Mining Shared Social Media Links to Support Clustering of Blog Articles

Darko Obradović
German Research Center for AI (DFKI)
& University of Kaiserslautern
Kaiserslautern, Germany
Email: darko.obradovic@dfki.uni-kl.de

Fernanda Pimenta
German Research Center for AI (DFKI)
& University of Kaiserslautern
Kaiserslautern, Germany
Email: fpimenta@dfki.uni-kl.de

Andreas Dengel
German Research Center for AI (DFKI)
& University of Kaiserslautern
Kaiserslautern, Germany
Email: andreas.dengel@dfki.uni-kl.de

***Abstract*—When monitoring blog articles for the tracking of a certain personality or product, the automatic identification of topic clusters is of high interest. Clustering by textual content is a popular method to accomplish this. In this paper we investigate how links between individual blog articles can be used to support this clustering with another dimension of information. Given the existing component structure of these networks, we focus on the extension with links based on shared social media resources. We show that the component structure extended in this way is of very high use for supporting textual clustering algorithms, and may be used for a new type of hybrid clustering algorithms in the future.**

I. Introduction

A. The Blogosphere

Blogs (weblogs) are an interesting phenomenon that arised with the Web 2.0. Commonly defined as "dynamic Internet pages containing articles in reverse chronological order", they can be utilised for various purposes by their authors, including private online diaries, journalism, corporate information delivery, etc.

Blogs also offer rich possibilities for interaction. Authors can include not only textual and multimedia content, but also link to original content, refer to articles in other blogs, maintain a collection of links to other blogs, or let visitors post comments to their articles, which often also contain links.

Thus blogs can and do link to each other. The resulting network of blogs forms the *blogosphere*.

B. Topics in the Blogosphere

Previous studies on the inner workings of the blogosphere have shown that bloggers tend to write in *bursts* about specific topics [1]. These can be topics about personalities, corporations, products or technologies. Furthermore, these topics are sometimes discussed in *conversations* across different blogs, where another blog article is mentioned, replied to or criticised [2].

This usually happens within communities in the blogosphere, which focus on specific domains, like politics, technology, culture etc. Obradovic et al. [3] have presented a methodology for the monitoring of specific domains, which can be a personality, a technology or any other entity that can be expressed in a search keyword. This methodology enables a good coverage of the blogosphere, along with a good measure of authority for the individual articles.

Based on such datasets, Schirru et al. have presented a textual approach for the identification of topics and trends in specific domains [4]. By using a NMF-clustering algorithm, the articles of the domain were clustered by the content of their titles, with the goal to identify the different topics and events the bloggers are talking about.

The evaluation revealed a good performance, but there still is room for improvement, and especially the right number of clusters to use is very difficult to decide.

C. Motivation

Given the goal to cluster blog articles in order to identify topics, results with a purely textual approach are already good, but could be improved. Considering the inner workings of the blogosphere with respect to linking behaviour and conversations, as described before, we follow the idea to utilise the link structure of the blog articles to support the clustering of articles by topic.

In [3] it was shown that the article network of a keyword-based domain is forming a series of independent components, as illustrated in Figure 1. This is against all statistical odds, as you would expect the links to form a giant component quickly, if they were set randomly [5]. This leads us to the hypothesis, that links are motivated by a common topic, and that the resulting components can be used as an additional input for a hybrid clustering algorithm, which uses textual and social information.

As a first step towards this challenging combination, we try to increase the information value of these components by extending them with shared links to social media resources. These links should also be motivated by a common topic, thus integrating these relations into the article network should provide a better set of components to start with.

The rest of the paper is organised as follows. In Section II we present the datasets we are using for our experiments. Our methodology for extending the blog article networks with social media links if presented in Section III and then evaluated in Section IV, before the paper is concluded in Section V.

978-1-4577-1133-6/11/$26.00 ©2011 IEEE

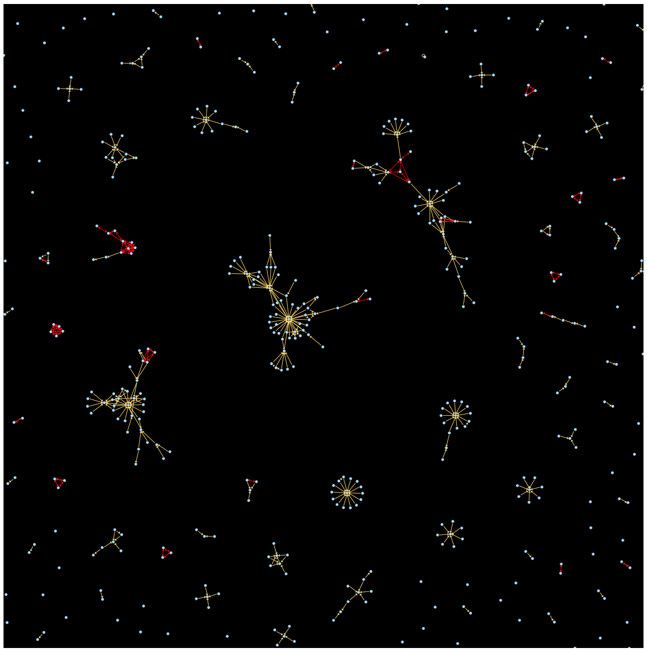

TABLE III
SOCIAL MEDIA RESOURCES PER DOMAIN

Domain	Resources	Shared Resources	Added Links
1	787	10.7%	146
2	198	2.0%	4
3	213	17.4%	321
4	463	18.4%	264
5	1,167	7.8%	402
6	135	0.0%	0
7	21,719	11.9%	17,170

The fact that two articles link to the same social media entity, relates them in a certain way. Looking at the top shared resources and the content of the articles that shared them, we can find good illustrations of this relation. For example, in domain 7, there are six articles that share the same Youtube video[1]. Reading the content of these articles, we notice that all of them talk about the same topic, the Nokia N900 and how it was able to run Google Wave. In another example, now for domain 5, five articles shared a link to a Facebook resource [2]. In this example, all articles write about the addition of Robin Williams songs to the "Band Hero" video game.

We now modify our article network by adding an additional link between two articles, if they share the same social media resource, i.e. they both link to it. Considering the social media resources and the articles as two classes of nodes in a two-mode network, we transform this network into its one-mode projection [7] by linking two articles together if they have at least one social media resource in common. The result of this projection are 18,307 new links between articles in all seven domains. We then added these new links to our static article network, if the nodes were not already connected by article citations before. Table III lists the shared resources per domain, and the number of new links that were added by the described extension of the network.

For larger groups of articles that share the same resource, this procedure always turns them into a clique, as all articles are then related to each other. This can lead to a very high number of links if many articles often share the same resource.

IV. EVALUATION

For evaluating the resulting component structure of the extended article network, we present two approaches, which show that the extended component structure meets the requirements for supporting the clustering of blog articles even better than the original article network does.

A. Probabilistic Evaluation

In Section III-C we presented some examples, which indicate a common topic of two articles when they share a social media resource. Given the assumption that links from article citations do so, and thus the resulting component structure reflects the topics to some degree, we will evaluate how the social media links correlate to the original component structure with a probabilistic approach.

TABLE IV
RANDOM AND REAL PROBABILITIES FOR THE ADDITION OF LINKS

	Case	Random	Real
1	within component	0.5%	6.1%
2	isolated ↔ component	22.6%	27.8%
3	isolated ↔ isolated	75.7%	63.4%
4	component ↔ component	1.2%	2.7%

Considering the existing original components, four types of cases may occur in theory, when adding a social media link to the article network:

1) it connects two articles within the same component
2) it connects an isolated article with a component
3) it connects two isolated articles
4) it connects two articles of two separate components

In order to evaluate these four types of cases, we compare the real results with the probabilities of the cases when adding a link between two articles by random. Given a network $G = (V, E)$ with a set V of n vertices and a set E of m edges, it consists of a series of z disjoint weakly connected components $C_j, 0 \leq j < z$ with $c_j \geq 2$ nodes each, and a set I of i isolated nodes. The probability p_x for a randomly added edge to be of case x out of the four defined cases can be calculated as follows based on maximally possible undirected connections.

$$p_1 = \frac{\sum_{C_j \in C} c_j \cdot (c_j - 1)}{n \cdot (n-1)} \tag{1}$$

$$p_2 = \frac{i \cdot (n-i)}{n \cdot (n-1)/2} \tag{2}$$

$$p_3 = \frac{i \cdot (i-1)}{n \cdot (n-1)} \tag{3}$$

$$p_4 = \frac{\sum_{C_j \in C} c_j \cdot (n - i - c_j)}{n \cdot (n-1)/2} \tag{4}$$

Table IV lists the expected probability for a random addition of a link for each case along with the real percentages obtained by the addition of links according to shared resources. The most important observation here is that shared resources have a twelve times higher probability to add links inside existing components, compared with raqndom additions. This is a very strong indicator supporting our hypothesis that shared social media resources relate articles writing about the same topic. Thus they can be used to support topic-clustering algorithms.

This indication implies that the other three cases should also do something meaningful, i.e. connect articles about the same topic. This would mean that previously isolated articles join a corresponding topic component, new components of a topic are formed, or two existing components of the same topic are connected.

[1] http://www.youtube.com/user/MobileDeveloperTV
[2] http://www.facebook.com/bandhero

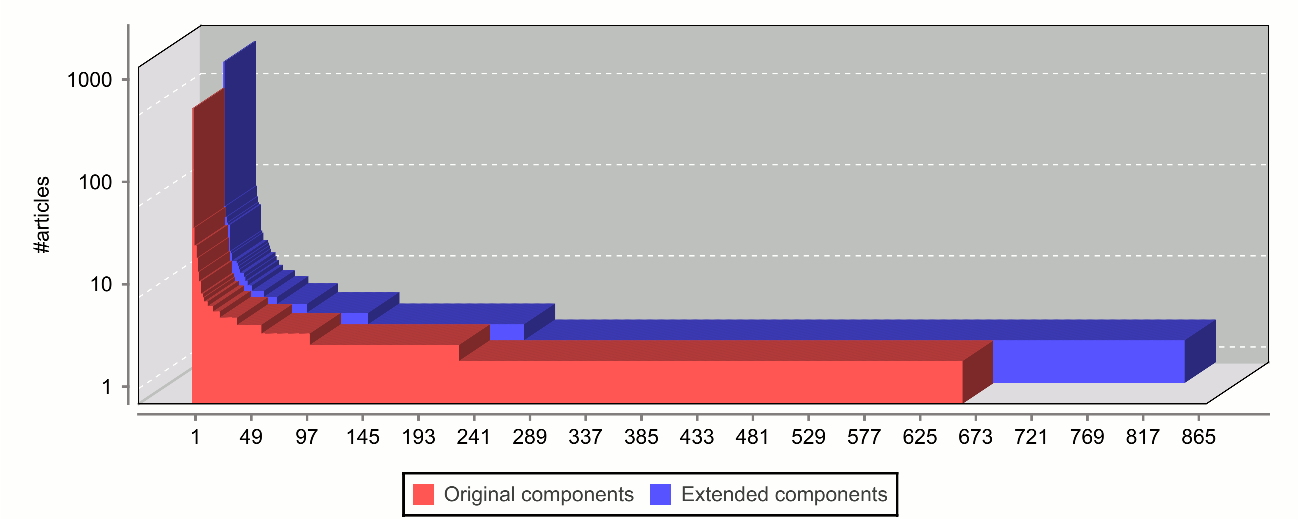

Fig. 2. Number and size of components before and after the extension in direct comparison

TABLE V
ARTICLE NETWORKS BEFORE AND AFTER SOCIAL MEDIA EXTENSION

Domain	Connected Articles		Components	
	before	after	before	after
1	145	188	94	105
2	48	51	34	37
3	99	161	91	106
4	73	123	63	72
5	64	143	58	78
6	9	9	8	8
7	664	1,915	358	474

B. Comparing the Original and Extended Components

Ultimately, the addition of all the new links should not result into a global connection of the graph, or the emergence of a giant component, as expected in a random network [5], instead the fragmentation of the graph should be preserved. Table V shows the differences between the original article network, and the social media extended network in terms of connected (i.e. non-isolated) articles and the number of weakly connected components. Indeed the number of components even increases in every single domain, while there are relatively few articles removed from isolation. This means that overall, existing components are densified and new ones created, which is an enormous improvement for supporting topic-based textual clustering algorithms.

Figure 2 provides a direct comparison of the component structures between the original article network and the network extended with shared social media. The nature of the curve is not changed at all by the extension, both curves show a typical power law with "fat had and long tail", but there are more and slightly larger components in the extended network, which signifies a larger and thus more robust input for the clustering, while preserving the original fragmented structure.

V. CONCLUSION

In this paper we have observed that articles in a domain tend to form components of specific topics by citation and conversation. Furthermore we observed that links to social media resources are often shared when writing about the same topic. We exploited this for extending the original article networks with these shared links, and obtained very good evidence that this extension will be a valuable input for a hybrid text-component-clustering algorithm. Developing such a new algorithm will be our goal for future work.

The mining also enables a further application. Once the clusters are found, the shared social media can be used to illustrate the clusters with the social media content in the visualization component of the final monitoring tool.

REFERENCES

[1] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," *World Wide Web*, vol. 8, no. 2, pp. 159–178, 2005.

[2] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu, "Conversations in the blogosphere: An analysis "from the bottom up"," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

[3] D. Obradovic, S. Baumann, and A. Dengel, "A social network analysis and mining methodology for the monitoring of specific domains in the blogosphere," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2010)*. IEEE, 2010, pp. 1–8.

[4] R. Schirru, D. Obradovic, S. Baumann, and P. Wortmann, "Domain-specific identification of topics and trends in the blogosphere," in *Advances in Data Mining. Applications and Theoretical Aspects. Industrial Conference on Data Mining (ICDM-10)*, ser. LNAI, P. Perner, Ed., vol. 6171, 2010, pp. 490–504.

[5] M. Molloy and B. Reed, "The size of the giant component of a random graph with a given degree sequence," vol. 7, p. 295305, November 1998.

[6] E. Adar, "A language and interface for graph exploration," in *CHI 06: Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM Press, 2006, pp. 791–800.

[7] M. Latapy, C. Magnien, and N. D. Vecchio, "Basic notions for the analysis of large two-mode networks," *Social Networks*, vol. 30, no. 1, pp. 31 – 48, 2008.