Topigraphy: Visualization for Large-scale Tag Clouds

Ko Fujimura† Shigeru Fujimura† Tatsushi Matsubayashi‡
Takeshi Yamada‡ Hidenori Okuda†

†NTT Cyber Solutions Laboratories
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa
239-0847 Japan

‡NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto
619-0237 Japan

ABSTRACT
This paper proposes a new method for displaying large-scale tag clouds. We use a topographical image that helps users to grasp the relationship among tags intuitively as a background to the tag clouds. We apply this interface to a blog navigation system and show that the proposed method enables users to find the desired tags easily even if the tag clouds are very large, 5,000 and above tags. Our approach is also effective for understanding the overall structure of a large amount of tagged documents.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering, selection process.*

General Terms
Algorithms, Design, Experimentation.

Keywords
Tag clouds, search, blog, clustering, topography, topigraphy.

1. INTRODUCTION
Tag clouds, i.e., a navigation and retrieval interface for tagged contents, have become very popular and many Web sites such as Flickr, del.icio.us, and Technorati now use this interface. It is, however, hard to identify the desired tag if the number of tags is very large, e.g., 5000, especially from the typical alphabetical order layout, in which relations between the tags are not presented.

To resolve this issue, we propose a new method for displaying tag clouds called *topigraphy* (topic + topography). Topigraphy uses a topographic image as the background against which the tag clouds are displayed; tag "height" represents the centrality of the concept of the related tags while the 2-dimensional layout addresses tag similarity.

We developed the blog navigation system called *BLOGRANGER TG* [5], hereafter "TG", to evaluate the feasibility and usability of topigraphy. A screenshot of TG is shown in Figure 1. TG automatically extracts about 5,000 major and informative tags by analyzing 10,000,000 Japanese blog entries collected within the last five weeks. TG regenerates the topigraphy weekly. It enables the user to grasp what is happening now in the blogosphere.

Although the topigraphy screen generated by TG large, i.e., 10,000 x 10,000 pixels, the proposed display method enables the user to find interesting tags easily. Kaser and Lemire [2] proposed a space-efficient algorithm for drawing tag clouds. It reduces the wasteful white space between tags which is important for small-display devices, such as PDAs and cell phones. Another approach

WWW2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

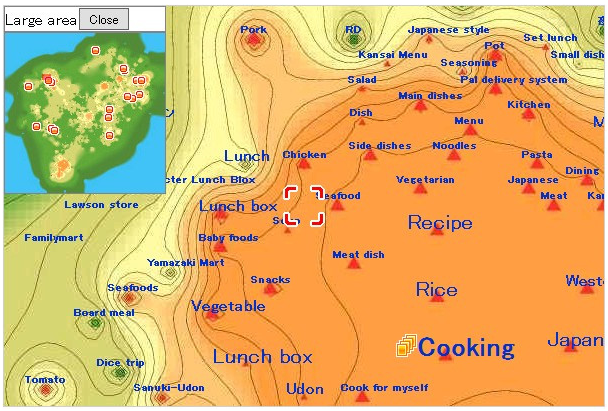

Figure 1. Screen capture of topigraphy around "Cooking" [1]

for such devices is to provide topigraphy via a scrollable map.

Section 2 elucidates the algorithm used to generate topigraphy from a given set of tagged documents. Section 3 describes our implementation experience of TG. A summary is presented in Section 4.

2. ALGORITHM
Topigraphy is automatically generated from undirected graph $G = (V, E)$ where each vertex v in V is a tag and $D(v)$ is a set of documents tagged as v and each edge e_{ij} in E is the similarity between $D(v_i)$ and $D(v_j)$.

To measure the similarity, we use the cosine similarity of the *tag feature vectors* which contain terms and their weight generated from $D(v)$. In this process, term weighting is the most important factor. For this we adopted *residual document frequency* (RDF) presented in our prior work [1]. RDF measures the difference between document frequency of term t within the documents tagged v ($df_{t,v}$) and document frequency predicted by assuming a Poisson distribution:

$$rdf_{t,v} = df_{t,v} - df_v(1 - e^{-\frac{f_t}{n}})$$

where df_v is document frequency tagged v, f_t is the total frequency of term t, and n is the total number of documents.

After generating E, weak edges with low similarity are removed by thresholding.

Given graph G, we can proceed to the (x, y) position calculation. In topigraphy, tags with high similarity are located near each other. To satisfy this requirement, various methods have been

[1] Tag labels are translated from Japanese into English. See original: http://ranger.labs.goo.ne.jp/TG/.

proposed in the area of the graph layout problem. Our prior work [4] also proposed an efficient method for drawing very large-scale graph data based on a force-directed method.

The conventional technique used to indicate the importance or frequency of tagged documents is to alter the font size of the tags. Most conventional algorithms compute node position by treating nodes as "point," i.e., tag size is not considered the size, which can cause node overlapping. Some tools, e.g., graphviz [7], provide a mode that generates a layout without node label overlap, but computation cost of these methods is high and difficult to apply to a large-scale graph.

To resolve this issue, we have developed a new algorithm that gives a different ellipsoidal potential to each node depending on its label (tag) size as well as a common point-symmetric potential and the combination of these two potentials defines a repulsion force. This is a generic algorithm that can solve any graph layout problem and can be applied to various purposes as well as the key component for generating topigraphic images. We will thus report details and tests of this algorithm in a separate paper.

Next step is to calculate tag z position (height). Topigraphy introduces a topographic image to express tag height; it represents the abstraction level of each tag. In Figure 1, for example, "Cooking" is more abstract than "Seafoods" or "Lunch box" and should be given a higher score. This enables the user to grasp the relationship among tags intuitively and to find related topics easily by tracking the ridges of the topigraphy. To calculate such a score, we use the centrality score of *G* since a generic tag has a lot of edges to similar tags and thus has a high centrality score. We adopted k-dense [3] for determining the centrality score.

The final step is to generate the topigraphic image. The (x, y, z) positions generated by the above steps are only the position of each vertex (tag); the surface of the topigraphy is not given. To generate a smooth surface, we used GMT [6], which has a tool to handle fine-grained grid points as well as the position of the tags.

3. EXPERIMENTS

TG is a blog navigation system developed to evaluate the usability and user acceptance of the new navigation interface provided by topigraphy. By clicking a tag in the topigraphy, blog entries of that topic are narrowed down by the tag and displayed in the right pane of the window. This interface is effective for users who want to know about topics in a certain area without having a specific search intention.

TG also provides a keyword search interface. By entering a keyword in the query box, tags related to the keyword are extracted and shown. Icons are also located on the positions of the tags in the topigraphy. By clicking a tag, we can get the keyword search results narrowed down by the tag's topic. This function is effective for preventing topic drift. TG does not use tags manually assigned for narrowing down search results. Instead, TG uses tags re-assigned by the auto-tagging function when blog entries are collected by the crawler since 40% of entries are not tagged, there are many orthographic or synonymous tag variations, and not all tags are informative. See [1] for the details of this auto-tagging function.

The topigraphy corpus and auto-tagging data are re-generated every week. For generating this corpus, TG collects about 10,000,000 blog entries from major Japanese blog providers over the last five weeks. In this collection, about 6,000,000 entries are typically manually tagged by the authors and TG uses them as the corpus. In this collection about 500,000 varieties of tags are collected. From this tag collection, TG extracts about 5,000 varieties of tags based on frequency, informativeness, and uniqueness. To remove non-informative tags such as "Diary" or "Daily life," we use the Euclidean distance of the tag feature vector described in Section 2 as the measure of informativeness and tags with low score are removed by thresholding. We also use the title words of Wikipedia to remove orthographic or synonymous tag variations.

The size of the topigraphy generated by TG is 10,000 x 10,000 pixels and the image is provided as a scrollable map. Although this image is huge, trials confirmed that the method presented in this paper enables the user to find interesting tags easily.

Our approach is also effective in understanding the most recent hot topics mentioned in the blogosphere at a glance. Figure 2 shows the overall structure of Japanese blogs on January 2008.

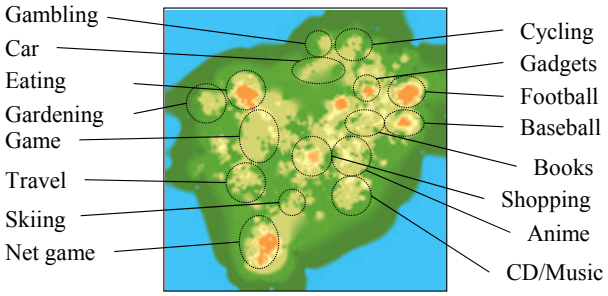

Figure 2. Japanese blogosphere on January 2008.

4. CONCLUSION

Although a quantitative analysis of usability is future work, our trials on TG show that the techniques of improved layout and the use of a background image are effective in displaying large-scale tag clouds. As far as we can determine, no other system can visualize such large-scale tag clouds in a practical service. We are going to apply topigraphy-based navigation to areas other than the blogosphere such as news, movies, shopping malls, books, and Web.

5. REFERENCES

[1] Fujimura, S., Fujimura, K., and Okuda, H. Blogosonomy: Autotagging Any Text Using Bloggers' Knowledge, In *International Conference on Web Intelligence*, 205-212, 2007.

[2] Kaser, O. and Lemire, D. Tag-Cloud Drawing: Algorithms for Cloud Visualization, In *WWW Workshop on Tagging and Metadata for Social Information Organization*, 2007.

[3] Saito, K., Yamada, T., and Kazama, K. Extracting Communities from Complex Networks by the k-dense method, In *ICDM Workshop on Mining Complex Data,* 2006.

[4] Matsubayashi, T. and Yamada, T. A Force-directed Graph Drawing based on the Hierarchical Individual Timestep Method, International Journal of Electronics, Circuits and Systems, Vol.1, No.2, 116-121, 2007.

[5] BLOGRANGER TG, http://ranger.labs.goo.ne.jp/TG/

[6] The Generic Mapping Tools, http://gmt.soest.hawaii.edu/

[7] Graphvis, http://www.graphvis.org/