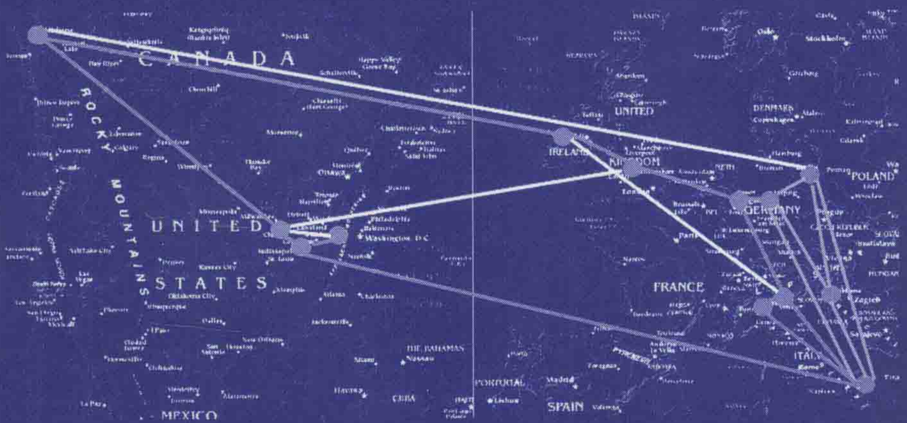
Bettina Berendt
Andreas Hotho
Dunja Mladenič
Giovanni Semeraro (Eds.)

LNAI 4737

From Web to Social Web: Discovering and Deploying User and Content Profiles

Workshop on Web Mining, WebMine 2006
Berlin, Germany, September 2006
Revised Selected and Invited Papers



Springer

Bettina Berendt Andreas Hotho
Dunja Mladenič Giovanni Semeraro (Eds.)

From Web to Social Web: Discovering and Deploying User and Content Profiles

Workshop on Web Mining, WebMine 2006
Berlin, Germany, September 18, 2006
Revised Selected and Invited Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Bettina Berendt
Institute of Information Systems
Humboldt University Berlin, Germany
E-mail: berendt@wiwi.hu-berlin.de

Andreas Hotho
KDE Group at the University of Kassel, Germany
E-mail: hotho@cs.uni-kassel.de

Dunja Mladenič
J. Stefan Institute, Ljubliana, Slovenia
E-mail: Dunja.Mladenic@ijs.si

Giovanni Semeraro
Department of Informatics
University of Bari, Italy
E-mail: semeraro@di.uniba.it

Library of Congress Control Number: 2007934911

CR Subject Classification (1998): H.2.8, H.3-4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-74950-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-74950-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12123469 06/3180 5 4 3 2 1 0

Preface

The World Wide Web is a rich source of information about human behavior. It contains large amount of data organized via interconnected Web pages, traces of information search, user feedback on items of interest, etc. In addition to large data volumes, one of the important characteristics of the Web is its dynamics, where content, structure and usage are changing over time. This shows up in the rise of related research areas like communities of practice, knowledge management, Web communities, and peer-to-peer. In particular the notion of collaborative work and thus the need of its systematic analysis become more and more important. For instance, to develop effective Web applications, it is essential to analyze patterns hidden in the usage of Web resources, their contents and their interconnections. Machine learning and data mining methods have been used extensively to find patterns in usage of the network by exploiting both contents and link structures.

We have investigated these topics in a series of workshops on Semantic Web Mining (2001, 2002) at the European Conference on Machine Learning / Principles and Practice of Knowledge Discovery from Databases (ECML/PKDD) conference series, in the selection of papers for the post-proceedings of the European Web Mining Forum 2003 Workshop, published as the Springer LNAI volume 3209 "Web Mining: From Web to Semantic Web" in 2004, as well as in the Knowledge Discovery and Ontologies workshop in 2004 and in the selection of papers for the post-proceedings of the ECML/PKDD 2005 joint workshops on Web Mining (European Web Mining Forum) and on Knowledge Discovery and Ontologies, published in 2006 as the Springer LNAI volume 4289 "Semantics, Web and Mining".

In 2006, we organized a workshop on Web mining that continues the aforementioned series of workshops on these topics. The workshop attracted a number of submissions and the highest-quality selected research papers, as well as the invited talk on "Web Usage Mining and Personalization in Noisy, Dynamic, and Ambiguous Environments" by Olfa Nasraoui (University of Louisville), fostered stimulating discussions among the participants. Specifically, the move from Web to Social Web (or Web 2.0) was an "emergent phenomenon" during the development of the workshop. The distinguishing mark of Social Web is user-generated content, which can play a key role if properly processed through advanced semantic technologies, such as text mining, natural language processing and image processing.

In fact, user-generated content represents a valuable source of information on users, in order to extract from content objects (bookmarks, blogs, photos, interaction logs, ...), relevant information about users (profiles) and the specific context in which they are interacting with a system, as well as to automatically annotate the content objects themselves and bootstrap the Semantic Web.

These topics were also investigated in the workshop "Ubiquitous Knowledge Discovery for Users" (UKDU) at ECML/PKDD 2006, which discussed the Web as one of today's most important ubiquitous environments. As the topics of that workshop complement the topics of our Web Mining workshop, this book also includes three invited and extended papers from the UKDU workshop.

Selected authors submitted expanded versions of their workshop papers. Those papers were reviewed again and the results of the selection were the eight papers chosen for this book.

The emergent phenomenon of Social Web and the widespread use of technologies such as Web logs, social bookmarking, wikis, RSS feeds are producing a significant change in Web usage. Understanding the dynamic of the relationship between topics and users in blogs, with the aim of constructing a plausible explanation for blogger behavior, is the main subject of the paper by Hayes, Avesani and Bojars. The paper proposes a set of measures to track topic and user drift, and shows how these measures can be used to explain user behavior. Collaborative environments are the basis of the Social Web. Flasch, Kaspari, Morik and Wurst consider the distributed organization of data employed in collaborative-filtering systems, which support users in searching and navigating media collections. They present Nemoz, a distributed media organizer based on tagging and distributed data mining.

The incorporation of semantics into the mining process is studied in two papers about Web usage mining. The invited contribution by Nasraoui and Saka provides a review of the recent efforts to incorporate content and other semantics to obtain a deeper representation of Web usage data, generally represented as a bag of clicks or URLs visited by a user. The paper examines the incorporation of simple cues from a Web site hierarchy in order to relate clickstream events that would otherwise seem unrelated. Facca concentrates on conceptual Web logs, that are XML documents enriched with information about the structure and content of the Web site. The paper shows how these logs can be automatically generated starting from a proper logging facility and a conceptual application model, and how this richer log representation allows one both to support the data mining process at different levels of abstraction and to analyze more easily the results of the mining process.

User profiles, as models of users' interests, play a key role in the recommendation of relevant content on the Web. Semeraro, Basile, de Gemmis and Lops describe a semantic recommender system able to provide the most interesting scientific papers to users according to their interests. The system learns semantic user profiles from documents represented using WordNet synsets. The hypothesis is that replacing words with synsets in the indexing phase helps learning algorithms to infer more accurate semantic user profiles. Anand and Mobasher, inspired by models of human theory developed in psychology, distinguish between users' short- and long-term interests; defining a recommendation process that exploits these two different models of users' interests. Often, the process of building user profiles relies on the analysis of digital data created or accessed by the users. The paper by Berendt and Kralisch focuses on other dimensions for understanding users' behavior: how language and culture may influence the way people access data and knowledge, and how these factors can be integrated into Web mining. A shift from technological to human aspects is needed for user-centered knowledge discovery, which deals with the ubiquity of people.

In the paper by Probst, Ghani, Krema, Fano and Liu, the authors propose an approach in which Web content (product descriptions) is processed in order to extract relevant attributes which can be used to describe items. The advantage of the approach is that it dynamically extracts attribute-value pairs, thus it differs from the classical information extraction task, in which a static template is filled in with relevant facts extracted from the text.

We thank our reviewers, the conference organizers, and the KDubiq project for sponsoring and support.

July 2007

Bettina Berendt
Andreas Hotho
Dunja Mladenic
Giovanni Semeraro

Organization

Web Mining (WebMine) 2006 was organized as part of the 17th European Conference on Machine Learning (ECML) and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD).

Workshop Chairs

Bettina Berendt	Institute Information Systems Humboldt University Berlin, Germany
Andreas Hotho	KDE Group at University of Kassel Kassel, Germany
Dunja Mladenic	J. Stefan Institute Ljubljana, Slovenia
Giovanni Semeraro	Department of Informatics University of Bari, Italy

Program Committee

Sarabjot Anand	University of Warwick, UK
Mathias Bauer	DFKI, Germany
Janez Brank	J. Stefan Institute, Slovenia
Michelangelo Ceci	University of Bari, Italy
Marco de Gemmis	University of Bari, Italy
Miha Grcar	J. Stefan Institute, Slovenia
Marko Grobelnik	J. Stefan Institute, Slovenia
Pasquale Lops	University of Bari, Italy
Bamshad Mobasher	DePaul University, USA
Ion Muslea	Language Weaver, Inc., USA
Myra Spiliopoulou	Otto-von-Guericke-Univ. Magdeburg, Germany
Gerd Stumme	University of Kassel, Germany
Maarten van Someren	University of Amsterdam, Netherlands

Additional Reviewers

P. Basile
I. Palmisano
M. A. Torsello
D. Truemper

Table of Contents

An Analysis of Bloggers, Topics and Tags for a Blog Recommender System

Conor Hayes[1], Paolo Avesani[2], and Uldis Bojars[1]

[1] Digital Enterprise Research Institute,
National University of Ireland, Galway, Ireland
conor.hayes@deri.org, uldis.bojars@deri.org,
[2] ITC-IRST,
Via Sommarive 18
38050 Povo (Trento), Italy
avesani@itc.it

Abstract. Over the past few years the web has experienced an exponential growth in the use of weblogs or *blogs*, web sites containing journal-style entries presented in reverse chronological order. In this paper we provide an analysis of the type of recommendation strategy suitable for this domain. We introduce measures to characterise the blogosphere in terms of blogger and topic drift and we demonstrate how these measures can be used to construct a plausible explanation for blogger behaviour. We show that the blog domain is characterised by bloggers moving frequently from topic to topic and that blogger activity closely tracks events in the real world. We then demonstrate how tag cloud information *within* each cluster allows us to identify the most topic-relevant and consistent blogs in each cluster. We briefly describe how we plan to integrate this work within the SIOC[1] framework.

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 [14] 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). Tag clouds refer to aggregated tag information, in which a taxonomy or 'tagsonomy' emerges through repeated collective usage of the same tags.

[1] www.sioc-project.org — Semantically-Interlinked Online Communities.

B. Berendt et al. (Eds.): WebMine 2006, LNAI 4737, pp. 1–20, 2007.

In previous work we presented an empirical evaluation of the role for tags in providing organisational support for blogs [6]. 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.

We concluded 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 produced multiple topic-related or *local* tag clouds, which could provide discriminating secondary information to further refine and confirm the knowledge produced by the clustering. Furthermore, local tag clouds established topic-based relationships between tags that were not observable when considering the global tag cloud alone.

This work was motivated by the need to build a blog recommender system in which a registered blogger would be regularly recommended posts or tags by other bloggers with similar interests. In such systems a key decision is how often the neighbourhood set or clustering needs to be calculated [12]. If similar users at time t are no longer similar at time $t+1$, models derived from data at time t may become obsolete very quickly.

We suggest a set of measures to track topic and user drift and we provide an explanation of topic evolution with reference to independently observed news events during the clustering period. Our initial results would suggest that many bloggers tend to have a short-lived attachment to a particular topic, which means that the neighbourhood relationships produced by each clustering cycle are relevant for a short period of time.

We then refine this analysis using information derived from the tag usage in each cluster. We find that blogs that contribute to the local tag definition of each cluster tend to be the most relevant in each cluster and, importantly, tend to be clustered together for extended periods. This behaviour suggests that topics uncovered by clustering have a core of relevant blogs surrounded by blogs that move between topics on a regular basis. In terms of defining a recommendation strategy, clustering followed by tag analysis allows us to define topics and potential authorities for those topics.

We briefly describe our current work which involves allowing the knowledge produced by automated learning techniques to be exported and reused using the SIOC (Semantically-Interlinked Online Communities) framework.

In the next section we give an overview of related work. Section 3 describes the datasets we use in this paper. Section 4 introduces our clustering method and the criteria we use for assessing cluster quality. In Section 4.2 we summarise our work on refining clusters using tag analysis. In Section 5 we introduce our experiments for tracking the relationship of users to topics as clustering is carried out on 6 data sets, each representing a week's worth of blog data. In Sections 5.1 and 5.2 we suggest a set of measures to track user and topic drift, and using these measures we provide an explanation of topic evolution in a cluster with reference to independently observed news events. In Section 6, we demonstrate how relevant sources of consistent topic-relevant information can be identified using simple tag analysis. We briefly describe our future work in Section 7 which involves integrating the information produced using knowledge discovery techniques with the SIOC framework. We present our conclusions in Section 8.

2 Related Work

The Semantic Web project has facilitated several initiatives concerned with linking and integrating topic-related material on the Web. For example, the SIOC framework facilitates the connection and interchange of information from Internet-based discussions and forums such as blogs, mailing lists, newsgroups and bulletin boards [1].

Tagging is a 'grassroots' solution to the problem of organising distributed web resources, with emphasis on ease of use. Quintarelli [10] 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' or inferring or describing relationships between tags.

Although tagging is widely used by blog users, its effectiveness as a primary organising mechanism has not been demonstrated [2,6]. 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[2], the photo-sharing site, Flickr[3], and the social bookmarks manager, Del.icio.us[4], 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 [6]. Golder and Huberman [5] 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

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

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

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

the majority on the Web [7]. Link analyses on our datasets have produced the same results. For this reason we do not consider links between blogs in this paper.

3 Blog Data Sets

Our blog data set is based on data collected from 13,518 blogs during the 6-week period between midnight January 15 and midnight February 26, 2006[5]. All blogs were written in English and used tags. We found that 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 spammers ('sploggers') or community blogs. We selected data from 7209 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 6-week period.

For each blog we selected the posts from the most frequently used tag during the 6-week period. This allowed us to associate a single topic (as defined by the blogger's tag) with each of the 7209 blogs. We chose to examine one topic per blog because blog topics from a single blog are often similar, as the blogger may use multiple tags for each post. Thus each of the 7209 blog 'documents' constitutes a single topic from a single blogger from the 6-week period.

The data was divided up into 6 data sets, each representing post data from a single week. As all 7209 bloggers do not post every week, the data sets have different sizes and overlap in terms of the blog instances they contain (see Table 1). Each instance in a data set 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 data set for week 4. As shown in Table 1, on average, 71% of the blogs present in the data set win_t will also be present in the data set win_{t+1}.

We processed each data set 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 condsidered at this point. Each word was weighted according to the standard TF/IDF weighting scheme and the document vector normalised by the L^2 norm. This created a feature set of approximately 3,500 words for each data set. Table 1 gives the window period, size and overlap with the subsequent window.

[5] The blog URLs were kindly supplied by Natalie Glance of www.blogpulse.com

Table 1. The periods used for the windowed blog data set. Each period is from midnight to midnight exclusive. User overlap refers to the overlap with the same users in the data set for the next window.

data set	Dates (2006)	Size	# Feat.	Mean Feat.	Overlap win_{t+1}	%
win_0	Jan 16 to Jan 23	4163	3910	122	3121	75
win_1	Jan 23 to Jan 30	4427	4062	123	3234	73
win_2	Jan 30 to Feb 6	4463	4057	122	3190	71
win_3	Feb 6 to Feb 13	4451	4124	122	3156	71
win_4	Feb 13 to Feb 20	4283	4029	122.	2717	63
win_5	Feb 20 to Feb 27	3730	4090	121	-	-
mean	-	**4253**	**4043**	**122**	**3084**	**71**

4 Clustering and Tags

The blog domain contains many 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 [15]. 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 [11], browsing large document collections [3], organising search engine return sets [16] and grouping similar user profiles in recommender systems [13,9,8].

As our objective was to analyse user behaviour using a clustering solution, we implemented the *spherical k-means* algorithm, a well understood variation of the k-means clustering algorithm that scales well to large document collections and produces interpretable cluster summaries [4]. Spherical k-means produces k disjoint clusters, the centroid of each being a concept vector normalized to have unit Euclidean norm.

4.1 Clustering Quality

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 [17], 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 data set, C (see Equation 1).

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

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 [6]

4.2 Partitioning by Tags or Clustering

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. This is an approach used in a global tag cloud view of the blog domain. Tag clouds refer to aggregated tag information, in which a taxonomy or 'tagsonomy' emerges through repeated collective usage of the same tags.

Part A of Figure 1 illustrates this view of our blog data set. By clicking on a tag, the recent posts labelled with that tag are retrieved.

However, in any system where tags are aggregated, few tags are used very frequently and the majority of tags are used infrequently. This Zipfian tag-frequency

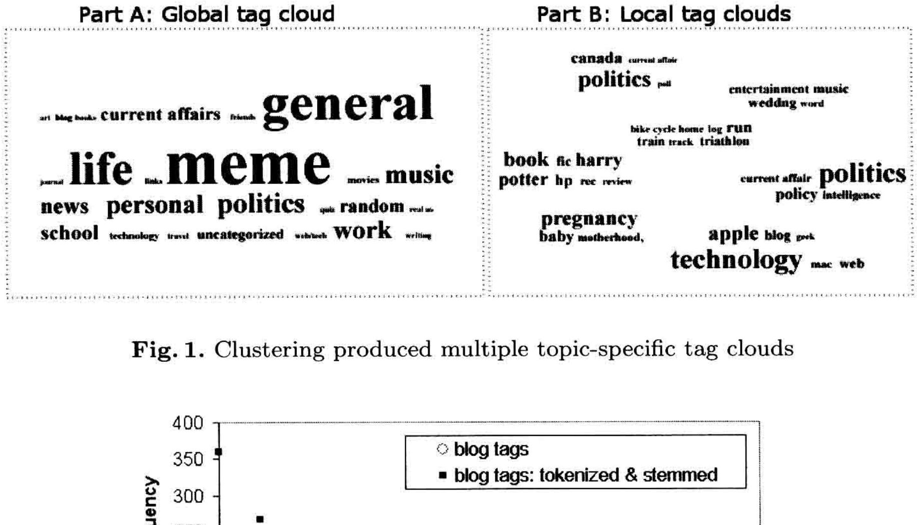


Fig. 1. Clustering produced multiple topic-specific tag clouds

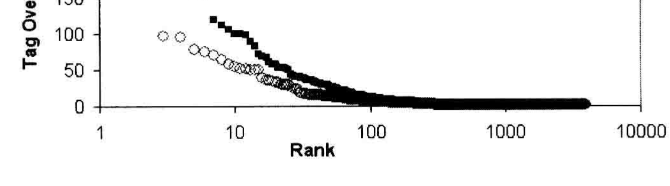


Fig. 2. Tag frequency vs. tag rank by frequency for the set of blog tags and blog tag tokens

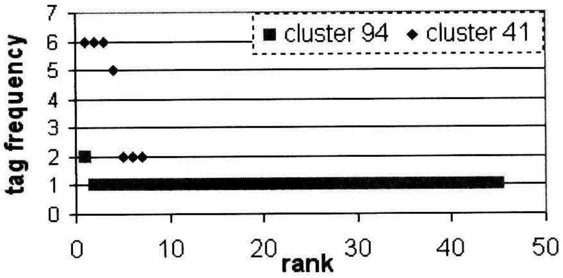


Fig. 3. Tag token frequency distribution for cluster 41 (high $\mathcal{H}_r$) and cluster 94 (low $\mathcal{H}_r$)

distribution means that only a small proportion of tags at any time can be used for retrieval purposes. Out of the 7209 documents in our data set only 563 (14%) out of 3934 tags were used 2 or more times, meaning that 86% of tags were useless for retrieval using an exact matching approach. This distribution is illustrated in Figure 2 where the circle icons represent 'raw' tag data and the square icons represent tags that have been tokenised and stemmed.

In previous work we demonstrated that tags generally performed poorly in comparison with clustering by content in identifying coherent topics in our blog corpus [6]. Furthermore, clustering by content partitioned the global tag space, producing multiple topic-related tag clouds as illustrated by Part B of Figure 1. In this view, the aggregated tag data in each cluster produced relationships between tags, which were not visible in the global view, and produced topic descriptions in the form of local tag clouds.

A key observation was that the tag frequency distribution per cluster varied according to cluster strength ($\mathcal{H}_r$). Weak clusters tended 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 tended to contain many high-frequency tags and a shorter tail.

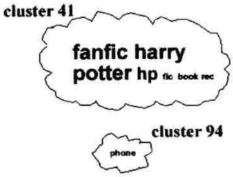


Fig. 4. The tag clouds for cluster 41 (high $\mathcal{H}_r$) and cluster 94 (low $\mathcal{H}_r$)

Figure 3 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 4 illustrates the tag cloud for each cluster based on these distributions. The tag cloud description of Harry Potter fan fiction shown in Figure 4 could not have been identified within the typical global tag cloud.

We refer to tag tokens that are not repeated by any other user in the cluster as **C-tags**. These tags are represented by the long tail of the frequency distribution and are not represented in the tag cloud view. **B-tags** are tag tokens with a frequency ≥ 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.

Combining clustering with subsequent tag analysis has allowed us to automatically identify and remove semantically weak clusters and to produce interpretable topic descriptions using local tag clouds [6].

5 Tracking User and Topic Drift

However, using clustering and tags on a static data set ignores the dynamic nature of the blogging domain. Blog data should be viewed as a stream of information, which we need to categorise and from which we need to extract the most relevant sources of information. The clustering solution we have described clusters blogs together by virtue of their similarity at a particular point in time. As bloggers continue to add new posts to their blogs, a key question is whether the relationships established by a clustering solution will be valid in the next time frame. Another key question is how the most relevant and consistent blogs associated with a particular topic can be identified.

In the following sections we attempt to make these questions clearer by measuring user and topic drift in our blog data over time. In the final section, we will turn again to tag analysis to allow us to identify bloggers that are consistently relevant to a given topic.

In these experiments 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 interseed 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

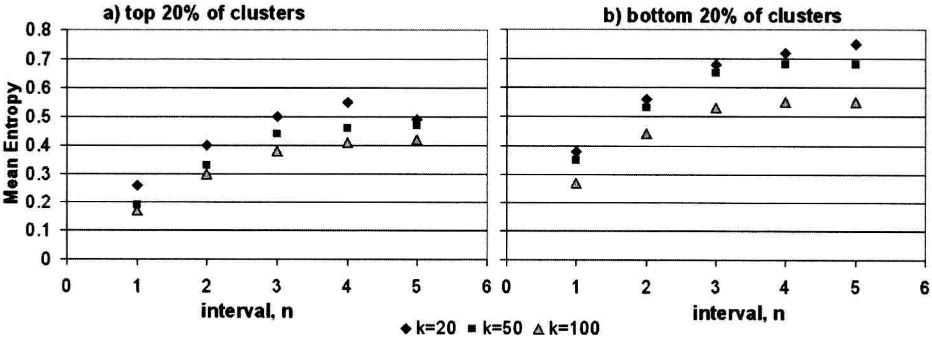


Fig. 5. Mean user entropy recorded where the intervals between windows vary from 1 to 5. The diagram on the left gives the entropy recorded for the top 20% of clusters according to $\mathcal{H}_r$. The diagram on the right gives the user entropy for the bottom 20% of clusters.

of the current week. In each pair of adjacent windows, the feature set overlap between windows is greater than 95%. The feature values for each seed are the feature weights from the corresponding centroid in the previous week.

In order to compare clustering in adjacent windows we define the following measures: *user entropy* per cluster, $\mathcal{U}_r$, and *interwindow similarity* per cluster, $\mathcal{W}_r$. 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 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 $|c_{r,t} \cap c_{i,t+1}|$, the number of users from cluster $c_{r,t}$ contained in $c_{i,t+1}$.

The interwindow score, $\mathcal{W}_r^{t+1}$, for a cluster r in window win_t, is the similarity between the centroid of cluster r and the centroid of the corresponding cluster r in window win_{t+1}. Likewise, $\mathcal{W}_r^{t-1}$ is the similarity between the centroids of cluster r at windows win and win_{t-1}. Intuitively, $\mathcal{W}_r^{t+1}$ is a measure of the drift of the centroid concept, C_r, at win_t, where C_r is also the seed for cluster r at win_{t+1}.

$$\mathcal{W}_r^{t+1} = \cos(C_{r,t}, C_{r,t+1}) \tag{3}$$