2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, May 27-31, 2014

BlogSphere – A Topical Map of the Blogosphere

Philipp Berger[1], Patrick Hennig[1], Stephan Detje[2],
David Eickhoff[2], Daniel Taschik[2], Bjoern Wagner[2], Christoph Meinel[3]
Hasso-Plattner-Institute, University of Potsdam, Germany
[1]{philipp.berger, patrick.hennig}@hpi.uni-potsdam.de
[2]{stephan.detje, david.eickhoff, daniel.taschik, bjoern.wagner}@student.hpi.uni-potsdam.de
[3]office-meinel@hpi.uni-potsdam.de

Abstract

BlogSphere is a visualization of topical diversity and information flows of the blogosphere. By showing the entirety of blogs, it enables the user to navigate and explore blogs using the map metaphor. We describe our mapping of visual parameters like node size, position, background color, etc. that enable the user to easily understand the overall structure of the blogosphere. We then dive into observations made through BlogSphere and the first user reactions to our prototype.

1 Introduction

Social networks, especially weblogs, are drastically growing over the past years in the World Wide Web, making them a valuable source of information. As the amount of information exponentially grows, current research focuses on the meaningful and easy-to-use presentation of this massive amount of interconnected data. Thus, making it feasible to navigate, explore, and monitor the public knowledge and emotions of the complex social networks.

We focus on the blogosphere as an example for complex networks that becomes tedious to navigate, filter, and search. During our research we often encountered the problem that it is not possible to get an overall picture of the whole social network space. Thus, we combined different visualization techniques with multi-dimension reduction to create a overview map for a complex social network that enables user to easily navigate and understand the information space as a whole.

The interface of our tool, BlogSphere, is shown in Figure 1. This map shows the entirety of all blogs from our test data set. Its purpose is to visualize the whole topical space of the blogosphere by uniquely positioning every blog according to its topics. The map enables users to navigate the landscape and easily find and remember specific blogs by linking it to their own mind-map [11, 8] of the information space.

The goal of our prototype is to fulfill nowadays expectations on a visualization using metaphors and common control patterns to give the user a flawless experience while exploring the blogosphere on a desktop computer as well as on tablet device.

In the following section, we present the related work in this area concerning map-like visualization of information

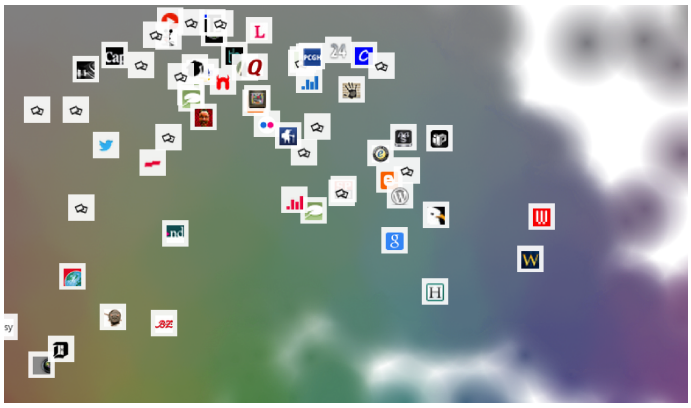

Figure 1: The topical map of the entire blogosphere

flows and topics. Consecutively, we give a short introduction to weblogs, their structure, and linking behavior. Section 4 elaborates on different aspects of visualizing the blogosphere and give insights in our findings for a solid, user-centered and easy-to-use user interface.

The blogosphere and it's linking behavior

The blogosphere is the entirety of all blogs on the web. A blog is a journal like website that consists of reverse-chronological ordered articles called posts. A post is the main information unit of the blogosphere. It has various fields like title, publication date, and content. The content is unstructured, formatted HTML text that can contain links. Through the post's links the actual network structure of the blogosphere evolves. An example of network is shown in Figure 2.

For our visualization we abstract from the actual post links and derive uni-direct blog links that result from a post link in any direction from one blog to another blog. Thereby, we help the user to understand the overall connections of blogs. The detailed view on post links can afterwards be revealed by visualizations with a higher detail resolution like Bross et al. [4].

2 Related Work

BlogSphere offers a combination of a map and a graph visualization. Thus, we can divide our related work in two

ISBN: 978-1-62561-000-3

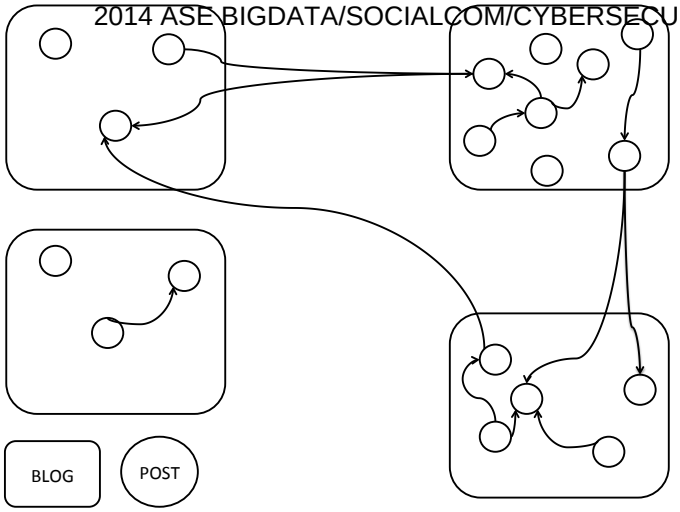

Figure 2: Linkstructure of the blogosphere

areas: graph visualizations of social networks and map visualization of social networks.

The visualization of social network has a long history beginning in the 1930s with the first graph describing social relationships by Jacob Moreno [18]. One typical example is Vizster by Heer et al. [9]. The authors present a tool that shows every social network user and the connections among them. Further it offers diverse filter criteria and a community visualization that groups nodes together which belong to the same cluster. Similarly, Verbert et al. [20] created a graph to better understand recommendations. They also show groups as colored areas to better recognize communities. The recent work from Viegas et al. [21] present Google+ Ripples. Ripples shows the sharing graph of Google+ by visualizing the information flow through combined bubbles. Thereby the authors create a very space-saving method to visualize the complex information flow graph as a whole. In all these approaches, the position of a document or user is non-deterministic calculated. To give meaning to the position of an entity related work introduces the metaphor of an information map.

One examples for a map visualization is described by Nocaj et al. [15]. They visualize search results in a map with deterministic position preserving the mental map of the user. The positioning is based on a hierarchical clustering of search results in combination with a Voronoi treemap. Another example is the work of Bross et al. [3]. As contrast, to other solutions, their positioning is rather random, but they introduce the size and coloring of entities as additional visual variables for their blog visualization.

As contrast to related work, BlogSphere aims to visualize an entire social network eg. the blogosphere. Therefore, we introduce a topical clustering combined with a dimensional reduction for blog. Further, BlogSphere also shows connectivity and importance of the shown blogs.

3 System Overview

To deliver the necessary data for our visualization, we use a scalable blog crawler implementation described by Berger et al. [1]. We deployed it on a 20 node cluster and run it for a period of 3 weeks resulting in over 40.000 blogs with over 3 million posts. This data set is clearly to hard to understand without the help of a sophisticated visualization. Further, it is interesting to reveal specifics of the crawler if it may prefer certain topics of the blogosphere or end up in islands.

The dataset is stored in a central in-memory database called SAP HANA. That serves as input for the positioning, clustering algorithms and as backend for the web app and IPad app.

The information flows in the blogosphere are of special interest to this work. Usually, blog system offer integrated concepts of referencing other weblogs called trackbacks. Users also use explicit links to posts of other blogs to refer to its content. Further, users can simply copy and paste content of other blogs [13]. As already discussed, we abstract from the different link types and show them as unidirectional links between blogs. Future work will address the distinction between those different types.

When we started to design the system architecture for the application, we conclude that our application shall be served via web browser and tablet platforms. On first glance this will be supported by HTML5, but native mobile applications are still superior in terms of performance and usability. E.g. Facebook decided in 2012 to switch from a HTML5 based app to a native app, because of performance and synchronization issues [5]. The high-level architecture is illustrated in figure 3.

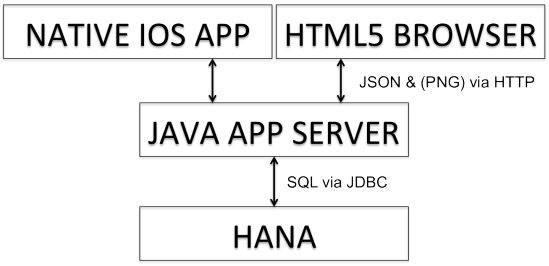

Figure 3: High-level Architecture

4 Concepts and Visualization Techniques

Our visualization as shown in Figure 1 utilizes six concepts to enable the user to efficiently navigate and understand the structures of the blogosphere.

4.1 Topical Map

Each entity in the information space, e.g. blog, has a specific position on the map. We have the choice between different aspects that we can encode in the position of an entity. The most common approach is to use the linkage of an entity. Thus, one has to define links and strength of links for each blog and than apply a force directed layout to position the blog by linkage. This enables the user to rapidly find influential entities, but the position of the satellite entities

is rather random. Therefore, we decide to use the text content of each entity as indicator for their position. Here, we can identify two approaches.

Exact Positioning based on TF-IDF vectors Our first approach is to use nearly the exact content of an entity as indicator for its position. Therefore, we build a vector for each text by segmenting it into terms and calculate for each term the tfidf value. We call the resulting vector term-tfidf-vector (similar to Pazzani et al. [17]). This vector has as many dimensions as terms in the total data set. Thus, we get even on our comparable small blog corpus over 200 million terms. The challenge is to construct from this n-dimensional space a two dimensional space. The PCA analysis is one approach to reduce dimensions. It identifies the so called eigenvectors of an set of n-dimensional vector. We actually just search the most significant eigenvectors with highest eigenvalues. We are using the Lanczos algorithm to identify these vectors. Finally, we project the original vectors to the two dimension space that results from the two eigenvectors. This results in an ellipse-like structure as shown in Figure 4. The ellipse is caused by a small variance in the corpus among the term usage. We see that the rectangular space of the map is not optimally used.

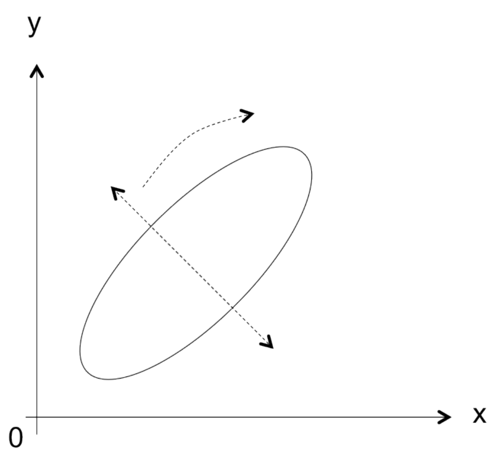

Figure 4: Result of the PCA projection

Moreover different entities may be positioned so narrowly that they overlap each other. Thus, we apply a force-directed layout by Fruchtermann et al. [7] to the result of the PCA projection. This layout preserves the relative distance of the n-dimensional vectors but minimizes the overlapping and maximizes the usage of the rectangular map. For this purpose each point is assigned an attractive as well as a repulsive force as shown in Figure 5. The attractive force is scaled linear to the distance of the tfidf vectors of blogs. Whereas the repulsive force increases with the entities importance. The impact of an entities forces on another entity declines with increasing distance between them. Having determined these forces we can calculate an overall force for a specific point by combining all of those which have an influence on it. Subsequently we adjust the points position by adding the overall force vector. This process is executed repeatedly until the error between the actual visual distance and the calculated similarity converges. Thereby,

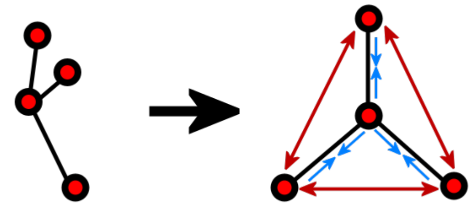

Figure 5: left: random graph format, right: after applying force directed layout, red arrows represent repulsive force

the overlapping gets minimized because the points slightly shake until each point has enough space. Further, a negative gravity is applied to the edges of the map. In case a point moves towards an edge it gets repelled from the gravity field of the edge. Finally, we create a link between two nodes only if the similarity exceeds a specific threshold e.g. *0,5*. We linearly map the similarity to link strength and run the layout algorithm that also incorporates the links. We choose this threshold to reduce the computational complexity by looking only at the best 20% of links.

Positioning based on topic vectors Although the positioning using the exact tfidf vectors is rather intuitive, it has its disadvantages. First of all, the calculation complexity is very high due to the extreme length of vector which is caused by the high number of words in the corpus. Further, the force-directed layout blurs the exact position to ease readability.

A first approach is to remove words with a low tfidf value and thus a low significance in the document set in scope. This removes usual words like normal stop-words (e.g. have, like, etc.), but creates only a minor reduction of words. To solve this issue, we compute a clustering that groups words together according to their similarity. We use k-means [22] to categorize the words in buckets. K is empirically set to 100000 producing clusters of nearly the same size. This very high h is meant to reduce the dimensionality of the data as a preprocessing without loosing too much information about the single word usage.

Each of this clusters now contains topically highly similar words because only words that co-occur very often will be in the same cluster. By creating topical vectors that assign each cluster the sum of all word tfidf values per document, we run the positioning procedure again. This results in an equal position with the advantage that small groups now gather in big visual clusters.

4.2 Node size

Instead of only positioning equally sized nodes on the map, we like to use the size to indicate the importance of a node in the information space like cities and villages in a traditional map. Therefore, we have to define importance in the context of a topical map for the blogosphere. Importance can be defined in different ways. First, we can apply the traditional PageRank [16] that indicates the importance of a webpage based on the linkage behavior to and from other pages. As already discussed in [6], PageRank is not suffi-

cient for 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, May 27-31, 2014 of a blog that define the importance. The BI-Impact is a specialized ranking for blogs, which incorporates additional factors like publication frequency, number of comments per post, number of posts, etc.

We use the BI-Impact as preset ranking an thus resizing the blog node accordingly. Furthermore, we admit that this is not the only valid ranking and we also offer the PageRank as well as an custom expert rank. The expert rank is hereby based on the topical consistency of authors and can easily reveal experts on the topical continents of the map. The adaptability of ranking algorithm is one of the design goals. This enables us to later on add other algorithms as well as nearly fully user-defined rankings.

The standard node size is quiet small, because we only use the node as unit for the background coloring. Thus, the size mainly influences the visibility of the blog "dot" on the map that has to be adjusted with a growing data set. We decide to use the favicons of blogs as interaction elements to enable the user to click on blogs and recognize them faster instead of using the blurred blog "dots" of the background. Thereby, the user is able to quickly find known blogs. With a higher zoom-level more and more blogs get visible until each blog is represented by its favicon. During the zooming, the size of favicons stays constant whereas the blog "dot" in the background gets zoomed. Please note, that small dot are rendered in the foreground to ensure that the user can estimate its importance e.g. its ranking.

4.3 Background Coloring

The coloring of the map is, beside the positioning, one of the main features of our visualization. Hereby, each entity has a color corresponding to its topic. There are multiple ways to define the topic of an entity.

First, one can simply translate the tfidf-vector or the topical vector to the HSV color model. By scaling the h-value we achieve a continuous coloring according to the color spectrum. The scaling is done by applying a PCA to reduce to the one dimensional space.

Another method is to apply an additional document clustering. For instance, a k-means clustering will result in k color areas on the map that look to the user like continents on a political map. One can hereby vary k. A small k will produce huge continents with small islands. In contrast, a large k will result in an equally continuous coloring as produced by the application of topical vector color mapping. To easily switch between different degrees of cluster granularity, we also implement a hierarchical single-linkage clustering that results in a binary tree. The granularity varies between different depths of the tree.

For our first prototype we stick to the continuous coloring resulting of the topical vectors. This gives us for each entity a specific color and prevents the creation of large indistinguishable blobs with the same color.

As shown in Figure 9, the background is dominated by high-ranked blogs. Through the combination of low and high ranked blogs we can observe a nearly continuous color flow with slightly bigger bubbles in the center of a topic region. Smaller blogs are placed in the foreground to ensure clustering of blogs into topical areas (see Section 5.1).

4.4 Linkage

To better understand the influence of a blog, we also present the interlinkage of a blog to the user. We collect the links during the crawling process and aggregate links of each document of a blog (post) to one undirected link between two blogs (as shown in Figure 2).

Drawing links between connected blogs has been implemented with Bezier curves. Curves start at the center of a blog item and have the first control point 50 pixel outside the item with an angle of 0, 90, 180 or 270 degrees depending on the position of the other blog. The endpoint and the second control point is the center of the connected blog item as seen in Figure 6.

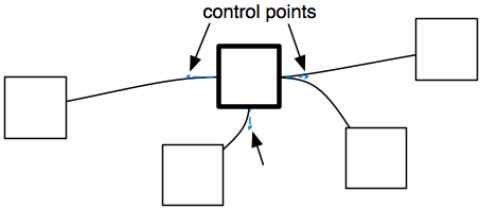

Figure 6: Drawing links with Bezier curves

The resulting view in the prototype is shown in Figure 7. In this example the blog *bildblog.de* is in focus. This entity has close interlinkage with over five other blogs. It is remarkable that although the newspaper blog *blid.de* has a lower score, it can get in the attention of the user by the visible link of the high-ranked discussion blog.

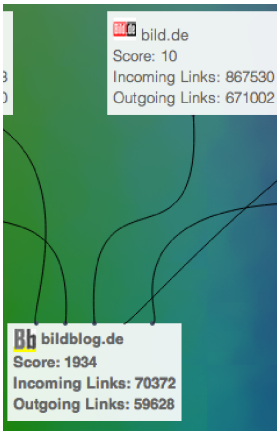

Figure 7: Links help to understand the influence of a blog and its nearest relatives

4.5 Rendering

Todays web technologies like SVG and HTML cannot handle the immense amount of entities which exists in a state-of-the-art document corpus. We circumvent this problem by tiling the map into pieces and calculating the actual background image, which has to incorporate every blog, on the server side. In our current test data set, we crawled

over 40.000 blogs. By the standard zoom-level only the top 100 ranked blogs are visible, but the whole data set can be roughly estimated via the bubbles in the background.

With higher zoom-level the number of visible blogs increases. A blog item is usually presented by the websites favicon and the hostname of the blog 8b. Since the number of drawn blog item can easily reach a count in which they would overlay each other, only the favicons will be drawn once a certain threshold is exceeded (see 8a). For blogs which do not offer a favicon, a default icon is used instead.

Detailed information about a blog, can be retrieved with clicking a blog item. A selected blog increases its dimensions as shown in 8c and presents information about blogs score and the number of incoming and outgoing links. All other blogs are hidden and only connected blog items are displayed.

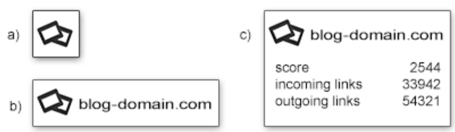

Figure 8: Blog items in different zoom levels

4.6 Interaction Hooks

As necessary to support the user's understanding process, our visualization is interactive. It supports the following interaction hooks: searching, filtering, zooming, and panning. As shown in Figure 9, our web app consists of a sidebar that tilts up on hovering and gives the user access to the interaction hooks and additional statistics of the blog like number of incoming links etc..

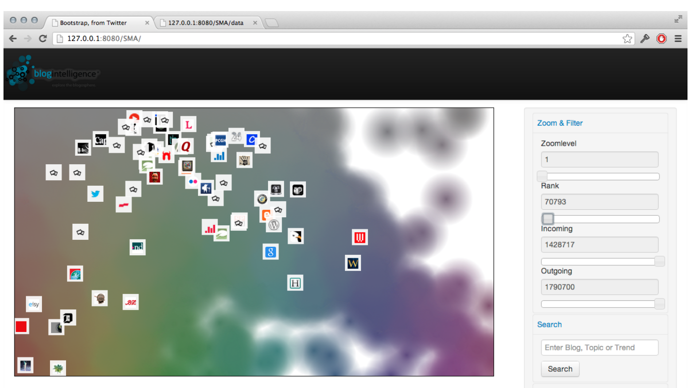

Figure 9: The web app offers zooming, panning, search, and additional filters for blog KPIs

4.6.1 Zoom

As discussed in Section 4.5, the visualization consists of a semantic zooming feature that reveals lower-ranked blogs with increasing zooming level. Thus, the user can step-wise dive deeper into the whole data set. Nevertheless, the user is able to estimate the total number of blogs by the background coloring where each blog actually has at least one dot. We implemented the standard zoom&pane behavior [10].

4.6.2 Filtering

To further decrease the overlapping and the visual clutter, we enable the user to filter the displayed blog objects by blog KPI's. Our current prototype is limited to the blog rank, number of incoming links, and number of outgoing links. The set of KPI's is in no means fix and shall be extended in future version of the prototype. Hereby, the inclusion of other ranking criteria like expert rankings [2, 19]. On limiting a KPI to a specific range all blog objects that do not fulfill the requirement fade out. The background coloring is not changed to keep the awareness of the dataset size and the size of the different visual clusters.

4.6.3 Search

We implement a search bar that enables the user to search for specific keywords. These key words are searched in the post contents of each blog and only matching blogs stay visible. The feature helps the user to identify areas of the blogosphere that talk about interesting terms. Thus, improve the understanding of discussions among different topics.

5 Revealed Features

BlogSphere enables us to draw different conclusion to structural features of the blogosphere and to the used ranking mechanisms.

5.1 Visual Clustering

Although we try to equally distribute the blogs offer the rectangular space on the map, we can still observe clusters. This is caused by the applied force-directed layout where similar entities tend to drift together. These clusters correspond to topical clusters that can also be derived by common clustering algorithms like described in [23]. The visual representation on a map has two advantages. On the one hand, the clusters are overlapping because a blog can lay between two clusters. One has to mention that if a blog belongs to very distinct clusters its position will be in the middle of both, which can be the middle of another cluster (handling a mixed topic). This case is rather rare because if two clusters have more blogs in common they will have a lower distance. On the other hand, the visual clustering can indicate an implicit hierarchical structure. One can observe clusters that form big fragmented islands where each subarea represents a sub-cluster of the overall island.

5.2 Ranking Implications

By observing the ranking of blogs that is represented by the node size (see Section 4.2), we are actually able to understand the underlying assumption of the ranking mechanisms. High-ranked blogs have many links, but are not limited to one topical cluster. Instead, we are able to observe

 ISBN: 978-1-62561-000-3

that each high-ranked blog is always linked very frequently from relatively unrelated areas of the blogosphere. This is also a indicator for the common assumption that the top dogs of the blogosphere tend to link each other to save their high position.

5.3 Common network structures

Further, we can observe the usual social network structures like hub-and-spoke[12], where the high-ranked blogs have lots of relatively lower-ranked satellite blogs around them. Equally, one can observe information cascades through the blogosphere by following the links. The observed information cascades in our data set tend to have equal variations as described by Leskovec et al.[14].

6 Reactions to BlogSphere

Our application enables the user to explore the BlogoSphere farther. Thus, we collected first user feedback by questioning IT students and social network researches. We present the four most promising use cases.

6.1 Feeling for the data set

There are numerous visualization methods for blogs as well as for other social networks. Most of them specify on presenting particular aspects of the data. Hence, they often lack giving the user a feeling for the size of the presented data set. Anyway we think this is an important issue. In BlogSphere every blog is either presented by an icon or, if it is not ranked high enough for the current zoom-factor, by simply a colored point in the background. So each colored point on the map represents a blog. This technique enables the user to experience the massive amount of data without distracting him/her from the higher ranked blogs.

6.2 Overview and important topics

When using our application the user first gets an overview of the entire data set. Only high ranked blogs are displayed with name and description at the beginning. Thus, it is an easy task to find the currently most important blogs in the BlogoSphere. Furthermore the most important topics can be detected by simply looking at the high ranked blogs, since the blogs are positioned according to their topic. If there is a narrowly positioned group of high-ranked blogs on the map, then obviously the topic those blogs are about is an important one in the Blogosphere too.

6.3 Serendipity

Whilst exploring the application the user may search for a particular term to narrow the shown data sets down. This way he can find blogs of interest to him. By highlighting the position of the search blogs, the user actually gets aware of the surrounding blogs as well and can explore them in more details. Those similar blogs discuss the same or related topics. Hence, it is very likely that they are relevant for the user. So it is very likely for users of our application to stumble upon interesting blogs that were unknown to them before by chance.

6.4 Borderlines and topic borders

Despite BlogSphere arranging the blogs according to their topic there are no hard borderlines between them. The transitions between different topics are fluent just as the transitions between the different colors in the visualization. Due to this technique one can detect related topics by looking at the blogs. If there are a lot of blogs in the space between two topics those are obviously related. Furthermore one can reason that these blogs contain both of the topics.

7 Conclusion

BlogSphere is a visualization of the topical diversity and information flows of the blogosphere. As contrast to related solutions, BlogSphere emphasize the size of the data set and enables the user to see the blogosphere as a whole. The background coloring of BlogSphere is used to communicate the existence and topic of each blog in the data set. To remove visual clutter we implement semantic zooming, searching, and filtering. Thus, the user is able to dive deeper into various areas of the blogosphere and navigate from the important blogs to the lower ranked blogs. We show the viability of our approach by implementing a first prototype.

Further, we observed that the map metaphor also helps researchers to understand the mechanisms of rankings and information flows. We observed that the typical pattern of information flows could also be found via our visualization and may lead to a better understanding of the overall information flow and to new research questions. By collecting reactions to our prototype, we identified different use cases. The identification of important topics was very frequently mentioned and helps the user to understand the discussions in the underlying data set. We also observed that users like to find known blogs and explore their surroundings in the map leading to the exploration of yet unknown but interesting new blogs to the user.

The navigation of BlogSphere is currently limited to simple panning and zooming. Nevertheless, further research question include the integration of advanced map features into the information space. For example, the next question is whether a routing feature like GPS navigation can be applied to blogs, shown in our visualization, as mini map in the left corner and enable the user to travel from one blog to another via the most "scenic" routing.

Finally, we hope to integrate our work into search engines like Technorati [1] or BlogIntelligence [2]. To serve a entry point to the blogosphere and foster the understanding of the information space as a whole.

[1] http://technorati.com/

[2] http://blog-intelligence.com/

References

[1] P. Berger, P. Hennig, J. Bross, and C. Meinel. Mapping the blogosphere–towards a universal and scalable blog-crawler. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 672–677. IEEE, 2011.

[2] P. Berger, P. Hennig, and C. Meinel. Identifying domain experts in the blogosphere–ranking blogs based on topic consistency. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 252–259. IEEE, 2013.

[3] J. Broß, P. Schilf, M. Jenders, and C. Meinel. Visualizing the blogosphere with blogconnect. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 651–656. IEEE, 2011.

[4] J. Bross, P. Schilf, and C. Meinel. Visualizing blog archives to explore content-and context-related interdependencies. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 647–652. IEEE, 2010.

[5] J. Dann. Under the hood: Rebuilding facebook for ios. `https://www.facebook.com/notes/facebook-engineering/under-the-hood-rebuilding-facebook-for-ios/`, Aug. 2012. [Online; accessed 1-May-2014].

[6] M. Elbashir and S. Williams. Bi impact: The assimilation of business intelligence into core business processes. *Business Intelligence Journal*, 12(4):45, 2007.

[7] T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[8] W. G. Griswold, J. J. Yuan, and Y. Kato. Exploiting the map metaphor in a tool for software evolution. In *Software Engineering, 2001. ICSE 2001. Proceedings of the 23rd International Conference on*, pages 265–274. IEEE, 2001.

[9] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39. IEEE, 2005.

[10] K. Hornbæk, B. B. Bederson, and C. Plaisant. Navigation patterns and usability of zoomable user interfaces with and without an overview. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(4):362–389, 2002.

[11] B. Kuipers. The" map in the head" metaphor. *Environment and Behavior*, 14(2):202–220, 1982.

[12] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, 2004.

[13] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM, 2007.

[15] A. Nocaj and U. Brandes. Organizing search results with a reference map. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2546–2555, 2012.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[17] M. J. Pazzani, J. Muramatsu, D. Billsus, et al. Syskill & webert: Identifying interesting web sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.

[18] J. Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[19] K. Sriphaew, H. Takamura, and M. Okumura. Cool blog identification using topic-based models. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 402–406. IEEE, 2008.

[20] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 351–362. ACM, 2013.

[21] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren. Google+ ripples: a native visualization of information flow. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2013.

[22] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584, 2001.

[23] R. Xu, D. Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.