Journal of Universal Computer Science, vol. 14, no. 4 (2008), 546-556
submitted: 3/10/07, accepted: 11/2/08, appeared: 28/2/08 © J.UCS

Success Factors in a Weblog Community

Christian Safran
(Institute for Information Systems and Computer Media
Graz University of Technology, Austria
csafran@tugraz.at)

Frank Kappe
(Institute for Information Systems and Computer Media
Graz University of Technology, Austria
frank.kappe@iicm.edu)

Abstract: User generated content published via weblogs (also known as blogs) has gained importance in the last years, and the number of globally available weblogs increases. However, a large fraction of these show low publishing activity and are rarely read. This paper is a quantitative analysis of success factors in a community of over 15.000 weblogs, hosted by a local Austrian newspaper. We looked at publishing activity by content type, community activity and writing style. Also, the interconnectedness of the community was analyzed.

Keywords: User Generated Content, Weblog Analysis, E-Communities, Blogging
Categories: H.3.5, H.4.3, H.5.1, M.6

1 Introduction

In the age of the so-called Web 2.0, user generated content is credited increasing importance, as *participation* is one of the key characteristics of this concept [O'Reilly 05]. In the context of journalism the generation of news content by users, also called *citizen journalism* or *grassroots journalism* [Gillmor 04], is gradually complementing and, in some cases, even contesting information provided in a classic, publishing environment. Especially in the news domain, weblogs (also known as blogs) have become an important source of information beside newspaper websites. This paper will focus on such *user generated content* published via weblogs hosted by an Austrian newspaper.

A weblog is a “log of the web”, a term coined by Barger in 1997 [Paquet 03]. The entries in such a weblog, the so called posts are presented in reverse chronological order. Weblog software usually provides the possibility for comments and trackbacks. A *trackback* is a technology which allows weblogs to be automatically notified when other weblogs link to individual blogs and create backward links [SixApart 04]. In such a way a distributed, collective and interlinked blogosphere is created [Kumar 04].

The key question we asked ourselves in this context is: What makes a weblog successful? In the context of this research, success is measured by the number of visits to a weblog. For the use in this analysis, visits were counted as unique IP addresses, counting a new visit after 20 minutes of inactivity. So, the question can be reformulated as: what are the factors that lead to a high number of visits of the content created by the users?

Another interesting aspect of user generated content is the community of the involved users. Questions in this context concern the structure of this community. Is the community divided into smaller communities or is there a central group of active users?

2 Related Work

Quite a lot of research on weblogs has been published in the recent years. Interesting in this context are publications on ranking weblogs like [Kritikopoulos 06]. The authors present a modification of the PageRank algorithm [Page 98] designed to take into account the links between the weblogs and the similarity of the users, as well as links to non-weblog URLs. The target is to rank weblogs by importance.

[Du 2006] tries to answer the question for success factors of weblogs from a technology perspective. The study analyzed the impact of technology used on the success of 126 blogs taken from the top 100 listings of the Technorati[1] website. In the case of Technorati, success is measured by the number of inbound links to a weblog.

[Choi 2007] describes a model for evaluating similarities among a larger number of weblogs. A similarity matrix is computed by using Jaccard's coefficient, an adjacency matrix, the inverse shortest path algorithm or the sum of inverse distances. This similarity matrix is used to categorize weblogs and find clusters of topically related weblogs.

As far as the more general research on weblogs and communities therein is concerned, [Zhoe 07] defines weblog communities based on social network theories as groups of vertices that have high density edges within them, with low density edges between the groups. Moreover a specialised web crawler is proposed to prepare data for community detection.

As far as the analysis of weblogs communities is concerned, [Cohen 06] counts connections by hyperlinks and relation by type or topic as possible relations forming communities of weblogs. This point of view focuses on the relations provided in the content of the weblog. A complementing approach, as presented in [Li 07] is to take into account the information available from the guest comments to the entries.

To the best of our knowledge, our work is the largest quantitative analysis of success factors for weblogs to date.

3 The Analysis

The "Meine Kleine" Weblogs[2] of the local Austrian "Kleine Zeitung" newspaper offer a promising possibility to take a closer look at a large number of weblogs in a relatively closed environment. Users of this environment have the ability to publish text and images (photos) to their own weblogs. In addition, comments as guestbook entries can be written to the weblogs of other users.

[1] http://www.technorati.com

[2] http://www.meinekleine.at

By November 21st 2006, the blogspace consisted of 15702 active blogs, ranging from topical weblogs by the newspaper's editors to private diaries of individual readers. In our research, we had access to the servers log files as well as the database holding the weblog entries. The log files and entries from Oct 7th to Nov 21st 2006 (about 6 weeks) were analyzed in the course of this project.

3.1 Basic Statistics

The community of Kleine Zeitung readers consist of a large number of more than 118,000 registered users. Of those, 15702 have activated the weblog for their account and had at least one visit in the analyzed 6-week period. 560 of those bloggers were active publishers in this period, meaning they added content to their weblog in these 6 weeks. This, together with the fact that only the top 1730 weblogs had five or more visits in this period, resulted in our decision to take only the 2000 most visited weblogs into account for all further examinations (see Figure 1).

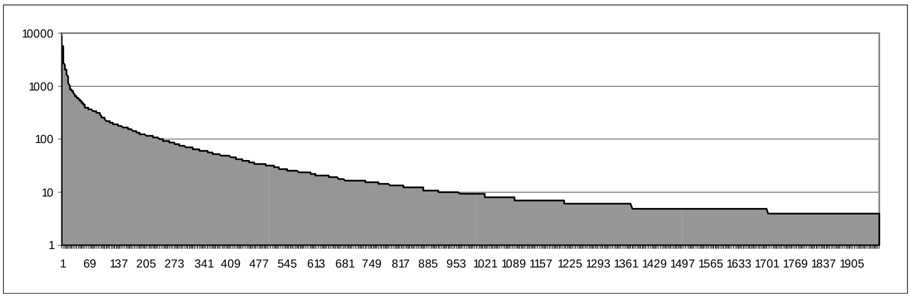

Figure 1: Distribution of visits of the top 2000 weblogs

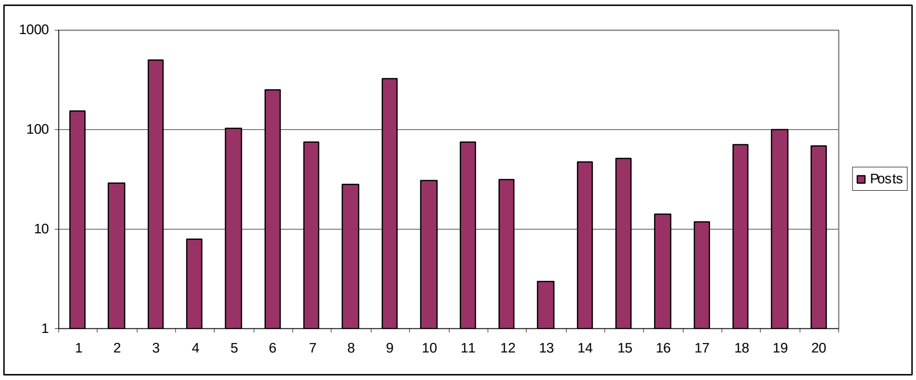

Figure 2: Publishing Activity of the 20 most-visited weblog users

The publishing activity of the examined users was quite incoherent. The number of (text) entries published ranged from 0 to 45 for the individual blogs. The most active poster of images published 469 images in the examined period. As far as the activity in the community is concerned, the most active users posted 80 comments and 137 guestbook entries in other users' weblogs. Figure 2 shows the summary of publishing activity for the users with the 20 most-visited weblogs.

3.2 Activity of Authors vs. Readers by Time of Day

Beside the basic statistics of the weblogs, an investigation of the user activity in the course of a day has been carried out. This focussed on the access of readers to the weblogs as well as on authoring activity by the users.

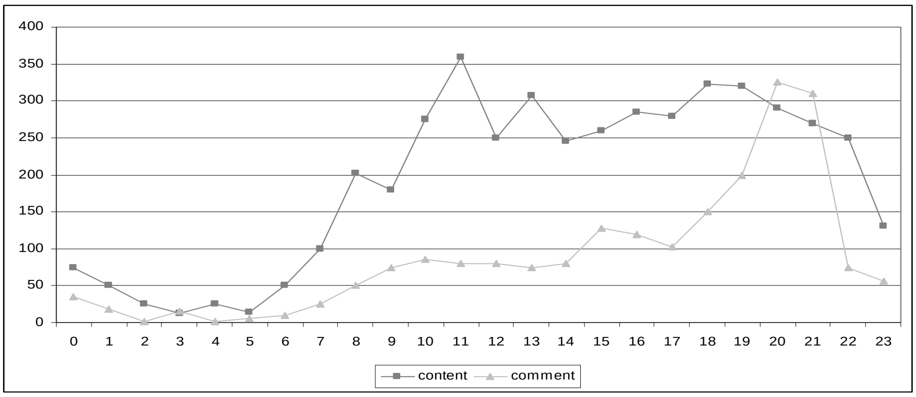

Figure 3: Creation of Content and Comments in the course of a day

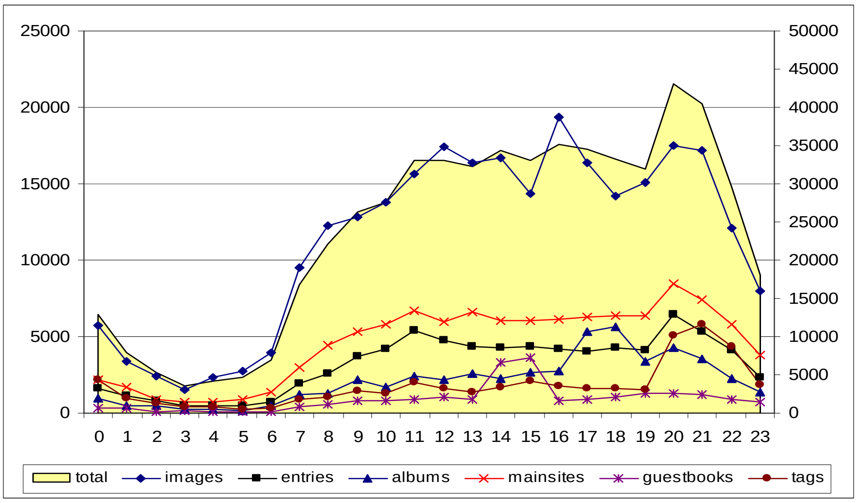

Figure 4: Views of different content types in the course of a day

On the one hand the activity of the content authors, those users creating text entries or uploading photos, shows an almost even distribution throughout the day, with one peak at noon and one in the early evening. This is in contrast to our assumption that most users would be contributing content to their blogs in the evening and in a home environment. This assumption is true, however, for the posting of comments, which mainly occurs between 19:00 and 22:00.

Figure 3 shows the significantly different graphs for publishing own content versus activity in the community (i.e. writing comments and guestbook entries in other users' weblogs).

On the other hand, the activity of the visitors of the blogs in the course of a day was also investigated and viewed separately for the different types of content. As with the authoring activity, reading is almost evenly distributed from 7:00 to 23:00. Peak usage is from 19:00 to 22:00, which corresponds to the peak in commenting. The visits of guestbook entries vary from this general observation, as they have a peak in the early afternoon and none in the evening (see Figure 4).

4 Influences on Popularity

The main focus of the research on the "Kleine Zeitung Digital" blogspace was to find and verify factors for the success of weblogs. We decided to analyze nine possible criteria, arranged into three groups. These criteria were chosen based on the available data from the server logs and after intense discussion with those editors responsible for the blogspace of "Kleine Zeitung Digital" concerning their experience and assumptions. Together with these editors nine hypotheses were formulated. These hypotheses were tested by calculating the Pearson's correlation coefficients for the variables concerned and verifying the significance with a Student's t-test at a confidence level of 0.99

H_1	Authors who write more posts attract more visitors.
H_2	Authors who provide more images attract more visitors.
H_3	Authors who provide new content more often attract more visitors.
H_4	Authors who actively comment other weblogs attract more visitors.
H_5	More frequently visited authors receive more comments.
H_6	Authors who actively post in other guestbooks attract more visitors.
H_7	More frequently visited authors receive more guestbook entries.
H_8	Authors whose postings are similar to the writing style of newspaper editors attract more visitors.
H_9	The posts of frequently visited authors are similar in writing style.

Table 1: Hypotheses

4.1 Influence of Content Types provided

In a first step the different types of content composing the weblog were analyzed on their impact on the popularity. The visits of the individual blogs were compared to the number of textual entries (H_1), images (H_2) and the days the user was active (H_3) in the period investigated. *Textual entries* are all posts in the weblog, regardless of the fact whether images were also included or not. *Images* are access to all image files and photos in the weblog. This results in the fact that an entry with photos is counted as textual entry as well as image. Figure 5 depicts the activity for the 20 top-scoring weblogs in these categories.

The value of *active days* is meant to give an overview of the activity of the author within the six week period of data collection. It encompasses all days the user has actively contributed to his weblog.

The highest correlation of content type to popularity could be found with the user's active days (H_3), being 0.68. The correlation to the entries (H_1) and images (H_2) were lower, at 0.60 and 0.53, respectively.

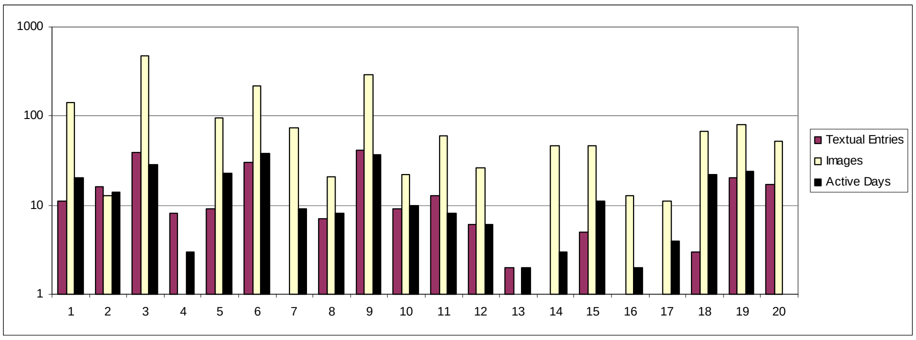

Figure 5: Number of posts and images provided by 20 most visited bloggers

4.2 Influence of Community Activity

Secondly, the influence of community activity was investigated. We decided to analyze the correlation of comment and guestbook activity, outbound as well as inbound. Trackbacks were not considered due to the fact that this technology is not available for "Meine Kleine" weblogs. Figure 6 depicts the activity for the 20 top-scoring weblogs in these categories.

The highest value was found for the obvious correlation of received comments and visits to the weblog with 0.77 (H_5). The number of guestbook entries received correlates with 0.69 (H_7). For own comments (H_4) and guestbook entries in other blogs (H_6) the correlations are 0.70 and 0.68, respectively (see Table 2).

It should be noted that these correlations coefficients are higher than those of publishing content. In other words, in order to have a highly visible weblog, it is even more important to be active in the community than to publish own content regularly! This is true for the individual correlations as well as for the summary of content provided respectively own community activity. There is a total correlation of 0.61 of content provided to the number of visits, while the correlation of community activity to number of visits is 0.71.

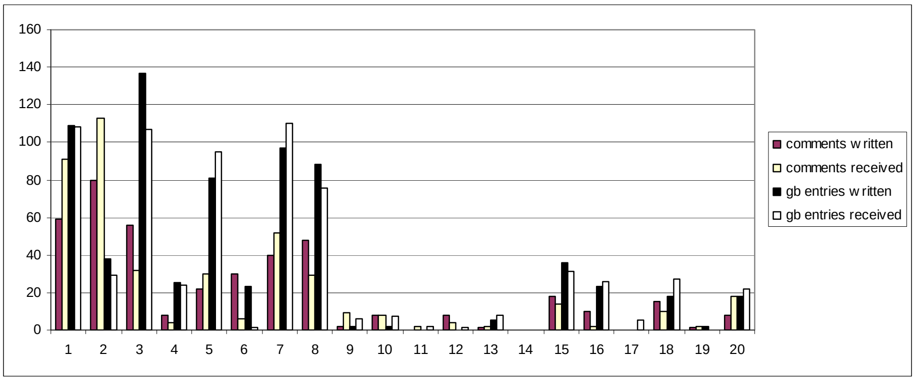

Figure 6: Comments and guestbook entries given and received by 20 top visited bloggers

4.3 Influences of Writing Style

A further aspect of the research investigated upon the contents of the Kleine Zeitung blogspace is the influence of the author's writing style. For this purpose, the similarity computation of the Autonomy Search Engine, which is based on Bayesian Inference [Autonomy 07], was used.

Individual weblog entries of the top 2000 blogs were compared to the editorial blogs of the Kleine Zeitung (written by professional journalists) and to the top 5 blogs. This first approach led to no significant results.

In a next step we compared the similarity of a summary document containing all weblog entries of one user with the top scoring editorial weblog and the top scoring user weblog. With these conditions we found a correlation of 0.38 for the similarity to the editorial weblog (H_8) and a correlation of 0.25 for the similarity to the top scoring user weblog (H_9). Both correlations are lower than those for the content type and user activity. Figure 7 shows the similarity to the top scoring editorial weblog and the top scoring user weblog for the 20 most visited weblogs.

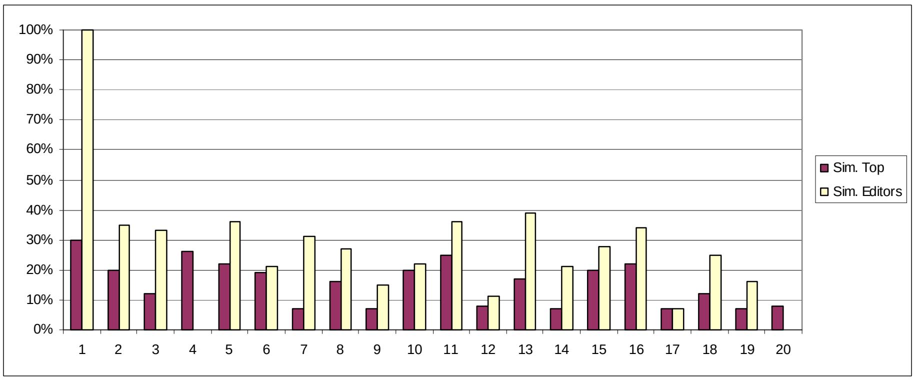

Figure 7: Similarity of 20 Top-Ranked Weblogs to Editorial and Top-Ranked Weblog

4.4 Conclusions

All of the success factors investigated have a positive correlation with the number of visits, in each case statistically significant with a degree of freedom of 1998 (2000 samples) and a confidence interval of 0.99. This means all nine hypotheses can be confirmed. The highest correlations can be found for the activity in the blogging community. Table 2 illustrates the correlations of the individual factors.

number of visits correlated to	*hypotheses*	*correlation*	*t-test*	*confidence interval*
active days	H_1	0.68	41.46	0.99
# of posts written	H_2	0.60	33.52	0.99
# of images uploaded	H_3	0.53	27.94	0.99
# of comments given	H_4	0.70	43.81	0.99
# of commetns received	H_5	0.77	53.94	0.99
# of guestbook entries given	H_6	0.68	41.46	0.99
# of guestbook entries received	H_7	0.69	42.61	0.99
similarity to top rated user	H_8	0.25	11.54	0.99
similarity to editorial blogs	H_9	0.38	18.36	0.99

Table 2: Summary of correlations

5 Communities in "Meine Kleine" Weblogs

As the previous chapter showed, the activity within the community is a crucial factor for the success of a weblog. In a final step of our research, we tried to visualize the communication between the community members by comments and guestbook entries. We used the *prefuse*[3] framework to visualize a GraphML representation [Brandes 02] of users and their connections.

In the resulting graphs, the members of the community are depicted as nodes and links resulting from comments or guestbook entries as edges. A Fruchterman-Reingold force directed placement algorithm [Fruchterman 91] was chosen for the graph for a comprehensive visualisation of the interconnectedness of the community. An initial version of the community graph containing all communications turned out to be too complex.

[3] http://prefuse.org/

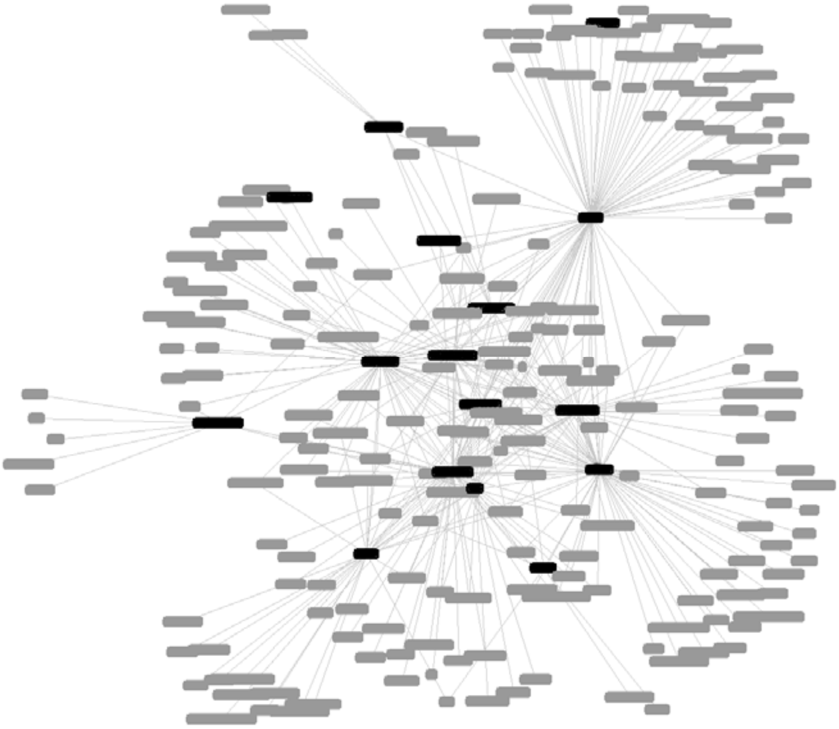

Figure 8: Force Directed Placement of the 20 most-visited Users and their Contacts[4]

In order to produce clearly arranged graphs, the number of community members taken into account was reduced. As the twenty users with the highest numbers of visits were responsible for 91% of the community activity, only these twenty and the corresponding conversational partners were taken into account.

In the resulting Figure 8 those nodes that represent the 20 most-visited weblogs are presented in darker shade than the others. Four of these top-twenty users have did not give any comments or create guestbook entries in the analyzed period and were thus removed from the graph. The remaining 16 top-scorers form a tight network with several communities unique to individual users.

As a next step to clarify the social network of the Kleine Zeitung blogspace, only those edges were taken into account that represented three or more communication activities. The resulting graph shows a tight network of eleven of the top score users, while the remaining 9 have no connections to the graph (Figure 9). Four of these eleven users also build their own sub graphs of community activities.

[4] For privacy reasons the user names have been blackened in the community graphs.

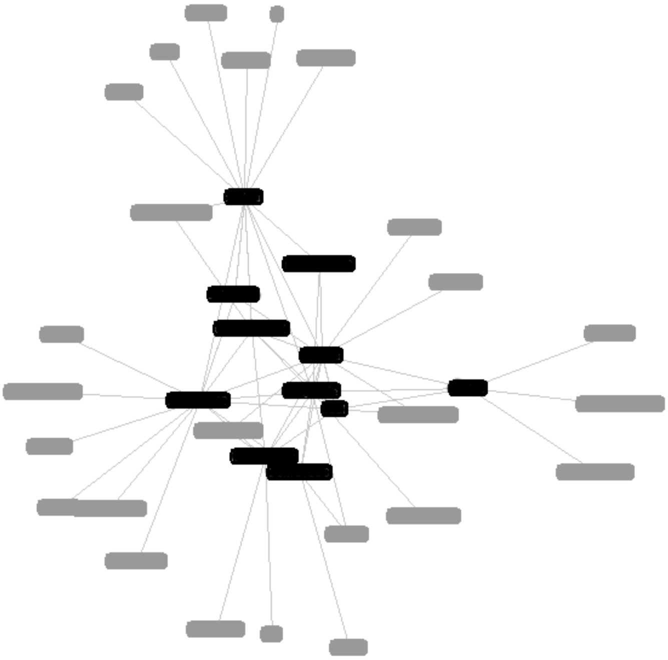

Figure 9: The Community of the most-visited users

6 Conclusions

The analysis on the blogspace of the Kleine Zeitung online community was conducted to find key factors for success of weblogs. The factors activity, number of textual entries, number of images, comments given, comments received, guestbook entries given and guestbook entries received were analyzed in this context. The comparison of the influence of these nine factors showed that the most important of these factors are the community activities of the authors, i.e. writing comments and guestbook entries in other blogs.

Acknowledgements

Financial support of this research by Styria Media AG, in conjunction with the Endowment Professorship for Innovative Media Technology at Graz University of Technology, is gratefully acknowledged.

References

[Autonomy 07] Autonomy Inc., Autonomy Technical Overview, 2007, http://www.autonomy.com/content/Technology/index.en.html

[Brandes 02] Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M. and Marshall, M.S., GraphML Progress Report: Structural Layer Proposal. *Proc. 9th Intl. Symp. Graph Drawing (GD '01), LNCS 2265, Springer-Verlag,* 2002, 501-51

[Choi 07] Choi, H. and Krishnamoorthy, M., Categorization of Blogs through Similarity Analysis, In: *Intelligence and Security Informatics, IEEE*, 2007, 160-165

[Cohen 06] Cohen, E. and Krishnamurthy, B., A short walk in the Blogistan. *Comput. Networks* 50, 5, Apr. 2006, 615-630.

[Du 2006] Du, H. S. and Wagner, C., Weblog success: Exploring the role of technology. *Int. J. Hum.-Comput. Stud.* 64, 9, Sep. 2006, 789-798.

[Fruchterman 91] Fruchterman, T. M. and Reingold, E. M. 1991. Graph drawing by force-directed placement. *Softw. Pract. Exper.* 21, 11, Nov. 1991, 1129-1164.

[Gillmor 04] Gillmor, D., We the Media. Grassroots Journalism by the People, for the People, *O'Reilly Media*, ISBN 0596007337, 2004

[Kritikopoulos 06] Kritikopoulos, A., Sideri, M., and Varlamis, I., BlogRank: ranking weblogs based on connectivity and similarity features. In *Proceedings of the 2nd international Workshop on Advanced Architectures and Algorithms For internet Delivery and Applications* (Pisa, Italy, October 10 - 10, 2006).

[Kumar 04] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A., Structure and evolution of blogspace. *Commun. ACM* 47, 12, Dec. 2004, 35-39.

[Li 07] Li, B., Xu, S., and Zhang, J. 2007. Enhancing clustering weblog documents by utilizing author/reader comments. In *Proceedings of the 45th Annual Southeast Regional Conference* ,Winston-Salem, North Carolina, March 23 - 24, 2007

[O'Reilly 05] O'Reilly, T., What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software, 30.09.2005, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[Page 98] Page, L., Brin, S., Motwani, R. & Winograd, T., The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, USA. 1998

[Paquet 03] Paquet, S., Personal knowledge publishing and its uses in research. Knowledge Board, 10., 2003, http://www.knowledgeboard.com/item/253

[SixApart 04] Six Apart, TrackBack Technical Specification. Developer Documentation, Version 1.2, 2004, http://www.sixapart.com/pronet/docs/trackback_spec

[Zhou 07] Zhou, Y. and Davis, J., Discovering Web Communities in the Blogspace, In Proceedings of the 40th Annual Hawaii international Conference on System Sciences (January 03 - 06, 2007). HICSS. IEEE Computer Society, 2007