Topigraphy Project

Tatsushi Matsubayashi†, *Takahide Hoshide, and Ko Fujimura*

Abstract

This article introduces topigraphy, which is a novel method for displaying a large-scale tag cloud as a contour map of related tags. It uses a topographic image as a background picture on which the tag cloud is displayed. The Topigraphy project is supported two weblog (blog) navigation systems: BLOGRANGER TG and BLOGRANGER QA. We have also developed a three-dimensional visualization system and applications for smartphones using the Android operating system (Android OS).

1. Introduction

The *tag cloud* interface has recently become popular as another search interface, and many Web sites such as Flickr [1], delicious [2], and Technorati [3] use it. Tags are meaningful descriptors of objects and are usually provided manually by the large number of users. A tag cloud is a list of the most popular tags, usually displayed in alphabetical order, and visually weighted by font size. By just clicking on the tag of interest, a user can find relevant target objects. The tag cloud interface is useful when the list is sufficiently small; however, if it is too large, it becomes difficult to identify individual tags.

Instead of the conventional tag cloud layout scheme, we have proposed a method for displaying a tag cloud called topigraphy (derived from topic + topography) [3]. Topigraphy uses a topographic image as the background on which a large-scale tag cloud (in excess of 5000 tags) is displayed as a contour map. The two-dimensional (2D) tag layout addresses tag similarities as semantically similar tags placed close to each other. The tag *height* or *altitude* represents the abstractness of the concept represented by the tag, so it provides a visual cue to the user about which tags are organized into semantic hierarchies. A user can enjoy spending time freely browsing and clicking through the 2D tag landscape and successfully discover objects of interest.

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan

2. From *searching* to *exploring*

Almost all search engines display search results as an ordered list ranked by their relevance to the query. However, to get an optimal result with such systems, we must initially provide suitable key words. Therefore, we sometimes cannot obtain the best information.

For example, to search for an interesting film, we usually enter the query *film* in the search engine. Then, the high ranks of the search results are filled with the definition of a film or information about major films. In some cases, we already know about these films and really want more information about films that are similar to our favorites.

The topigraphic map (topigraph) of the film network generated from Wikipedia data enlarged around the *Superman* tag is shown in **Fig. 1**. If you like Superman or Batman and are searching for a similar film, you can spot some films related to them by exploring the map. In typical search tasks, high speed and accuracy of the search algorithm are of the greatest importance. In the Topigraphy project, however, we attach greater importance to *exploring*, i.e., discovering unanticipated information, than *searching*.

3. Topigraph construction

A flowchart for constructing a topigraph is shown in **Fig. 2** and the steps are described in detail below.

Fig. 1. Local image of topigraphic map around the *Superman* tag. The inset shows the whole landscape.

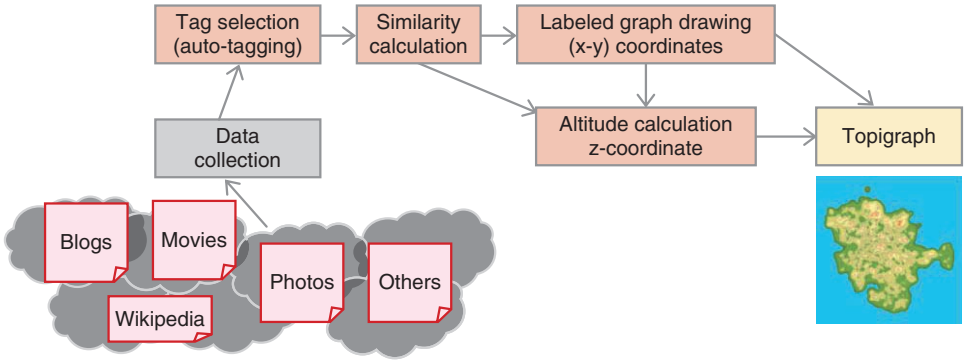


Fig. 2. Topigraph construction flowchart.

3.1 Data collection

Topigraphy can use any relational data such as weblog (blog) entries, movie contribution sites like YouTube [5], photo sharing sites like Flickr [1], and Wikipedia [6] data. Even if the content does not contain any tag data, we can still generate tags from entry texts, comments, or titles by using text mining and auto-tagging techniques.

3.2 Tag selection

Next, we extract significant tags from the data. While some users directly register a lot of meaningful tags, there are too many meaningless tags that are generated mechanically like registration time or weather information. Furthermore, even among the user-registered tags, there are many uninformative or irrelevant tags such as *read later* or *interest*. In BLOGRANGER TG (see section 4.1), to remove insignificant tags, we use an auto-tagging technique

Fig. 3. Screenshot of BLOGRANGER TG.

called the residual document frequency method [7] and use the titles of Wikipedia entries as a white list.

3.3 Similarity calculation

To measure the similarity of the extracted tags, we use several co-occurrence measures, such as the cosine and Jaccard coefficient. In this process, we also use Fisher's exact test, which is a statistical hypothesis testing method, to remove noisy edges and select only significant edges from the large similarity network generated.

3.4 Labeled graph drawing

In topigraphy, tags with high similarity are located near each other. On the basis of the similarity, we can proceed to the (x, y) position calculation using the labeled graph drawing method that we recently proposed [8]. Most conventional algorithms compute node (tag) positions by treating nodes as *points*, i.e., tag size is not considered; consequently, there can be overlapping nodes. To remove this overlapping both efficiently and effectively, we propose a fast labeled graph drawing method called the individual ellipsoidal potential method, which takes the size and shape of each node label into account and avoids label overlap when generating a topigraph from tag data. Its parallel implementation on a GPGPU (general-purpose computing on graphic processing unit) lets us efficiently generate a large-scale high-quality tag cloud representation of 5000 tags taken from a tag co-occurrence network within 30 minutes.

3.5 Altitude calculation

The final step is to calculate the z-position (height) of a tag. Topigraphy introduces the tag height as a topographic expression. The tag height represents the abstraction level of each tag. For example, *sports* is more abstract than *baseball* or *football* and should be given a higher score. This lets the user grasp the relationship among tags intuitively and find related topics easily by tracking topigraphy ridges. To display the features, we have proposed a centrality score method [9], [10] based on the document frequency, user frequency, similarity, and Euclidean distance of the (x, y) position. The coordinates generated in the above steps are used to generate a smooth topigraphic surface through the use of the GMT application [11].

4. Applications of topigraphy

In this section, we introduce some of the main applications of topigraphy.

4.1 BLOGRANGER TG

We have developed a blog navigation system called BLOGRANGER TG (hereinafter TG) to evaluate the feasibility and usability of topigraphy. We opened TG as a test run for one year from December 2007 at goo labs [12], which is an online laboratory. A screenshot of TG is shown in **Fig. 3**. TG automatically extracted about 5000 major and informative tags by analyzing 22 million Japanese blog entries collected during a period of four weeks. The topigraph of TG was updated weekly. It enabled users to grasp what was

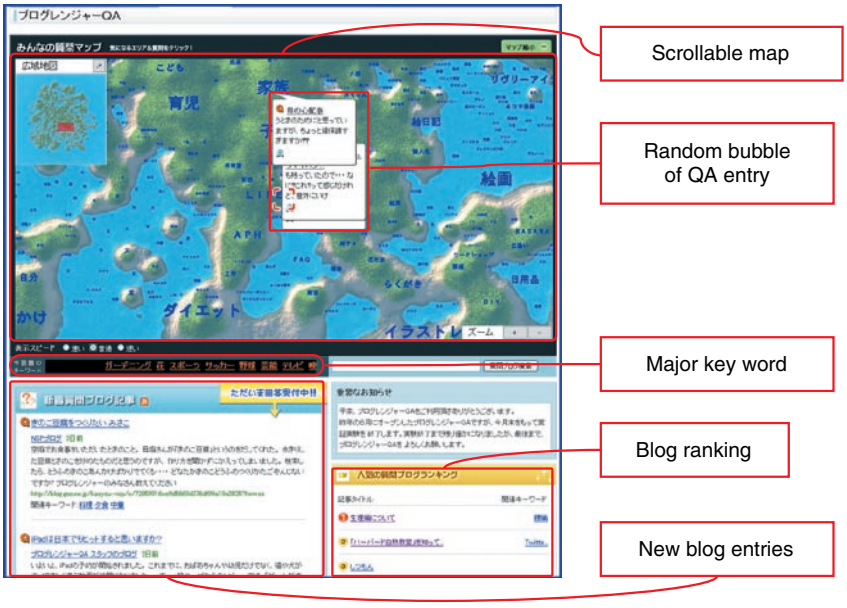


Fig. 4. Screenshot of BLOGRANGER QA.

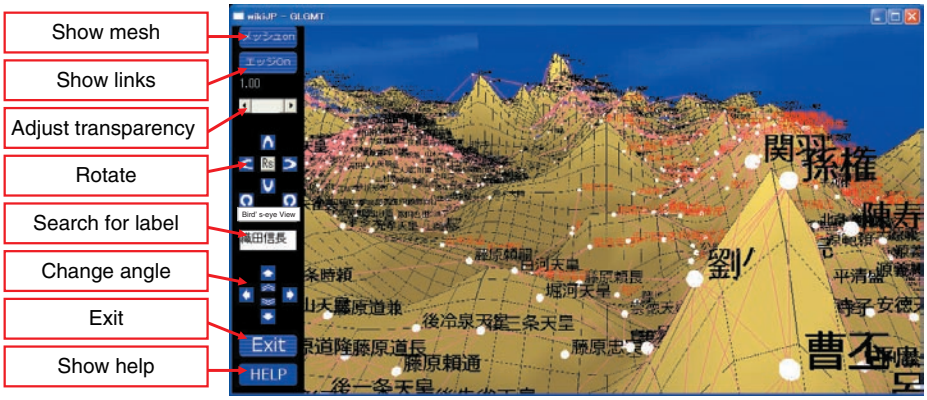


Fig. 5. Screenshot of OpenGL-version topigraphy application.

happening in the blogosphere at that time. Furthermore, TG's application programming interface and blog parts were opened to blog users. These customizable options won high praise from users*.

4.2 BLOGRANGER QA

BLOGRANGER QA (hereinafter QA) is a question-and-answer navigation system based on blog entries. It also had a one-year test run from July 2009 at goo labs. A screenshot of QA is shown in **Fig. 4**. QA tags were generated by the same method as in TG, but the tag score simply corresponded to the number of blog entries concerned with the particular tag. QA lets the reader directly answer a question entry by posting a blog entry through this system.

4.3 OpenGL & 3D display

By using OpenGL, we developed a three-dimensional (3D) application of topigraphy (**Fig. 5**). The OpenGL version topigraph has additional functions.

* TG has won several prizes, including the grand prix of the best of show award in the application field at Interop 2008 and the best poster award of WWW2008.

Fig. 6. Demonstration of OpenGL version of topigraphy.

Fig. 7. Cell phone version of topigraphy running on Android OS.

Fig. 8. Screenshots of Android OS version of topigraphy.

Show links can represent more specific relationships between tags. *Rotate* and *Change angle* let us view the data from various angles in 3D space. This application can also work in stereo. We can see a real 3D topigraph using a 3D monitor and a pair of 3D glasses. A demonstration at Open House 2010 of NTT Communication Science Laboratories is shown in **Fig. 6**. The display system, which ran on an NVIDIA 3D Vision device and used a 3D projector, let ten people experience 3D reviewing at the same time.

4.4 Topigraphy for Android

A cell phone version of topigraphy for the Android operating system (Android OS) is shown in **Fig. 7**. Its functions include voice retrieval, zoom in/out, category retrieval, and a mini map guide, as shown in **Fig. 8**. This topigraph was created from a Japanese person name list obtained from Wikipedia data. If the tag is clicked, the summarized Wikipedia page will appear.

5. Future work

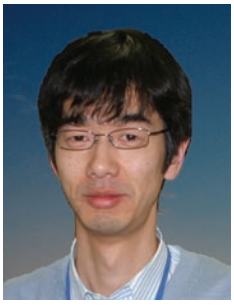
We are planning to launch the Android OS version and open the OpenGL version as free software. We will develop topigraphy-based navigation for a user-customizable interface for devices such as smartphones.

References

[1] http://www.flickr.com/
[2] http://delicious.com/
[3] http://technorati.com/
[4] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda, "Topigraphy: Visualization for Large-scale Tag Clouds," Proc. of the World Wide Web 2008, pp. 1087–1088, Beijing, China.
[5] http://www.youtube.com/
[6] http://ja.wikipedia.org/ (in Japanese).
[7] S. Fujimura, K. Fujimura, and H. Okuda, "Blogosonomy: Autotagging Any Text Using Bloggers' Knowledge," Proc. of the Web Intelligence 2007, pp. 205–212, Silicon Valley, CA, USA.
[8] T. Matsubayashi, T. Yamada, S. Fujimura, and K. Fujimura, "Labeled Graph Drawing Based on the Individual Ellipsoidal Potentials," IPSJ, TOM21, Vol. 1, No. 1, pp. 88–101, 2008 (in Japanese).
[9] K. Fujimura, "Blog Mining and Visualization Technologies Equipped with BLOGRANGER TG," IEICE, Technical Report, AI2008-10, pp. 57–62, 2008 (in Japanese).
[10] T. Matsubayashi and T. Yamada, "Topigraphy: Tag Mountains," IPSJ, NetecoSymp 2009, pp. 117–124, Okinawa, Japan (in Japanese).
[11] http://gmt.soest.hawaii.edu/
[12] http://labs.goo.ne.jp/ (in Japanese).

Tatsushi Matsubayashi
Research Scientist, Innovative Communication Laboratory, NTT Communication Science Laboratories.
He received the B.E. degree in theoretical astrophysics from Kyoto University in 2000 and the M.E. and Ph.D. degrees in theoretical astrophysics from Tokyo Institute of Technology in 2002 and 2006, respectively. He joined NTT Communication Science Laboratories in 2005. Since then, he has been engaged in R&D of graph drawing. He is a member of the Information Processing Society of Japan (IPSJ).

Takahide Hoshide
Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in information engineering from Kyushu University, Fukuoka, in 1991 and 1993, respectively. He joined NTT Information and Communication Systems Laboratories in 1993. He is currently engaged in R&D of web mining. He is a member of IPSJ.



Ko Fujimura
Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in electrical engineering and the Dr.Eng. degree in information engineering from Hokkaido University in 1984, 1986, and 1989, respectively. He joined NTT Information Processing Laboratories in 1989 and engaged in R&D of transaction processing systems and electronic commerce systems. Since 2003, he has been engaged in research on web and social media mining. He is also a visiting professor at the University of Electro-communications, Tokyo. He is a member of IPSJ and the Institute of Electronics, Information and Communication Engineers of Japan.