An Analysis of the Use of Tags in a Blog Recommender System

Conor Hayes Paolo Avesani Sriharsha Veeramachaneni
ITC-IRST,
Via Sommarive 18 - Loc. Pantè, I-38050 Povo, Trento, Italy
$\{$*hayes,avesani*$\}$@*itc.it*

Abstract

The Web is experiencing an exponential growth in the use of weblogs or *blogs*, websites containing dated journal-style entries. Blog entries are generally organised using informally defined labels known as *tags*. Increasingly, tags are being proposed as a 'grassroots' alternative to Semantic Web standards. We demonstrate that tags by themselves are weak at partitioning blog data. We then show how tags may contribute useful, discriminating information. Using content-based clustering, we observe that frequently occurring tags in each cluster are usually good meta-labels for the cluster concept. We then introduce the $\mathcal{T}_r$ score, a score based on the proportion of high-frequency tags in a cluster, and demonstrate that it is strongly correlated with cluster strength. We demonstrate how the $\mathcal{T}_r$ score enables the detection and removal of weak clusters. As such, the $\mathcal{T}_r$ score can be used as an independent means of verifying topic integrity in a cluster-based recommender system.

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 [Sifry, 2006] 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. The solutions to this problem can be characterised by two positions: the *tagging* approach and the Semantic Web approach. 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. Conversely, researchers on the Semantic Web project have proposed frameworks in which blogs are marked up using machine-readable meta-data, written in a language such as RDF, which would facilitate cross-blog indexing [Cayzer, 2004; Karger and Quan, 2005].

Tagging has achieved widespread acceptance by blog users and advocates argue this is due to its simplicity [Quintarelli, 2005]. Bloggers can easily categorise resources without having to learn the precise vocabulary of a formally defined classification scheme. The drawback is that a tag is defined locally and there is no mechanism for indicating that two tags mean the same thing.

As tagging has been so widely adopted, we present an empirical evaluation of how tags perform in organising and annotating blogs. Using a simple clustering approach, we firstly demonstrate that tags perform poorly compared to a standard content-based approach. We then introduce a supporting role for tags. We observe that a power law probability distribution exists for the frequency of tag usage *within* clusters. Using measures of intra- and intercluster distance, we demonstrate that 'strong' clusters are more likely to contain high proportions of high-frequency tags than weak clusters. We observe that high-frequency tags are usually good meta-labels for the cluster concept. We introduce, $\mathcal{T}_r$, a score based on the proportion of high-frequency tags in a cluster, and demonstrate that it is strongly correlated with cluster strength. Finally, we demonstrate how the $\mathcal{T}_r$ score enables the detection and removal of clusters that appear coherent but are in fact weak and impure. As the $\mathcal{T}_r$ score is produced by the aggregated tag data from a set of independent users, it can be viewed as an independent means of verifying cluster topic integrity.

In Section 2, we describe recent work on tagging. In Section 3, we describe our datasets: a blog dataset and a labelled newsgroup dataset that we use for comparison purposes. We compare clustering using content and tags in Section 4. In Section 5, we show that tags can be used as cluster meta-labels. We introduce the $\mathcal{T}_r$ score and demonstrate empirically that it can be used to automatically identify clusters with poor semantics. We present our conclusions in section 6.

2 Background

The Semantic Web vision for the blog domain is typified by prototype applications in which a RDF-based data model allows sophisticated, inference-enabled querying of blogs [Cayzer, 2004; Karger and Quan, 2005]. In contrast, tagging is a 'grassroots' solution to the problem of organising distributed web resources, with emphasis on ease of use. Tags are flat propositional entities and there are no techniques for specifying 'meaning', inducing a hierarchy or inferring or describing relationships between tags. Despite the obvious weakness of this approach, tagging appears to be part of a trend toward information sharing on the Internet. Sites like the blog aggregator, Technorati[1], the photo sharing site, Flickr[2], and the social bookmarks manager, Del.icio.us[3], all rely upon tags to allow users to discover blogs, photos and websites tagged by other people. Quintarelli [2005] proposes that tag usage engenders a *folksonomy*, an emergent user-generated classification. Although the mechanisms of such a self-organising process are not clearly outlined, we do share the view that aggregated tag data can provide discriminating features for partitioning purposes.

Brooks and Montanez [2005] 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 this paper we will show 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 matching to retrieve a larger proportion of tagged blogs. Golder and Huberman [2006] 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 [Herring *et al.*, 2005]. Link analyses on our datasets have produced the same results. For this reason we do not consider links between blogs in this paper.

3 Data

Our blog dataset 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. We removed stop-words and stemmed each document. Low-frequency words appearing in less than 0.2% of the documents, and high-frequency words occuring in more than 15% of the documents were then removed. Documents with less than 15 tokens were discarded at this point. The features in each instance were weighted using the TF/IDF weighting scheme. In order to account for documents of different lengths, each instance was normalised so that it was of unit length $\|d_{tfidf}\|_2 = 1$. This produced a dataset, **blog_c**, based on blog content with 7209 instances and 3857 tokens (see Table 1).

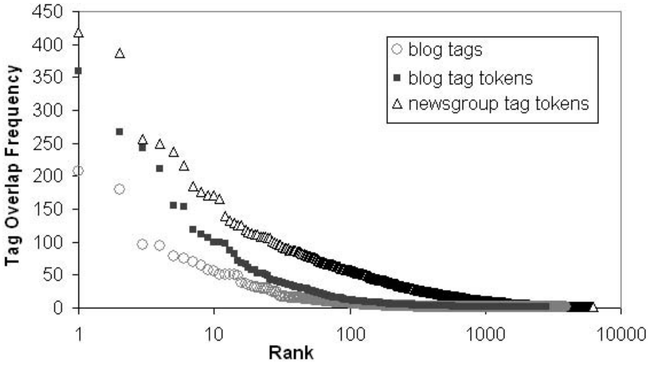

Figure 1: Tag frequency vs. tag rank by frequency for the set of blog tags, blog tag tokens and newsgroup tag tokens.

We created a second view of the data based on the *same* set of documents as blog_c but using the tag data of each document instead of the content. We tokenised each tag and performed stop-word removal and stemming as before. Furthermore, we removed 24 very commonly occurring tag tokens such as 'general' and 'random' (see Section 5.1). We then removed instances that did not share at least one token with at least one other instance in the dataset. This produced a dataset, **blog_t**, based only on tag tokens with 6064 instances and 1003 tokens (see Table 1). Each document in blog_t represents the tokens from a single tag used by a single user. The set of users in this dataset is a subset of the users in the 'content' dataset.

3.1 Newsgroup Data

The 20 newsgroup dataset is a popular *labelled* dataset for experiments in text classification and clustering[4]. In the newsgroup domain users post dated messages in a public place to which on-line readers may reply. However, newsgroups are logically organised into topics and contributors must select a relevant forum before writing, whereas blogs are blogger-centred and their authors write without fear of being censured for non-relevance. For comparison purposes we performed a parallel set of experiments on this dataset.

We based our dataset on 92% of the 18,826 posts, which were dated during a 4-week period from April 5 to May 3 1993. We extracted 11,139 email addresses from the headers of the posts, each of which acted as a unique user id. As

[1] http://www.technorati.com

[2] http://www.flickr.com

[3] http://www.del.icio.us

[4] http://people.csail.mit.edu/jrennie/20Newsgroups/

Dataset	# Instances	# Features	Mean # Feat.
blog_c	7209	3857	187.82
blog_t	6064	1003	1.27
news_c	7183	3248	80.6
news_t	7183	3152	5.5

Table 1: A summary of the datasets used for partitioning.

with the blog dataset, we created a document for each user containing the largest number of posts made to a single newsgroup during the 4-week period. Stop-words were removed and the remaining words stemmed. Too frequent and infrequent words were removed. At this point documents with less than 15 words were discarded. Each word was weighted according to the standard TF/IDF weighting scheme and the document vector was normalised so that it was of unit length $\|d_{tfidf}\|_2 = 1$. This produced a dataset, **news_c**, with 7183 instances and 3247 features (see Table 1). Each instance is associated with 1 out of 20 newsgroup class labels.

Unlike the blog dataset, the newsgroup dataset does not have a user-defined tag associated with the posts a user makes in a single newsgroup. Instead, each post has a short subject header designed to summarise the content of the post. We created a synthetic tag for the posts of a single user by aggregating the tokens from the subject header of each post, removing duplicate tokens and stop-words, and stemming. The remaining tokens were then weighted using TF/IDF and normalised. This produced a dataset, **news_t**, with 7183 instances and 3152 features (see Table 1).

4 Partitioning by Tags or Content

A simple way to recommend new blog posts would be to use the tag label of each post to retrieve posts by other bloggers with the same tag, an approach used by Technorati and analysed by Brooks and Montanez [2005]. However, a problem with this approach is illustrated by Figure 1: the frequency distribution for blog tag overlap follows a power law. From 7209 documents, there were 3934 tags of which only 563 (14%) were used 2 or more times, meaning that 86% of tags were useless for retrieval using an exact matching approach. However, by allowing a partial match between tag tokens the overlap is much greater, allowing us to make at least 1 match based on shared tag tokens for the 6064 documents in the blog_t corpus.

For our first experiment we use a text-clustering approach where we compare the partitioning produced using tag tokens, content tokens and random assignment. For the blog datasets we use only the 6064 blog instances found both in blog_c *and* blog_t. We implement the spherical *k*-means algorithm because of its efficiency in partitioning large corpora and its ability to produce cluster concept descriptions [Dhillon *et al.*, 2001]. In these experiments we do not address the issue of an optimal value of *k* as we are interested simply in the partitioning ability of tags compared to content. We cluster the content data and the tag data for both the blog and newsgroup datasets at values of *k* from 5 to 100. 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. As a baseline we create a partition of *k* clusters by randomly assigning instances.

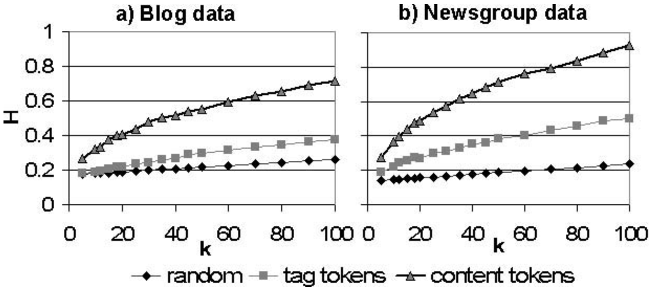

Figure 2: $\mathcal{H}$ scores at *k* for the blog and newsgroup datasets. $\mathcal{H}$ scores were measured based on the *content* tokens of the document instances.

Typically, clustering 'goodness' on unlabelled datasets is measured using criterion functions based on intra- and intercluster distance. Following Zhao and Karypis [2004], we use the ratio of intra- to intercluster similarity, $\mathcal{H}_r$. Intracluster similarity, $\mathcal{I}_r$, is the average of the cosine distances of each instance in the cluster to 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). $\mathcal{H}$ is the summation of the $\mathcal{H}_r$ contributions of each cluster where each $\mathcal{H}_r$ score is weighted by the fraction of the overall instances contained in the cluster. $|S_r|$ is the size of cluster *r*.

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

$$\mathcal{H} = \sum_{r=1}^{k} \frac{n_r}{n}\mathcal{H}_r \quad (2)$$

4.1 Discussion

Using the random clustering as a baseline, we can see from Figure 2a that clustering by blog tag tokens achieves 22% of the area under the curve achieved by the content clustering. Figure 2b shows results from the same experiment on the newsgroup dataset. Clustering using tag tokens performs considerably better for the newsgroup dataset, achieving 37% of the area achieved by clustering using content tokens. Our results would suggest that any recommendation service based on blog tag matching is likely to recommend/retrieve posts that are only marginally more relevant than a selection of posts chosen at random. We suggest that the better performance of the newsgroup tags is due to the fact that the subject header of a post is generally adopted by other users who choose to make a reply, whereas the distributed architecture of the blog domain means that users must select tags in an arbitrary fashion.

In order to confirm that the $\mathcal{H}_r$ is a good indicator of topic cohesiveness, we conduct a simple experiment on the labelled newsgroup dataset (news_c). At each value of *k* we measure the *purity* of each cluster, $\mathcal{P}_r$ i.e. the fraction of the cluster

Blogs				Newsgroups			
Partition	**auc**	**dfr**	**%**	**Partition**	**auc**	**dfr**	**%**
content	51.9	31.1	100	content	65.1	47.2	100
tags	27.6	6.8	21.9	tags	35.6	17.7	37.5
random	20.8	0	0	random	17.9	0	0

Table 2: A summary of the information in Figure 2. auc = area under curve. dfr = difference from random.

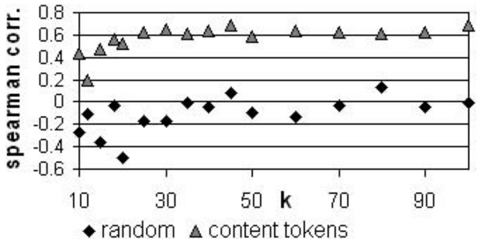

Figure 3: The Spearman correlation of $\mathcal{H}_r$ against $\mathcal{P}_r$.

made up of instances of a single class. At each value of *k* we correlate $\mathcal{P}_r$ against $\mathcal{H}_r$. Figure 3 illustrates a high positive correlation for all values of *k*, suggesting that $\mathcal{H}_r$ is indeed a good indicator of whether a cluster contains documents of a single class. Figure 3 also illustrates the low correlations achieved for clusters generated randomly.

5 Cluster Meta Labels and the $\mathcal{T}_r$ Score

In the previous section we suggested that blogs tags cannot usefully partition our datasets. In this section, we propose a *supporting* role for tags where clustering is carried out using content. We cluster the content datasets as before but this time we examine how the tag token data is distributed per cluster. For example, we cluster blog_c and then examine how the associated tokens in blog_t are distributed throughout the clusters. Figure 1 demonstrates that few tags are used very frequently and the majority of tags and tag tokens are used once. Partitioning the data using content clustering, we observe a tag token frequency distribution per cluster which 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 4 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%. We propose that the probability of distributed users independently using the same tag is higher in a cluster with a well defined topic (high $\mathcal{H}_r$) than in a cluster where the topic is weakly defined (low $\mathcal{H}_r$). In the next section, we define a score, $\mathcal{T}_r$, based on the proportion of high-frequency tag tokens in a cluster and we test the correlation of $\mathcal{T}_r$ against $\mathcal{H}_r$. We show that clusters with low $\mathcal{T}_r$ scores have poor semantic description in terms of the most highly weighted terms in the cluster centroid. More importantly, we demonstrate that $\mathcal{T}_r$ allows us to identify (and remove) clusters with spuriously high $\mathcal{H}_r$ scores.

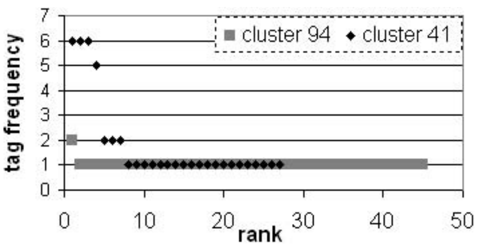

Figure 4: Tag token frequency distribution for cluster 41 (high $\mathcal{H}_r$) and cluster 94 (low $\mathcal{H}_r$).

id	Centroid keywords	$\mathcal{H}_r$	A-tags	$\mathcal{T}_r$
42	knit, yarn, sock, stitch, pattern	1.55	knit, sock, main, olympics, project	0.67
37	loan, estate, mortgage, credit, debt	1.50	loan, credit, real, estate, mortgage	1.11
44	interior, hotel, toilet, decoration, bathroom	1.50	interior, design	1.45
46	**jen, golf, rug, club, patent**	**1.01**	-	**0**
11	**www, http, br, similar, jan**	**0.9**	**link, web, business, blog, tech**	**0.16**
16	wordpress, upgrade, spam, fix, bug	0.89	blog, technology, tech	0.49
3	israel, iran, hamas, al, nuclear	0.87	politics, affairs, current, america, israel	0.84
24	muslim, cartoon, islam, danish, prophet	0.87	politics, religion, war, current, affair	0.59

Table 3: The top 8 clusters in terms of $\mathcal{H}_r$ scores from the blog_c dataset where *k*=50.

5.1 Tag Types

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 frequency distribution. **B-tags** are tag tokens with a frequency ≥2 that occur in ≥2 clusters. 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. Our experiments on the blog dataset identified a set of 24 b-tags which remained consistent even when we clustered the data on a weekly basis. **A-tags** are the remaining high-frequency tags. As Table 4 demonstrates, the average percentage of b-tag tokens in a cluster varies little for different values of *k*. On the other hand, as *k* increases there is a trade-off between a-tags and c-tags. As clusters become smaller, fewer co-occurring tags are found, decreasing the a-tag count and increasing the c-tag count. At *k* =50 on the blog dataset, only 24% of tags in a cluster, on average, are a-tags. Clearly, a-tags should contribute to cluster interperatability as they represent an independent description of the cluster topic by 2 or more users.

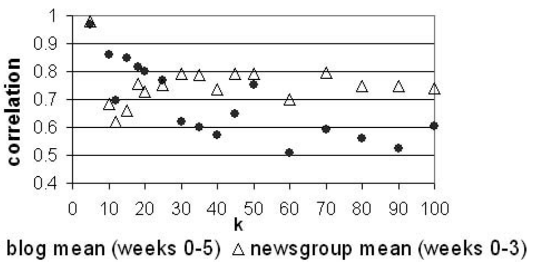

Figure 5: The mean correlation between $\mathcal{H}_r$ and $\mathcal{T}_r$ for the 6 blog and 4 newsgroup datasets.

Table 3 lists the top 8 clusters of the blog_c dataset ranked in terms of $\mathcal{H}_r$ where *k* = 50. Clusters are described in terms of the top 5 most highly weighted terms in the cluster centroid and in terms of their most frequent a-tags. In many cases the a-tag descriptors for each cluster provide clear meta-labels for each cluster. For example, cluster 16, which contains many documents on a recent upgrade to the Wordpress blog software, is described as ‘blog, technology, tech’. In some cases, such as cluster 42, a-tags provide contextual information not included in the keyword descriptions. The term ‘olympics’ in the a-tag list is due to the fact that 8 of the documents in this cluster refer to knitting while watching the Winter Olympics.

Two rows in Table 3, representing clusters 46 and 11, are marked in bold. Although the $\mathcal{H}_r$ scores of these clusters lie within the top 20%, the centroid keywords for these clusters do not impart a clear description of a topic. This is problematic as we would expect a high $\mathcal{H}_r$ score to imply a coherent topic description. In the next section we describe the $\mathcal{T}_r$ score, which is based on the proportion of a-tags in each cluster. We show how these problematic clusters can be created due to noise in the feature set and how the $\mathcal{T}_r$ score allows us to identify them.

5.2 The $\mathcal{T}_r$ score

The $\mathcal{T}_r$ score is the fraction of a-tags in a cluster, scaled by the cluster size. A_r, B_r and C_r are the (disjoint) sets of a-, b- and c-tags respectively found in cluster *r*. *n* is the number of instances in cluster *r*.

$$T_r = \frac{1}{|n|} \frac{\sum_{i=1}^{|A_r|} |a_i|}{\sum_{i=1}^{|A_r|} |a_i| + \sum_{i=1}^{|B_r|} |b_i| + \sum_{i=1}^{|C_r|} |c_i|} \quad (3)$$

In order to test that the correlation between $\mathcal{T}_r$ and $\mathcal{H}_r$ is consistent over time, we created 6 datasets, each dataset representing a week’s worth of data from the blogs represented in the blog_c dataset. Likewise, we divided the newsgroup dataset into 4 weekly datasets. We measured the correlation between $\mathcal{T}_r$ and $\mathcal{H}_r$ for each dataset at values of *k* from 5 to 100. We found that the correlation was high for all datasets and all values of *k*, suggesting a strong relationship between cluster strength and tag frequency. Figure 5 illustrates the mean correlation between $\mathcal{T}_r$ and $\mathcal{H}_r$ for the 6 blog datasets and the 4 newsgroup datasets at different values of *k*.

	Blog tags			**Newsgroup tags**		
k	**A**	**B**	**C**	**A**	**B**	**C**
5	0.47	0.17	0.36	0.78	0	0.22
10	0.38	0.20	0.42	0.73	0	0.27
20	0.33	0.20	0.47	0.62	0	0.38
50	0.24	0.21	0.55	0.57	0	0.43
100	0.17	0.21	0.62	0.4	0	0.6
250	-	-	-	0.34	0	0.66

Table 4: The table gives the mean fractions of a-tags, b-tags and c-tags per cluster at different values of *k*. The means are measured over the 6 windowed blog datasets.

However, examining the top 8 (out of 50) $\mathcal{H}_r$ scores shown in Table 3 we can see that the $\mathcal{T}_r$ and $\mathcal{H}_r$ scores appear to disagree strongly in several places. In fact, clusters 46 and 11 have $\mathcal{T}_r$ scores that lie within the bottom 20% of $\mathcal{T}_r$ scores for this clustering. This is not an anomaly: it can also be observed in each clustering of the 6 weekly blog datasets. Clusters 46 and 11 are also clusters whose top-weighted centroid key words provide us with no clear concept description. As text clustering is based on token-based matching of sparsely featured instances, without consideration for semantics, it is possible for clusters with high $\mathcal{H}_r$ scores but with poor semantics to emerge. This poses problems for an application such as a blog recommender system, where a misleadingly high $\mathcal{H}_r$ score would cause blogs to be recommended that have nothing in common.

One source of error is noise in the feature set. For example, on inspecting the instances in cluster 11 we found that the high $\mathcal{H}_r$ score was due to the matching of noisy token features such as ‘http’ and ‘www’, which were extracted from lists of URLs posted as free text. We suggest that the $\mathcal{T}_r$ score, because it measures human consistency in assigning descriptors to documents in a cluster, facilitates the identification of clusters that may in fact be meaningless.

5.3 Using $\mathcal{T}_r$ to Remove Noisy Clusters

As the blog dataset is unlabelled, our assessment of whether a cluster is meaningful or not is overly subjective. In this section we describe a more rigorous experiment using the labelled news_c and news_t datasets, where we measure the effect on cluster purity *after* we remove clusters that have a high $\mathcal{H}_r$ and low $\mathcal{T}_r$.

We limit the number of tokens per instance in news_t to the top TF/IDF weighted tag token. Thus, the average token frequency is comparable to the mean of 1.27 for the blog_t dataset (see Table 1). Secondly, because news_t has a much greater proportion of a-tags per cluster than the blog data, we raise the value of *k* to 250, which decreases the mean proportion of a-tags per cluster to 0.34, which is comparable with the mean observed in the blog dataset when *k* is between 20 and 50 (See Table 4).

We select the top 20% of clusters ranked according to $\mathcal{H}_r$ in descending order, which we term $\hat{H}$. We then select the bottom 20% of clusters ranked according to $\mathcal{T}_r$ in descending order, which we term $\check{T}$. The intersection of $\hat{H}$ and $\check{T}$ provides us with a set of candidate ‘weak’ clusters, $\mathcal{W}$. To de-

k = 250		**Purity**				
Test	$\|\mathcal{W}\|$	$\hat{H}$	$\mathcal{W}$	$\mathcal{R}$	$\hat{H}$-$\mathcal{W}$	$\hat{H}$-$\mathcal{R}$
1	9	0.58	0.35	0.72	0.62	0.55
2	6	0.65	0.32	0.85	0.68	0.64
3	4	0.62	0.44	0.86	0.63	0.61
4	8	0.63	0.39	0.71	0.66	0.62
5	4	0.64	0.28	0.66	0.66	0.64

Table 5: The purity results from 5 tests at k = 250 on the news_c dataset.

		Mean Purities				
k	$\|\mathcal{W}\|$	$\hat{H}$	$\mathcal{W}$	$\mathcal{R}$	$\hat{H}$-$\mathcal{W}$	$\hat{H}$-$\mathcal{R}$
250	6.3	0.63	0.38	0.76	0.65	0.61
300	11	0.63	0.39	0.75	0.66	0.61
350	15	0.64	0.42	0.72	0.67	0.63

Table 6: The mean purities for each value of k.

termine whether these candidates are really weak we use the cluster purity score, $\mathcal{P}_r$, described in Section 4.1. Firstly, we measure the mean purity of the clusters in $\hat{H}$. Then, we measure the mean purity score of $\hat{H}$ again, excluding the clusters in $\mathcal{W}$. If our hypothesis holds, we should observe an increase in purity as 'impure' clusters are removed from $\hat{H}$. We compare this to an experiment in which we remove a set of randomly selected instances, $\mathcal{R}$, from $\hat{H}$, where $|\mathcal{R}| = |\mathcal{W}|$. This experiment is conducted 150 times, 50 times each at k =250, 300 and 350 using randomly selected seeds for each clustering. Table 5 shows 5 results from the experiment conducted at k = 250. Table 6 shows the mean purities recorded for k =250, 300 and 350.

5.4 Discussion

From Table 6 we can see that the mean purity values for $\mathcal{W}$, the set of 'weak' clusters identified using $\mathcal{T}_r$, are considerably lower than the mean values for $\hat{H}$, the top 20% of clusters. At k=250, the mean purity of $\mathcal{W}$ is 0.6 of the mean purity of $\hat{H}$. By removing these clusters from $\hat{H}$ we see an overall increase in mean purity (as shown by the purity for $\hat{H}$-$\mathcal{W}$). This increase in purity is generally small because, on average, the number of clusters in $\mathcal{W}$ is small. However, we note that the increase was observed in all 150 tests. We tested the difference using a 1-tailed t-test and found that it was significant at the 0.05 alpha level. The key result of this experiment is that the $\mathcal{T}_r$ allowed us to automatically identify a subset of clusters with poor purity scores, which the standard $\mathcal{H}_r$ score could not identify. The straw man in this experiment was the random selection technique. In all tests $\hat{H}$-$\mathcal{W}$ recorded higher purity scores than $\hat{H}$-$\mathcal{R}$ and this difference was found to be significant at the 0.05 level.

6 Conclusions

Increasingly, tagging is being proposed as a decentralised alternative to Semantic Web standards. In the blog domain, however, we find that tags are rather poor at partitioning blog data. Using content-based clustering, we observe that a small proportion of users in every cluster have independently used the same tag tokens to describe his/her posts. The key observation is that the $\mathcal{T}_r$ score, a score based on tag frequency in a cluster, is an independent measure of agreement on the cohesiveness of a cluster topic and can be used to automatically identify weak clusters not identifiable using standard distance-based measures. Our future work in this area involves tracking and measuring topic drift in the blog domain. In particular we are devising tag-based scores which indicate the growth and decay of blog topics.

Acknowledgements

We would like to thank Natalie Glance from www.blogpulse.com for providing us with an initial set of blog URLs. This work is part of the PeRec project which is funded by a grant from the Provincia Autonoma di Trento.

References

[Brooks and Montanez, 2005] Christopher H. Brooks and Nancy 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.

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

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

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

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

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

[Quintarelli, 2005] Emanuele Quintarelli. Folksonomies: power to the people. *paper presented at ISKO Italy-UniMIB Meeting, Mi*, June 2005.

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

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