A Self Organizing Document Map Algorithm for Large Scale Hyperlinked Data Inspired by Neuronal Migration

Kotaro Nakayama and Yutaka Matsuo
Center for Knowledge Structuring
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
nakayama@cks.u-tokyo.ac.jp, matsuo@biz-model.t.u-tokyo.ac.jp

ABSTRACT

Web document clustering is one of the research topics that is being pursued continuously due to the large variety of applications. Since Web documents usually have variety and diversity in terms of domains, content and quality, one of the technical difficulties is to find a reasonable number and size of clusters. In this research, we pay attention to SOMs (Self Organizing Maps) because of their capability of visualized clustering that helps users to investigate characteristics of data in detail. The SOM is widely known as a "scalable" algorithm because of its capability to handle large numbers of records. However, it is effective only when the vectors are small and dense. Although several research efforts on making the SOM scalable have been conducted, technical issues on scalability and performance for sparse high-dimensional data such as hyperlinked documents still remain. In this paper, we introduce MIGSOM, an SOM algorithm inspired by a recent discovery on neuronal migration. The two major advantages of MIGSOM are its scalability for sparse high-dimensional data and its clustering visualization functionality. In this paper, we describe the algorithm and implementation, and show the practicality of the algorithm by applying MIGSOM to a huge scale real data set: Wikipedia's hyperlink data.

Categories and Subject Descriptors

I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning

General Terms

Algorithms

Keywords

SOM,Wikipedia,Visualization,Clustering,Link Analysis

1. INTRODUCTION

Although a significant number of Web document clustering researches have been conducted in the past, Web document clustering is still a major challenge. Much research is left to be done for applications such as ad matching systems and recommendation systems. Since a set of Web documents usually has variety and diversity in terms of domains, content and quality, one of the technical difficulties is to find a reasonable number and size of clusters. SOM (Self Organizing Map), one of the most popular unsupervised machine learning algorithms that maps high-dimensional vectors into low-dimensional data, has a powerful functionality of clustering visualization. In the past, a number of researches on SOMs for Web documents such as WebSOM [2] have been conducted. However, there still exist two difficulties on this research area; scalability and performance. In the previous researches, dimension reduction was the common practice to make SOMs scalable. We do not disallow the effectiveness of dimension reduction, but it is clear that dimension reduction reduces information which affects the result accuracy (esp. of clustering), so that scalability and performance were a trade-off. In order to satisfy these two demands at the same time, we need an efficient novel computational model.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

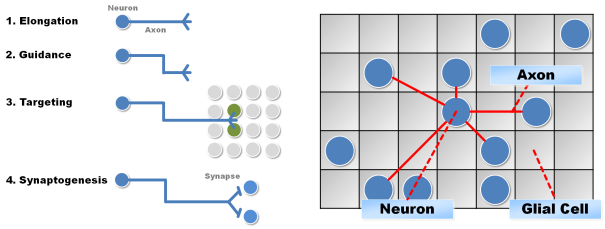

Figure 1: Translocation and MIGSOM
(Movements of neurons are the principle of training. Neurons use their axons to find similar cells. The connectivity is based on Gaussian selection)

There has been a renewal of interest in biomimicry in computer science recently caused by dramatic progress of brain science. Researchers have unveiled more about the brain in the last 10 years than in all previous centuries, due to new analytical technologies such as fMRI. As a result, the mechanism of the central nervous system (CNS) including brain and spinal cord has been uncovered in detail. *Neuronal migration* is one of the research topics that has advanced greatly in this trend. Neuronal migration describes how neurons move from their birth place to their final location in the nervous system. Changing connectivity is widely known as a principle of learning in CNS and recent neuroscience researches have proved the important role of neuronal migration in addition to changing connectivity.

In this paper, we propose *MIGSOM*, a new SOM algorithm inspired by neuronal migration. MIGSOM is also an interactive learning algorithm like Kohonen's, but the data structure and process are different and novel. The main characteristic of MIGSOM is the scalability in terms of both number of documents and dimensionality.

2. MIGSOM

MIGSOM is a new SOM algorithm inspired by neuronal migration. In neuronal migration, neurons find a suitable place and move using their own axons. We show the detailed process of neuronal migration in Figure 1. First of all, neurons start to expand their axons after they are born.

Algorithm $train()$:
1 Randomly select g from G
2 $\vec{m_g}=\vec{O}$ #Initialize by null vector
3 $N = GaussianSelection(g)$
4 **for each** $n \in N$
5 $dist = distance(n, g)$ #Distance in map
6 $power = tanh(dist)$
7 $\vec{m_g}=\vec{m_g} + U(n, g) \cdot power \cdot Sim(n, g)$
8 if $| \vec{m_g} | > t$
9 $Translocate(g, \vec{m_g})$

Figure 2: MIGSOM's Training Algorithm

Then, they explore their peripheral area by using their axons to decide the direction to expand the axons. After finding a suitable place to establish, they create synaptic connections to the neurons around the area. Finally they move to the location by shortening the axons.

MIGSOM, like Kohonen's SOM, has a map (2D grid) to represent geometric relationships among records. This means that similar records are deployed (mapped) closer to each other in the grid. Major differences between Kohonen's SOM and MIGSOM are its representation and learning process. Traditional SOMs including Kohonen's have a grid in which each node has its own vector without reference to the input data (records). MIGSOM, in contrast, has a grid in which each node corresponds to an input record. For instance, for document categorization based on a document-term matrix, each node corresponds to a document vectorized by terms. Precisely, there are two types of nodes; neuron cells and glial cells. Each neuron cell corresponds to a document (input record) and each glial cell has a randomly generated vector. Glial cells work as intermediators to guide neurons. Glial cells had been thought as a trivial cells which just bring energy to neurons but recent researches brought more and more evidences to prove the important rolls of Glial cells in learning process.

MIGSOM's learning process is another difference from traditional SOMs. SOMs iteratively modify the vectors of nodes on the grid but MIGSOM iteratively *move* (migrate) the nodes to organize the map. We illustrate the detailed process of iteration (training) in Figure 2.

MIGSOM begins the training procedure by randomly selecting a node g from the set of nodes in the grid G and adopts a neuron (or glial cell) on the node as training data. Then, initialize $\vec{m_g}$, a motion vector of a neuron on g, by null vector. Next, select a set of random nodes around g by a function that selects nodes following Gaussian distribution $GaussianSelection(g)$. For each n, calculate Euclidean distance by $distance(n, g)$ and similarity by $Sim(n, g)$ to modify $\vec{m_g}$. Finally, migrate the g to direction of $\vec{m_g}$ by using $Translocate(g, \vec{m_g})$ if the $| \vec{m_g} |$ (norm of $\vec{m_g}$) is larger than t. $tanh(dist)$ is a hyperbolic tangent function of distance to calculate the power of reeling in. Farther nodes have stronger power. $U(n, g)$ represents a unit vector from g to n.

3. A CASE STUDY: WIKIPEDIA SOM

In order to prove the practicality of MIGSOM, we applied MIGSOM to Wikipedia's hyperlink data (dumped in April 2009, 3+ million articles, 128+ million internal hyperlinks) and visualized the data for clustering analysis. A Wikipedia article corresponds to a conceptual entity and articles are connected by a number of hyperlinks. Wikipedia's dense link structure has been widely proved in previous researches and it allows researchers to extract relations among articles (entities). An adjacency matrix generated from Wikipedia's hyperlink network is a typical large scale sparse matrix, thus it is suitable for testing our proposed method.

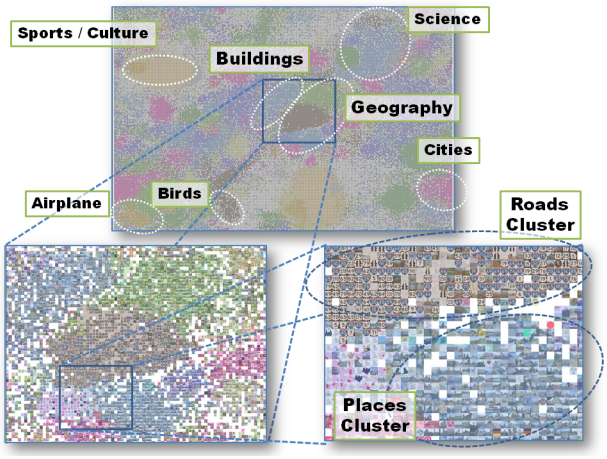

Figure 3: Asynchronous Visualization Interface
(It enables users to analyze global clusters and local clusters seamlessly thanks to the asynchronous interface. See http://sigwp.org/wikisom)

We have developed a cluster visualization system (Figure 3 right) for more detailed analysis. It displays pictures corresponding to articles (we adopted the first image file in each article) and enables users to navigate through the space interactively via Ajax zoom interface. The UI allows users to analyze both the global structure (inter-cluster) and local structure (inter-subcluster or inter-record) seamlessly. This approach makes SOM based clustering more practical, since traditionally the results of SOMs use rasterized static 2D bitmaps and interpreting the results of SOMs for huge data was a technical issue.

4. DISCUSSION AND CONCLUSION

By using this interface, we can understand that various cluster like Sports / Culture, Science, Nature and Cars are generated. In addition to this, we can see that similar clusters are also located closer in the map. Further, magnification for the Geography cluster gives us a detailed view of subclusters such as roads, places and landmarks. SOM's one of the most important applications is cluster visualization for data where both number and size of the clusters are unclear. In other words, the SOM is used to understand the structure of the input data, and in particular, to identify clusters of input records that have similar characteristics. Classification of tumors based on gene expression patterns is a typical example that SOMs are suitable to be adopted [1] because of the difficulties on identification of new cancer classes. In this context, MIGSOM is promising as a tool for cluster visualization because of the MIGSOM's scalability and the visualization functionality.

5. REFERENCES

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[2] K. Lagus, S. Kaski, and T. Kohonen. Mining massive document collections by the websom method. *Information Science*, 163(1-3):135–156, 2004.