Identifying Representative Textual Sources in Blog Networks

Karen Wade
Humanities Institute of Ireland
University College Dublin

Derek Greene, Conrad Lee, Daniel Archambault, Pádraig Cunningham
Clique Research Cluster, School of Computer Science & Informatics
University College Dublin

University College Dublin
Technical Report UCD-CSI-2011-01
February 2011

Abstract

We apply methods from social network analysis and visualization to facilitate a study of the Irish blogosphere from a cultural studies perspective. We focus on solving the practical issues that arise when the goal is to perform textual analysis of the corpus produced by a network of bloggers. Previous studies into blogging networks have noted difficulties arising when trying to identify the extent and boundaries of these networks. As a response to calls for increasingly data-led approaches in media and cultural studies, we discuss a variety of social network analysis methods that can be used to identify which blogs can be seen as members of a posited "Irish blogging network" (and hence sources of textual material). We identify hub blogs, communities of sites corresponding to different topics, and representative bloggers within these communities. Based on this study, we propose a set of analysis guidelines for researchers who wish to map out blogging networks.

1 Introduction

A major challenge for researchers setting out to analyze blogs and other social media is the volume of data that needs to be considered. In this paper, we describe a case study that attempts to map out the Irish blogosphere in order to identify representative textual sources.

This study is part of a larger research project that draws on gender studies, cultural studies, and new media research to analyze the interactions between online community, Irish culture, and gendered identity. The objective is to understand how this community of bloggers use language and online manifestations of identity to relate to current issues of gender identity and sexuality. This work is intended as a contribution to the burgeoning research on ethnographic and literary analysis of blogs and other social media [Van Dijck, 2004; Yu, 2007; Herring *et al.*, 2007]. While there is other research on identifying representative blogs in the the blogosphere [Hassan *et al.*, 2009], a particular contribution of the work presented here is an assessment of the usefulness of the results from a cultural studies perspective.

We consider a set of 614 blogs that forms the largest connected component in the Irish blogosphere. The interactions between these blogs can be characterized via two distinct types of connection: explicit *blogroll links* and normal hyperlinks occurring within the HTML content of blog posts

Figure 1: Largest connected component from the Irish blogroll network, consisting of 614 distinct blogs. Nodes are scaled in proportion to in-degree.

(henceforth referred to as *post-links*). An illustration of the former view of the data is given in Figure 1. It is also possible to identify topical groupings of bloggers based on the similarity of the textual content that they produce. In our case study, the set of 614 blogs account for close to two hundred thousand unique posts over a thirteen year period. The availability of different data views and the scale of the text content makes the task of identifying representative blogs for detailed textual analysis daunting.

In the first part of this paper, we describe related work and the data on which the analysis is based[1]. We then describe the analysis performed on that data under three categories: network characterization, text content clustering, and link-based community finding. We identify distinct communities in the blogosphere based on blog content similarity and inter-blog linking patterns. The paper concludes with a discussion of the findings from this case study, comments on the generality of these findings and a set of analysis guidelines.

2 Related Work

2.1 Textual Analysis in the Humanities

Humanities and social sciences scholars have undertaken much research into blogging in recent years. We argue that some methods of data collection and analysis used in prior studies of blogs result in impoverished analyses and limit the ultimate utility of the studies performed. Small-scale qualitative analyses of a handful of blogs are relatively frequently seen in the literature, particularly where the study involves interviews or ethnographic surveys, or requires prior consent from participants for ethical reasons; the study by Yao *et al.* [2009] of uses and gratifications associated with blogs by Filipino women working in the UK, based on in-depth interviews with eight bloggers, is an example of this type of study.

The intent here is not to critique studies based on small, demographically restricted research groups, which may provide insights into specific communities, but problems occur when the results of such work are assumed to reflect wider communities of blogs, or even blogs in general. Herring *et al.* [2004] argued that "the fundamental nature of the weblog phenomenon" has been misrepre-

[1]The data set is available at `http://mlg.ucd.ie/blogs`

sented by biases in the media and in blogging scholarship that failed to take into account the demographics of the wider blogosphere, through focusing primarily on political blogs. Nardi *et al.* [2004] base their ethnographic analysis on 23 bloggers based around Stanford, California, and although acknowledging that this is a small and "fairly uniform" group, they do make generalizations about blog users based on their responses. Over-generalizing observations about blogging based on a study of 23 Californian blogs is problematic; as of the time this paper was written, the BlogPulse blog tracking website cites the existence of over 150 million individual blogs, although not all are currently active. Some work in the humanities does indicate a need for more data-driven methods of research into blogging.

Recently, some studies have adopted random sampling techniques to select blogs for research purposes, either making use of randomizers built into blogging platforms themselves [Jones *et al.*, 2008], or selecting blogs randomly from directories such as Technorati or Truth Laid Bare [Hindman, 2009]. Researchers working on social interaction and communities between blog users are also increasingly using social network analysis techniques in order to identify and define their groups of research subjects. For example, approaching the issue of how to identify a community of disparate blogs, Efimova *et al.* [2005] utilize the metaphor of "studying the life between buildings", drawn from archaeological work, as a way of conceptualizing the study of blogging communities by social network analysis.

Some recent studies in blogging in the humanities utilize both large-scale quantitative and small-scale qualitative methods of analysis. For example, Van Doorn *et al.* [2007] begin their study of the online presentation of gender identity by analyzing a randomly selected set of 100 bloggers for demographic data and then performing close readings on a smaller subset of these. In this study we hope to demonstrate a similarly two-pronged approach for the study of a large, distributed network or community of blogs, making use of both quantitative and qualitative methods.

2.2 Social Network Analysis

With regard to choosing how to represent the blogosphere as a network, Marlow considered two alternative criteria for linking blogs: explicit blogroll links or post-links (*i.e.* hyperlinks appearing within the content of a blog post). [Marlow, 2004]. He finds that blogroll networks are more static, and accentuate the "rich get richer" phenomenon, whereas networks based on post-links are more dynamic. Hassan *et al.* addressed the problem of finding a representative set of influential blogs [Hassan *et al.*, 2009]. They propose a method that utilizes random walks on a blog graph where edges are created between any pair of blogs whose text similarity is greater than some threshold; their algorithm also includes measures to ensure diversity among the resulting set of blogs.

Community finding plays an important role in our analysis. Dozens of community finding algorithms have been proposed in the last few years. Fortunato [2010] offers the most recent and comprehensive review. It has recently been emphasized that although the majority of these methods produce a non-overlapping partition of nodes, communities in many empirical networks have highly overlapping communities [Ahn *et al.*, 2010]. In the present work we therefore employ methods capable of detecting overlapping communities.

2.3 Visualization

A number of visualization techniques and systems exist for drawing and interacting with graphs of a variety of sizes [Herman *et al.*, 2000; von Landesberger *et al.*, 2010]. As the number of nodes and edges for data sets in this work are relatively small, standard techniques can be used to produces drawings of the networks and their full details can be displayed. For the node link diagrams used in this paper Tulip [Auber, 2003] and Gephi [Bastian *et al.*, 2009] were used to filter down the node set and produce force-directed layouts of the network diagrams.

In order to represent sets of items and their intersections, we use the Euler-like diagram drawing technique of Simonetto *et al.* [2009]. Euler diagrams are similar to the well-known Venn diagrams except they are not constrained, by definition, to show all intersections between the sets. Rather, Euler diagrams show only non-empty overlaps. To further enhance regions of overlap in the diagram, colors and textures interfere in patterns that indicate the sets participating in the intersection. Figure 4 shows an example of an Euler diagram representation for an overlapping clustering.

3 Data Collection

To produce an initial seed set of blogs, we began with a set of 21 popular blogs within the Irish blogosphere, as indicated by winners of the "2010 Irish Blog Awards"[2]. Starting with the seed set, we manually extracted blogroll links where a blogroll was available. We repeated this process for two steps out from the seed set. We subsequently manually filtered out blogs that were either password-protected, inactive during the year 2010, aggregation sites, or whose geographical designation did not correspond to Ireland. Finally, we removed blogs with fewer than two in-links (*i.e.* links coming from blogrolls on other blogs, representing popularity). This yielded a verified set of 635 blogs.

3.1 Text Content View

We identified all posts archived by Google Blog Search[3] for the set of 635 blogs. At the time of collection, the archive was limited to approximately 1,000 posts per blog – this provided us with a total 179,015 unique posts. While the vast majority of entries (93%) were published during the period 2007–2011, entries in the collection date back as far as 1997. We retrieved the complete set of posts and extracted text content from the raw HTML. After removing dead links and posts that did not contain any text content, (*e.g.* posts containing only videos or images with no associated text), a total of $\approx 176k$ posts remained from 614 different blogs – blog posts were no longer available for the 21 other blogs. Therefore, the core set of 614 blogs is the focus of the case study here.

Since we were interested in grouping blogs and bloggers, rather than individual posts, we chose to represent each blog by its *content profile* – this is defined as the concatenation of the text content from all available posts for that blog. We then applied standard stop-word filtering, stemming, and term frequency filtering (we removed terms appearing in < 5 blogs). Using a vector space model representation, this yielded an extremely sparse, high-dimensional space, with 614 profile vectors of length $\approx 153k$ dimensions (terms) – $\approx 4\%$ of the values in the term-profile matrix were non-zero.

3.2 Tagging Data

To provide an additional source of validation for groups identified on the various Irish blogosphere views, we also collected all tag records available for the set of 614 blogs on the *Del.icio.us* social bookmarking portal. At the time of collection (January 2011), this process yielded a total of 13,887 tagging records covering 410 blogs (67% of the complete set). Tags contained in a set of fifteen highly-frequent stop-words (*e.g.* "blog","blogger") were removed, leaving 1,914 unique user-assigned tags.

3.3 Network Views

There are many reasonable, yet distinct, criteria one could use to decide whether two blogs should be considered connected to each other in the network of blogs. This freedom of choice for the "edge criterion" casts the following doubt on any findings based on network analysis: if the researcher had used a different criterion for creating edges, would the same results have followed? To alleviate such doubts, we use two different edge criteria to create network representations: the *blogroll network* and the *post-link network*. Although the networks constructed by these criteria are quite different, our main findings are similar in both networks, indicating that they are robust to the choice of edge criterion.

The blogroll network is unweighted and consists of permanent or nearly-permanent links between one blog and another – these are generally located in the template section of the blog or on a dedicated blogroll-page. The post-link network, on the other hand, is weighted and meant to measure the number of non-permanent post links between blogs.

For each blog, we first concatenate all posts together and extract all URLs that point to other blogs in the core set. If a specific URL appears in at least half of that blog's posts, or appears at least 100 times, we add it as an unweighted edge to the blogroll graph. If the URL appears fewer than 20 times and in less than 20% of all posts, then we add it as a weighted edge to the post-link graph, where the

[2] `http://awards.ie/blogawards`

[3] `http://blogsearch.google.com`

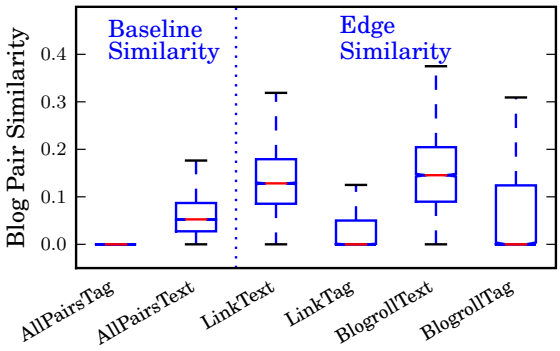


Figure 2: On the right, we see that the similarity of nodes connected in the blogroll network is higher both in terms of text similarity (BlogrollText) and *Del.icio.us* tag similarity (BlogrollTag) than edges in the post-link graph. Connected blogs in both graphs are more similar than the baseline similarity, displayed on the left.

weight equals the URL count. We find that 98.9% of all URLs between posts fall into one of those two categories – we ignore the remaining links. We add all of the manually collected blogroll links from the initial crawl to the blogroll graph. After removing isolated nodes, the resulting blogroll network contains 604 nodes and 4,640 edges, while the post-link network contains 588 nodes and 9,927 edges and a total edge weight of 23,005.

We can in one sense compare the quality of the blogroll and post-link networks by looking at the similarity of connected blogs. Figure 2 indicates that in this regard the blogroll network is of higher quality because connected blogs tend to use more similar vocabularies and tend to share more similar *Del.icio.us* tags. This finding suggests that if one wishes that connected blogs be more similar to each other, than the blogroll graph provides somewhat better results.

4 Finding Central Blogs

Given the objective of identifying representative sources for detailed textual analysis in a blogging community of significant size ($\approx 180k$ posts), an obvious starting point is to identify significant blogs based on network centrality measures. For this analysis, we chose in-degree centrality because we would like our sample of the blogosphere not only to be representative, but also include the most influential blogs. The in-degree of a blog in the blogroll and post-link networks is a straightforward proxy for the popularity of a blog – although there may be exceptions, we assume that the blogs that attract the most links are the most read. We did not employ the popular betweenness centrality metric because global path length is not a clearly meaningful concept in this network.

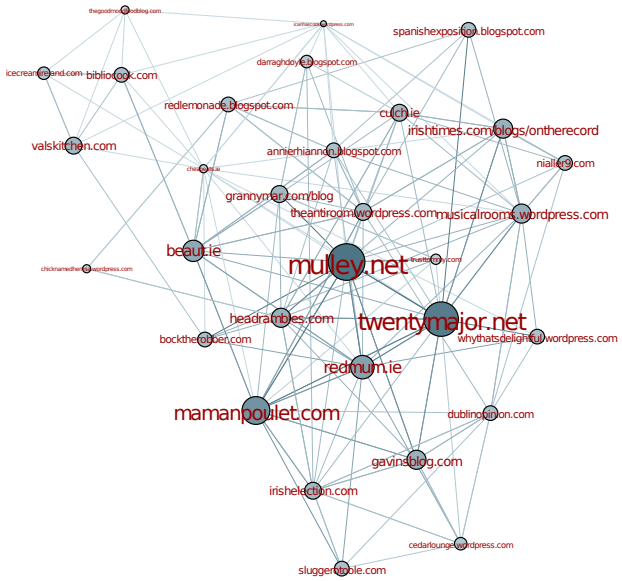
Figure 3 shows the blogs with the highest in-degree in both the blogroll and post-link networks. We found good agreement between these networks: they shared fifteen of their top twenty blogs. One naïve solution to the problem addressed in this paper – finding a good sample of blogs – would be to simply take this set. However, while these blogs are influential, it is clear from Figure 1 and the analysis below that they do not provide good coverage of the wider Irish blogosphere.

5 Blog Content Clustering

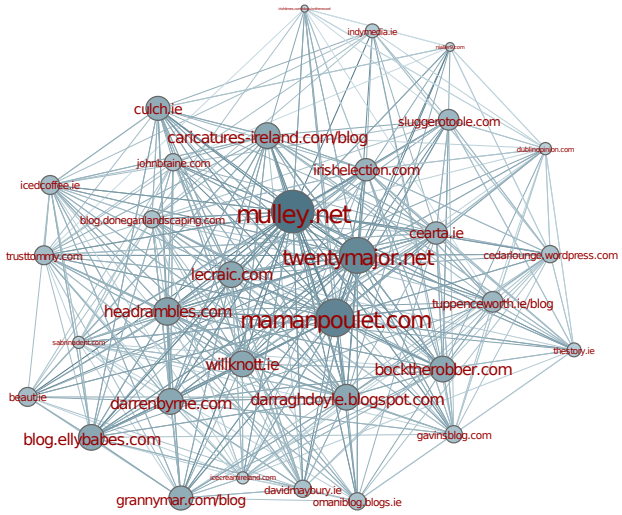
In this section we explore the text content view described previously in an attempt to identify groups of blogs pertaining to coherent topics.

5.1 Experimental Setup

A natural approach to identifying the topical groupings in a blog network is to apply cluster analysis techniques. We may be interested in discovering whether the data contains a hierarchy, with different of levels of topical granularity. However, a distinct drawback of many hierarchical methods lies in the fact that each entity can only reside in a single branch of the tree at a given level, and can only



(a) Blogroll Network



(b) Post-link Network

Figure 3: Top blogs based on in-degree for (a) the blogroll (in-degree $\geq$ 25), and (b) post-link (in-degree $\geq$ 50) network views of the Irish blogosphere dataset.

belong to a single leaf node. In contrast, we might expect that some bloggers will discuss a range of different subjects, and so will belong in several topical communities.

Therefore to support both objectives when clustering blog content text, we employ an implementation of the ensemble non-negative matrix factorization (NMF) algorithm[4], which was previously used to explored cluster structures in biological data [Greene *et al.*, 2008]. This algorithm involves combining an ensemble consisting of a diverse collection of factorizations generated on a given data set. The information from the ensemble members is aggregated into a consensus solution. This solution takes the form of a *soft hierarchical clustering*, where items are organized into a binary tree such that they can be associated with multiple clusters in the tree to varying degrees.

[4]Available at `http://mlg.ucd.ie/nmf`

Based on the content view of the Irish blogosphere described previously, we applied standard log-based TF-IDF term weighting to the blog profile vectors, and normalized the vectors to unit length. From these we computed a 614×614 symmetric cosine similarity matrix. Entries in this matrix indicate the level of similarity between pairs of blog content profiles. The ensemble NMF algorithm was applied to this matrix. We experimented with a number of sets of values for the range $[k_{min}, k_{max}]$ to produce the ensemble members, with a range $[10, 15]$ yielding the most stable results. An ensemble of 1,000 factorizations was generated, to yield a robust final clustering. Based on the range for the ensemble members, we cut the final tree at $k = 12$. It is interesting to note that we observed little or no cohesive hierarchical structure above this point in the tree. We use a membership threshold of 0.1 (as opposed to > 0) when producing discrete, overlapping cluster assignments from weighted cluster memberships, to emphasize precision over recall. Note that this still permits blogs to be assigned to multiple clusters, but also means that blogs having weak associations with clusters in terms of content similarity will not be assigned to any cluster.

5.2 Discussion of Results

Among the twelve clusters that were identified, discrete cluster assignments were provided for 419/614 (68%) of the full set of blogs. A visual representation of these clusters using Euler diagrams is shown in Figure 4. Each set corresponds to a different thematic cluster, where the set size is roughly proportional to the size of the cluster, and regions where two or more clusters overlap are indicated via cross-hatching.

Figure 4 also shows a selection of highly descriptive keywords for the clusters. Cluster keywords were automatically identified by ranking the terms for each cluster based on their Information Gain (IGain) [Yang and Pedersen, 1997]. Given a cluster of blog profiles, the ranking of terms for the cluster is performed as follows: firstly the centroid vector of the cluster is computed; subsequently, we compute the IGain between the cluster centroid vector and the data set centroid vector. Terms that are more indicative of a cluster will receive a higher score, thereby achieving a higher ranking in the list of keywords for the cluster. Figure 4 also includes putative cluster names that have been manually selected based on all of the top terms.

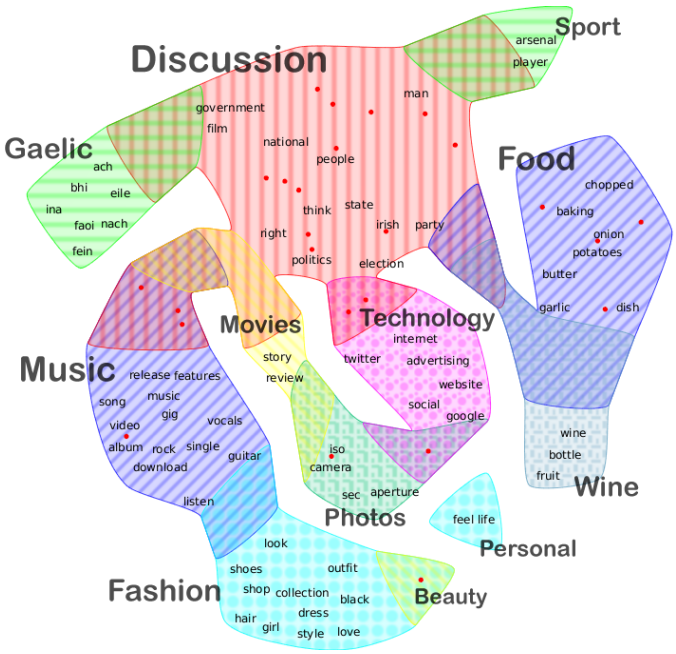


Figure 4: Euler diagram showing twelve blog clusters identified via content clustering. The clusters have been manually annotated with putative labels. Frequent terms from the posts are indicated in the sets. The blogs corresponding to the union of the 20 highest in-degree node sets from the *blogroll network* and the *post-link network* are indicated in red.

Figure 5: Tag clouds for six sample thematic clusters identified on the text content view of the Irish blog data set, corresponding to groups of blogs primarily discussing fashion, food, wine, photography, movies, and the Irish language (Gaelic) respectively.

In Figure 5 we illustrate sample tag clouds for six of the content clusters – these represent clusters of blogs annotated as "Fashion", "Food", "Wine", "Photos", "Movies", and "Gaelic" respectively. The clouds were produced using Wordle [Viégas *et al.*, 2009], where the sizes of the terms in the tag clouds are proportional to the weights calculated using IGain as described above – the top 25 highest-weighted terms are shown for each of the clusters. We observe that all of the terms appearing in the fourth tag cloud originate from the Irish language.

5.3 Cluster Validation

To verify the set of text content clusters, we propose the use of the tagging data retrieved from *Del.icio.us*, in combination with traditional internal cluster validation techniques. The motivation is that a good clustering on the content view should lead to a high internal cluster validation score for the same cluster assignments on the tagging data. Following the approach for evaluating tag-based clusters outlined by Hayes *et al.* [2007], we employ the cosine-based validation criterion proposed by Zhao and Karypis [2004] to evaluate and rank the quality of individual clusters. This criterion calculates the ratio of intra-cluster similarity relative to inter-cluster similarity. The former is defined by the mean of document-centroid similarities, while the latter is defined as the similarity between a cluster centroid and the centroid of the entire data set $\bar{\mu}$. Formally, for a cluster C_i with centroid vector μ_i, we calculate

$$\mathcal{H}_r(C_i) = \left(\frac{1}{|C_i|} \sum_{x_j \in C_i} cos(x_j, \mu_i) \right) \cdot \frac{1}{cos(\mu_i, \bar{\mu})} \quad (1)$$

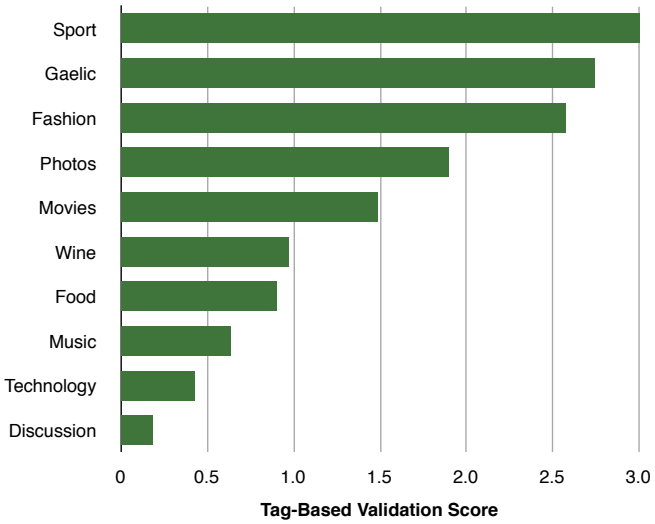


Figure 6: Internal cluster validation scores for ten blog clusters identified in the text content data, illustrating their relative consistency with the *Del.icio.us* tagging data.

where a larger value is indicative of a more cohesive cluster. Although this criterion has previously been used to cluster and validate disjoint partitions, the use of the data centroid in the inter-cluster similarity term (as opposed to a "competing" cluster centroid) makes the criterion suitable for use in cases where blogs are potentially assigned to more than one cluster.

To provide a suitable representation for calculating the cosine similarity values used by the internal validation index, we follow the approach described for processing tagged blog data by Hayes *et al.* [2007]. For each tagged blog we create a tag frequency vector based on its associated tags and tag counts. The resulting vectors are used for cluster validation.

Figure 6 shows results for ten of the twelve text clusters previously shown in Figure 4, as ranked by their validation scores. Note that tagging information was unavailable for the blogs contained in the other two small clusters ("Personal" and "Beauty"), so they are excluded from the analysis. Despite the apparent disparity between the two sources of data, the plots indicate that many of the content-based clusters are strongly-consistent with the tagging data. Only in the case of the large "Discussion" group do we find a cluster that fits poorly with the tag information, suggesting that this cluster is not particularly cohesive in terms of its members. We examine this particular group in more detail later in the paper.

To further investigate the relationship between the text clusters and the corresponding user-assigned tags, we identify the top tags for each of the ten content clusters. We do this by taking tags corresponding the highest-weighted entries from cluster centroids computed on the unit normalized

Annotation	*Top Tags*
Sport	football, soccer, paris, sport, barcelona
Gaelic	gaeilge, language, irish, bloganna, blag
Fashion	fashion, style, fashionblog, design, revenuemodels
Photos	photography, photoblogs, photo, photoblog, photos
Movies	writing, law, film, reviews, copyright
Wine	wine, food, cooking, recipes, gastronomie
Food	food, foodblog, recipes, cooking, baking
Music	music, mp3, reviews, mp3blog, dublin
Technology	technology, tech, marketing, pr, advertising
Discussion	politics, culture, humour, dublin, news

Table 1: Top five tags from the *Del.icio.us* tagging data, for ten blog clusters identified in the blog text content data.

document vectors, as proposed by Dhillon and Modha [2001]. The top five tags for each of the ten clusters are shown in Table 1. We observe that in many cases, the highest-ranked *Del.icio.us* tag relates closely to the cluster labels based on the top terms in the text data (see Figure 4). However, it is apparent again from the tags in Figure 6 that the "Discussion" cluster appears to consist of blogs covering a diverse range of topics.

6 Sub-Communities in Document Clusters

The largest cluster detected in the last section was manually annotated with the label "Discussion" because the top words in this cluster's tag cloud, which is not displayed, were "politics," "people," and "Ireland". However, this cluster was the least coherent in terms of the tagging data, and it is apparent from both the text content and tags that this group includes blogs that discuss issues pertaining to a number of topics.

To test whether the discussion cluster consists of distinct subgroups, we first selected all nodes in the cluster, and then extracted the subgraphs induced by these nodes in both the blogroll and post-link networks. We then searched for *network communities* – a network community is a group of nodes that is densely connected to each other while relatively sparsely connected to the surrounding network. We ran three different community detection algorithms on the data: InfoMap [Rosvall *et al.*, 2010], MOSES [McDaid and Hurley, 2010], and GCE (with $\alpha = 1.5$, $\epsilon = 0.25$) [Lee *et al.*, 2010]. These three algorithms produced widely varying results, leaving us initially without any definitive set of communities.

We assume that if a network community is pronounced, then an algorithm should be "stable" with regard to various network representations. Thus, for each algorithm, we accepted only those communities that were found in both the post-link and blogroll subgraphs. We considered two communities to be the same if their memberships overlap considerably. Specifically, we computed the Jaccard set similarity between all pairs of communities in both subgraphs, and identified matching pairs with a similarity score ≥ 0.4. A stable community was formed from each of these pairs by taking the intersection of the pair's membership sets.

GCE produced the best set of overlapping communities of reasonable size – these are visualized in the Euler diagram shown in Figure 7, where the communities have been projected onto the blogroll subgraph. Table 2 shows the corresponding highest ranked *Del.icio.us* tags for the sub-communities, which we use to manually annotate these groups with appropriate labels. These tags were generated using the same centroid-based methodology used to produce the results in Table 1. We observed distinct communities related to education and legal matters, humor and satire, and music. The latter community overlaps substantially with the "Music" text cluster, as shown in Figure 4.

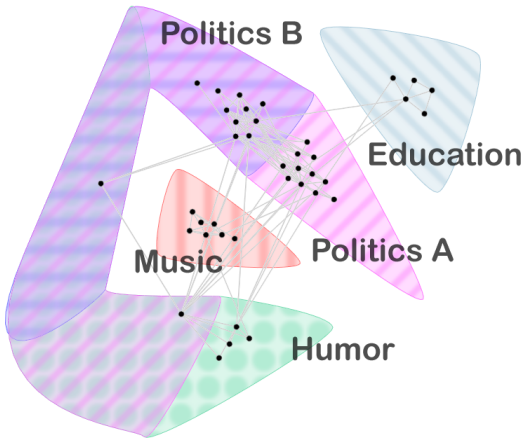


Figure 7: Euler diagram illustrating the overlap of five sub-communities identified on the subgraphs defined by the "Discussion" text cluster, with links from the blogroll graph.

Annotation	*Top Tags*
Education & Law	law, education, legal, universities, rights
Humor	comics, humour, inspiration, history, funny
Music	music, industryexpert, journalist, reviews, irishtimes
Politics A	politics, anarchism, sluggernet, media, dublin
Politics B	sluggernet, politics, blogaly, history, northernireland

Table 2: Top five tags from *Del.icio.us* tagging data, for five blog sub-communities identified in the subgraphs defined by the "Discussion" text content cluster.

The stable communities included two closely-related groups of political blogs, where the smaller group is a subset of the larger group. On inspecting the tagging data for these blogs and the actual blogs themselves, we find that the smaller community ("Politics B") contains blogs largely concerned with Northern Ireland, while the larger community ("Politics A") consists of political blogs covering both Northern Ireland and the Republic of Ireland (*e.g.* sluggerotoole.com).

7 Analysis and Interpretation

7.1 Selection of Representatives

Table 3 shows the final set of fourteen blogs chosen as "representative" members of various thematic communities within the Irish blogosphere. This set was chosen based on the union of the eleven cohesive text content clusters shown in Figure 4 and the "Discussion" sub-communities shown in Table 2. For each group, we identified the node with the highest in-degree in the subgraph of the blogroll graph induced by the group membership. We use this "local popularity" approach because global in-degree tends to favor globally popular blogs, even if they are not very representative of a particular community. Only five of the locally popular nodes have a global in-degree that appears in the top twenty of either the blogroll or link graphs. Also, the representative blogs for "Beauty" and "Personal" were selected based on global degree due to insufficient local in-links within their subgraphs. Not all sub-communities of Table 2 appear in Table 3 as *splinteredsunrise.wordpress.com* was identified as representative of "Politics A" and "Politics B". Also, "Music" is not considered, as it is a subset of "Music" in Figure 4.

The blogs listed in Table 3 are evenly balanced along authorship gender lines – six male, six female, and two for which the gender was not clear. Represented in this set are a variety of different blog types and authorial stances: there are group blogs, dual-authored blogs and solo blogs, and they range in style from the "personal journey" (icanhascook.wordpress.com) to traditional print media journalism (irishtimes.com/blogs/ontherecord). This indicates that our proposed analytics methodology achieves the kind of *diversity* that has been the goal of previous work in blog data mining research [Hassan *et al.*, 2009].

7.2 Interpretation of Results

This study is part of a larger research project that draws on gender studies, cultural studies and new media research. The objective of the study is to understand how this community of bloggers engages in the "doing" of Irishness, with a particular focus on the use of Hiberno-English in their texts, and these online manifestations of Irish identity relate to or impact on current issues of gender identity and sexuality.

Our analysis has provided a previously unavailable understanding of the size and topical composition of the network of bloggers who are strongly embedded in the Irish blogosphere. We have identified sub-communities complete with annotations of topics based on textual content and tags from social media. These results should facilitate smaller studies of sub-communities and provide future research with an understanding of the contexts of individual blogs. This is of particular importance for researchers interested in online manifestations of community and personal identity. For example, the identification of nested political sub-communities within the "Discussion" cluster will aid researchers interested in Irish politics. It is interesting to contrast these groups of political

Theme	*Representative Blog*
Beauty	** beaut.ie
Education & Law	** cearta.ie
Fashion	blanaid.com
Food	** icanhascook.wordpress.com
Gaelic	miseaine.blogspot.com
Humor	counago-and-spaves.blogspot.com
Movies	scannain.com
Music	** irishtimes.com/blogs/ontherecord
Personal	anonomousangel.wordpress.com
Photos	slkav.com
Politics	splinteredsunrise.wordpress.com
Sport	dangerhere.com
Technology	** mulley.net
Wine	firstpress.blogspot.com

Table 3: Set of fourteen blogs chosen as "representative" members of various thematic communities within the Irish blogosphere. Blogs prefixed with "**" are in the set of top 20 in-degree nodes in either the blogroll or the link network.

blogs – which are not clearly divided along party lines – with the segregated US political blogosphere [Adamic and Glance, 2005].

The ability to identify a short list of representative blogs is invaluable to a researcher attempting to understand community of over 600 blogs. In contrast to simply picking top blogs, such a list allows one to avoid "cherry-picking" by providing a sample that one can reasonably claim to represent each sub-community of the larger network.

The prevalence of sub-communities in this network is of specific relevance to the study of genre in blogging. Previous studies [Herring *et al.*, 2005] have made distinctions between the blog genres of "personal", "filter" (which provide commentary on external or current events) and "k-logs" (which list the author's expertise on a particular subject) and have found personal blogs to be the most prevalent in their samples. Our "Discussion" cluster appears to contain examples of all three genres, with some blogs combining aspects of two or more varieties; for example, the author of the law blog lexferenda.com discusses events relevant to his field of expertise (media and Internet law) as well as his own current work, while humor blogger emeraldbile.blogspot.com posts about expatriate life and current events such as the Wikileaks controversy. Further investigation of this cluster could reveal whether Irish bloggers make strong distinctions between these three genres of blogging or whether Herring *et al.*'s fourth category, "mixed", is most prevalent; it may be the case that *theme* (*e.g.* "Music", "Fashion") is more of a concern for blog users than Herring *et al.*'s categories.

Additionally, relatively few personal blogs were identified in our analysis; the "Personal" subcategory consisted of just three blogs which were textually similar but unconnected via linking. This does not exclude the possibility of the existence of Irish personal blog sub-communities within this network, but it is interesting that such networks are not as easily identified by network and textual analysis. We observed during the course of the study that textual analysis methods had difficulty identifying humor or irony within texts, which is unsurprising – *Del.icio.us* tags were helpful for categorizing blogs dealing with these complex concepts.

A further issue of note is the comparison of our results to previous discussions of blogroll and content post-link networks. Marlow [2004] suggests that blogrolls links are likely to be less current than post-links. Efimova *et al.* [2005] choose to build their network based on post-links rather than blogrolls on similar grounds, stating that while blogrolls represent how bloggers "self-identify their close connections", links made within blog posts are more accurate representations of the networks they inhabit. However, our analysis found significant correspondence between the two networks – there is a 75% overlap between the top 20 blogs in the two networks. Furthermore, we found that pairs of blogs connected in the blogroll network tended to be more similar, both in terms of text content and user-assigned tags.

8 Conclusions

In this paper, we present a case study on identifying representative sources in a blogging community of significant size ($\approx 180k$ posts) for detailed textual analysis. We believe that the procedures followed and the main lessons learned are generally applicable. First, we see from Figure 4 that centrality analysis is not adequate to identify representative sources as it over-samples dense regions of the network and misses more peripheral communities. To address this, clustering based on text similarity is effective for organizing the blogosphere into general themes. We found that the output of this clustering process is more easily interpreted than community finding based on post-link and blogroll networks. However, community finding within specific themes was found to be useful – largely due to the more high-level themes identified by content-based clustering. Given that many blogs belong to more than one thematic community, it is advisable to use overlapping text clustering and community finding algorithms.

Finally, since text clustering and community finding algorithms will always return a result, it is important to validate the communities uncovered. This validation should be done both by evaluating against external sources of data, such as user-assigned *Del.icio.us* tags, and by engaging one or more experts in cultural studies.

Acknowledgments

This work is supported by Science Foundation Ireland Grant No. 08/SRC/I140 (Clique: Graph and Network Analysis Cluster). Karen Wade acknowledges the support of the IRCHSS GREP scholarship programme for Gender, Identity and Cultural Change.

References

[Adamic and Glance, 2005] L.A. Adamic and N. Glance. The political blogosphere and the 2004 US election: Divided they blog. In *Proc. 3rd International Workshop on Link Discovery*, pages 36–43, 2005.

[Ahn *et al.*, 2010] Y.Y. Ahn, J.P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[Auber, 2003] D. Auber. Tulip : A huge graphs visualization framework. In *Graph Drawing Software*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[Bastian *et al.*, 2009] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proc. of 3rd ICWSM Conference*, 2009.

[Dhillon and Modha, 2001] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, January 2001.

[Efimova *et al.*, 2005] L. Efimova, S. Hendrick, and A. Anjewierden. Finding 'the life between buildings': An approach for defining a weblog community. *Internet Research*, 6:1997, 2005.

[Fortunato, 2010] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[Greene *et al.*, 2008] Derek Greene, Gerard Cagney, Nevan Krogan, and Pádraig Cunningham. Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

[Hassan *et al.*, 2009] A. Hassan, D. Radev, J. Cho, and A. Joshi. Content based recommendation and summarization in the blogosphere. In *Proc. of 3rd ICWSM Conference*, pages 34–41, 2009.

[Hayes *et al.*, 2007] C. Hayes, P. Avesani, and S. Veeramachaneni. An analysis of the use of tags in a blog recommender system. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 2772–2777, 2007.

[Herman *et al.*, 2000] Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Trans. on Visualization and Computer Graphics*, 6:24–43, 2000.

[Herring *et al.*, 2004] S.C. Herring, I. Kouper, L.A. Scheidt, and E. Wright. Women and children last: The discursive construction of weblogs. *Into the blogosphere: Rhetoric, community, and culture of weblogs*, 2004.

[Herring *et al.*, 2005] S.C. Herring, L.A. Scheidt, E. Wright, and S. Bonus. Weblogs as a bridging genre. *Information Technology & People*, 18(2):142–171, 2005.

[Herring *et al.*, 2007] S.C. Herring, J.C. Paolillo, I. Ramos-Vielba, I. Kouper, E. Wright, S. Stoerger, L.A. Scheidt, and B. Clark. Language networks on LiveJournal. In *Proc. 40th Annual Hawaii International Conference on System Sciences*, 2007.

[Hindman, 2009] M.S. Hindman. *The myth of digital democracy*. Princeton Univ Press, 2009.

[Jones *et al.*, 2008] S. Jones, S. Millermaier, M. Goya-Martinez, and J. Schuler. Whose space is MySpace? A content analysis of MySpace profiles. *First Monday*, 13(9), 2008.

[Lee *et al.*, 2010] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. In *SNA-KDD 2010*, pages 33–42. ACM, 2010.

[Marlow, 2004] C. Marlow. Audience, structure and authority in the weblog community. In *International Communication Association Conference, May, 2004, New Orleans, LA*. Citeseer, 2004.

[McDaid and Hurley, 2010] A. McDaid and N. Hurley. Detecting highly overlapping communities with Model-based Overlapping Seed Expansion. In *Proc. Int. Conference on Advances in Social Networks Analysis and Mining*, pages 112–119. IEEE, 2010.

[Nardi *et al.*, 2004] B.A. Nardi, D.J. Schiano, and M. Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? In *Proc. ACM conference on Computer supported cooperative work*, pages 222–231, 2004.

[Rosvall *et al.*, 2010] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, April 2010.

[Simonetto *et al.*, 2009] Paolo Simonetto, David Auber, and Daniel Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum*, 28(3):967–974, 2009.

[Van Dijck, 2004] J. Van Dijck. Composing the self: Of diaries and lifelogs. *Fibreculture*, 3, 2004.

[Van Doorn *et al.*, 2007] N. Van Doorn, L. van Zoonen, and S. Wyatt. Writing from Experience. Presentations of Gender Identity on Weblogs. *European J. of Women's Studies*, 14(2):143–159, 2007.

[Viégas *et al.*, 2009] F.B. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE Trans. on Visualization and Computer Graphics (InfoVis '09)*, 15(6):1137–1144, 2009.

[von Landesberger *et al.*, 2010] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs. In *EG 2010 - State of the Art Reports*, pages 37–60, 2010.

[Yang and Pedersen, 1997] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning*, pages 412–420, 1997.

[Yao, 2009] A. Yao. Enriching the migrant experience: Blogging motivations, privacy and offline lives of Filipino women in Britain. *First Monday*, 14(3-2), 2009.

[Yu, 2007] H. Yu. Blogging Everyday Life in Chinese Internet Culture. *Asian Studies Review*, 31(4):423–433, 2007.

[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, 2004.