Soc. Netw. Anal. Min. (2013) 3:221–232
DOI 10.1007/s13278-012-0075-7

ORIGINAL ARTICLE

A social network analysis and mining methodology for the monitoring of specific domains in the blogosphere

Darko Obradović · Stephan Baumann · Andreas Dengel

Received: 15 November 2010 / Revised: 21 March 2012 / Accepted: 29 May 2012 / Published online: 20 June 2012

Abstract Whenever the question arises of how a product, a personality, a technology or some other specific entity is perceived by the public, the blogosphere is a very good source of information. This is what usually interests business users from marketing or PR. Modern search services offer a rich set of tools to monitor or track the blogosphere as a whole, but the analysis with respect to a certain domain is very limited. In this paper, we lay some foundations to aggregate blog articles of a specific domain from multiple search services, to analyze the social authorities of articles and blogs, and to monitor the attention articles of the domain receive over time. These are the building blocks required for a monitoring application that presents users the topics and trends in a specific domain along with the currently most interesting articles. This methodology has been instantiated and combined with additional textual analysis methods to create highly automated business intelligence application in the context of the Social Media Miner project.

D. Obradović (✉) · A. Dengel
German Research Center for AI (DFKI),
University of Kaiserslautern, Kaiserslautern, Germany
e-mail: obradovic@dfki.uni-kl.de;
darko.obradovic@dfki.uni-kl.de

A. Dengel
e-mail: andreas.dengel@dfki.uni-kl.de

S. Baumann
German Research Center for AI (DFKI),
Berlin, Germany
e-mail: stephan.baumann@dfki.de

1 Introduction

1.1 The blogosphere

The blogosphere contains a huge amount of information created by a multitude of sources. According to the "Technorati State of the Blogosphere"[1] there are at least 900,000 articles per day published, with an upward trend. Whenever the question arises how a product, a brand, a personality, an institution, a technology or some other specific entity is perceived by the public, the blogosphere is a good source of information. For this paper, we define such an entity as a *domain*.

These specific domains usually interest professionals in marketing and PR businesses the most, as opposed to the broader interests of sociologists and blogosphere researchers. The latter try to answer general research questions about structure (Marlow 2004), influence (Agarwal et al. 2008), or community relations (Chau and Xu 2007) in the blogosphere, to name a few examples.

Modern search services offer a rich set of tools to monitor or track the blogosphere, but the analysis with respect to a specific domain is very limited. For example, Icerocket Blog Trends[2] can plot the number of articles per day for a specific query. It plots a static, non-interactive curve, but there is neither an explanation of this curve nor access to further information. It has to be post-processed manually with different tools by the market researcher.

From our experiences we know that there is a strong demand for business-oriented social media monitoring, with the ultimate goal of making better decisions thanks to better information. That demand cannot be served by

[1] http://technorati.com/state-of-the-blogosphere/.

[2] http://trend.icerocket.com/.

search services yet; hence, we want to create a blogosphere-specific methodology to bootstrap such business intelligence systems.

1.2 The Social Media Miner project

The goal of the "Social Media Miner" (SMM) research project is to enable domain-specific blogosphere monitoring, which will then enable business intelligence applications. These applications can then perform clustering, trend detection, information extraction, sentiment analysis, or other content-based blog-specific mining technologies (Hassan et al. 2009) on top of this data.

Figure 1 shows the workflow realized in the SMM project. The collected blog articles are post-processed by a topic clustering component, which gives a timely overview of the activities inside a domain for a given timeframe. The information access per topic is then supported by a relevance ranking of the blog articles.

The focus of this paper is limited to describing the foundational social network analysis and mining aspects. We will justify all of our decisions, and provide empirical evidence where possible.

1.3 Rationale

In this paper, we pursue four concrete goals to enable domain-specific blogosphere monitoring, which will then enable business intelligence applications. These applications can then perform clustering, trend detection, information extraction, sentiment analysis, or other content-based blog-specific mining technologies (Hassan et al. 2009) on top of this data. The focus of this paper is limited to describing the foundational social network analysis and mining aspects. We will justify all of our decisions, and provide empirical evidence wherever possible, but a thorough evaluation is not possible before a concrete business intelligence application is built on top of this methodology.

As a first goal, we try to aggregate as many articles of the domain as possible. Kumar et al. (2005) have shown that in blogspace information evolves in bursts. This has been successfully modeled by Goetz et al. (2009). In consequence, there is a repeater effect for information, and the more articles we have at hand, the better the extent of this effect can be observed and exploited in textual processing methods. A selection of relevant articles can still be made afterwards, when presenting results to the user.

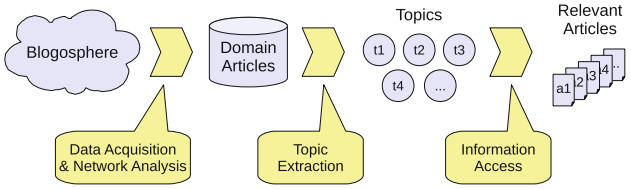

Fig. 1 The SMM workflow

In order to enable this selection, it is our second goal to derive a meaningful measure of social authority, based on links among blogs and articles. The more articles we have at hand, the better the interconnectivity between them. And the more accurate the social authority derived from these links, the better the filtering and ranking that can be presented to the user in the end.

Third, we will enable the approach to work over very long time periods of monitoring. Therefore, we need a metric of attention for articles that can find the "hot" articles and blogs in our evolving domain at any given point of time.

Last but not least, we want to identify the most important blogs of an observed domain after a relatively short initial observation period.

The rest of the paper is organized according to these four goals, including a final conclusion and an outlook to future work in this area.

2 Data aggregation

In order to find blog articles of our domains, we define the keywords for an appropriate search query and aggregate the search results from multiple blog search services. Thus, we do not have to set up a complete search engine infrastructure by ourselves, and we can reach more articles than a single search service can provide, as our experiments will show.

2.1 Existing experiences

An indicator for the hypothesis, that search engines obviously have very different indexes, is given by Herring et al. (2005), who noticed huge differences when comparing different top 100 lists with each other.

In a preliminary experiment (Wortmann et al. 2009), the quality and reach of five popular blog search services was analyzed manually to validate this hypothesis. These services were Technorati,[3] Google Blogsearch,[4] Bloglines,[5] Icerocket,[6] and BlogPulse.[7] The domain of this test was represented by the keyword "Henrietta Hughes", which unequivocally refers to an event on February 10th 2009, where this homeless person talked to US president Barrack Obama. The event had a

[3] http://www.technorati.com/.
[4] http://blogsearch.google.com/.
[5] http://www.bloglines.com/.
[6] http://www.icerocket.com/.
[7] http://www.blogpulse.com/.

noticeable impact on broadcast media, as well as on social media, especially the blogosphere.

None of the services delivered more than 50 % of all the articles found, and concerning the validity of the search results, there was a number of non-blog articles and pages not even mentioning the lady's name. Google Blogsearch had a comparatively high false positive rate of 50 %, and consequently, we left this service out of the final aggregation component. With these experiences, we implemented a number of heuristics to detect non-blogs, based on the URL, meta-data and the site content, in order to filter out as many of the invalid results as possible.

2.2 The aggregation component

For our analyses, we need the URL of each blog article along with the date of publication, the title and the textual content. As the methodology is intended to monitor a domain over a very long period of time, the crawler is implemented as a permanently running service that regularly queries the search services for the latest articles, and adds these to the data set.

All search services allow to return the query results unfiltered and sorted by date, enabling us to quickly fetch all the latest results. Each search result is listed with the notion of the article's age. In a second step, each result is validated and, if an RSS entry is available on the blog site, the more accurate date and the textual content is saved from it.

Another important aspect of our data sets is the link structure among these articles. We want to track all links, where the textual content of an article is citing another blog article in the domain. These links are used later as a social assessment of the authority of articles, as widely known from PageRank (Page et al. 1998) and similar algorithms.

We impose some requirements on these article links, in order to include only expressive ones. First of all, links between articles on the same blog are ignored, since their expressiveness of authority is doubtful at best. These often appear in a "Related Articles" section at the end of an article. Links from articles that contain dozens of references are also ignored, as these are usually spam articles trying to manipulate PageRank and other ranking algorithms.

In a next step, we extract the underlying blog URLs out of the article URLs and gain a second type of data, the blogs. We then collect the blogroll links between these blogs, according to our method presented in Obradovic et al. (2008). These will serve as supplementary authority indicators in the following network analyses.

2.3 Example data

We have chosen a number of different domains, from products over services up to personalities, to test our methodology on them. All seven domains have been observed during October 2009, and the data are available on the author's homepage[8] as a zipped MySQL dump file. Table 1 lists the seven domains along with the number of articles, blogs and links.

Table 1 Example data per domain

#	Domain	Articles	Article links	Blogs	Blog links
1	Android G1	2,511	416	1,319	460
2	VW Golf	1,328	99	806	136
3	Toyota Hybrid Car	2,719	138	1,521	246
4	Angela Merkel	2,150	103	1,415	1,057
5	Robbie Williams	3,595	84	2,253	517
6	Fraunhofer	348	10	289	23
7	Google Wave	15,836	2,017	10,594	4,793

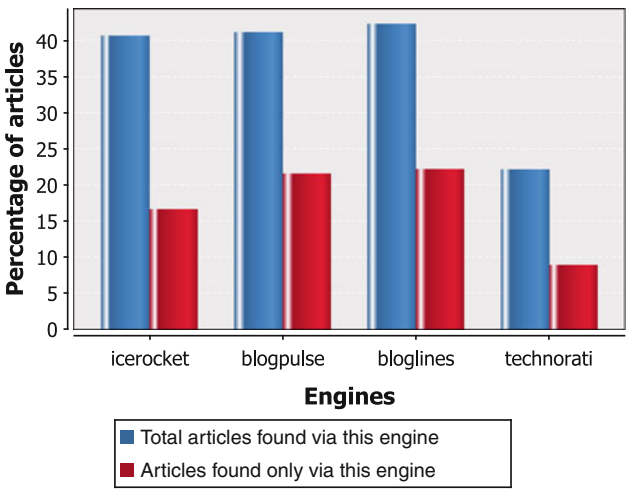

Fig. 2 Performance of the search engines

Based on this data, we have analyzed the performance of the four search engines that we used. Figure 2 depicts each search engine with two values. The blue bar denotes the percentage of articles of the aggregated set that was found via this engine, the red bar denotes the percentage of articles of the aggregated set that was found only via this engine. For our extensive data sets, none of the search engines was able to find more than 50 % of all articles, but each one contributed a significant share of articles that was not known to any of the other three engines.

This is in principle what we had expected and why we have chosen a meta-search approach, but the extent of the

[8] http://www.dfki.uni-kl.de/~obradovic/data.

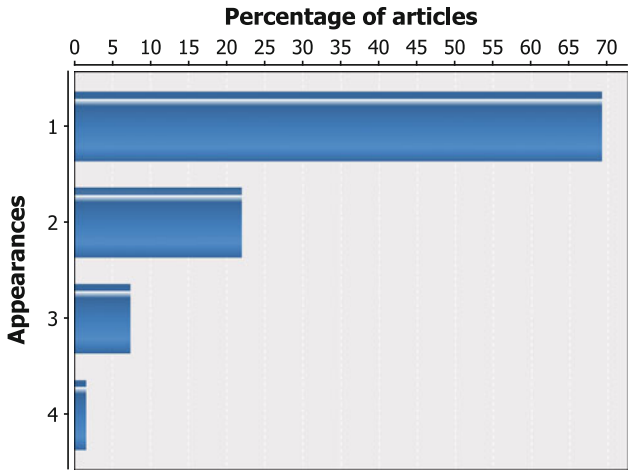

Fig. 3 Article ratios based on appearances

effect was not foreseen. It becomes more apparent when looking at the ratios of articles based on the number of engines they were found in. Figure 3 plots this data and reveals that only 1.5 % of all articles were found by all four search engines, the remaining 98.5 % were unknown to at least one of the engines, and nearly 70 % of the articles were found only via one engine.

With this characteristic number, that we will call the *appearances* of an article, we have another independent measure of article popularity available. Later, Fig. 6 will reveal that there is a high correlation between the number of appearances of an article and its number of citations.

3 Determining social authorities

Social authority can be defined as the measurement of centrality, importance or relevance induced by inbound links in social networks. There are many different metrics for authority in the field of Social Network Analysis (SNA) (Wasserman et al. 1994), which are all based on graph algorithms.

3.1 Notations

First of all, we summarize the terms and notations we will adhere to in the rest of this paper. In general we define a directed graph as $G = (V, E)$ with a set of vertices V and a set of directed edges $E = (V \times V)$. Given a node $v \in V$, the function $indeg(v)$ returns the in-degree of v, i.e. the number of incoming edges. The function $succ(v)$ returns the set of all successor nodes of v, and the function $pre(v)$ returns the set of all predecessor nodes of v.

3.2 Authority values

In this paper we do not focus on a specific metric for the measurement of authority. The presented methodology is intentionally designed to work with an abstract authority metric, with some constraining assumptions. We assume to have an abstract authority function *auth* returning normalized authority values for a given node.

$$auth : V \rightarrow [0, 1]. \quad (1)$$

An important property of this function for our reasoning in this paper is the direct dependency on the indegree of a node, as defined below:

$$\forall v \in V(indeg(v) > 0 \leftrightarrow auth(v) > 0). \quad (2)$$

All popular authority metrics like PageRank (Page et al. 1998), HITS (Kleinberg et al. 1998) or the more blog-specific iRank (Adar et al. 2004) comply to this condition and can be used with our methodology.

3.3 Networks from data aggregation

In the example data that we aggregated we have two separate social networks, the article network G_{articles} with citation links and the blog network G_{blogs} with blogroll links, as defined below:

$$G_{\text{articles}} = (V_{\text{articles}}, E_{\text{articles}}) \quad (3)$$

$$G_{\text{blogs}} = (V_{\text{blogs}}, E_{\text{blogs}}). \quad (4)$$

There also exist links between articles and blogs due to the containment of each article in a specific blog. This is a *two-mode network* on its own (see Wasserman et al. 1994, p. 39f). Looking at all three networks at once, we have a construct which we decide to call a *hybrid network*, which

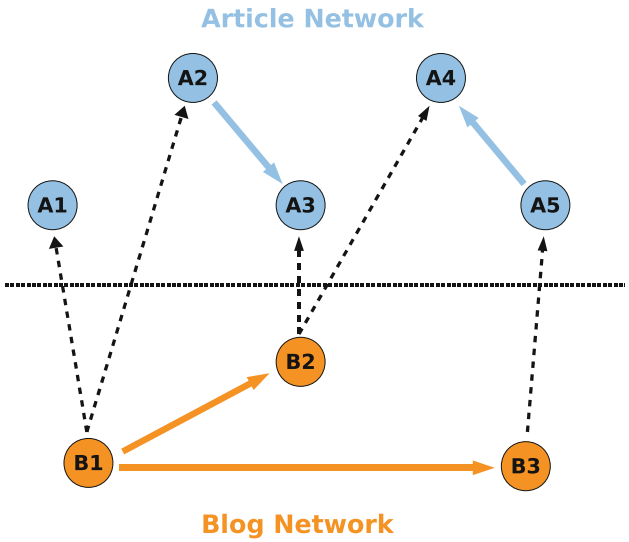

Fig. 4 A hybrid article/blog-network

is the starting point for our analyses. A simple example of such a network is given in Fig. 4.

3.4 Original article authority

Using the plain network G_{articles}, we can compute the authority values for articles from this network. We define $auth_{\text{article}}(v)$ to be the *original article authority*, as derived from G_{articles}. However, the empirical data showed that articles are very sparsely connected in specific domains (see Table 1), and therefore we decided to use a more sophisticated method for calculating social authorities, which will give us more articles with non-zero authority values in the end.

For the determination of our social authorities we use a mutually dependent measure. The authority of an article depends on the authority of its blog, and the authority of a blog depends on the authorities of its articles. We present the derivation of the two measures in the following two sections. We will use the original article authority later, to compare if the final social authority of articles indeed gives less non-zero authority values than the original article authority does.

3.5 Blog authority

To realize these mutually dependent metrics, we first map the article links into the blog network. This is possible with a function returning the hosting blog for a given article.

$$blog : V_{\text{articles}} \rightarrow V_{\text{blogs}}. \tag{5}$$

So we can map each egde $(a_1, a_2) \in E_{\text{articles}}$ from the article network to an edge $(blog(a_1), blog(a_2))$ in the blog network with another function.

$$map : E_{\text{articles}} \rightarrow (V_{\text{blogs}} \times V_{\text{blogs}}). \tag{6}$$

As we have excluded links between articles of the same blog in the data aggregation, this cannot introduce loops in the new graph. However, this can introduce parallel edges, and turn our blog network into a *multi-graph* G_{multi}, i.e. a graph with multiple sets of differently typed or colored edges (see Wasserman et al. 1994, pp. 145f).

$$G_{\text{multi}} = (V_{\text{blogs}}, E_{\text{blogs}}, \{map(e), e \in E_{\text{articles}}\}). \tag{7}$$

Figure 5 illustrates the resulting multi-graph G_{multi} for the example hybrid network from Fig. 4.

In order to compute the authorities of blogs with standard algorithms, which are not designed to operate on multi-graphs, we have to perform one last transformation, the unification of parallel edges.

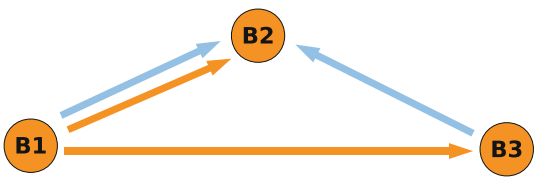

Fig. 5 Blog multi-graph from the hybrid network

All multi-edges are transformed to normal weighted edges, with a weight equivalent to the number of original edges in the multi-edge. This results in a weighted directed network, which is the most complex form that can be analyzed by standard algorithms without major modifications. In the example multi-graph from Fig. 5, the multi-edge ($B1$, $B2$) would be transformed to an edge with a weight of 2 while the remaining two edges have a weight of 1 each.

As a result of this, we assume to have an authority function $auth_{\text{blog}}$, derived from the multi-graph transformed in such a way.

3.6 Validation of article link mapping

While the idea behind the mapping is relatively straight forward, we still want to validate this decision by comparing the resulting graphs $G_{i,\text{multi}}$ for a domain i with the original graph $G_{i,\text{blogs}}$ with respect to their goodness-of-fit to the well-known model of *scale-free networks* (Barabasi and Albert 1999). Series of previous work found out that the degree distributions of a lot of typical real-world networks, especially those concerned with popularity and Internet like our blogroll network (Shirky 2003), follow a power law distribution of the form $p(x) = x^{-\alpha}$. We try to demonstrate the validity of our suggested mapping by checking whether the resulting multi-graphs $G_{i,\text{multi}}$ fit better to the expected power law than the original blog networks $G_{i,\text{blogs}}$ without the mapping.

Fitting networks to power-laws is a very active statistical research issue. In this work, we mainly adhere to the goodness-of-fit method for discrete values as suggested by Clauset et al. (2009), which is based on a *maximum likelihood estimation* and the *Kolmogorov–Smirnov statistic D* for the goodness-of-fit. The constant α is usually in the range of [1.5, 3.5] for most real-world data. The fit is done with a constraining parameter x_{min}, which limits the degree values considered for the fitting. The background here is that low degree values tend to introduce a bias for most empirical data. On the other side, a high x_{min} increases the uncertainty in the estimation of α. For our data, we strictly use $x_{\text{min}} = 2$ for all fits as a fair compromise given our relatively small samples.

Table 2 shows the fitting results for our domains 1 and 7, which are affected the most by mapping article links into the blog network (see Table 1). For both domains it compares the fit of the original blogroll network opposed to the fit of the multi-graph after the mapping. The estimated values for α are within the expected range. The standard deviation and D are smaller for the multi-graphs, which is an indicator for a better fit, but the differences are only small in domain 1 and negligible in domain 7. Overall, we cannot state that our multi-graphs improve the goodness-of-fit

Table 2 Comparison of Power-Law Fits

Network	Samples	Samples $\geq x_{\min}$ (%)	α	D
$G_{1,\text{blogs}}$	134	52.2	2.05 ± 0.13	0.1022
$G_{1,\text{multi}}$	163	55.8	1.86 ± 0.09	0.0659
$G_{7,\text{blogs}}$	1,267	47.9	2.10 ± 0.05	0.0471
$G_{7,\text{multi}}$	1,516	49.4	2.03 ± 0.04	0.0466

significantly, but they clearly do not impair it. Furthermore, the fits of the multi-network consider a slightly larger fraction of the samples, which makes them statistically more certain.

3.7 Combined article authority

We calculate the final article authority by combining two factors. The first one is the original article authority $auth_{\text{article}}$, as described in Section 3.4. The second factor is the authority of the blog the article was published in, using the function $auth_{\text{blog}}$, as described in Sect. 3.5. Additionally, we need a function $auth_{\text{comb}}$ that returns the final *combined authority* value in the interval [0,1] for a given article a. In the simplest form, such a function looks as follows.

$$auth_{\text{comb}}(a) = \frac{auth_{\text{article}}(a) + auth_{\text{blog}}(blog(a))}{2}. \quad (8)$$

Any other form of combination can be used with this methodology, but the suitability depends on the exact requirements of the final application.

Through this procedure for the derivation of the combined article authority, we achieve to compute meaningful authority values for substantively more articles than by using the original article authority. We provide some empirical evidence for both claims in the following sections, for the increase in non-zero authoritative articles, and for the validity of the new measure.

3.8 Increase of authoritative articles

Table 3 lists the number of authoritative articles per domain for both metrics, when using the original article authority, and when using the combined article authority metric. Along with the absolute number we also provide the percentage with respect to all articles contained in the domain data set. Based on these two numbers we present the increase factor, calculated as the number of authoritative articles using $auth_{\text{comb}}$ divided by the number of authoritative articles using $auth_{\text{article}}$.

The increase achieved by this method is between 2.2 and 10.4 in our example domains. It directly depends on the structure of the hybrid blog/article network. The better the blogs are connected and the more articles a blog contains on average, the higher the increase. What we cannot explain yet is the impact of the domain on that structure. In the domains number 2 and 3, which both deal with cars, we have despite different sizes a highly similar structure and thus a nearly identical increase factor. This could be generally true for car domains, or a coincidence; at the least, it calls for further investigation.

Table 3 Authoritative articles per domain

Domain	$auth_{\text{article}} > 0$		$auth_{\text{comb}} > 0$		Increase
1	145	6 %	510	20 %	3.5
2	48	4 %	165	12 %	3.4
3	99	4 %	343	13 %	3.5
4	73	3 %	670	31 %	9.2
5	64	2 %	663	18 %	10.4
6	9	3 %	20	6 %	2.2
7	664	4 %	2,920	18 %	4.4

3.9 Evaluation of combined article authority

We justified our combined authority measure from a theoretical network perspective, proposing that a blog's authority also influences an article's authority. We are able to cross-check it with the authorities expected from the number of appearances of an article in the different search engines (see Sect. 2.3). Figure 6 plots for each class of appearances the percentage of articles with that number of appearances, that have a non-zero authority value. The red squares joined by a red line refer to the original article authority measure $auth_{\text{article}}$, the blue circles joined by a blue line refer to the combined article authority measure $auth_{\text{comb}}$.

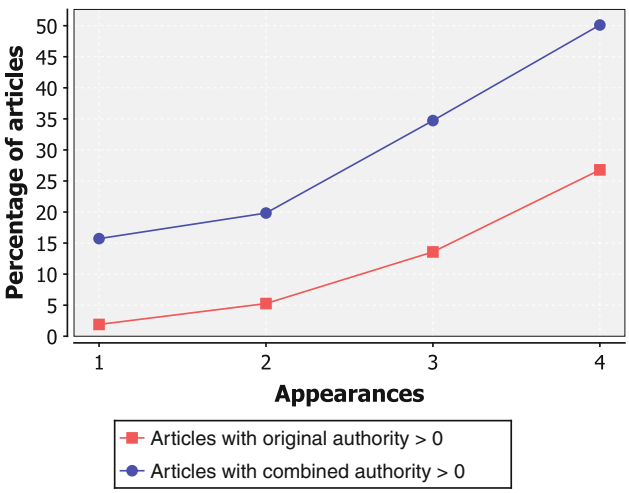

Fig. 6 Original and combined article authorities based on appearances (colour figure online)

The original authority of an article is obviously highly correlated to its appearances (red line), the more appearances an article has, the higher the probability to have a non-zero authority. We can also see that our combined authority measure does not only increase the number of articles with authority, but does so in a highly consistent way with respect to the appearances. There is the same correlation to the number of appearances (blue line), which is a strong indicator for the validity of our method.

4 The flow of time

Since it is our third goal to monitor specific domains over a long period of time, we have to consider the time dimension as well. In SNA, *dynamics* is usually interpreted as evolving networks, in which new nodes and edges are added over time (Berger-Wolf et al. 2006; Skyrms et al. 2000). The intent is to identify patterns of behaviour.

Both our original networks are evolving networks as well, but for business intelligence we are not interested in the patterns of behaviour in the first place. We are more interested in a measurement of *attention*, that reveals which articles are cited most often at a certain point of time.

The blog network with its blogroll links remains a static network in that case. Blogroll links do not change often, a regular update of each blog along with an update of the network is enough.

However, the article network is not only evolving, but is also a highly *time-sensitive* network. Each article has a timestamp and a link between two articles is characterized by the time difference between its two end points. During the monitoring of a domain, new articles are constantly added, new links are discovered and old links lose expressiveness for measuring the current attention. For example, an article that has been referenced a hundred times three months ago is not as relevant for the current situation of the domain as an article that has been cited two dozen times in the last 48 h. In contrast, we have seen articles being referenced during our observation, which were published 6 months ago. Thus, these still get a good share of attention months after their publication, and make them relevant for the current point of time. This makes clear that it is not enough to consider the articles of the last n days only, but that we need a more sophisticated measure instead.

4.1 Ages of links

To analyze this phenomenon, we first look at the occurring time differences of links in our example data. We first introduce some notations to handle this cleanly. Assume the current point of time is t_{cur}. Given a function $time(a)$ that returns the point of time an article a was published at, and a subtraction operator that returns the time difference between two points of time, we can define a function age for a directed edge from article a_s to article a_t as follows.

$$age((a_s, a_t)) = time(a_s) - time(a_t). \tag{9}$$

Figure 7 illustrates the ages of the links found in our example data sets, rounded down to full days. Using a log-scale for the number of links of a certain age, we can observe that the vast majority of links to an article is set right after publication, but there are still a number of links set several days after publication. So there is good reason

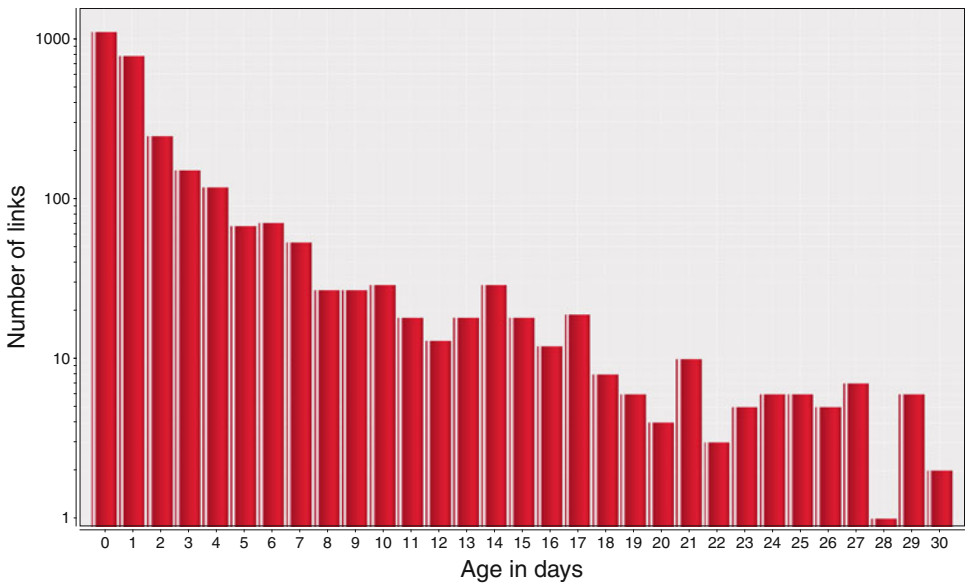

Fig. 7 Distribution of link ages

to respect this time difference when monitoring a specific domain over a long period of time.

4.2 A time-sensitive network model

Consequently, we extend our methodology to consider the age of links for the determination of an article's attention. This will allow articles to have high attention values, even if they were published long time ago. We choose an approach of link decay realized via edge weights.

We can define a time-sensitive weight function for an edge $e = (a_s, a_t)$, which can be implemented in various ways. For simplicity, we present an example with a linear decay that is parameterizable with a maximum lifetime of $\Delta t_{\max}$ for an edge. The resulting weight function looks as follows.

$$weight((a_s, a_t)) = 1 - \min\left(\frac{t_{\text{now}} - time(a_s)}{\Delta t_{\max}}, 1\right). \quad (10)$$

With this weight function, a time-sensitive attention can be computed exactly like in a simple static weighted network. For the time-sensitive network, we define the indegree of a node n at the point of time t_{cur} as the sum of the weights of all incoming links as follows.

$$indegree(n) = \sum_{s \in pre(n)} weight((s, n)). \quad (11)$$

Figure 8 illustrates the resulting effect for two articles. We have chosen two popular articles from domain number 7, which both have 31 incoming links in the static article network. The first one was published on the first day of October, the second one on the ninth day. With t_{cur} moving from day 1 to day 31, we plot the current indegree of the articles with $\Delta t_{\max}$ set to 10 days.

While the two articles had the same indegree in the static network, it is now visible how the attention is spread over time. There are articles that receive a lot of attention for a short period of time, and articles that receive less attention, but for a longer period of time.

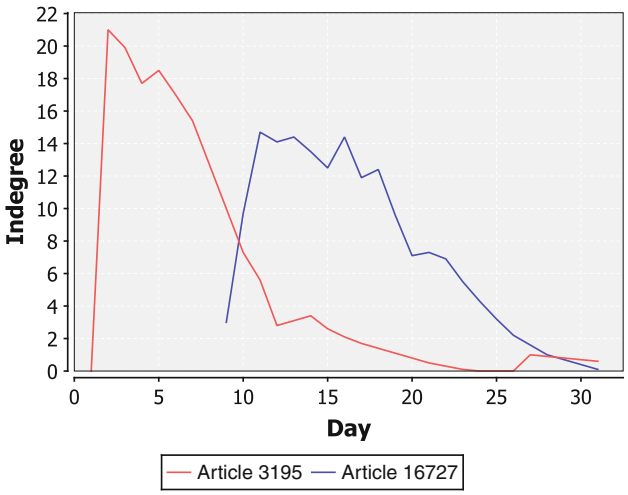

Fig. 8 Indegrees over time for two articles

Thanks to a model based on a standard weighted directed network, we can calculate the attention of an article with any standard algorithm. We assume to have a metric $att(a)$ that returns the attention of an article for the current point of time calculated from the time-sensitive network.

4.3 Time-sensitive relevance

With the new dimension of attention, the selection and ranking of presumably relevant articles at a certain point of time can be performed with a combination of article authority and attention. With authority only, we had to rely on articles around the given point of time to make a time-sensitive selection. Combined with attention, we can now consider the whole data set and an accordant scoring function will find the currently relevant articles, independently from their date of publication. In the simplest form, such a scoring function can look as follows:

$$relevance(a) = att(a) \cdot auth_{\text{comb}}(a). \quad (12)$$

4.4 Attention for blogs

With this model at hand, we can also provide an attention metric for blogs. Using the same mapping as for the calculation of blog authorities, we can construct a time-sensitive blog network. The fusion of multi-edges has to be done by adding up the weights of the mapped edges, time-insensitive blogroll links have to be omitted for attention calculation. In this resulting weighted network, we can calculate attention values in the same way as done for the articles.

Having this blog attention metric and the blog authority metric, these two can be combined to a time-sensitive relevance metric in the same way as done for the articles.

4.5 Enabling retrospection

With the extensions from the last section, we are now capable of monitoring blog article relevances over long periods of time. But currently, the calculation of metrics always refers to the current point of time t_{cur}. Often it is interesting to retrieve metrics or make calculations for points of time in the past, especially when there is a demand for a comparison of the current state with states in the past.

We therefore extend our network structure with retrospection capabilities. This means that for any given point of time from the past, we want to enable all network calculations. In other words, we want the network to be easily revertible to any point of time t_{net} in a single instance.

Duplicating network structures with snapshots and the like is considered too expensive and not expected to scale satisfiably.

We define the network structure valid for a point of time t_{net} as follows:

$$G(t_{\text{net}}) = (V(t_{\text{net}}), E(t_{\text{net}})) \tag{13}$$

$$V(t_{\text{net}}) = \{a \in V \mid time(a) \leq t_{\text{net}}\} \tag{14}$$

$$E(t_{\text{net}}) = \{(s,t) \in E \mid time(s) \leq t_{\text{net}}\}. \tag{15}$$

Such a network structure can be easily incorporated into a network data structure with a time attribute for the network. Therefore, we have to override some basic methods to respect this attribute as defined in the formulas 14 and 15. These basic methods are the default network methods for getting nodes and edges, the node methods for getting incoming and outgoing edges, and the edge method for getting its weight. With these changes, all subsequent methods based on these basic methods will behave in the correct way without further modifications. This is no problem in modern object-oriented languages, and we implemented this in plugins for the Perl `SNA::Network` package located at CPAN.[9]

5 The evolution of domain blogs over time

While the blog articles are being aggregated over time, we constantly see articles published in previously unknown blogs, as well as articles published in blogs known from previous articles observed in the domain.

5.1 Article publications per blog

To get an idea of the relation of these two phenomenons, Fig. 9 plots the daily updated average number of articles per blog over all of our seven domains.

After a very steep increase in the first days, when most blogs of a domain are found with their first article, the curve is becoming less steep over time, which means that we see more and more articles published by the same blog. In consequence, this leads us to the idea that after some time of observation, the opinion-leading group of blogs for this specific domain should emerge in the structure of the blogroll network. This fact is of very high relevance in the given context of media monitoring for marketing or business intelligence, since it gives the user a hint where the blogspace of interest can be influenced in the most effective way. Such an influence could be the placement of advertisings, the distribution of comments, incentives for featured articles, and so on.

[9] http://www.cpan.org/.

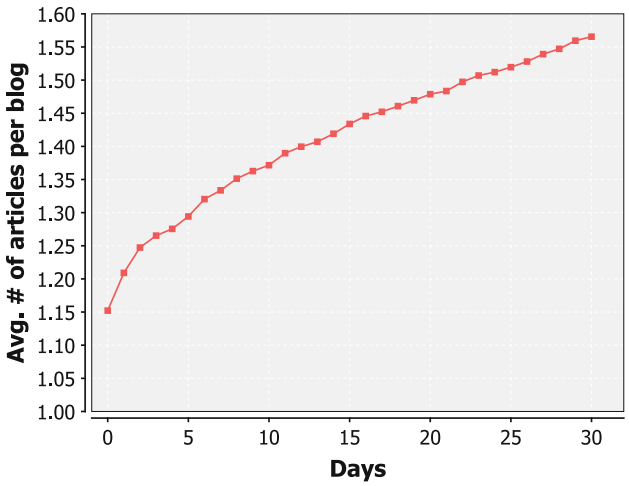

Fig. 9 Average number of articles per blog over time

5.2 Identifying the most authoritative blogs of a domain

In order to detect the opinion-leading group for a domain, we use the method of identifying *k-in-cores* as shown in Obradovic et al. (2010). This relatively simple concept, based on Seidman's *k-cores*, returns a nested sequence of subgroups, where each member has at least k incoming links from the rest of the subgroup. In contrast to cliques and clans, there are no dozens of overlapping groups, but a straight sequence of nested cores, with the most inner $k_{\max}$-core being the most important group.

For all of our blog networks we observe the emergence of a *giant component* over some time. This is a weakly connected component that contains the majority of nodes in a graph, while the rest of the nodes is either isolated or connected in multiple small weakly connected components. Table 4 lists the number of weakly connected nodes V_{conn} in the domain's blog network as opposed to the number of nodes in the giant component $|V_{\text{GC}}|$. Furthermore, it lists the highest value $k_{\max}$ for the detected *k-in-core*, which is a cohesive subgroup in which each member receives at least

Table 4 Emergence of in-cores in the blog networks

Domain	$\lvert V_{\text{conn}}\rvert$	$\lvert V_{\text{GC}}\rvert$	$k_{\max}$	Members
1	279	231	4	6
2	128	55	2	10
3	215	103	3	4
4	462	429	3	20
5	385	194	2	23
6	26	8	0	8
7	3,066	2,603	6	7

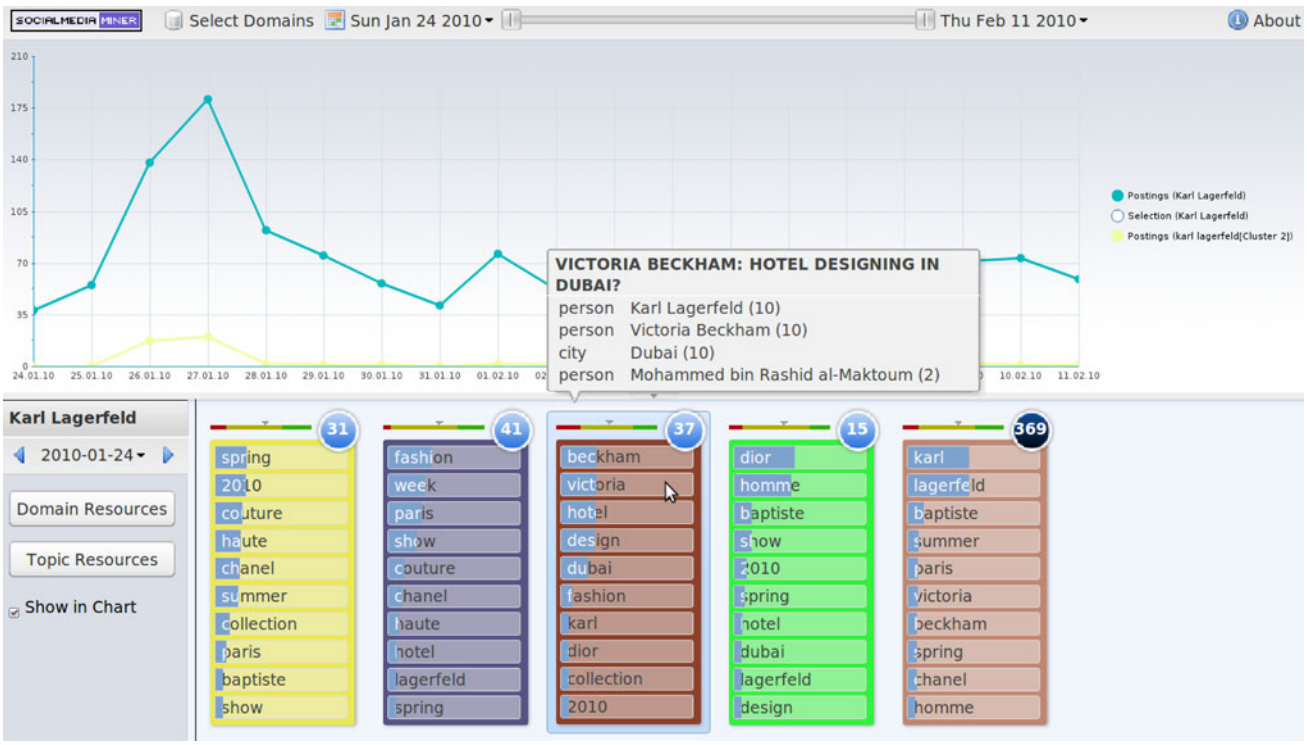

Fig. 10 The SMM main view for the domain "Karl Lagerfeld"

k incoming links from the other members of the *k-in-core*. The number of members is also listed in the table.

We tried to evaluate this by comparing the resulting $k_{\max}$ value with the expected value from randomly generated networks. These were generated according to the *configuration model* (see Section IV.B.1 in Newman 2003, p. 22), based on the degree distribution of the giant components for each domain. But with our networks being very sparse, more than 99 % of all sampled networks in each domain had a resulting $k_{\max} \leq 1$, which means that the emergence of high k for the most inner *k-in-core* is extremely improbable. The emergence of high k values in our networks are thus a significant indicator for an authoritative subgroup according to the *A-List* theory, as outlined in Obradovic et al. (2010). Defining a threshold $\Delta t_{\max}$ for active blogs, this method can constantly provide the user with a list of the most influential blogs for the domain.

Looking at our largest example data set, the "Google Wave" domain, we have a 6-in-core with 7 members. As k-cores are a nested structure, we look at the larger 5-in-core with 20 members, and find all the famous technology blogs in there, especially Engadget[10] and TechCrunch[11] for example, which confirms very clearly that this method is working well.

6 The final tool

Having implemented the aggregation component and the authority/relevance measurements described in this chapter, this has been implemented together with a textual topic-clustering component (Schirru et al. ICDM-10) and a sentiment analysis component (Pimenta et al. 2010). The result is the prototype of the SMM project, a web-based graphical interface realising the architecture defined in Sect. 1.

6.1 Domain overview

Figure 10 shows the starting screen for the observation of the German star fashion designer "Karl Lagerfeld". The upper part plots the volume of articles aggregated during the observation, as described in Sect. 2.

The lower part shows the detected topics in the selected time interval. This includes a list of the 10 most characteristic keywords of the topic, the volume of the topic, the overall sentiment in the topic, and by highlighting, a key phrase of the topic along with detected semantic entities.

6.2 Articles overview per topic

When accessing a topic, for example the Dubai Design hotel project planned together with Victoria Beckham, the interface lists all blog articles relevant to the topic ranked

[10] http://www.engadget.com.

[11] http://www.techcrunch.com.

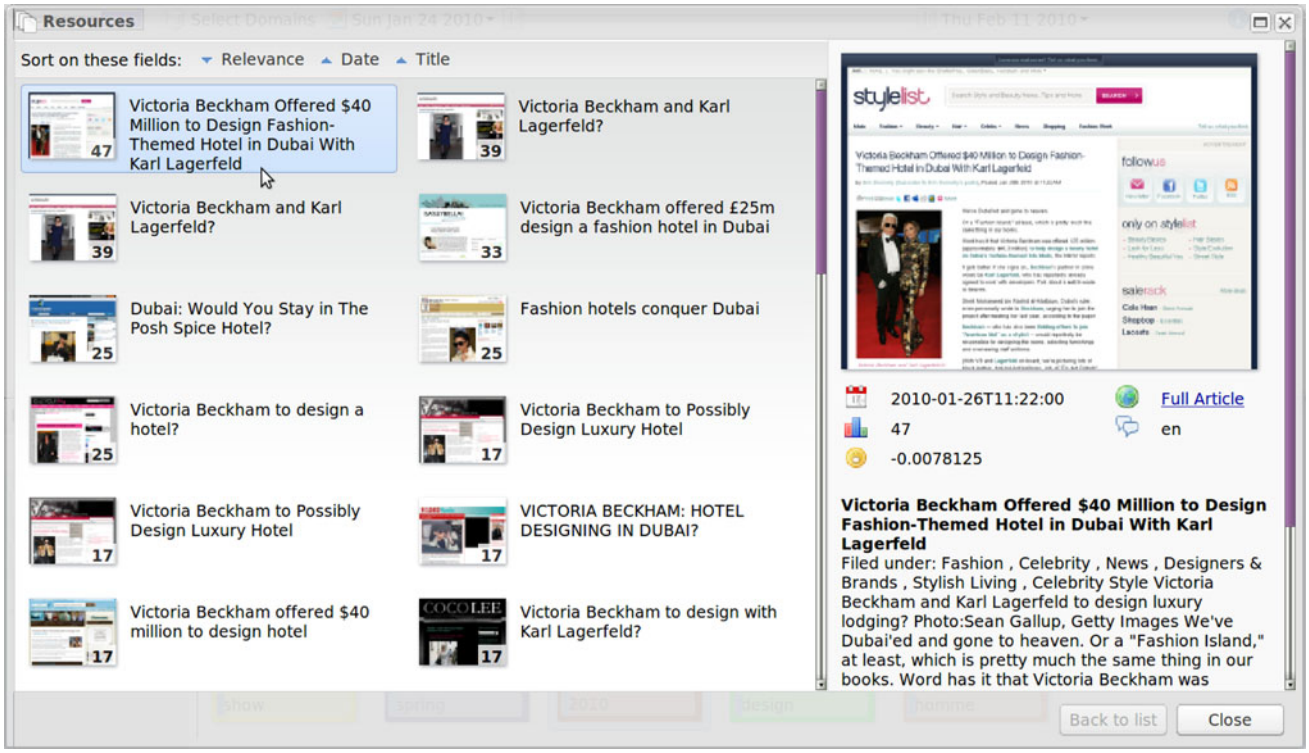

Fig. 11 The article list for the topic around the "Dubai Design Hotel"

by authority as shown in Fig. 11. Authority values have been computed using the HITS algorithm, globally normalized to rounded numbers between 0 and 100. Thanks to the combined article authority, we can list several really authoritative articles at the top of each topic in all cases.

Here, the user can select the articles of interest from the left part, and see a thumbnail of the article page, some meta data and the full text on the right part.

7 Conclusion

In this paper, we have presented a methodology that enables the aggregation of blogosphere data of specific domains, the analysis of authorities, attentions and relevances over long periods of time, and the detection of authoritative blogs after a short period of observation.

Depending on the specific interests, these steps can be instantiated with concrete parameters and algorithms suited best for the individual use-case, and provide an extensive data set for further textual processing, clustering, trend detection, topic tracking, sentiment analysis, and so on. Finally, it provides metrics for the selection and ranking of reading recommendations based on the results of these processing steps. Thus, it can be used to bootstrap business intelligence applications.

We have described and analyzed our data sets, and justified our decisions based on empirical data or by giving real-world examples. We also presented the final application, where this methodology has been partly realized and used in practice by project partners and industrial clients. The feedback has been very positive.

8 Future work

Taking into account the time dimension presented in Sect. 4, the reverse network flows in the components also reveal some information about the dependencies of events and about the role of an article, be it as an authority or as a hub.

The support of established textual processing algorithms with the component structures found in the networks, as illustrated in Fig. 12, is one of the logical next steps from this work. This already has been partially approached in Obradovic et al. (2011).

Considering business intelligence applications that monitor topics of a domain, we found many links to shared social media resources in the blog articles that can be easily used to illustrate the detected topics for an enhanced user experience.

Another entity found in the frequent external links are blog articles. Despite the very effective meta-search

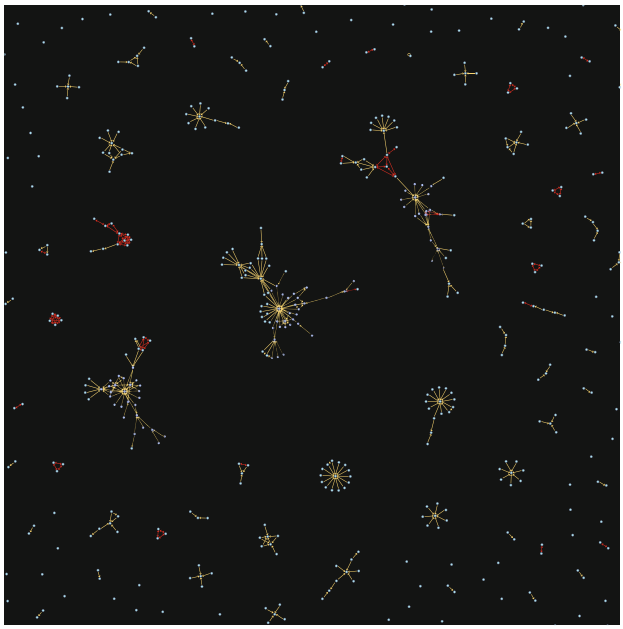

Fig. 12 Visualisation of article components

aggregation described in Sect. 2, there still exist dozens of articles that are referenced from our article set, which have to have a certain authority in consequence, but which were not listed by any of the search engines. Adding these articles with appropriate article extraction heuristics should further improve the data set quality.

Acknowledgments This research has been financed by the IBB Berlin in the project "Social Media Miner", and co-financed by the EFRE fonds of the European Union.

References

Adar E, Zhang L, Adamic LA, Lukose RM (2004) Implicit structure and the dynamics of blogspace. In: Workshop on the weblogging ecosystem, WWW2004, New York, NY

Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. In: WSDM '08: proceedings of the international conference on Web search and web data mining. ACM, New York, NY, USA, pp 207–218

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Berger-Wolf TY, Saia J (2006) A framework for analysis of dynamic social networks. In: KDD '06: proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 523–528

Chau M, XU J (2007) Mining communities and their relationships in blogs: a study of online hate groups. Int J Hum Comput Stud 65(1):57–70

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. arxiv.org, February 2009. http://arxiv.org/pdf/0706.1062

Goetz M, Leskovec J, Mcglohon M, Faloutsos C (2009) Modeling blog dynamics. In: International conference on weblogs and social media

Hassan A, Radev D, Cho J, Joshi A (2009) Content based recommendation and summarization in the blogosphere. In: International conference on weblogs and social media

Herring SC, Kouper I, Paolillo JC, Scheidt LA, Tyworth M, Welsch P, Wright E, Yu N (2005) Conversations in the blogosphere: an analysis "from the bottom up". In: Proceedings of the 38th annual Hawaii international conference on system sciences. IEEE Computer Society, p 107.2

Kleinberg JM (1998) Authoritative Sources in a hyperlinked environment. In: Proceedings of the 9th annual ACM-SIAM symposium on discrete algorithms. AAAI Press, pp 668–677

Kumar R, Novak J, Raghavan P, Tomkins A (2005) On the bursty evolution of blogspace. World Wide Web 8(2):159–178

Marlow C (2004) Audience, structure and authority in the weblog community. In: Proceedings of the international communication association conference

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Obradovic D, Baumann S (2008) Identifying and analysing Germany's top blogs. In: Proceedings of the 31st German conference on AI. Springer, pp 111–118

Obradovic D, Baumann S (2010) A journey to the core of the blogosphere. In: Memon N, Alhajj R (eds) From sociology to computing in social networks. Lecture Notes in Social Networks, vol 1. Springer, Berlin, pp 25–43

Obradovic D, Pimenta F, Dengel A (2011) Mining shared social media links to support clustering of blog articles. In: Proceedings of the 2011 international conference on computational aspects of social networks (CASoN 2011). IEEE, pp 181–184

Page L, Brin S, Motwani R, Winograd T (1998) The pagerank citation ranking: bringing order to the web. Stanford University, Technical Report. http://explorer.csse.uwa.edu.au/reference/browse_paper.php?pid=23328182

Pimenta F, Obradovic D, Schirru R, Baumann S, Dengel A (2010) Automatic sentiment monitoring of specific topics in the blogosphere. In: Workshop on dynamic networks and knowledge discovery (DyNaK 2010)

Schirru R, Obradovic D, Baumann S, Wortmann P (2010) Domain-specific identification of topics and trends in the blogosphere. In: Perner P (ed) Advances in data mining. Applications and theoretical aspects. Industrial conference on data mining (ICDM-10), LNAI, vol 6171. Springer, Berlin, pp 490–504

Shirky C (2003) Power laws, weblogs, and inequality. http://shirky.com/writings/powerlaw_weblog.html

Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Natl Acad Sci USA 97(16):9340–9346

Wasserman S, Faust K, Iacobucci D (1994) Social network analysis: methods and applications (Structural analysis in the social sciences). Cambridge University Press, Cambridge

Wortmann P (2009) Topic-based blog article search for trend detection. Technical University of Kaiserslautern, project thesis. http://www.dfki.uni-kl.de/obradovic/download/pa-wortmann.pdf