. We build an argument for this hypothesis by extensive empirical evaluation.

In Section 2, we describe recent work on tagging. In Section 3, we describe our datasets, which were collected over a 6-week period. Section 4 summarises our previous work and our clustering technique. In Section 5, we demonstrate that a-blogs form subclusters in each cluster that are more self similar and closer to the cluster centroid than c-blogs. In Section 6, we demonstrate the relevance of a-blogs to the cluster topic definition in a novel experiment in which we query Google using the cluster topic definition and then compare a-blog and c-blog similarity to the retrieved pages. In Section 7, we examine the likelihood of bloggers remaining together as clustering continues over time. We define a measure of blogger entropy and show that a-blogs have significantly lower entropy than c-blogs, suggesting that a-blogs will tend to be clustered together in later clusterings. We present a discussion and conclusions in Sections 8 and 9.

2. Related work

The Semantic Web vision for the blog domain is typified by prototype applications in which an RDF-based data model allows sophisticated, inference-enabled querying of blogs [3, 11]. In contrast, tagging is a 'grassroots' solution to the problem of organising distributed web resources, with emphasis on ease of use. Quintarelli [15] proposes that tag usage engenders a *folksonomy*, an emergent user-generated classification. However, tags are flat propositional entities and there are no techniques for specifying 'meaning', inducing a hierarchy or inferring or describing relationships between tags.

The Semantic Web approach has the disadvantage of being potentially too knowledge intensive, and risks being ignored by web users. Although tagging is widely used by blog users, its effectiveness as a primary organising mechanism has not been demonstrated [2, 8]. Despite its obvious weaknesses, tagging is firmly a part of the so-called web 2.0 trend toward information sharing and collaboration on the Internet, typified by sites like the blog aggregator, Technorati[1], the photo-sharing site, Flickr[2], and the social bookmarks manager, Del.icio.us[3], all of which rely upon tags to allow users to discover resources tagged by other people.

Brooks and Montanez [2] have analysed the 350 most popular tags in Technorati in terms of document similarity and compared these to a selection of similar documents retrieved from Google. In previous work we have shown that the most popular tags form a small percentage of the overall tag space and that a retrieval system using tags needs to employ *at least* token-based partial matching to retrieve a larger proportion of tagged blogs [8]. Golder and Huberman [6] provide a good introduction to the dynamics of collaborative tagging on the Del.icio.us social bookmarks site. However, the Del.icio.us site differs from the blog domain in that tags are applied in a centralised way to URLs generally belonging to other people. A Del.icio.us user can view the bookmark tags already applied to the URL he wishes to index and choose an existing tag or use another. This aggregating facility is not available to the blogger, who must tag a piece of writing he/she has just completed. Whereas a tag on Del.icio.us references the URL of a website, a blogger's tag often references a locally defined *concept*.

Although the popular collective term 'blogosphere' implies a type of social network, recent research suggests that less-connected or unconnected blogs are in the majority on the Web [9]. Link analyses on our datasets have produced the same results. For this reason we do not consider links between blogs in this paper.

[1] http://www.technorati.com
[2] http://www.flickr.com
[3] http://www.del.icio.us

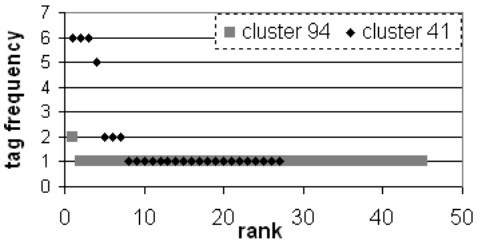

Fig. 1: *Tag token frequency distribution for cluster 41 (high $\mathcal{H}_r$) and cluster 94 (low $\mathcal{H}_r$)*

3. Dataset

Our blog dataset[4] is based on data collected from 13,518 blogs during the 6-week period between midnight January 15 and midnight February 26, 2006. All blogs were written in English and used tags. Blogging activity obeys a power law, with 88% of bloggers posting between 1 and 50 times during the period and 5% posting very frequently (from 100 to 2655 posts). On inspection, many of these prolific bloggers were either automated blog spammers or community blogs. We selected data from bloggers who had posted from 6 to 48 times during the evaluation period. The median for this sample is 16 posts. On average, each user posted at least once per week during the six-week period.

For each user we selected the tag that was used on the largest proportion of the user's posts during the evaluation period. We aggregated these posts to form a single document. Thus, each document represents the collective posts indexed under a single tag by one user during the evaluation period.

The data was divided up into 6 datasets, each representing post data from a single week. As all 7209 bloggers do not post every week, the datasets have different sizes and overlap in terms of the blog instances they contain (see Table 1). Each instance in a dataset is a 'bag of words' made up of the posts indexed under the most frequently used tag from a single blog during that week, *plus* the posts made in the previous 2 weeks (using the same tag). As the posts in a single week are often quite short and take the form of updates to previous posts, we include the previous 2 weeks to capture the context of the current week's updates. For example, if a blog is updated in week 3, the instance representing that blog in the dataset for week 3 is based on the posts in weeks 3, 2 & 1. If the blog is not updated in week 4, the instance representing the blog is excluded from the dataset for week 4. As shown in Table 1, on average, 71% of the blogs present in the dataset win_t will also be present in the dataset win_{t+1}.

We processed each dataset independently, removing stop words and stemming the remaining words in each document. We then removed low-frequency words appearing in less than 0.2% of the documents, and high-frequency words occuring in more than 15% of the documents. Documents with less than 15 tokens were not considered at this point. Each word was weighted according to the standard TF/IDF weighting

[4] The authors will release this data to interested researchers.

data	Dates	Size	Num. Feat.	Mean Feat.	O_{t+1}	%
win_0	Jan 16-Jan 23	4163	3910	122	3121	75
win_1	Jan 23-Jan 30	4427	4062	123	3234	73
win_2	Jan 30-Feb 6	4463	4057	122	3190	71
win_3	Feb 6-Feb 13	4451	4124	122	3156	71
win_4	Feb 13-Feb 20	4283	4029	122.	2717	63
win_5	Feb 20-Feb 27	3730	4090	121	-	-
mean	-	**4253**	**4043**	**122**	**3084**	**71**

Table 1: *The periods in January and February 2006 used for the windowed blog dataset. Each period is from midnight to midnight exclusive. O_{t+1} refers to the overlap with the same users in the dataset for the next window*

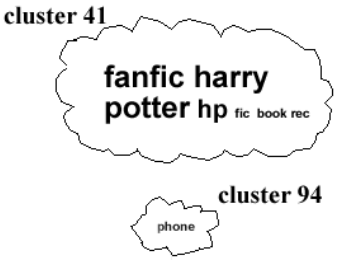

Fig. 3: *The tag clouds for cluster 41 (high $\mathcal{H}_r$) and cluster 94 (low $\mathcal{H}_r$)*

scheme and the document vector normalised by the L^2 norm. This created a feature set of approximately 3500 words for each dataset. Table 1 gives the window period, size and overlap with the subsequent window.

Each instance in each dataset is associated with a single 'tag'. There are no constraints on how many words a blogger may use in a tag. Many tags are made up of several words and would be unlikely to be aggregated with other tags. For this reason we tokenise each tag into its constituent words, we stem each word and we remove all non-alphanumeric characters and all stop words. Thus each instance is associated with a set of tag tokens that are more easily (partially) matched with other tag tokens. When we refer to 'tags' (a-tags, b-tags, c-tags) from this point onwards we are referring to tag *tokens*.

4. Clustering and tag analysis

The blog domain contains tens of millions of documents, constantly being updated. A reasonable goal would be to try to organise these documents by topic or type. Document clustering is a well established technique for organising unlabelled document collections [18]. Clustering has two goals: to uncover latent structures that accurately reflect the topics present in a document collection and to provide a means of summarising and labelling these structures so that they can be interpreted easily by humans. Clustering has been used for improving precision/recall scores for document retrieval systems [19], browsing large document collections [4], organising search engine return sets [14, 10] and grouping similar user profiles in recommender systems [16, 13, 12]. For this work we implemented the *spherical k-means* algorithm, a well understood variation of the k-means clustering algorithm, which scales well to large document collections and produces interpretable cluster summaries [5]. Spherical k-means produces k disjoint clusters, the centroid of each being a concept vector normalised to have unit Euclidean norm.

Given a set of data points, the goal of a clustering algorithm is to partition them into a set of clusters so that points in the same cluster are close together, while points in different clusters are far apart. Typically, the quality of a clustering solution is measured using criterion functions based on intra- and intercluster distance. Following [20], the quality of cluster r is given as the *ratio* of intra- to intercluster similarity, $\mathcal{H}_r$. Given S_r, the set of instances from cluster r, intracluster similarity, $\mathcal{I}_r$, is the average cosine distance between each instance, $d_i \in S_r$, and the cluster centroid, C_r. Intercluster similarity, $\mathcal{E}_r$, is the cosine distance of the cluster centroid to the centroid of the entire dataset, C (see Equation 1).

In previous work, we have confirmed that clusters with high $\mathcal{H}_r$ scores tend to be clusters with large proportions of documents of a single class [8]

$$\mathcal{H}_r = \frac{\mathcal{I}_r}{\mathcal{E}_r} = \frac{\frac{1}{|S_r|}\sum_{d_i \in S_r} \cos(d_i, C_r)}{\cos(C_r, C)} \quad (1)$$

In the experiments that follow we do not address the issue of selecting an optimal value of k and, as such, we cluster the data at several values of k. For each value of k, a random seed is chosen after which k-1 seeds are incrementally selected by choosing the seed with the greatest distance to the mean of the seeds already selected. In order to track user and topic drift from week to week, the seeds for the clusters in week t are based on the final centroids of the clusters produced in week t-1, except in the case of the first week where the seeds are chosen to maximise inter-seed distance.

In order to cluster data using the seeds based on the centroids from the previous week, we map the feature set from the previous week's data to the feature set of the current week. In datasets from adjacent weeks the feature set overlap is greater than 95%. The feature values for each seed are the feature weights from the corresponding centroid in the previous week.

4.1 Tag analysis after clustering

In the previous work we demonstrated that blogs tags are not useful as a primary means of partitioning our datasets. Instead we proposed a *supporting* role where tags identified weak clusters that could not be identified using standard techniques [8]. Typically, in any system where tags are aggregated, few tags are used very frequently and the majority of tags are used infrequently. When we partition the data using content clustering, we observe a tag frequency distribution per cluster that seems to vary according to cluster strength ($\mathcal{H}_r$). Weak clusters tend to have a long flat distribution, that is, few or no high-frequency tags (tokens) and a long tail of tags that have been used only once. Strong clusters tend to contain many high-frequency tags and a shorter tail.

Figure 1 illustrates the tag distribution for 2 clusters where k=100. Clusters 41 and 94 contain 47 and 43 instances per cluster respectively. Cluster 41 is in the top 20% of $\mathcal{H}_r$ scores and cluster 94 is in the bottom 20%. Figure 3 illustrates the tag cloud for each cluster based on these distributions.

We can qualify the tag frequencies per cluster. **C-tags** are tag tokens not repeated by any other user in the cluster. These tags are represented by the long tail of the fre-

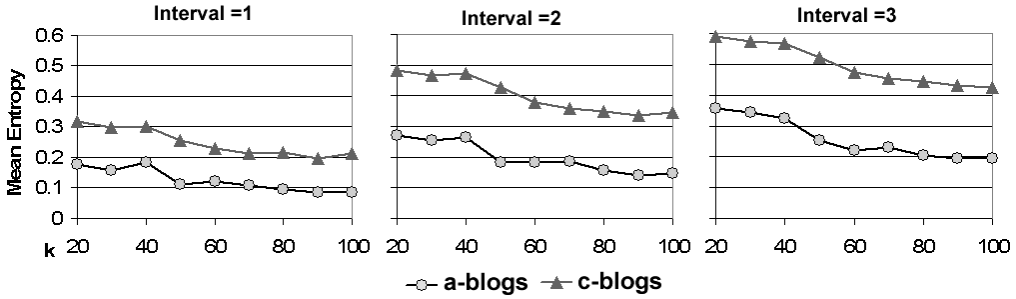

Fig. 2: *Mean a-blog vs. c-blog entropy at interval = 1,2 & 3*

quency distribution. **B-tags** are tag tokens with a frequency $\geq$2 that occur in several clusters at once. B-tags are analogous to stop-words, words that are so common that they are useless for indexing or retrieval purposes. Furthermore, b-tags also tend to be words with non-specific meaning, such as 'assorted', 'everything' and 'general'. As such, they do not contribute to cluster interpretation and are disregarded. **A-tags** are the remaining high-frequency tags. Clearly, a-tags are an important indicator of the semantics of the cluster as they represent an independent description of the cluster topic by 2 or more bloggers. For a more detailed description of tag types see [8].

5. A-blogs as relevant sources of information

In the previous section we described how each cluster can be described by a tag token cloud made up of a-tags. As the tag frequency distribution in each cluster follows a power law, only a portion of the blogs in each cluster will have contributed tag tokens to the tag description. For the sake of convenience, these blogs are termed *a-blogs*. The remaining blogs, which contribute single tag tokens to the long tail of the frequency distribution, are termed *c-blogs*. In this section we examine the characteristics of a- and c-blogs, keeping in mind our goal to automatically identify blogs that are most relevant to the topic definition produced by the cluster description.

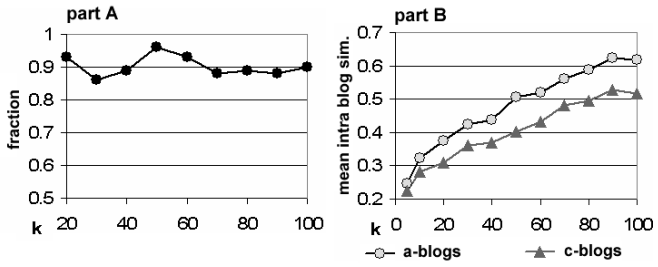

Fig. 4: *Part A: mean fraction of clusters where a-blog IBS > c-blog IBS. Part B: mean IBS for a-blogs and c-blogs*

The following experiments are based on clustering of each of the 6 blog datasets at values of k from 20 to 100. For each dataset and each value of k we chose the top 40% of clusters according to the clustering criterion $\mathcal{H}$. From this set, we removed any clusters identified as potentially weak or noisy by the cluster $\mathcal{T}_r$ score [8]. For each of the remaining clusters in each dataset, we measured the *intra-blog similarity* (IBS) of the a-blogs and the c-blogs. The IBS of a group of blogs is the mean pairwise similarity of all the blogs in the group, where similarity is measured using the cosine measure.

For the sake of space, the results presented in Figure 4 are averaged over the 6 datasets. Part A of Figure 4 gives the fraction of clusters at each value of k in which the IBS of the a-blogs was greater than the IBS of the c-blogs. Part B gives the mean IBS at each value of k. For each of the 6 datasets we found the difference between the means of the a-blog and c-blog scores to be significant at 0.05 alpha level. Part A of the figure provides evidence that in a high fraction of clusters a-blogs are generally 'tighter', that is, more similar to each other than c-blogs. Part B then illustrates the mean difference in IBS between a-blogs and c-blogs in each cluster at each value at k. From $k = 50$ upwards the difference is approximately 0.1.

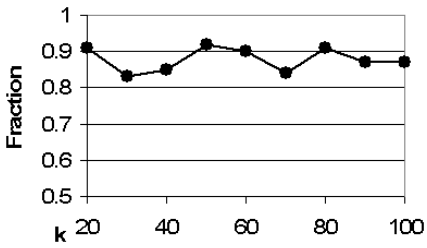

Fig. 5: *The mean fraction of clusters where a-blog similarity to the cluster centroid > c-blog similarity to the cluster centroid. The mean is calculated based on the fractions obtained for each dataset at each value of k*

In the second experiment we tested whether a-blogs were closer to the cluster centroid than c-blogs. The cluster centroid defines the 'concept' induced by the clustering process. The Spherical k-means algorithm produces a weighted term vector where the weights reflect the normalised summation of the term weights contributed by the documents in the cluster. The documents in a cluster will have differing degrees of similarity to the cluster centroid. Document vectors close to the centroid are more likely to contain highly weighted terms that are also highly weighted in the centroid vector. As such we would expect documents close to the centroid to be highly relevant to the concept description. Using the same set of clusters from each dataset, we measure the mean similarity of the a-blogs and c-blogs to each cluster centroid. Figure 5 presents the fraction of clusters where the similarity of a-blogs to the cluster centroid is greater than the similarity of c-blogs. The fraction shown here is the mean based on the fractions obtained from each dataset at each value of k. The

figure indicates that the a-blogs in each cluster are more likely to be closer to the cluster centroid than c-blogs. Figure 6 illustrates mean similarities to the cluster centroid for a-blogs and c-blogs for each of the 6 datasets. For each dataset the difference between the means of the a-blog and c-blog results was found to be significant for each dataset at an alpha level of 0.05.

The results from these first two experiments lead us to conclude that within each cluster a-blogs tend to form tight subgroups, which are generally more similar to the cluster centroid than the remaining c-blogs in the cluster. A key question is whether a-blog documents are more *relevant* to the cluster concept than c-blog documents. In information retrieval the cluster hypothesis [19] posits that documents that are more similar to each other are more likely to be relevant to a particular information requirement than less similar documents. The information requirement in this case is the concept summary presented by the cluster. In application terms, this is a synopsis of the topic presented to the user based on selection of key words and the retrieval goal is to suggest a set of blogs that are most relevant to the concept summary. The conventional way of measuring the ability of any IR algorithm to retrieve relevant documents is to measure its precision and recall abilities over a labelled relevant set of documents. However, as our blog dataset is unlabelled, we do not have a direct way of measuring the precision or recall scores for a-blogs or c-blogs. In the next section we describe an alternative technique to test the potential relevance of a-blog documents to the cluster concept description.

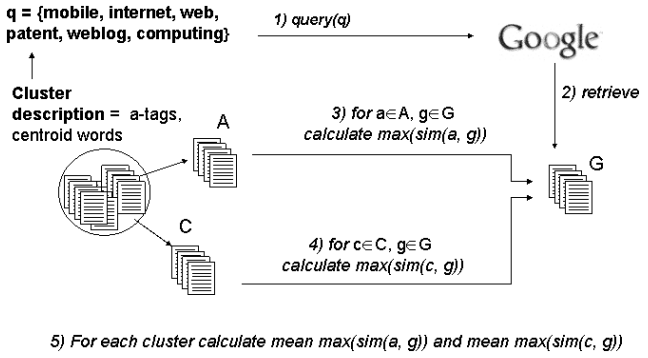

Fig. 7: *The figure illustrates the steps in the Google retrieval experiment*

6. Querying Google

The key idea of this experiment is to retrieve an independently defined relevant set of documents using the concept description of each cluster and to then measure the similarity of the a-blogs and c-blogs in the cluster to the retrieved documents. To do this we rely upon the Google search engine, which uses keyword-matching and the PageRank algorithm to retrieve relevant documents to a submitted query [1]. Google is the most widely used search engine today and we rely upon the documents it returns as a type of 'gold standard' of relevance. Although this is an extremely naive idea, it does allow us to retrieve an independently defined set of documents based on a query extracted from the cluster key words.

Using the same clusters from each dataset, we extract a concept description from each cluster using the top 5 highest weighted key words in the cluster centroid. We generate a query string from the key words and submit it to Google using the Google SOAP API. As each key word has been stemmed, we perform reverse stemming before submitting the query. We enable the Google search filter to remove duplicate pages from the search result set. Each query returns 10 ranked URLs from the Google search engine. We then retrieve the page associated with each URL. We process each page into a bag of words by stripping away the HTML, removing stopwords and stemming. We then save the bag of words to a database.

query	# queries	# pages
centroid	954	9213
a-tags	883	8633

Table 2: *The number of queries generated and pages returned using the centroid and a-tag query methods*

When a set of pages has been retrieved for each cluster concept query in each dataset, we then apply a feature mask of the relevant dataset to each bag of words and produce a set of document vectors weighted according to normalised TF/IDF. The IDF scores are calculated using the total set of pages retrieved for queries associated with a single dataset. Each cluster is thus associated with a set of Google-retrieved document vectors, $\mathcal{G}$.

Then, for each cluster we select the set of a-blogs, $\mathcal{A}$, and a set of c-blogs, $\mathcal{C}$. Usually the number of a-blogs in a cluster is less than the number of c-blogs. As such we make a random selection of c-blogs so that $|\mathcal{C}| = |\mathcal{A}|$. If the number of c-blogs is greater we select randomly from $\mathcal{A}$ so that $|\mathcal{A}| = |\mathcal{C}|$. For each cluster we calculate the similarity of each a-blog $\in \mathcal{A}$ to each document vector $g \in \mathcal{G}$. For each a-blog we select the maximum similarity score achieved with the documents in $\mathcal{G}$. Likewise, for each c-blog we calculate the similarity to each document vector $g \in \mathcal{G}$ and we select the maximum similarity score achieved. Thus, for each cluster we calculate the mean of the maximum similarity scores achieved for the blogs in $\mathcal{A}$ and the mean of the maximum similarity scores achieved for the blogs in $\mathcal{C}$. This is carried out for the clusters in each dataset at values of k from 20 to 100.

We carry out a parallel set of experiments by extracting a concept description in terms of the a-tag cloud for the cluster. We select the top 5 most frequently used tag tokens and perform the same experiment using these tokens as a query. Being based on categorical information, the a-tag descriptions of the concept are generally more abstract than the descriptions extracted from the centroid. Furthermore, the a-tag descriptions are often more coherent and interpretable than the centroid descriptions. We would therefore expect the result sets returned from Google to be less noisy in the case of the a-tag queries. Figure 7 summarises the main steps in this experiment and Table 2 gives the number of queries generated and pages returned for the centroid and a-tag query generation methods.

6.1 Results

Figure 8 presents the mean fraction of clusters where the mean a-blog similarity (to retrieved Google pages) was greater than the mean c-blog similarity. The mean is based on the fraction for each of the 6 datasets. The fraction is high at almost all values of k, with a more consistent performance being

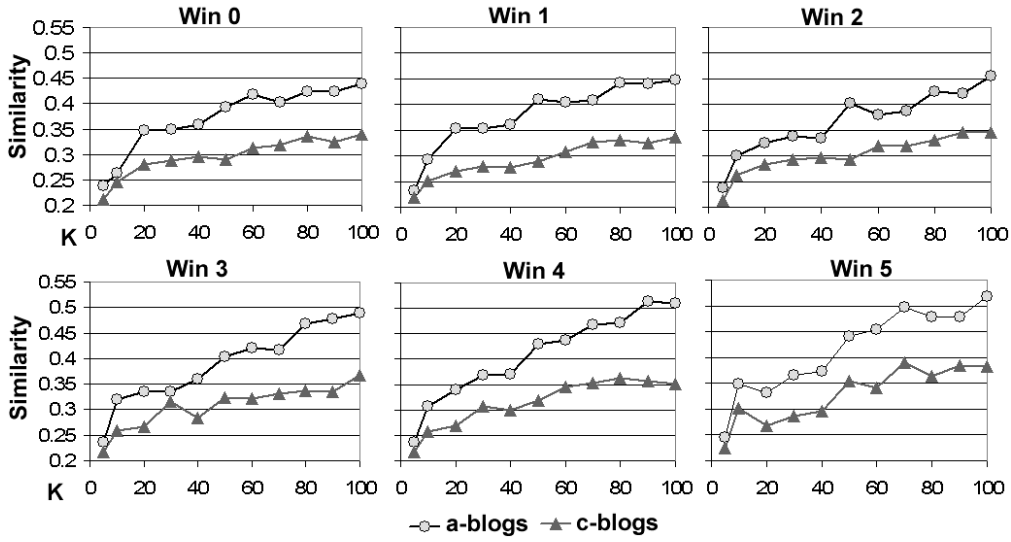

Fig. 6: *The similarity to the cluster centroid for a-blogs and c-blogs*

recorded for a-tag queries than centroid-generated queries. For both types of query, the fraction is always above 0.5, indicating that a-blogs are likely to be more similar to pages retrieved from Google than c-blogs. This appears to hold for all values of k. Figure 9 indicates the mean a-blog and c-blog similarity to the retrieved pages for the first 3 datasets ($win_0 win_2$) at different values of k. The difference between the a-blog and c-blog means for each of the 6 datasets was found to be significant at the 0.05 alpha level, even in the case of dataset win_2, where the means appear to be close. The results confirm that a-tag blogs in each cluster are more similar to a set of pages retrieved using the cluster to seed a query. This is an interesting result because tag tokens themselves are not used in the clustering process, yet they consistently allow us to pinpoint those documents that appear to be most relevant to the cluster concept. In this case, we define relevance in terms of similarity to a set of documents retrieved through Google.

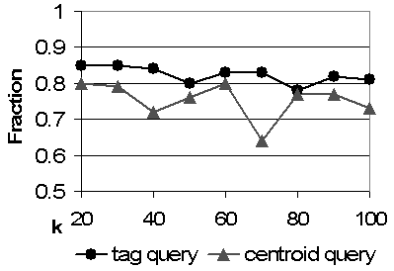

Fig. 8: *Mean fraction of clusters where a-blogs are more similar than a-blogs to the retrieved Google Pages. The mean is based on the fraction for each of the 6 datasets*

Table 3 gives a selection of a- and c-blogs from a cluster defined by the top 5 centroid keywords *mobile, internet, weblog, web* and *patent*. The descriptions of the web blogs are extracted from the blog title, except where the description is given in italics, in which case the author has given a summary. The a-blogs are clearly relevant to the cluster description, providing 3 blogs about mobile technology and 2 blogs about technology and intellectual property issues. The relevance of the c-blogs appear less convincing. Philips Brooks' patent blog is the most relevant, dealing with general patent issues. There are 2 personal blogs that offer a mixture of marginally relevant topics and a blog on religion, which is completely mis-clustered. This is probably due to the fact that its definition for the church matches almost exactly the definition that is often given for the so-called web 2.0 and, by extension, the mobile web 2.0.

7. Blogger entropy

In previous work [7] on the same datasets we described the phenomenon of *user drift*. This refers to the observation that, as the datasets are clustered from one week to the next, many blogs are often not clustered together again. This is problematic as it suggests that blog data requires constant re-clustering and that the relationships established between blogs based on shared topics in one week cannot be exploited for any length of time. It also suggests that (many) bloggers may be writing in a 'shallow' way i.e. they are not regularly using terminology that allows them to be strongly associated with a particular topic. We defined a measure to track this, which we term user entropy. User entropy, $\mathcal{U}_r$, for a cluster is a measure of the dispersion of the users in one cluster throughout the clusters of the next window. For a fixed value of k, if many of the users in a single cluster in win_t are also in a single cluster in win_{t+1}, then entropy will approach zero. Conversely, if the neighbourhood of users at win_t is spread equally among many clusters at win_{t+1}, entropy will tend toward a value of 1.

$$\mathcal{U}_r = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \tag{2}$$

$c_{r,t}$ is cluster r at win_t; $c_{i,t+1}$ is a cluster i at win_{t+1}, which contains users from $c_{r,t}$. S_{t+1} are all the instances in win_{t+1}. q is the number of $c_{i,t+1}$ (the number of clusters at win_{t+1} containing users from cluster $c_{r,t}$). $n_r = |c_{r,t} \cap S_{t+1}|$. n_r^i is

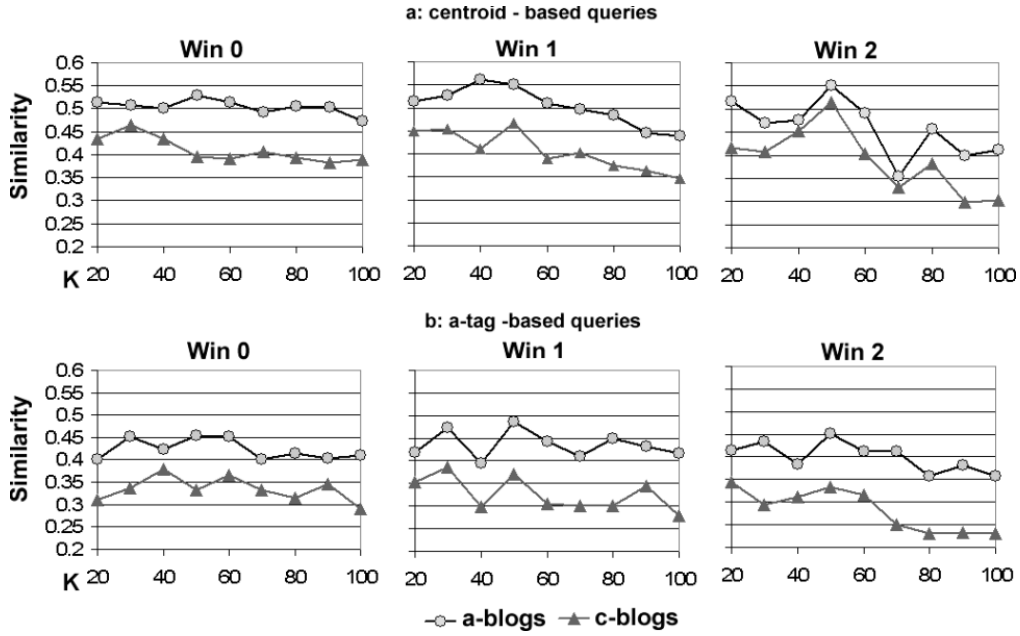

Fig. 9: *Part A : Similarity to documents retrieved from Google using* ***centroid****-based query for the first 3 datasets. Part B: Similarity to documents retrieved from Google using* ***a-tag****-based query for the first 3 datasets*

	Name	Description
a	**Communications**	Technology, Economic and Social Issues at the Intersection of Telecom, Mobility and the Internet
a	**IP Blawg**	*Technology and Intellectual Property Blog*
a	**Small business IP management blog**	Patent, Trademark, Copyright, Internet and Technology Law
a	**Open Gardens**	Wireless mobility, Digital convergence - Mobile web 2.0
a	**Mobile Enterprise Weblog**	the voice of enterprise mobility management
c	**Digital Music Den**	Digital Music, online music marketing
c	**icarusindie.com blog about nothing**	*General computing, programming and technology*
c	**Dunkie's Saga**	*Personal blog: personal, gaming, quizzes, some technology*
c	**Complex Christ**	A vision for church that is organic, networked, decentralized, bottom-up, emergent, communal, flexible, always evolving
c	**Philips Brooks patent infringement updates**	*Legal blog on general patent issues (some technology-related material)*

Table 3: *A selection of a-blogs and c-blogs and their descriptions taken from cluster 28, k = 50. The 5 highest weighted keywords from the concept centroid are:* ***mobile, internet, weblog, web, patent***

$|c_{r,t} \cap c_{i,t+1}|$, the number of users from cluster $c_{r,t}$ contained in $c_{i,t+1}$.

However, our previous analysis did not differentiate between blogs in each cluster and the entropy measure was calculated over both a-blogs and c-blogs. We return to this experiment and calculate the entropy for a-blogs and c-blogs separately in each cluster. Using the same clusters as before, the mean entropy is calculated at different values of k where the interval between datasets is increased from 1 to 3. For example, when the interval is 1 we calculate the mean entropy based on the entropy scores recorded between the following pairs of windows: (win_0, win_1), (win_1, win_2), (win_2, win_3), (win_3, win_4) and (win_4, win_5). When the interval is 3 the mean entropy score is based only on the following pairs : (win_0, win_3), (win_1, win_4) and (win_2, win_5). Figure 2 illustrates that a-blogs have much lower entropy than c-blogs at all values of k. As the distance between windows (and each clustering) increases, we would expect to see a rise in entropy. However, a-blogs have significantly smaller entropy scores and experience smaller increases in entropy than c-blogs as the interval increases.

This is an important observation because it suggests that not only do a-tags allow us to identify relevant sources of information about a topic, but that these sources tend to be *consistent* over time. In other words, we can identify bloggers that are consistently associated with topics and would be important candidates to consider in any topic-based recommendation strategy.

8. Discussion

In earlier work we employed tag information to refine the partitions produded by a simple clustering algorithm. An advantage of this approach is that it allows the tag space to be partitioned into smaller sub-clouds in which related tags are aggregated based on the similarity of the underlying con-

tent. This allows for a more meaningful aggregation of tag data. For example, the tag cloud description of Harry Potter fan fiction shown in Figure 3 could not have been identified within the typical global tag cloud. From a clustering perspective, tags provide discriminating information about the clustering solution, because the probability of two or more users choosing the same tag token for blogs in a well defined cluster is much higher than in a mixed or noisy cluster.

Identifying relevant sources of information is an important topic for the blog domain. The fact that a-tags are successful at identifying potentially strong sources of topic-based information may tell us something about the profile of a-bloggers. We suggest that a-bloggers are keen to cultivate readership of their blogs and therefore carefully select tags that are easily understood and representative of the topic. Furthermore, it appears that a-bloggers write in depth about fairly narrowly defined subjects. Thus, similar a-bloggers are regularly clustered together. In contrast, a large proportion of bloggers keep less formal blogs, posting short entries for friends and family and tagging carelessly. We believe that tag behaviour is one of a number of features we can identify to allow us to automatically classify new blogs. Future work will involve identifying and combining these features for classification purposes.

9. Conclusions

In this paper we extend earlier work where we used tag information to refine the output of a clustering solution. We suggest that a-bloggers, bloggers who contribute tokens to the cluster a-tag description, tend to be the most relevant sources of topic information. Our hypothesis is that these bloggers choose their tags carefully in consideration of their readers. Our evaluation found that a-bloggers tend to form the core of each cluster. We tested their relevance to the concept description by measuring their similarity to pages ranked by Google. We found that they were consistently more similar to these pages than c-bloggers. Furthermore, we demonstrated that these bloggers tend to be clustered together again in later periods.

Acknowledgments

This work has been funded by a grant from the Provincia Autonoma di Trento. We would also like to acknowledge the support of Science Foundation Ireland under grant number SFI/02/CE1/I131.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, 1998. Elsevier Science Publishers B. V.

[2] C. H. Brooks and N. Montanez. An analysis of the effectiveness of tagging in blogs. In *Proceedings of the 2005 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. AAAI, March 2005.

[3] S. Cayzer. Semantic blogging and decentralized knowledge management. *Commun. ACM*, 47(12):47–52, 2004.

[4] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *15th international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329, New York, USA, 1992. ACM Press.

[5] I. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In G. K. R. Grossman and R. Naburu, editors, *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, 2001.

[6] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[7] C. Hayes, P. Avesani, and S. Veeramachaneni. An analysis of bloggers and topics for a blog recommender system. In *Workshop on Web Mining (WebMine), 7th ECML/10th PKDD*, September 2006.

[8] C. Hayes, P. Avesani, and S. Veeramachaneni. An analysis of the use of tags in a blog recommender system. In *IJCAI-07, the International Joint Conference on Artificial Intelligence*. www.ijcai.org, January 2007.

[9] S. Herring, I. Kouper, J. Paolillo, and L. Scheidt. Conversations in the blogosphere: An analysis "from the bottom up". In *Proceedings of HICSS-38*, Los Alamitos, 2005. IEEE Press.

[10] http://www.clusty.com. Clusty search engine, 2006.

[11] D. R. Karger and D. Quan. What would it mean to blog on the semantic web. *Journal of Web Semantics*, 3(2):147–157, 2005.

[12] J. Kelleher and D. Bridge. An accurate and scalable collaborative recommender. *Artificial Intelligence Review*, 21(3 - 4):193 – 213, June 2004.

[13] M. O'Connor and J. Herlocker. Clustering items for collaborative filtering. In *ACM SIGIR Workshop on Recommender Systems*, Berkeley, CA, 1999.

[14] O. E. O. Oren Zamir. Grouper: A dynamic clustering interface to web search results. In *8th International WWW Conference*, Toronto, Canada, May 1999.

[15] E. Quintarelli. Folksonomies: power to the people. *ISKO Italy-UniMIB Meeting, Mi*, June 2005.

[16] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *5th International Conference on Computer and Information Technology*, 2002.

[17] D. Sifry. State of the blogosphere: Part 1 - on blogosphere growth. *http://technorati.com/weblog/2006/04/96.html*, April 2006.

[18] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *6th ACM SIGKDD, World Text Mining Conference*, Boston, MA, 2000.

[19] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.

[20] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311 – 331, 2004.