Web Page Organization and Visualization Using Generative Topographic Mapping – A Pilot Study

Xiao-Feng Zhang, Chak-Man Lam, William K. Cheung

Department of Computer Science
Hong Kong Baptist University
Kowloon Tong, Hong Kong
{xfzhang, johanna, william}@comp.hkbu.edu.hk

Abstract

Automatic Web page organization and visualization is an effective way for foraging information in a Web structure. Web pages contain both text (content) and links (structure), implying that content and structure analysis techniques should be adopted and properly integrated. In this paper, we take the probabilistic model-based approach and extend a topography-preserving model known as Generative Topography Map (GTM). The extended GTM provides a principled way to integrate Web pages and hyperlinks and project them into a low-dimension latent space (2D in our case) for visualization. The proposed extension has been applied to an artificially created dataset and also the WebKB dataset for performance evaluation. Based on the preliminary results obtained, we propose several directions for future research.

Keywords: Web page organization and visualization, Web content and structure analysis, Generative Topography Map

1. Introduction

The need of foraging information in the Web has long been identified and the use of keyword-based search engines is by far the most effective way approach. However, the Web has been developed to such a stage that the information embedded in it can no longer be well managed using only the conventional keyword-based approach. In response to the need, many Web search service providers also provide some pre-categorized information. Web pages contain both text (content) and links (structure). Algorithms that can take into account both Web content and structures for supporting the self-organization and visualization of the Web-based information become increasingly important. While many related algorithms have been proposed for analyzing Web related information with applications to Web page classification [1], topic distillation [2], Web page ranking [3], Web communities identification [4], etc, not many of them are taking the statistical model-based approach which is known to be effective in discovering hidden knowledge in the observed data via the underlying model estimation process.

Generative Topographic Mapping is a model-based non-linear dimension reduction technique that tries to project the observed data space (normally at a high dimension) onto a latent variable space (at a lower dimension) via non-linear mappings, with the topographic relationship of the data preserved. It has known to be effective for data visualization. Among the related techniques, such as self-organizing map (SOM) [6] and linear local embedding (LLE) [7], GTM has the merit of possessing a rigorous probabilistic interpretation and is thus chosen in this paper to facilitate the integration of Web content and hyperlinks in a principled manner. Also, while this kind of techniques have long been applied to content-based automatic organization and visualization of digital archives like text documents [6] and music files [8], we, in this paper, extend the GTM to take into account also the hyperlink structure for the organization and visualization of Web pages. Our work is inspired from the use of the latent class model for related analysis done by Cohen *et al.* [9] which, however, does not support the visualization feature as provided by GTM.

1.1 Paper Organization

The remaining of the paper is as follow. Section 2 gives the background of GTM. Section 3 describes how GTM can be extended for integrating both content and links of Web pages. Section 4 provides the results

obtained by applying the proposed GTM to the WebKB dataset. Limitations of this work as well as possible future research directions can be found in Section 5. Section 6 concludes the paper.

2. Generative Topographic Mapping (GTM)

GTM was first introduced in [5] as a probabilistic non-linear latent variable model. It was primarily designed for exploring the intrinsic dimension of a set of high-dimensional data by assuming that the data are generated due to a set of latent variables in a low-dimensional (usually 2D or 3D) latent variable space. Visualizing the latent variable space with the original data projected back to it can result in a *map* (for 2D case) that provides very intuitive understanding of the structure and organization of the high-dimensional data.

Let $\boldsymbol{t}_n = (t_{n1}, \ldots, t_{nD})$ denote an instance of the set of observed high-dimensional data $\boldsymbol{T}$, $\boldsymbol{z} = (z_i, \ldots, z_k)$ denote a finite set of sample points defined in the L-dimensional latent space, $y(z;\boldsymbol{W}) := \boldsymbol{W}\boldsymbol{\Phi}(z)$ maps in an non-linear fashion a point in the latent space onto a corresponding point in the data space, with the mapping governed by a generalized linear regression model $\boldsymbol{\Phi}$ weighted by $\boldsymbol{W}$. It is generative as it tries to find a parametric representation for explaining the data distribution in the data space, i.e., $p(\boldsymbol{T})$, based on $y(z;\boldsymbol{W})$. Thus, each sample point in the latent space would be mapped onto an L-dimensional non-Euclidean manifold in the data space. A Gaussian distribution in the data space is assumed for $\boldsymbol{t}$ given z_k in data space, given as

$$p(\boldsymbol{t} \mid z_k, \boldsymbol{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\{-\frac{\beta}{2} \| y(z_k;\boldsymbol{W}) - \boldsymbol{t} \|^2\} \quad (1)$$

The overall log likelihood function for GTM is given as

$$L(W, \beta) = \sum_i \ln\{\frac{1}{K} \sum_k p(t_i \mid z_k, W, \beta)\} \quad (2)$$

The EM algorithm is typically used for the GTM parameter estimation, where the E-step is given as

$$P(z_k \mid \boldsymbol{t}_i, W_{old}, \beta_{old}) = R_{ki}(\boldsymbol{W}_{old}, \beta_{old}) = \frac{p(\boldsymbol{t}_i \mid z_k, \boldsymbol{W}_{old}, \beta_{old})}{\sum_{k'} p(\boldsymbol{t}_i \mid z_{k'}, \boldsymbol{W}_{old}, \beta_{old})} \quad (3)$$

And the M-step is

$$\sum_i \sum_k R_{ki}(\boldsymbol{W}_{old}, \beta_{old}) \{\boldsymbol{W}_{new}\phi(z_k) - t_i\} \phi^T(z_k) = 0 \quad (4)$$

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_i \sum_k R_{ki}(\boldsymbol{W}_{old}, \beta_{old}) \| \boldsymbol{W}_{new}\phi(z_k) - \boldsymbol{t}_i \|^2 \quad (5)$$

GTM has been applied to visualizing high-dimensional data like images and documents. In this paper, we would like to extend it for applying to Web pages, where the modeling of hyperlink structure has to be incorporated.

3. Extending GTM for Modeling Web Content and Links

It is believed that the hyperlinks in a Web page provide further (non-linear) cues about how it should be related to other pages, instead of only relying on the Web page content. The question is how the link information should properly be taken into account in order to manifest its effect. Figure 1 shows one possible way for the integration which is inspired from [9].

Let d_j denotes the j^{th} document, c_l denotes the l^{th} cited document, and z_k denote the k^{th} sample point in the latent space, N_{ij} denote the number of occurrences of the feature vector $\boldsymbol{t}_i$ in the j^{th} document and A_{lj} denote the times that the j^{th} document links to the l^{th} document. Note that it is different from the one used in [9], where $\boldsymbol{t}_i$ stands for a word in [9] but a document feature vector in our case.

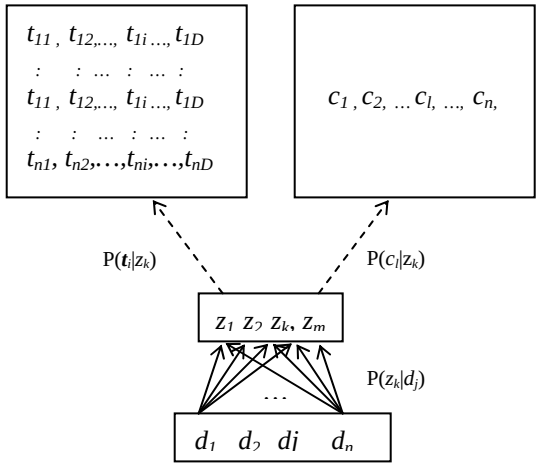


Figure 1 The extended Generative Topographic Mapping.

The whole likelihood function of this model can be written as:

$$L = \sum_i \sum_l \sum_j \log p(t_i, c_l \mid d_j)$$
$$= \sum_i \sum_l \sum_j \log[\, p(t_i \mid z_k, d_j)\, p(c_l \mid z_k, d_j)\, p(z_k \mid d_j)]$$
$$= \alpha \sum_i \sum_j \log[\, p(t_i \mid z_k)\, p(z_k \mid d_j)] + (1-\alpha) \sum_l \sum_j \log[\, p(c_l \mid z_k)\, p(z_k \mid d_j)]$$

By assuming that $\boldsymbol{t}_i$ and c_l are independent given d_j,

$$\begin{aligned} p(\boldsymbol{t}_i, c_l \mid d_j) &= p(\boldsymbol{t}_i \mid d_j) P(c_l \mid d_j) \\ p(\boldsymbol{t}_i \mid d_j) &= \sum_k p(\boldsymbol{t}_i \mid z_k) P(z_k \mid d_j) \\ P(c_l \mid d_j) &= \sum_k P(c_l \mid z_k) P(z_k \mid d_j) \end{aligned} \quad (6)$$

Consider first $p(\boldsymbol{t}_i \mid z_k)$.

The probability that a document feature vector $\boldsymbol{t}_i$ occurs in d_j can be computed, given as

$$\begin{aligned} p(\boldsymbol{t}_i \mid z_k) &= \int_{\beta} \int_{\boldsymbol{W}} p(\boldsymbol{t}_i, \boldsymbol{W}, \beta \mid z_k) d\boldsymbol{W} d\beta \\ &= \int_{\beta} \int_{\boldsymbol{W}} p(\boldsymbol{t}_i \mid \boldsymbol{W}, \beta, z_k)\, p(\boldsymbol{W}, \beta \mid z_k) d\boldsymbol{W} d\beta \\ &\approx p(\boldsymbol{t}_i \mid \boldsymbol{W}^*, \beta^*, z_k)\, p(\boldsymbol{W}^*, \beta^* \mid z_k) \end{aligned} \quad (7)$$

By assuming that $p(\boldsymbol{W}, \beta \mid z_k)$ is uniformly distributed, we get

$$p(t \mid z_k) \propto p(t_i \mid z_k, w, \beta) \quad (8)$$

The log likelihood function becomes

$$L(\boldsymbol{W}, \beta) = \sum_j \Big[\alpha \sum_i \ln\{\frac{1}{K} \sum_k N_{ij}\, p(\boldsymbol{t}_i \mid z_k, \boldsymbol{W}, \beta) P(z_k \mid d_j)\} + (1-\alpha) \sum_l \ln\{\frac{1}{K} \sum_k A_{lj} P(c_l \mid z_k) P(z_k \mid d_j)\} \Big] \quad (9)$$

where α here acts as a trade-off value to balance the weight of the content information and link information for the parameter estimation. Also, N_{ij} is diagonal in our case. In general, we can in fact use some kind of similarity measure for defining the value of N_{ij} (see Section 5 for more discussion).

Following the EM algorithm, one can then easily get corresponding E-step concerning contents as

$$\begin{aligned} P(z_k \mid t_i, d_j) = R_{ijk}(W_{old}, \beta_{old}) &= \frac{p(\boldsymbol{t}_i \mid z_k) P(z_k \mid d_j)}{\sum_{k'} p(\boldsymbol{t}_i \mid z_{k'}) P(z_{k'} \mid d_j)} \\ &\approx \frac{p(\boldsymbol{t}_i \mid z_k, w^*, \beta^*) P(z_k \mid d_j)}{\sum_{k'} p(\boldsymbol{t}_i \mid z_{k'}, w^*, \beta^*) P(z_{k'} \mid d_j)} \end{aligned} \quad (10)$$

and the M-step related to the contents as

$$\sum_i \sum_k \sum_l R_{ijk}(W_{old}, \beta_{old}) \{W_{new} \phi(z_k) - t_i\} \phi^T(z_k) = 0 \quad (11)$$

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_l \sum_j \sum_k R_{ijk}(W_{old}, \beta_{old}) \,\|\, W_{new} \phi(z_k) - t_l \,\|^2 \quad (12)$$

Similar calculation can be applied to links related parameters for deriving the corresponding E-step and M-step. Finally, the E-step and M-step for the extended GTM can be summarized as Figure 2. Note that the effect due to content and links are aggregated in the M-step for estimating $p(z_k \mid d_j)$, which will then affect the posterior probability estimation for z_k in the E-step.

E-step:

$$R_{ijk} = \frac{p(\boldsymbol{t}_i \mid z_k, \boldsymbol{W}, \beta) P(z_k \mid d_j)}{\sum_{k'} p(\boldsymbol{t}_i \mid z_{k'}, \boldsymbol{W}, \beta) P(z_{k'} \mid d_j)}$$

$$R_{ljk} = \frac{p(c_l \mid z_k) P(z_k \mid d_j)}{\sum_{k'} p(c_l \mid z_{k'}) P(z_{k'} \mid d_j)}$$

M-step:

For content:

$$\sum_i \sum_k \sum_l R_{ijk}(W_{old}, \beta_{old}) \{W_{new} \phi(z_k) - t_i\} \phi^T(z_k) = 0$$

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_l \sum_j \sum_k R_{ijk}(W_{old}, \beta_{old}) \,\|\, W_{new} \phi(z_k) - t_l \,\|^2$$

For links:

$$p(c_l \mid z_k) = \frac{\sum_j A_{lj} R_{ljk}}{\sum_{l'} \sum_j A_{l'j} R_{cdz}}$$

$$p(z_k \mid d_j) \propto \alpha \frac{\sum_i N_{ij} R_{ijk}}{\sum_k \sum_{i'} N_{i'j} R_{ijk}} + (1-\alpha) \frac{\sum_l A_{lj} R_{ljk}}{\sum_k \sum_{l'} A_{l'j} R_{ljk}}$$

Figure 2 The EM algorithm for the extended GTM.

4. Experiments

To evaluate the proposed extension, we have first applied the extended GTM model to a small dataset based on around 100 artificially created Web pages of two different categories. As the Web pages are not linking to each other, we randomly added hyperlinks to the pages of the same category. In addition, we have also evaluated the proposed model using three data subsets extracted from the WebKB dataset. In particular, we have prepared three different subsets which consist of 10, 30 and 182 examples from each of the 3 categories of WebKB: course, department and faulty. Thus, the three subsets contain 30, 90 and 546 examples and are labeled as D30, D90, and D546 respectively.

4.1 Preprocessing

Given the data subsets, each Web page has to be pre-processed and represented as a document feature vector to be used for the subsequent model-learning step. A number of pre-processing steps have to be performed. First, all the unnecessary HTML tags and scripts should be removed. Also, the typical stop words removal and stemming steps should be followed. Lastly, only terms with their document frequencies higher than a threshold are retained. The threshold for the artificial dataset is 6 and those for the three data subsets D30, D90 and D546 are 2, 5 and 20 respectively. The distinct terms (and thus the dimension of the document feature vector) obtained for them range from 109 to 551. Then, the content related data $\boldsymbol{T}$ can readily be used for the model learning. As it is well known that the document feature vector dataset can be quite sparse, an interpolation-like process is also performed in such a way that each feature vector is replaced by an averaged version of its neighbors in the feature space.

The link related data $\boldsymbol{A}$ can easily be prepared by following the anchor tags <A> of the Web pages. The hyperlinks existed in the datasets are quite sparse. Preliminary experiments show that the help due to the incorporation of the link information is not significant. In order to amplify their effect on organizing Web pages, we, for each category, have also added hyperlinks, while leaving the inter-class links (provided by the dataset) untouched.

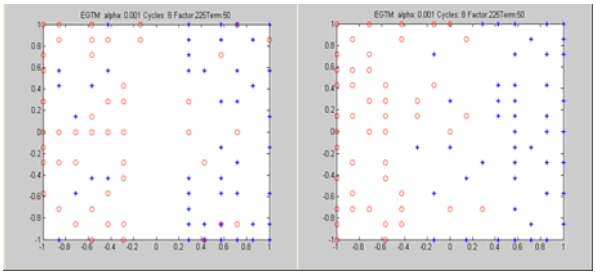
4.2 Experiment 1: Artificial Dataset

To illustrate how the proposed model can make use of within-category links to help differentiate Web pages with overlapping concepts, two FAQ Web pages on *Natural Language Processing* and *Neural Networks* are first chosen.[1] They are sub-fields under Artificial Intelligence. For each of the two pages, we randomly removed different parts to create a set of related documents forming two classes of data --- *NLP* and *NN*, 50 for each class. As expected, terms like "neural, networks, machine, artificial, computer, software" appear frequently in the class *NN*, and terms like "natural, language, linguistics, processing, speech" appears frequently in the class *NLP*. Those differences help the differentiation of the two classes. However, there are terms which are common to the two classes, such as "computer, software, science, project, university" which will sometimes confuse the classification.

In this experiment, altogether 10 latent sample points z_i are evenly selected along the horizontal and vertical directions of a 2D latent space. Given a trained GTM, the posterior probability $p(z_k \mid \boldsymbol{t}_n)$ given a Web page, and thus the mean value of z, can be computed and visualized as a point on the latent space. The visualization of the dataset with 50 terms used for each document feature vector is shown in Figure 3. It is supposed to be manifesting the intrinsic dimension of the high-dimensional data. By varying the value of α of the extended GTM, we would like to see how the incorporation of the link information could help exploring a better mapping, and thus a better data organization in the latent space.

In Figure 3a-f, the red circles and the blue asterisks correspond to the projections of the two categories of data in the latent space with the help of the extended GTM. For α=0.001 (so data organization solely based on content), Figure 3a shows the data organization results without the use of the feature vector averaging step. We noted that the red circles and blue asterisks overlap with each other. This is more or less the original GTM setting. For the same α value, Figure 3b shows that the averaging is an effective way to improve the organization performance.

Besides properly choosing the α value, we also noted significant improvement in data organization. From Figure 3c to Figure 3f, where α is changing from 0.3 to 0.75, it can be seen that the gap between the two class boundaries is getting wider and reaches its best performance at α=0.5~0.75. When α=0.99 (Figure 3f), which means hyperlink information is dominating, the organization quality dropped. In general, this reveals that proper integration of both content and structure information can achieve better data organization quality.



(a) α=0.001, no (b) α=0.001

[1] The URLs of the two pages are:
- http://www-2.cs.cmu.edu/Groups/AI/html/faqs/ai/nlp/ nlp_faq/faq.html
- ftp://ftp.sas.com/pub/neural/FAQ.html

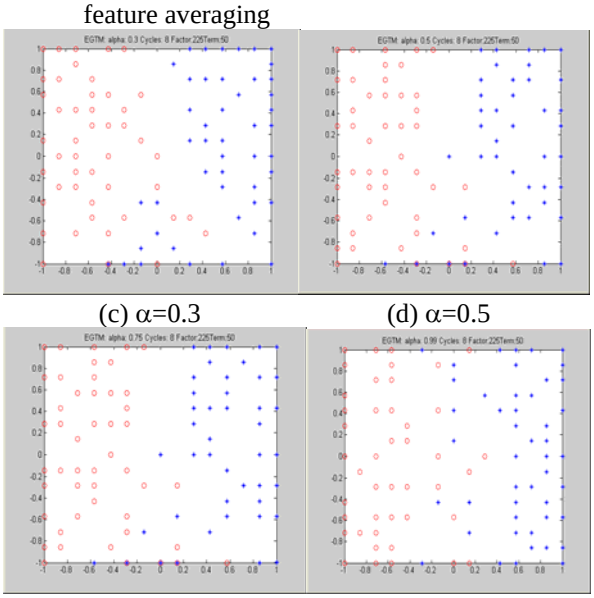


Figure 3 Performance comparison using the Artificial Dataset for the extended GTM with different values of α. The red circles and the blue asterisks correspond to labels for the classes NN and NLP, respectively. The dimension of the feature vector is 50 and the feature vectors are locally smoothed versions as mentioned in Section 4.1.

4.3 Experiment 2: WebKB Dataset

In this experiment, altogether 100 latent sample points z_i are evenly selected along the horizontal and vertical directions of the latent space. The visualization of the dataset D90 with 20 terms used for each document feature vector is shown in Figure 4.

In Figure 4a, it is noted that when α=0.01, which means the data organization is solely based on the content information, more than half of the red circles are in the bottom half of the map. Some red circles (corresponding to the student class) are mixed up with some green ones (corresponding to the faculty class), forming virtually a horizontal linear structure. So, this extended GTM fails to identify a latent space that can make the mixed-up part to be more uniformly distributed and separated. When the value of α is increased to 0.5 (Figure 4b), more emphasis is put on the link information. We can see that the linear structure spreads out and the blue circles (corresponding to the department class) move towards the edges of the map, when compared with Figure 4a.

When the value of α is further raised to 0.99, the projections of the Web pages are separated further apart but with the region of the red circles (student) and that of the green circles (faculty) overlapping with each other to a bit great extent.

We have repeated the experiments for different data subsets with different numbers of terms (10, 30, 50, 80) used for each document feature vector. Similar phenomena are observed when adjusting the value of α.

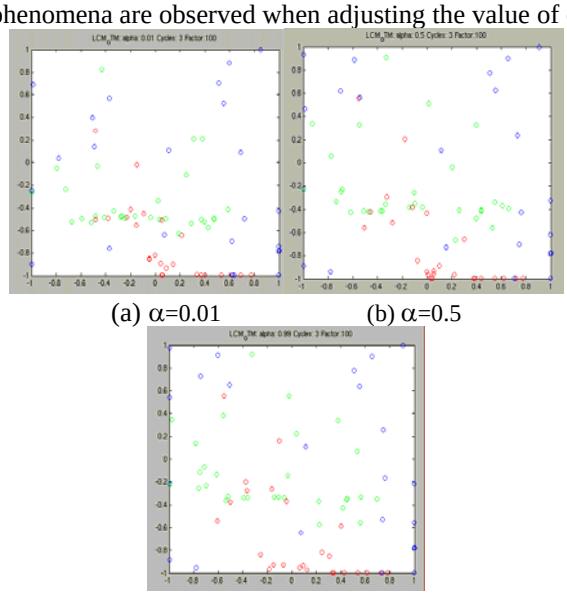
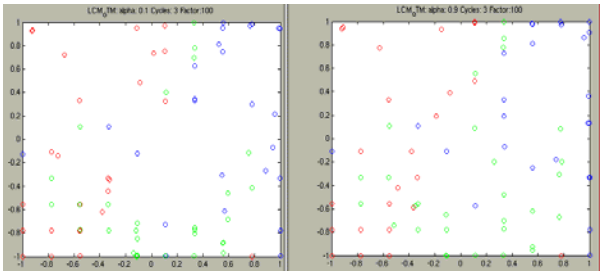


Figure 4 Performance comparison for the extended GTM with different values of α. The red, blue and green circles correspond to labels for the classes - student, department and faculty. The dimension of the feature vector is 20.

Figure 5 shows the result of the same dataset D90 but with 50 terms used for each document feature vector instead. With the introduction of the link information, it is noted that the Web pages feature vectors are further apart from each other in the latent space with α=0.9.

All the results obtained based on the WebKB assumes that artificial links are created to make Web pages of each class fully connected. We have also tested the cases with different densities of the within-class linkage. Preliminary experimental results show that our proposed model is not very sensitive to this type of variations.



(a) α=0.1 (b) α=0.9

Figure 5 Performance comparison for the extended GTM with different values of α. The dimension of the feature vector is 50.

5. Discussion and Future Work

To summarize, it is noted that by properly choosing the value of the trade-off parameter α (between 0 to 1), the extended GTM shows some (though not very significant) improvement in identifying a better latent space for spreading out the Web page projections, when compared with the pure GTM case (i.e., α=0.01 or 0.001). We argue the limited improvement by the fact that the proposed way of incorporating hyperlinks could only help the cases with marginally similar within-class documents to be assigned to the same sample point in the latent space, but failed in grouping together, via hyperlinks, within-class documents with significantly different use of terms.

Based on our preliminary experimental results, we believe that the extended model can further be enhanced in at least the following two ways:

1. A more accurate model for representing text and links in some intrinsic latent space is needed. One possibility is to introduce an additional latent variable corresponding directly to the possible classes of the data, with the hope that the link information can help not only one particular sample point in the latent space but the group of samples points which are supposed to be within the same class.

2. An orthogonal direction for further enhancement is to represent each document as an averaged version of the feature vectors of its own as well as those linked (or cited) by it. This approach is analogous to interpolating the missing values of the term frequencies of a document using the linked documents and to some extend similar to the feature-averaging trick used in the paper.

6. Conclusion

In this paper, we have extended the GTM for automatic organization and visualization of Web pages using both the content-based and link-based information. The preliminary results show marginal improvement based on the proposed extension when compared with pure GTM. A number of important intrinsic problems of the extension as well as possible enhancement have been discussed for future research.

Acknowledgment

This research is jointly supported by Hong Kong Baptist University via Faculty Research Grant FRG/03-04/I-27 and FRG/03-04/II-20.

References

1. Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor, "Composite kernels for hypertext categorization," *Proceedings of the 18th International Conference on Machine Learning*, pages 250--257, Williams College, US, 2001.
2. Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, 46(5):604--632, 1999.
3. Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual {Web} search engine," *Computer Networks and ISDN Systems*, 30(1--7):107--117, 1998.
4. Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans Coetzee, "Self-organization of the web and identification of communities," *IEEE Computer*, 35(3):66--71, 2002.
5. Christopher M. Bishop, Markus Svensen, and Christopher K. I. Williams, "GTM: The generative topographic mapping," *Neural Computation*, 10(1):215--234, 1998.
6. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self Organization of a Massive Document Collection," *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, volume 11, number 3, pages 574-585. May 2000
7. Sam Roweis & Lawrence Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, v.290, no.5500, Dec.22, 2000. pp. 2323--2326.
8. E. Pampalk and A. Rauber and D. Merkl, "Content-based Organization and Visualization of Music Archives," *Proceedings of the ACM Multimedia*, pp.570-579, Juan les Pins, France, December, 2002
9. David Cohn and Thomas Hofmann, "The missing link - a probabilistic model of document content and hypertext connectivity," *Neural Information Processing Systems*, 13, 2001.