Feeds That Matter: A Study of Bloglines Subscriptions

Akshay Java, Pranam Kolari, Tim Finin, Anupam Joshi, Tim Oates
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
{aks1, kolari1, finin, joshi, oates}@umbc.edu

Abstract

As the Blogosphere continues to grow, finding good quality feeds is becoming increasingly difficult. In this paper we present an analysis of the feeds subscribed by a set of publicly listed Bloglines users. Using the subscription information, we describe techniques to induce an intuitive set of *topics* for feeds and blogs. These topic categories, and their associated feeds, are key to a number of blog-related applications, including the compilation of a list of *feeds that matter* for a given topic. The site FTM! (Feeds That Matter) was implemented to help users browse and subscribe to an automatically generated catalog of popular feeds for different topics.

1. Introduction

Blogs have become a means by which new ideas and information spread rapidly on the Web. They discuss the latest trends and react to events as they unfold around the world. Protocols such as RSS, ATOM and OPML and services such as Blog search engines and ping servers have made it much easier to share information online. RSS and ATOM are XML-based file formats used for syndication. Outline Processor Markup Language (OPML) is a popular XML based format used to share an outline of the feed subscriptions.

Today, the feed infrastructure provided by RSS and ATOM is being used to serve a wide variety of online content, including blogs, wikis, mainstream media, and search results. All support different forms of syndication. Users can subscribe to feeds using reader such as Bloglines, Google Reader, NewsGator, etc. Typically, a user adds a feed in a feed reader when she came across it (perhaps, by chance) as a reference on another blog. This is not always the best way to find good feeds.

A number of blog search engines and some hand-crafted directories try to provide a high quality index of feeds. Blog search engines such as Technorati have introduced new features enabling people to find authoritative feeds on a given topic. The blog finder feature works by relying on the author of the blog to provide the tags. Further it ranks the blogs based on the number of inlinks. These problems make it insufficient in terms of finding topically authoritative blogs.

Hand-crafted directories have the disadvantage that they are based on the decision of the site creator. Additionally, there are only a limited set of sites that one can categorize manually. Recent efforts in tackling these problems have resulted in *Share your OPML*, a site where you can upload an OPML feed to share it with other users. This is a good first step but the service still does not provide the capability of finding good feeds topically.

An alternative is to search for blogs by querying blog search engines with generic keywords related to the topic. However, blog search engines present results based on the *freshness.* Query results are typically ranked by a combination of how well the blog post content matches the query and how recent it is. Measures of the blog's authority, if they are used, are mostly based on the number of inlinks. These factors make it infeasible to search for new feeds by querying blog search engines. As Michael Arrington points in a recent post [2], this can be misleading since a single post from a popular blogger on any topic may make him the top-most blog for that topic, even if his blog has little to do with the given subject.

Finding high-quality and topically authoritative feeds remains a challenge.

In this work we study the feed subscriptions of a large sample of Bloglines publicly listed users. Using this data, we first characterize the general feed usage patterns. Next, we identify the feeds that are popular for a given topic using folders names as an approximation for a topic. By merging related folders we can create a more appropriate and compact set of topics. Finally, we discuss some of the preliminary results in using this approach in support of a number of blog-related applications: feed browsing, feed recommendations, and searching for influential blogs in a different dataset.

The paper is organized as follows: Section 3 describes the dataset used in this study and the statistics gathered from the data. Section 4 discusses the techniques we used to rank blogs within topics and to group related topics together. Section 5 describes two simple applications – a feed recommendation system and an algorithm that uses the data to predict feed influence. In Section 6 we present the initial results of an evaluation of both the quality of a folder merging algorithm and the feed recommendations. Section 2 connects this work to related research in this area. Finally we conclude in Section 7 with discussions of the broader implications of this work and the direction of our future research.

2. Background and Related Work

Blog hosting tools, search services and Web 2.0 sites such as Flickr and del.icio.us have popularized the use of tags. Tags provide a simple scheme that helps people organize and manage their data. Tags across all the users, collectively, are termed as a *folksonomy* [21], a recent term used to describe

ICWSM'2007 Boulder, Colorado, USA

this type of user-generated content. Tags are like keywords used in the META tag of HTML. Adam Mathes [16] suggests that there are two reasons why people may use tags: to classify information for themselves or to help a community of users.

Brooks and Montanez [4] have studied the phenomenon of user-generated tags to evaluate effectiveness of tagging. Their study presents an analysis of the 250 most frequently used Technorati tags. Brooks et al. find that tagging can be helpful for grouping related items together but does not perform as well as text clustering. A text-based hierarchical clustering was used to group related tags together. We study similar problems, with the aim of finding important feeds for a topic. By using a dataset that is based on feed subscriptions rather than text in individual posts, we can group similar feeds together. Another study of a social bookmark tool, del.ici.ous, by Cattuto et al.[5], presents an analysis of collaborative tagging. Their research indicates that a common vocabulary emerges across user-generated tags and they also provide a stochastic model that approximates this behavior.

Shen and Wu [24] treat tags as nodes and the presence of multiple tags for a document as a link between the tags. According to Shen, the network structure and properties of such a graph resemble that of a scale-free network. In our analysis, we study the different usage and subscription characteristics of feeds and find that some of these features also follow a power law distribution. While such distributions would not be surprising anymore, it is interesting to note that while the total number of blogs are increasing, the feeds that matter are actually just a small portion of the Blogosphere.

Guy and Tonkin [9] discuss the issue of cleaning up the tag space. Their study of del.icio.us and Flickr tags found that a significant number of tags are misspelled. User enforced hierarchies created with tags separated by special characters accounts for a portion of the tag space. The biggest advantage of folksonomies is that it gives people the flexibility to label content using any terms that they find appropriate. Enforcing a set of rules or suggesting tag selection guidelines is helpful but not easy to implement. In this paper we propose an alternative, where variations of tag or folder name usage can automatically be inferred through merging related tags. This allows users to continue creating their own tags, while improving topical relevance of systems using this information.

An alternative suggested to improve the quality of tagging is AutoTagging [19]. Social bookmark tools like del.icio.us already provide suggestions for tagging a URL based on terms used to describe the same link by other users in the system. AutoTagging is a collaborative filtering based recommendation system for suggesting appropriate tags. The suggested tags are based on tags used for other posts that are similar in content. This work does not directly address AutoTagging but we describe a similar approach for a slightly different motivation - finding topically authoritative feeds and recommending new feeds based on tag usage similarity.

Dubinko et al. [7] describe tag visualization techniques by using Flickr tags. Their work concentrates on automatically discovering tags that are most 'interesting' for a particular time period. By visualizing these on a timeline they provide a tool for exploring the usage and evolution of tags on Flickr. In this work we take only a static view of feed subscriptions and folder usage. Feed subscriptions unlike flickr or technorati tag clouds evolve rather slowly and hence taking a static view of the data is not too unrealistic.

Marlow [15] compares blogroll links and permalinks (URLs of specific blog post) as features to determine authority and influence on the Blogosphere. The study suggests that permalink citations can approximate influence. Present blog search engines indeed use permalink citations or inlinks to a blog as a measure of authority. The disadvantage is that such measures do not work well when the goal is to find authoritative blogs in a particular topic. In our approach, we treat folder names as an approximation of topic and number of subscribers as an indication of the authority. We find that such measures are effective in finding topically authoritative blogs.

3. Analysis of Bloglines Subscriptions

Bloglines is a popular feed reader service. Using this tool makes it easy to monitor a large number of RSS feeds. Once a user subscribes to a set of feeds, this service monitors the subscriptions and allows the user to view unread posts from their subscribed feeds. The simple user interface and convenient feed monitoring ability have made Bloglines an extremely popular feed reader. Bloglines provides a feature wherein users may choose to share their subscriptions. We conduct a study of the publicly listed OPML feeds from 83,204 users consisting of a total of 2,786,687 subscriptions of which 496,879 are unique. These are essentially the *"feeds that matter"* [14] since they are feeds that people have actually subscribed to. Table 1 shows the distribution of the top domains in the Blogines dataset. In particular, there are a number of users who subscribe to Web 2.0 sites and dynamically generated RSS feeds over customized queries. It was also interesting to note that even though Blogspot has had serious splog issues [20, 13], based on the Bloglines dataset, it still contributes to a significant portion of the feeds that really matter on the Blogosphere.

Domain	Percentage	domain	Percentage
blogspot	24.36	hatena	1.07
livejournal	3.81	topix	0.89
flickr	2.89	technorati	0.75
msn	1.73	wretch	0.56
typepad	1.73	exblog	0.54
yahoo	1.71	wordpress	0.47
xanga	1.43	msdn	0.45
icio	1.24	blogs	0.45
google	1.22	rest	53.60
livedoor	1.10		

Table 1: *The distribution of domains in the Bloglines dataset*

According to Bloglines/Ask in July 2005 there were about 1.12 Million feeds that really matter, which is based on the feeds subscribed by all the users on Bloglines. A study of the feeds on Bloglines by McEvoy [17] in April 2005 showed that there were about 32,415 public subscribers and their feeds accounted for 1,059,140 public feed subscriptions. We collected similar data of the publicly listed users on Bloglines. From last year, the number of publicly listed subscribers had increased to 83,204 users (2.5 times that of last year) and there were 1,833,913 listed feeds (1.7 times) on the Bloglines site. Hence, even though the Blogosphere is almost doubling every six months [25], we found that the number of feeds that *"really matter"* doubles roughly every year. Inspite of this, popularly subscribed feeds are still only a small fraction of the entire Blogosphere. Following is a description of some of the usage patterns and interesting statistics obtained from our analysis.

Figure 3 shows the distribution of the number of subscribers for 496,879 unique feeds across 83,204 users. This graph indicates a typical power law behavior with a few feeds having a large number of subscribers while most having a small number of subscribers. The exponent of the curve was found to be about -2.1 which is typical in scale-free systems and WWW [1]. While the presence of a power law distribution across feed subscription is expected, it is interesting to observe that even across a large sample of users, the number of unique feeds subscribed is fairly small in comparison to the 53 Million blogs on the Blogosphere [25].

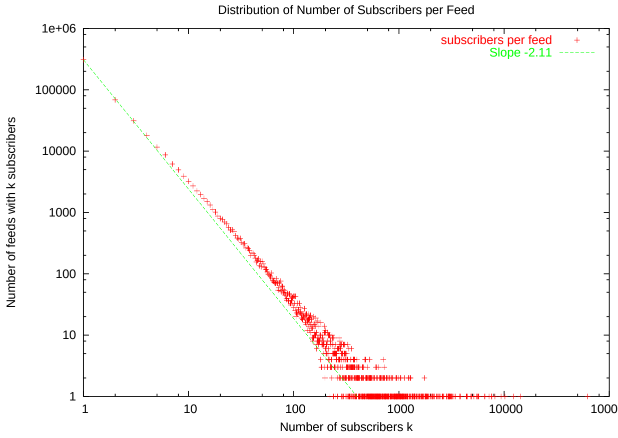

Fig. 1: *The number of subscribers for feeds follows a power law distribution.*

Next, we analyzed the number of feeds subscribed per user. The number of subscribers for a feed is an indication of its authority and influence over its audience. Figure 2 depicts the distibution of the number of feeds subscribed across all users. Almost 90% of the users have less than 100 subscriptions. It is possible that for most users there is an inherent limit on the amount of information that they can keep track of at any given time. This limits their attention and hence the number of feeds that they typically subscribe to.

Bloglines has a feature by which a user may organize their feeds into different folders. While only some (26,2436 or about 35%) of the public subscribers use folders, it provides a user generated categorization of feeds. Figure 3 shows the histogram of folder usage across all users. While the folder organization is not a very commonly used feature, most users who do use them have a relatively small number of folders. A vast majority of users had only one folder - named ‘subscriptions’ folder created by default for all users. Almost 90% of users have less than 10 folders and only roughly 100 users had more than 100 folders. Figure 4 shows a scatter plot of the number of folders compared to the number of feeds subscribed across all users. Although there is a very high variance, it can be observed from this graph that as the number of feeds subscribed increase, users generally organize them into greater number of folders.

Figure 5 shows the folder usage across all subscriptions. Each folder is ranked by the number of distinct feeds that have been categorized into that particular folder. It can be observed that the highly ranked folders are also those that are used by many subscribers. Thus the usage pattern suggests a consensus based on a folksonomy [1] emerges and a common vocabulary is being used to tag the feeds.

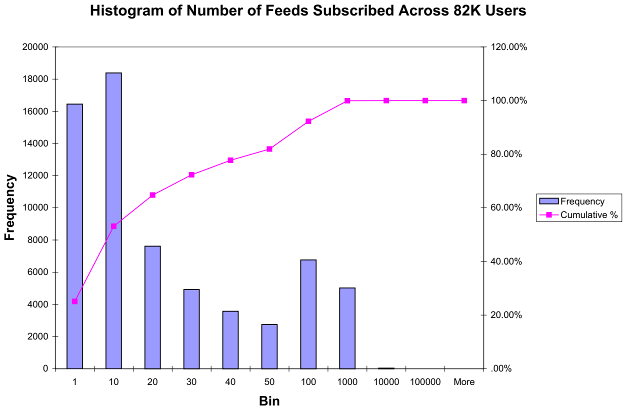

Fig. 2: *The Histogram of feeds subscribed per user*

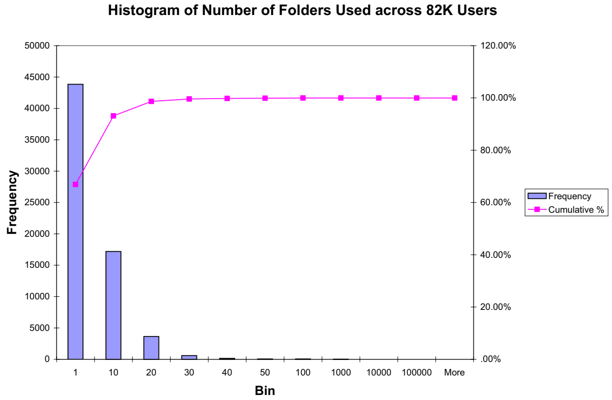

Fig. 3: *The histogram of folder usage across all users*

4. Grouping Related Topics

Folder names can be treated as an approximation of a topic. Folder names in Bloglines are used in a way that is similar to Folksonomies on the web. As shown in Figure 6, by aggregating folders across all users, we can generate a tag cloud that shows the relative popularity and spread of various topics across Bloglines users. The tag cloud shown here is based on the top 200 folders. Note that the tag cloud contains terms such as ‘humor’ and ’humour’, etc. These terms represent variations in which different users label feeds. By merging folder names that are *‘related’* we can generate a more appropriate and compact representation of the tag cloud. Automatic techniques for inferring concept hierarchies using clustering [23] WordNet [18] and other statistical methods [8] have been found to be effective in finding relationships between topics.

The following section describes an approach used to merge

[1] http://en.wikipedia.org/wiki/Folksonomy

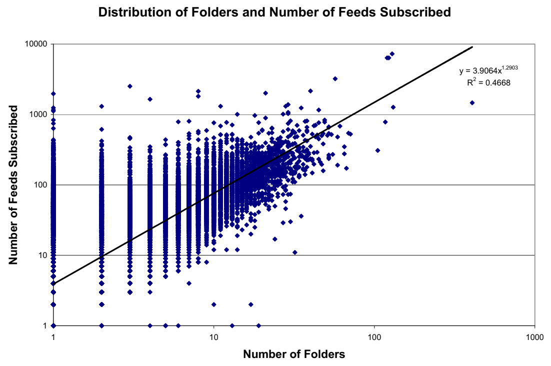

Fig. 4: *Scatter plot showing the relation between the number of folders and number of feeds subscribed. Note: This includes the feeds subscribed under the default folder labeled 'Subscriptions'*

related folders together. We were first tempted to use a morphologicial approach – merging the *blog* and blogs categories, for example. However, we soon discovered that folders with lexically similar names might actually represent different categorization needs of the users. For example, the folder 'Podcasting' consists of feeds that talk about how to podcast and provide tools. On the other hand 'Podcasts' refers to feeds containing actual podcasts. Other examples include 'Music' vs. 'Musica' (a topic with Spanish music blogs).

For each folder we construct a vector containing the feeds that have been categorized under that folder name and their corresponding counts. At this step we take only the top 100 most frequently occurring feeds per folder. This threshold was heuristically determined. Some folders, such as 'friends', were observed to consist of a large set of feeds for each of which there are only a handful of subscribers. On the other hand extremely popular folders like 'politics' contained a number of feeds that have many subscribers.

Two cases need to be considered for computing folder similarity: first is the case where feeds in one folder may either partially or completely subsume feeds present in another folder. Complete subsumption indicates that there is a broader category and the larger folder is more general while partial subsumption indicates that the two categories are related. For example the folder 'news' subsumes a number of folders that are more specific, such as 'tech news','IT news', 'general news', etc. For detecting the topics, it suffices to put these into a single category titled 'news'. To compute subsumption we first find an overlap factor. For all folder pairs i,j we maintain a score of the overlap of feeds in folder j with feeds in folder i as follows:

$$overlap = \frac{matches}{size_j}$$

Folder similarity can be described in terms of the feeds that are contained in the folders. Two folder names are considered to be similar if they contain similar feeds in them. For each pair of folder names we compute the cosine similarity as

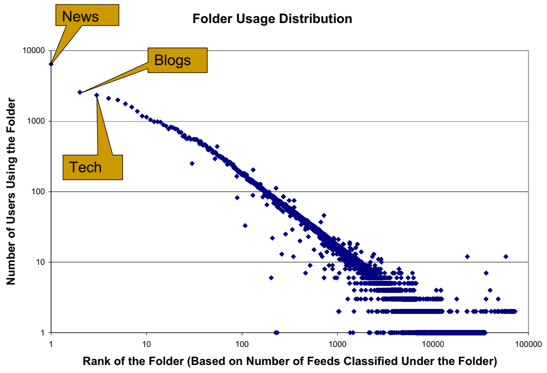

Fig. 5: *Folder usage distribution. The rank of the folder is computed by the number of distinct feeds that are categorized under that folder name.*

follows:

$$cos(i,j) = \frac{\sum_k W_{i,k} W_{j,k}}{\sqrt{\sum_k W_{i,k}^2 * \sum_k W_{j,k}^2}}$$

The weight $W_{i,k}$ is determined by a TFIDF score for each feed in the vector. The weights are computed using the following formula:

$$W_{folder}(feed) = freq_{folder}(feed) * log(\frac{folderCount}{|foldersContaining(feed)|})$$

First we start by ranking the folders based on the number of users using the folder. Next we go through this ranked list and merge related folders together. A lower ranked folder is merged into a higher ranked folder if $overlap > \theta$ or $cosine > \delta$. These thresholds were empirically set to 0.4. Setting it to smaller values leads to lowering the criteria for grouping, resulting in fewer topics and higher values resulted in fewer groupings resulting in more topics. Table 2 shows examples of folders names with the corresponding merged folders.

Once the merging of related folders is completed, a list of feeds relevant and authoritaive for a topic can be created. This task is similar to automatic resource compilation [6] which aims to find authoritative Web resources by analyzing the link structure. In our approach we say that a feed is topically relevant and authoritative if many users have categorized it under the given folder name. After merging related folders together, the total number of times each of the feeds appears across all the merged folders is added to obtain the final ranking. Tables 4 and 3 provide examples of feeds that matter for "Photography" and "Politics" that were found using this technique.

5. Applications

Our original motivation for this work was the need to classify blogs with respect to a set of topics for a study of influence on the Blogosphere [10]. We hoped that the folders common to many Bloglines users would provide an intuitive set of blog topics. Moreover, the feeds that Bloglines users assigned to

Folder	Merged Folders
comics	fun, humor, funny, humour, cartoons, fun stuff, webcomics, comix, comic strips
music	mp3, mp3 blogs
politics	political, political blogs
design	web design, web, web development, webdesign, webdev, css, web dev,web standards
programming	development, dev, technical, software development, code
culture	miscellaneous, random, misc. , interesting
productivity	gtd, lifehacks, getting things done

Table 2: *Example folders with corresponding merged sub-folders*

1	http://www.talkingpointsmemo.com
2	http://www.dailykos.com
3	http://atrios.blogspot.com
4	http://www.washingtonmonthly.com
5	http://www.wonkette.com
6	http://instapundit.com
7	http://www.juancole.com
8	http://powerlineblog.com
9	http://americablog.blogspot.com
10	http://www.crooksandliars.com

Table 3: *The Feeds That Matter for* **'Politics'**

1	http://wvs.topleftpixel.com
2	http://blog.flickr.com/flickrblog/
3	http://www.flickr.com/recent_comments.gne
4	http://www.east3rd.com
5	http://www.durhamtownship.com
6	http://www.digitalcamerawebsites.com
7	http://groundglass.ca/
8	http://www.photographica.org/
9	http://chromogenic.net/
10	http://www.backfocus.info/

Table 4: *The Feeds That Matter* **'Photography'**

the folders could be used to collect training data for the categories. This is in fact the case, but after working with the data, we recognized that it supports the needs of many other studies and applications.

5.1 Feed Recommender

Folder similarity allows us to compare how related two folder vectors are based on the feeds that occur in them. Feed similarity can be defined in a similar manner: two feeds are similar if they often co-occur under similar folders. Note that this definition of feed similarity does not use the textual content of the feed but is entirely based on the subscription data. This gives us an ability to compare two feeds and recommend new feeds that are like a given feed. For each feed there is a folder vector that maintains a count of the number of times the feed has been categorized under a folder name. For a pair of feeds i,j feed similarity is defined as:

$$cos(i,j) = \frac{\sum_k W_{i,k} W_{j,k}}{\sqrt{\sum_k W_{i,k}^2 * \sum_k W_{j,k}^2}}$$

The weight $W_{i,k}$ is determined by a TFIDF score for each folder in the feed vector. The weights are computed using the following formula:

$$W_{feed}(folder) = freq_{feed}(folder) * log(\frac{feedCount}{|feedsLabeled(folder)|})$$

The feed similarity measure could be further improved by using folder similarity as computed in Section 4. Two feeds are similar if they occur in similar folders (rather than identical folders).

5.2 Identifying Leaders using Influence Propagation

Consider a scenario where a user has a few blogs that she subscribes to or is familiar with a couple of extremely popular blogs for a topic. Now, she wishes to find other blogs that are also opinion leaders in this area. In this section, we present a simple method that is based on an influence propagation model using linear threshold. In the following section, we use the blog graph data from the 2006 Workshop on Weblogging Ecosystems (WWE). This dataset consists of posts from about 1.3M blogs and spans over a period of 20 days. Using a few authoritative blogs obtained from the Bloglines data, the technique identifies other topically authoritative blogs.

To propagate influence, starting from the seed set, we use the basic *Linear Threshold Model* [11, 12] in which each node has a certain threshold for adopting an idea or being influenced. The node becomes activated if the sum of the weights of the active neighbors exceeds this threshold. Thus if node v has threshold θ_v and edge weight b_{wv} such that neighbor w influenced v, then v becomes active only if

$$\sum_{w\ active\ neighbors\ of\ v} b_{wv} \geq \theta_v$$

and

$$\sum b_{wv} \leq 1$$

As described in Java et al. [10], we consider the presence of a link from site u to site v as evidence that the site u is *influenced by* site v. Using the above model, we rank each directed edge between u, v in the *Influence Graph* such that the presence of multiple directed edges provides additional evidence that node *node u influences node v*. If $C_{u,v}$ is the number of parallel directed edges from u to v the edge weight

$$b_{v,w} = \frac{C_{v,w}}{d_w}$$

where d_v is the indegree of node v in the influence graph.

The Identifying Leaders Using Influence Propagation (ILIP) algorithm described in Algorithm 1 finds a set of nodes that are influential for a given topic. As shown in figure 7, we start with some seed blogs for a given topic and induce a set of blogs that are termed as the *followers*. Followers are those

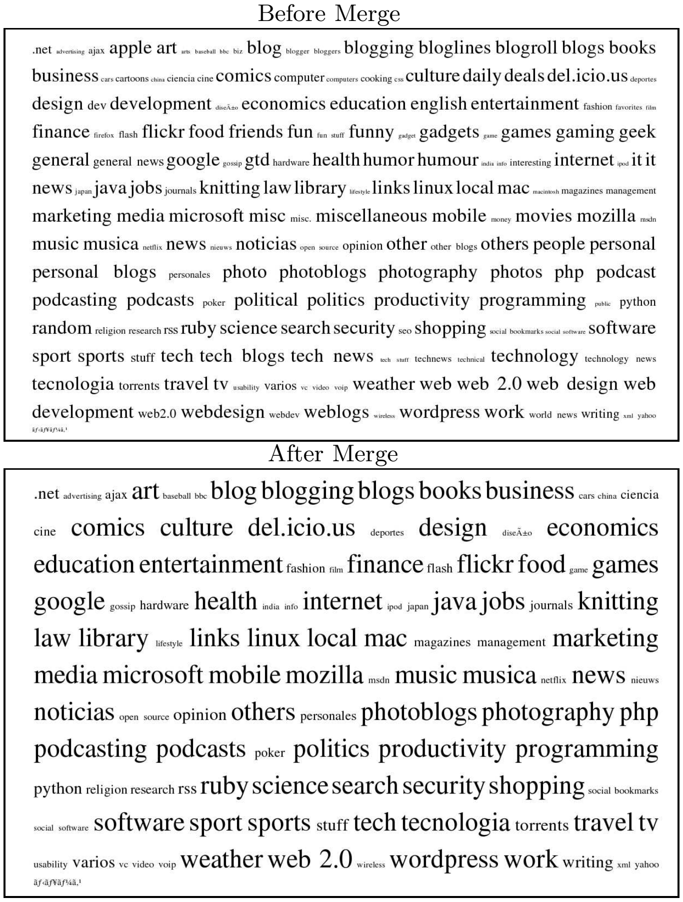

Fig. 6: *The tag cloud generated from the top 200 Folders before and after merging related folders. The size of the word is scaled to indicate how many users use the folder name.*

blogs that are often influenced by the seed set. The goal is to infer other authoritative blogs or *leaders* for the topic. By iterating the linear threshold influence propagation model over the entire blog graph, we can find other blogs that are topically similar to the seed set and are also authoritative. The pseudocode of the ILIP Algorithm 1 describes the various steps involved in identifying topical influential nodes. Starting with a few top ranked feeds from the Bloglines dataset for the folders 'Politics', 'Tech','Business' and 'Knitting' we use the ILIP algorithm to find other leaders in the blog graph. Table 5 to 7 show some of the results.

5.3 FTM! Feeds That Matter

FTM![2] is a site that was implemented out of a need to find a high quality listing or index of *topical* blogs and feeds. This site is based on the Bloglines dataset described in this paper and implements the algorithms presented here for merging folders and providing recommendations. For example if the user was interested in a topic, say photography, she could look at the tag cloud and quickly find feeds that are most often categorized under the folder name "photography". Next, the system allows users to subscribe to the popular feeds directly in their Bloglines or Yahoo RSS readers. Alternatively, one could start a known feed and FTM! would provide recommen-

Seed Blogs
http://www.dailykos.com
http://www.talkingpointsmemo.com
Top Leader Blogs
http://www.huffingtonpost.com/theblog
http://americablog.blogspot.com
http://thinkprogress.org
http://www.tpmcafe.com
http://www.crooksandliars.com
http://atrios.blogspot.com
http://www.washingtonmonthly.com
http://billmon.org
http://www.juancole.com
http://capitolbuzz.blogspot.com
http://instapundit.com
http://www.opinionjournal.com
http://digbysblog.blogspot.com
http://michellemalkin.com
http://www.powerlineblog.com
http://theleftcoaster.com
http://www.andrewsullivan.com
http://www.thismodernworld.com

Table 5: *Leaders found using ILIP for topic* **'Politics'**

Seed Blogs
http://slashdot.org
http://www.kuro5hin.org
Top Leader Blogs
http://www.boingboing.net
http://www.engadget.com
http://www.metafilter.com
http://www.c10n.info
http://www.makezine.com/blog
http://radio.weblogs.com/0001011
http://mnm.uib.es/gallir
http://www.mozillazine.org
http://weblogs.mozillazine.org/asa
http://www.gizmodo.com

Table 6: *Leaders found using ILIP for topic* **'Technology'**

Seed Blogs
http://www.yarnharlot.ca/blog
http://wendyknits.net
Top Leader Blogs
http://booshay.blogspot.com
http://mamacate.typepad.com/mamacate
http://www.thejonblog.com/knit
http://alison.knitsmiths.us
http://www.dioramarama.com/kmel
http://knittersofdoom.blogspirit.com
http://tonigirl.blogdrive.com
http://www.crazyauntpurl.com
http://www.januaryone.com
http://nathaniaapple.typepad.com/knit_quilt_stitch
http://www.knittygritty.net
http://www.katwithak.com
http://www.myblog.de/evelynsbreiwerk
http://nepenthe.blog-city.com
http://zardra.blogspot.com

Table 7: *The Leaders found using ILIP for topic* **'Knitting'**

[2] http://ftm.umbc.edu/

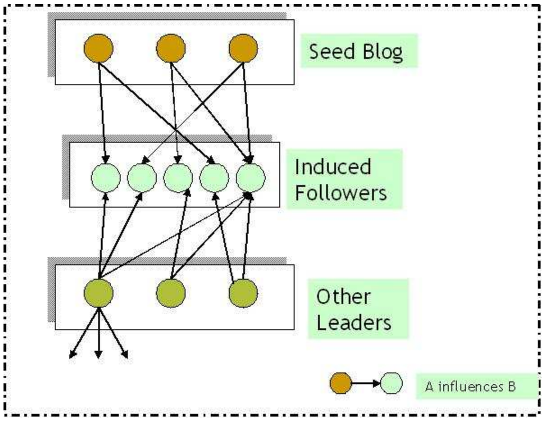

Fig. 7: *Identifying Leaders Using Information Propagation (ILIP): Starting with a few seed blogs on a topic a set of followers are induced and other leaders for this set are identified using the influence graph.*

dations based on the subscription information. "Feeds That Matter" has received a number of encouraging reviews especially from notable bloggers such as Micropersuation [22] and Lifehacker [3]. FTM! also has more than 500 bookmarks on delicious and our logs indicate that there is a steady stream of users who are actively using this service to find subscribe feeds in different categories.

6. Discussion

This section describes a preliminary evaluation of the system and its approach. We evaluate if folder similarity results in grouping related feeds together. We do this by comparing the folder similarity based on co-citations in URL vectors to text similarity of text obtained form the homepages of the feeds.

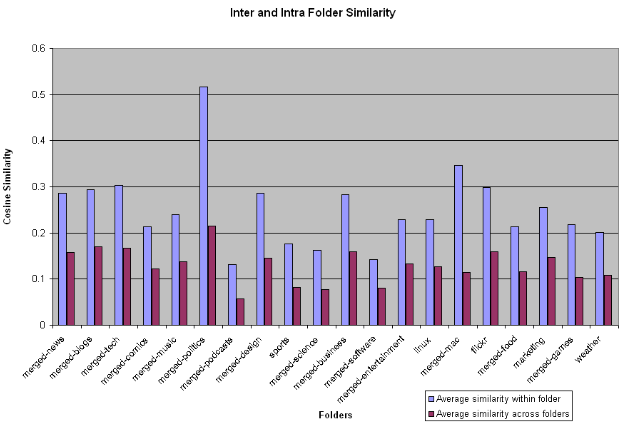

Fig. 8: *The average text similarity of the top 20 Folders. The chart shows the average similarity for the top 15 feeds within and across all the folders.*

Figure 8 shows a comparison of average text similarity of feeds in the top 20 folders. For all the folders it was found that the feeds shared a greater similarity within the folder rather than across other folders. While the scores may seem low, studies on Technorati data by Brooks [4] show cosine similarity of posts sharing the same tag to be around 0.3. According to their study, when the same posts were clustered using the high scoring TFIDF terms the average text similarity was around 0.7.

Algorithm 1 Identifying Leaders Using Influence Propagation (ILIP) Algorithm

```
$S \leftarrow SeedSet$
$F \leftarrow InfluencedFollowersSet$
$IG \leftarrow InfluenceGraph$
for all $i$ such that $0 \leq i \leq max\_iterations$ do
  Activate $S$
  for all $v \in IG$ do
    $\theta_v$ = random score
  end for
  for all $v \in IG$ do
    if $\sum_{w\ active\ neighbors\ of\ v} b_{wv} \geq \theta_v$ then
      Activate v
      add v to $F_i$
    end if
  end for
end for
$F = F_i \bigcup F_{i+1} \bigcup \cdots F_{max\_iterations}$
for all $k$ has inlinks to $F$ do
  $o_k$ = outlink count of k
  $n_k$ = number of nodes linked from k to F
  $leader\_score = \frac{n_k}{o_k} * log(o_k)$
end for
```

Table 8 shows some of the recommendations for a few blogs. The feed recommendations are obtained by comparing the feeds to find how often they co-occur in the same folder. To evaluate the effectiveness of this system, we use the text based cosine similarity as a measure of how related the feeds are. We find that many of the recommended feeds have a high similarity score with the feed submitted. The best way to evaluate such a system would be through user studies and human evaluation of the results. We hope to perform such a study in the near future.

7. Conclusions

A number of Web applications and services can benefit from a set of intuitive, human understandable topic categories for feeds and blogs. This is especially true if we can also have a good 'training' set of feeds for each category. We found that public data from the Bloglines service provides the data from which to induce a set of topic categories and to associate a weighted set of feeds and blogs for each. We have presented a study of the Bloglines subscribers and have shown how folder names and subscriber counts can be used to find *feeds that matter* for a topic. We have also described it's use in applications such as feed recommendations and automatic identification of influential blogs. We have also implemented FTM!, a prototype site based on the algorithms described in this paper. This site has received encouraging reviews from the blogging community and positive feedback from active users.

References

[1] L. A. Adamic, B. A. Huberman;, A. Barab'asi, R. Albert, H. Jeong, and G. Bianconi;. Power-law distribution of the world wide web. *Science*, 287(5461):2115a+, March 2000.

http://www.dailykos.com	**Similarity**
http://www.andrewsullivan.com	**0.496**
http://www.talkingpointsmemo.com	**0.45**
http://atrios.blogspot.com	0.399
http://jameswolcott.com	**0.466**
http://mediamatters.org	0.262
http://yglesias.typepad.com/matthew/	0.285
http://billmon.org/	0.343
http://digbysblog.blogspot.com	**0.555**
http://instapundit.com/	0.397
http://www.washingtonmonthly.com/	**0.446**
http://blog.fastcompany.com	
http://business2.blogs.com/business2blog	0.303
http://www.fastcompany.com	0.454
http://sethgodin.typepad.com/seths_blog/	0.374
http://www.ducttapemarketing.com/	0.028
http://customerevangelists.typepad.com	0.399
http://blog.guykawasaki.com/	**0.441**
http://www.tompeters.com	**0.457**
http://www.paidcontent.org/	0.351
http://slashdot.org	
http://www.techdirt.com/	**0.516**
http://www.theregister.co.uk/	0.1
http://www.geeknewscentral.com/	0.286
http://www.theInquirer.net	0.2
http://news.com.com/	0.24
http://www.kuro5hin.org/	0.332
http://www.pbs.org/cringely/	0.087
http://backword.me.uk/	-
http://digg.com/	0.165
http://www.infoworld.com/news/index.html	0.203
http://www.yarnharlot.ca/blog/	
http://wendyknits.net/	**0.419**
http://www.woolflowers.net/	0.139
http://zeneedle.typepad.com/	0.383
http://www.keyboardbiologist.net/knitblog/	0.297
http://alison.knitsmiths.us/	0.284
http://knitandtonic.typepad.com/knitandtonic/	**0.542**
http://www.crazyauntpurl.com/	**0.521**
http://www.lollygirl.com/blog/	**0.4**
http://ma2ut.blogspot.com	**0.423**

Table 8: *Example recommendations for blogs in bold.*

[2] M. Arrington. Finally! bloglines blog search. `http://www.techcrunch.com/2006/05/31/askcombloglines-launch-blog-search/`.

[3] W. Boswell. Find "feeds that matter". `http://lifehacker.com/software/bloglines/find-feeds-that-matter-209796.php`.

[4] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW*, 2006.

[5] C. Cattuto, V. Loreto, and L. Pietronero. Collaborative tagging and semiotic dynamics, 2006.

[6] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. M. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks*, 30(1-7):65–74, 1998.

[7] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW*, 2006.

[8] E. Glover, D. M. Pennock, S. Lawrence, and R. Krovetz. Inferring hierarchical descriptions. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 507–514, New York, NY, USA, 2002. ACM Press.

[9] M. Guy and E. Tonkin. Folksonomies: Tidying up tags. *D-Lib Magazine*, 12(1), 2006.

[10] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the Spread of Influence on the Blogosphere. Technical report, University of Maryland, Baltimore County, March 2006.

[11] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[12] D. Kempe, J. M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.

[13] P. Kolari, A. Java, and T. Finin. Characterizing the Splogosphere. In *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*. Computer Science and Electrical Engineering, UMBC, May 2006.

[14] J. Lanzone. Which feeds matter? `http://blog.ask.com/2005/07/what_feeds_matt.html`.

[15] C. Marlow. Audience, structure and authority in the weblog community. In *54th Annual Conference of the International Communication Association*, 2004.

[16] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`, 2004.

[17] C. McEvoy. Bloglines users are a load of knitters. `http://usability.typepad.com/confusability/2005/04/bloglines_user_.html`.

[18] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[19] G. Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 953–954, New York, NY, USA, 2006. ACM Press.

[20] C. Pirillo. Google: Kill blogspot already!!! `http://chris.pirillo.com/2005/10/16/`.

[21] E. Quintarelli. Folksonomies: power to the people. `http://www.iskoi.org/doc/folksonomies.htm`, 2005.

[22] S. Rubel. University study reveals rich data on bloglines feeds. `http://www.micropersuasion.com/2006/10/university_stud.html`.

[23] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA, 1999. ACM Press.

[24] K. Shen and L. Wu. Folksonomy as a complex network, Sep 2005.

[25] D. Sifry. State of the blogosphere, april 2006 part 1: On blogosphere growth. `http://www.sifry.com/alerts/archives/000432.html`.