

# Understanding weblog communities through digital traces: a framework, a tool and an example

Anjo Anjewierden<sup>1</sup> and Lilia Efimova<sup>2</sup>

<sup>1</sup> Human-Computer Studies Laboratory, University of Amsterdam, Kruislaan 419, 1098 VA Amsterdam, The Netherlands, [anjo@science.uva.nl](mailto:anjo@science.uva.nl)

<sup>2</sup> Telematica Instituut, PO Box 589, 7500 AN Enschede, The Netherlands, [Lilia.Efimova@telin.nl](mailto:Lilia.Efimova@telin.nl)

**Abstract.** Often research on online communities could be compared to archaeology [16]: researchers look at patterns in digital traces that members leave to characterise the community they belong to. Relatively easy access to these traces and a growing number of methods and tools to collect and analyse them make such analysis increasingly attractive. However, a researcher is faced with the difficult task of choosing which digital artefacts and which relations between them should be taken into account, and how the findings should be interpreted to say something meaningful about the community based on the traces of its members. In this paper we present a framework that allows categorising digital traces of an online community along five dimensions (people, documents, terms, links and time) and then describe a tool that supports the analysis of community traces by combining several of them, illustrating the types of analysis possible using a dataset from a weblog community.

## 1 Introduction

Although research on online communities has a long-standing history, the technological infrastructure and social structures behind them evolve over time. In this respect communities supported by weblogs is a relatively recent phenomenon.

A weblog is “a frequently updated web-site consisting of dated entries arranged in reverse chronological order” [22]. Weblogs are often perceived as a form of individualistic expression, providing a “personal protected space” where a weblog author can communicate with others while retaining control [11]. On one hand, a randomly selected weblog shows limited interactivity and seldomly links to other weblogs [13]. On the other hand, there is growing evidence of social structures evolving around weblogs and their influence on norms and practices of blogging. This evidence ranges from voices of bloggers themselves speaking about the social effects of blogging, to studies on specific weblog communities with distinct cultures (e.g. [23]), to mathematical analysis of links between weblogs indicating that community formation in the blogosphere is not a random process, but an indication of shared interests binding bloggers together [17].

Often research on online communities could be compared to archaeology [16]: researchers look at patterns in digital traces that members leave to characterise the community they belong to. In the case of weblog communities relatively easy access to these traces and a growing number of methods and tools to collect and analyse them make such analysis increasingly attractive (e.g. papers from the annual workshops on the Weblogging ecosystem at the WWW conference in 2004-06).

Many of the existing tools apply text or temporal analysis to large volumes of weblog data, often focusing on short bursts in time or popular topics (e.g. [1], [15], [21]). Others apply methods of social network analysis to identify and characterise networks between bloggers based on links between weblogs (e.g. [12], [19]). In our work we focus on combining both in order to go beyond currently available views on weblog data, aiming at developing tools that take into account existing community structures [14] and support the understanding of specific conversational clouds [18] and the “cloudmakers” behind them [20].

Although our work is based on the analysis of digital artefacts that weblog community members leave online, we find it important to articulate explicitly how studying the results points to more general questions about weblog communities: which digital artefacts and which relations between them are taken into account, and how the findings should be interpreted to say something meaningful about the community based on the traces of its members.

In this paper we present a framework that allows categorization of digital traces of an online community along five dimensions (people, documents, terms, links and time) and then describe a tool that supports the analysis of community traces by combining several of these dimensions, illustrating the types of analysis possible using a dataset from a weblog community.

## 2 Framework

In this section we present a simple framework that can assist in the analysis of online communities. The formulation of the framework is motivated by the perceived need to provide community researchers with a conceptual tool to focus on particular aspects of the community.

There is a strong relation between the framework we propose and ongoing research into the study of online communities. The field of social network analysis (SNA) can be characterised by studying the relations between persons and their links, sometimes taking into account time. The field of text mining from communities, sometimes called semantic social network analysis [4], mainly looks at the relation between terms and documents, largely disregarding the notion of the individual. Finally, the area of identifying trends in communities (e.g. [10], [8]) looks at documents, terms and time. The research that is closest to what we are trying to achieve is work on *iQuest* by Gloor and Zhao [9]. Their tool supports studying communities by making it possible to look at the community as a whole (topics discussed) and the contribution of members (who says what and when).

When thinking about online communities there are, therefore, at least five dimensions that play an important role and are possibly of interest for investigation:

**Document** A self-contained publication by a member in the community. Examples of documents are a web page, email or weblog post.

**Term** A meaningful term used by one or more members of the community. These terms occur in documents.

**Person** A member of the community.

**Link** A reference from one document to another document, and implicitly between the persons who authored the documents.

**Time** The date, and possibly time, of publication of a document.

The framework thus focuses on communities that leave digital traces in the form of documents, and derives the other dimensions from the metadata (person, time) and content (terms, links). Given a dataset represented along these dimensions the researcher can navigate through it by specifying one or more initial dimensions, fixating a particular dimension (e.g. focusing on a particular term, person, or time period). Navigating along multiple dimensions makes it possible for the researcher to obtain both an overall view (what are the most frequent terms used in the community) and more detailed views (term usage of a particular member over time). The more dimensions involved, the more detailed, and maybe also the more interesting are the results of the analysis.

## 3 Tool

The framework has been implemented on top of a tool called tOKo [7]. tOKo is an open source tool for text analysis, with support for ontology development and, given the extensions described in this paper, exploring communities.

The only input tOKo requires is a corpus of (HTML) documents. Applied to community research, we assume a corpus represents the documents of the community under study. To be able to apply the framework it is necessary that for all documents the author (person) and the date of publication (time) is provided as metadata. Terms are automatically extracted by the *Sigmund* module of tOKo [2] which clusters lexical variants (abbreviations, inflections and alternate spellings) into a single, possibly compound, term. For example the term *community* has the lexical variants *community* and *communities*. Links between documents are by default extracted assuming the documents are HTML (anchor element). The implicit links between persons are inferred by considering the author of a document. After this extraction process is completed we have the data along all five dimensions for a given community represented by a corpus.

Fig. 1 shows the user interface of tOKo with the community research extensions.<sup>3</sup> The main difference between the base version of tOKo and the extension are additional methods and visualisations to cater for the links and time dimensions. The community research functions are available through the *Community* popup in the menubar and by clicking inside the five browsers with community data at the top.

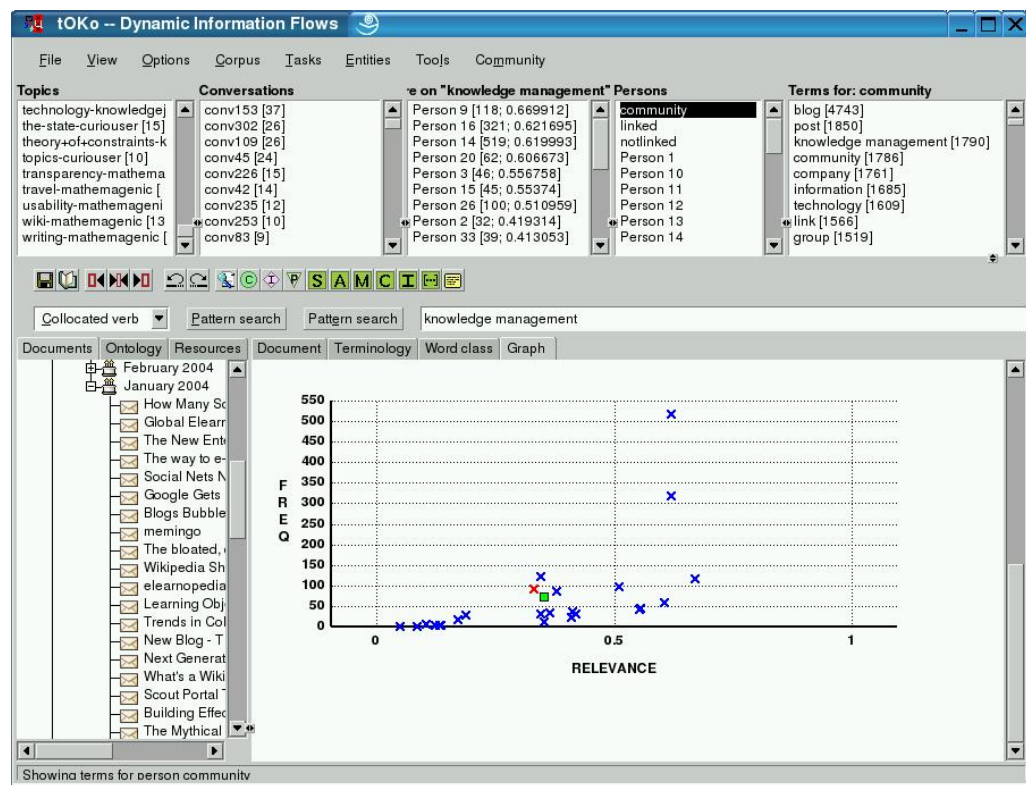


Fig. 1. Screenshot of the tOKo user interface with community research extensions

## 4 Examples

Our research focuses on a cluster of weblogs in areas of knowledge management and social software. This is a dense social network of weblog authors, and may be classified as a community,

<sup>3</sup> All other figures in this paper just show the content of the *Graph* sub-window.

given the many bonds and interactions between participants (see [6] for a discussion). The goal in this research is to understand how knowledge develops in a weblog community and to develop tools to monitor those processes. Although knowledge flows in a community are difficult to grasp with automatic analysis we focus on inferring them from combined investigation of patterns in language use and linking patterns between community members.

In the rest of this section we provide an example how the framework can be applied to address questions emerging in our study. In the analysis the tool is used to analyse a corpus of 6329 documents (weblog posts), over a period of a single year (2004), written by 24 persons (bloggers).

A researcher who wants to study a community asks herself questions. Answering such questions requires several transformations. First, a question will be related to a particular *slice* of the complete dataset according to the framework and the selection of the appropriate subset, which may result in less than five dimensions, is necessary. Second, on the resulting subset, some method of computation has to be applied. The particular method of computation obviously depends on the question asked, and in general the scientific community offers a wide variety of choices here. Finally, the results of the computation must be presented to the researcher. This is also the order we use in answering the questions below: relation to the framework, methods used and visualisation of the results.



Fig. 2. Network of terms that have a high co-occurrence with “instant messaging” in the community

#### 4.1 Q: What is the topical focus of the community?

This is a question that can be answered by considering a single dimension: terms. The rightmost browser in Fig. 1 shows the most frequent terms for the community at hand, low-content and high-frequency terms have been filtered out by *Sigmund*. A researcher might, based on this list conclude that this is a “knowledge management community”. The more frequent terms blog and post are the result of the publication medium.

A simple example of restricting the dataset is to only look at the terms of a single person. Some examples of the most frequent terms for individual members of the community are: **P1**: community, technology, virtual community, role, process; **P2**: knowledge management, organization, information, blog, article; **P3**: company, community, blog, corporate, knowledge management.

## 4.2 Q: How are community topics related to each other?

One approach to answering such a question is to consider an operationalisation of “related to” as “occurring in the same document”. This, first of all, requires two dimensions: document and term. And secondly, we need a statistical method to compute the relevance of what it means for two terms to be in the same document. For the latter we use the co-occurrence metric defined in [3]:

**Definition:** Let  $n(B | A)$  (respectively  $n(B | \neg A)$ ) be the number of occurrences of the term  $B$  in documents that contain the term  $A$  (respectively do not contain the term  $A$ ), and likewise let  $n(* | A)$  (respectively  $n(* | \neg A)$ ) be the total number of terms in the documents that contain the term  $A$  (respectively do not contain the term  $A$ ). Then the **co-occurrence degree**  $c(B | A)$  is defined as

$$c(B | A) = \frac{n(B | A)/n(* | A)}{n(B | \neg A)/n(* | \neg A)}, \quad 0 \leq c(B | A) \leq \infty$$

We say that  $B$  co-occurs with  $A$  to at least degree  $k$  if  $c(B | A) \geq k$ . Note that  $c(B | A) = 1$  if  $B$  is as frequent in documents containing as it is in documents not containing  $A$ , i.e. that term  $B$  and  $A$  seem to be unrelated.

Fig. 2 shows a graph of the co-occurrence network of the term “instant messaging” for the entire community. Edges between terms denote high co-occurrence (degree  $> 1.5$ ). For example, “instant messaging” has a high co-occurrence with both “telephony” and “bandwidth”. In addition, “telephony” and “bandwidth” themselves also have a high co-occurrence with each other.

The co-occurrence metric thus provides a device that enables finding related terms in the community. Obviously, the same method can be applied to a member of the community by fixating a single point on the person dimension.

## 4.3 Q: Do community topics change over time?

This question involves the dimensions term and time. And the obvious way to answer it might be to plot the frequency of a term over time, similarly to what BlogPulse [8] does. There are, however, technical, social and perhaps principle reasons for not (only) using plain frequency over time. The technical reason is that frequency of term usage in a particular (small) community generally results in an irregular sequence of spikes that makes it difficult to identify trends. Trending frequency for very large communities (e.g. the entire blogosphere as with BlogPulse) does not have this problem: when a large event occurs, within days a significant proportion of the blogosphere will mention it. A social reason is that, because we are studying a community, there is a significant difference when a term is resonated in the community, compared to when it is not.

The social issue is addressed by considering that a community can be defined in terms of who-links-to-who (along the person dimension), but that the real discussions in the community in all likelihood consists of the linked documents. Put another way, there is a principle reason to view the community not just as the collection of all documents produced by the community, but to also look at the cross-section of the documents that are linked. In order to investigate this we split the dataset in two sub-sets depending on whether a link exists within the community as follows:

$$D_{linked} = \{d_i \in D \mid link(d_i, d_j), person(d_i) \neq person(d_j)\}$$

That is,  $D_{linked}$  is the set of all documents linked in the community excluding self-links.  $D_{unlinked}$  is the difference between  $D$  and  $D_{linked}$ .

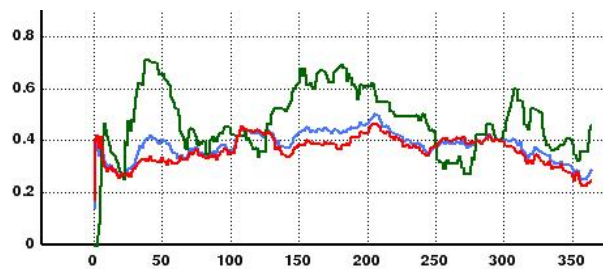
The technical issue is addressed by using *tf.idf* (term frequency vs. inverse document frequency) rather than frequency and by computing *tf.idf* over a sliding time-window to identify

trends. The formula we use for determining the relevance of a term  $i$  for a document  $j$  is one of the many variants of  $tf.idf$ :

$$weight(i, j) = (1 + \log(tf_{i,j})) * \log(N/df_j)$$

where  $tf_{i,j}$  is the frequency of term  $i$  in document  $j$ ,  $df_j$  the number of documents that contain term  $i$  and  $N$  the total number of documents.

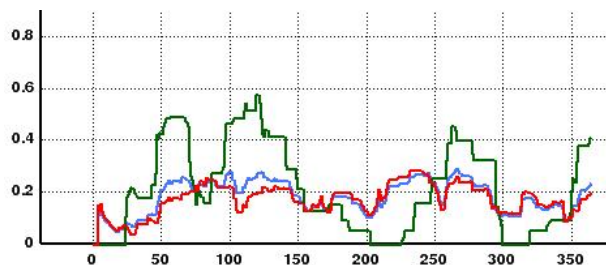
The time-window can be defined by the researcher, the default is a period of two weeks which takes into account that discussions in the blogosphere have a lag that is measured in terms of days rather than hours (as compared to discussions over email).



**Fig. 3.** Trend of “knowledge management” for all, linked and unlinked posts

Fig. 3 shows an example of the trend of the term “knowledge management” used in the community. The x-axis represents time in days and the y-axis is the moving  $tf.idf$  average over a period of two weeks. The three lines represent the values for the community as a whole (blue), linked documents ( $D_{linked}$ , green) and unlinked posts ( $D_{unlinked}$ , red). A conclusion that might be drawn from this visualisation is that “knowledge management” is more specific for linked posts than it is for unlinked posts and that this is the case continuously. Therefore, KM is one of the key terms of the community.

A graph for the term “Skype”, drawn using the same method, is shown in Fig. 4. The pattern which emerges is very different than the graph of KM. Skype is used in bursts from time-to-time (green peaks), then dies away (unlinked usage is above linked usage). According to the terminology used by Gruhl et al. [10], in this community KM is a *chatter* term, whereas Skype occurs in *spikes*. Gruhl et al. look at all documents in their community and make no difference between whether links exist between documents. This corresponds to the average trend in the graphs and it is interesting to observe that both KM and Skype would be chatter topics using Gruhl’s approach, whereas the inclusion of the link dimension reveals that in our community the trend patterns are different between how KM and Skype are used in discussions.



**Fig. 4.** Trend of “Skype” for all, linked and unlinked posts

#### 4.4 Q: What makes a specific community member different from others?

We address this question by looking at term usage, other answers might come from studying linking practices (SNA). Intuitively, we can define that a person is aligned with the community when she uses the same terms at the same rate compared to the community as a whole. Terms that are used more often by a person (compared to the community), might point to expertise or a certain passion for such terms (and the topics these terms point to).



Fig. 5. “Fingerprint” of term usage by a person with respect to the community

The method we use to find terms that significantly identify a person compared to the community is to compute for each term the following:

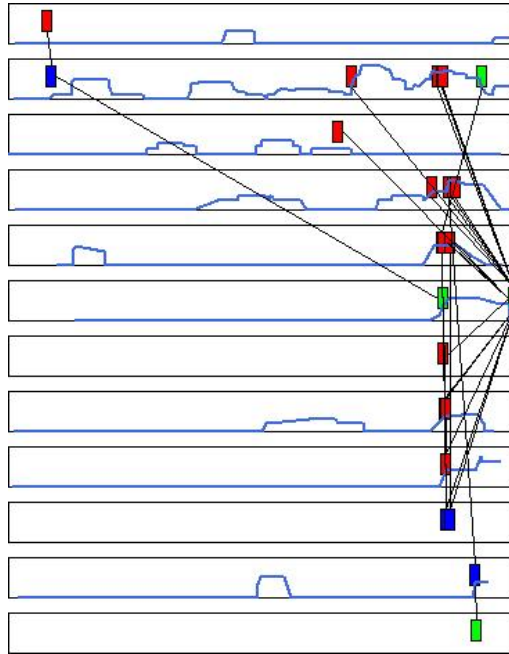
$$IDF(i) = \log(N/df_i)$$

$$sig(i, p_j) = average((1 + \log(tf_{i,j})) * IDF(i)), d_j \in D_{p_j}$$

where  $D_{p_j}$  is the set of documents by person  $p_j$ . In other words, we use the inverse document frequency (IDF) of the entire community and compare it to the average term frequency of a particular person. The results can be visualised by a list of the values for all terms. Fig. 5 shows a visualisation as a “fingerprint”, sometimes referred to as a *Zeitgeist*. In this kind of visualisation the font size of a term increases with significance.

#### 4.5 Q: What conversations occur in the community?

We define a conversation in a community as a set of documents that are linked directly or indirectly: if  $d_1$  links to  $d_2$  and  $d_3$  links to  $d_2$  all three documents are part of a single conversation [5]. Fig. 6 gives an example of the visualisation of such a conversation. Left to right is the time dimension, top to bottom are the persons involved in the conversation (in the tool the names are printed to the left of the boxes). The small coloured rectangles are the documents making up the



**Fig. 6.** An example of visualising a conversation

conversation, and the lines between them are the links (it is only possible to link backward in time, so no directional arrow is required).

A related question that could be asked is: what is this conversation about? For this we use the same technique as with fingerprinting, and for this conversation, it results in the following terms being the most significant (compared to all documents in the community): KM Europe, personal knowledge management [workshop], keynote, workshop, Amsterdam, etc. The conversation was about a personal knowledge management workshop at KM Europe in Amsterdam. The researcher can guess this from the above list of terms or by inspecting the documents making up the conversation in more detail.

The wavelike lines for each person shows the use of the term “personal knowledge management” over time (*tf.idf* for the given person). As can be observed the second person mentions this term regularly, whereas some of the other participants only pick it up when entering the conversation. Such visualisations can perhaps be used to identify when the community picks up a new term and study the stickiness of terms over time.

Finally, we note that Fig. 6 contains datapoints from all five dimensions from the framework in a two-dimensional visualisation.

## 5 Conclusions

In this paper we presented a framework that allows categorising digital traces of an online community along five dimensions (people, documents, terms, links and time) and illustrated how it can be applied to translate general questions about a weblog community and its members into specific questions that could be answered with specific sub-sets of the data and specific methods for analysis.

In addition to supporting research on communities the approach we propose might be useful for community members, facilitators or sponsors. For example, newcomers are often overwhelmed by a wealth of information available in a community and find it difficult to navigate

between multiple digital artefacts or find the right people to address. Dimensions of our framework could provide an additional way to navigate those, for example, by discovering documents or people associated with their own topic of interest. Community moderators or sponsors might be interested in identifying trends over time, to monitor patterns of activity (e.g. dynamics conversations) or to identify emergent “hot” topics or thought-leaders in order to adjust necessary support.

We find the framework and its implementation useful in several respects. First, it supports translating research questions into questions that could be answered based on a definite number of dimensions directly connected to the data available from the community interactions. Second, it inspires thinking of alternative ways to answer research questions by looking at combinations of the dimensions of data which might otherwise be missed. Third, it allows us to quickly compare alternate methods or research questions proposed in the literature. Finally, although our examples in this paper refer to a weblog community, our experience elsewhere suggests that the framework could be useful in other online community cases, for example for analysing forum-supported communities such as communities of practice.

Apart from extending the tool and its functionality we see two major challenges going forward. The main challenge is that we desperately need an “automatic” method to derive topics. Some researchers avoid this issue by simply stating term equals topic. Others are seriously addressing the topic issue, but report that automatically extracting them from weblog data is non-trivial [4]. Another challenge is to provide a user interface that allows a researcher to enter a question directly. Currently, the user interface is extended with a new control for each type of question.

**Acknowledgements.** The authors wish to thank Robert de Hoog and Rogier Brussee for their contributions to the underlying work and the reviewers for their constructive comments. This work was partly supported by the Metis project<sup>4</sup>. Metis is an initiative of the Telematica Instituut, Basell and Océ. Other partners are CeTIM, Technical University Delft, University of Amsterdam, University of Tilburg and University of Twente (all in The Netherlands).

## References

1. E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. HP Information Dynamics Lab, 2004.
2. A. Anjewierden, R. Brussee, and L. Efimova. Shared conceptualizations in weblogs. In Thomas N. Burg, editor, *BlogTalks 2.0: The European Conference on Weblogs (July 2004)*, pages 110–138, Vienna, February 2005. Danube University of Krems.
3. A. Anjewierden, R. de Hoog, R. Brussee, and L. Efimova. Detecting knowledge flows in weblogs. In Frithjof Dau, Marie-Laure Mugnier, and Gerd Stumme, editors, *Common Semantics for Sharing Knowledge: Contributions to 13th International Conference on Conceptual Structures (ICCS 2005)*, pages 1–12, Kassel, July 2005. Kassel University Press.
4. B. Berendt and R. Navigli. Finding your way through blogspace: using semantics for cross-domain blog analysis. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
5. L. Efimova and A. de Moor. Weblog conversations. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005)*, Los Alamitos, 2005. IEEE Press.
6. L. Efimova, S. Hendrick, and A. Anjewierden. Finding “the life between buildings”: An approach for defining a weblog community. In *Internet Research 6.0: Internet Generations (AOIR)*, Chicago, October 2005.
7. Anjo Anjewierden et al. tOKo and Sigmund: text analysis support for ontology development and social research. <http://www.toko-sigmund.org>, 2006.
8. N. Glance, M. Hurst, and T. Tomoyioko. Blogpulse. In *WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York, 2004.

---

<sup>4</sup> <http://metis.telin.nl>

9. P. A. Gloor and Yan Zhao. Analyzing actors and their discussion topics by semantic social network analysis. In *10th Conference on Information Visualization*, London, July 2006.
10. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explorations*, 6(2):43–52, 2004.
11. M. Gumbrecht. Blogs as “protected space”. In *WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York, 2004.
12. S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, and M. Tyworth. Conversations in the blogosphere: An analysis “from the bottom-up”. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005)*, Los Alamitos, 2005. IEEE Press.
13. S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS 2004)*, Los Alamitos, 2004. IEEE Press.
14. M. Hodder. Link love lost or how social gestures within topic groups are more interesting than link counts. <http://napsterization.org/stories/archives/000513.html>, 2005.
15. M. Hurst. 24 hours in the blogosphere. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
16. Q. Jones. Virtual communities, virtual settlements, and cyber archaeology: a theoretical outline. *Journal of Computer-Mediated Communication*, 3(3), 1997.
17. R. Kumand, J. Novak, P. Raghaven, and A. Tomkins. Structure and evolution of blogspace. *CACM*, 47(12):35–39, 2004.
18. A. Levin. Conversation clouds. <http://alevin.com/weblog/archives/001692.html>, 2005.
19. C. Marlow. Investment and attention in the weblog community. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
20. M. Ratcliff. Cloudmakers r us. [http://www.ratcliffeblog.com/archives/2005/08/cloudmakers\\_r\\_u.html](http://www.ratcliffeblog.com/archives/2005/08/cloudmakers_r_u.html), 2005.
21. M. Thelwall. Blogs during the London attacks: Top information sources and topics. In *WWW Workshop on the Weblogging Ecosystem*, Edinburgh, 2006.
22. J. Walker. Weblog. Routledge Encyclopedia of Narrative Theory, 2005.
23. C. Wei. Formation of norms in a blog community. In S. Antonijevic, L. Johnson, C. Ratliff, and J. Reyman, editors, *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*. University of Minnesota, 2004.