Using Structured Self-Organizing Maps in News Integration Websites

Ivan Perelomov, Arnulfo P. Azcarraga, Jonathan Tan and Tat Seng Chua
PRIS Group, School of Computing, National University of Singapore,
Singapore 117543

john@ip4206.spb.edu; {arnulfo+jtan+chuats}@comp.nus.edu.sg

ABSTRACT

The Bveritas system integrates and organizes news articles from English news websites based in Singapore, Malaysia, Philippines, and Thailand, plus news stories from CNN and Reuters International. The main features of the system are the following: 1) automatic clustering of news documents into themes; 2) ordering of these news clusters in a theme map; 3) extraction of meaningful labels for each cluster of news articles; 4) use of extracted labels and title words to retrieve ranked news article lists based on query words; 5) summarization of news articles; and 6) automatic generation of links to related news articles. The novel features of the system result directly from the use of a structured self-organizing map as the underlying logical structure for the news archive.

Keywords

Self-organizing maps (SOM), text classification, web-based systems, news integration system

1. INTRODUCTION

With the shift of many mundane activities over to the Web, most of the major dailies in the world, aside from news agencies like CNN and Reuters International, now have online web-based versions of their printed newspapers. Our *Bveritas* system is a single portal to a number of online dailies. The system downloads news articles from several news websites and automatically organizes the news stories into themes. The title *Bveritas* is a play of two words – the Malay word "berita" which means "news", and the Latin word "veritas" which means "truth".

The idea of a news integrating Web service is not unique. There are a number of existing portals that provide access to news downloaded from numerous online sites. Like most news portals, our system provides both topic and regional categorization, with a combination of headlines and late-breaking news. It also supports retrieval of news stories based on query words. Differing from other systems, however, the Bveritas system employs a self-organizing map (SOM) to automatically organize the news articles in a theme map. Similarity between documents for clustering purposes is based on the frequency distribution of the words used. Through the theme map, the system supports a wide range of services such as retrieval of news articles by query words, intuitive browsing of news archive by theme inspection, integration of searching and browsing, automatic generation of hypertext links to related news stories, and the personalization of the cluster layout to suit the browsing style and interests of readers. Furthermore, our system provides such useful services as news summarization, where a user can choose the desired summarization level of news articles.

2. SYSTEM ARCHITECTURE

The use of a structured Self-Organizing Map as the logical underlying structure of the news archive is the unique feature of the *Bveritas* system. The structured SOM groups together related documents into clusters and organizes the clusters of documents in a rectangular grid in such a way that clusters that are near each other have news documents that are similar.

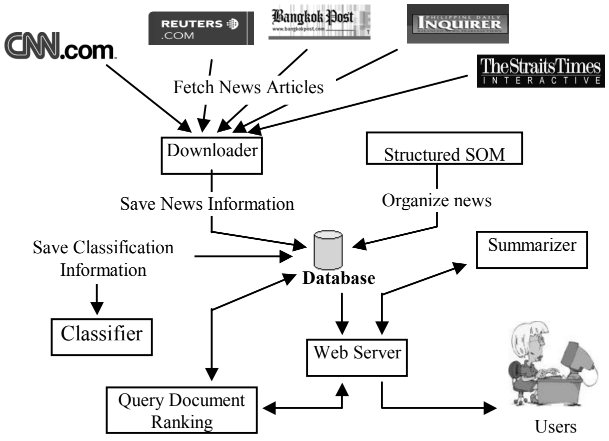

Figure 1. Bveritas system architecture

Aside from the structured SOM, the other components of the Bveritas system are the downloading module, database, news parsing module, classifier, summarizer, query ranking module and user interface (figure 1). The downloading module is concerned with analyzing homepages and retrieving news articles from URLs picked-up up from the news sites. It is designed to run in sessions, with downloading statuses that keep track of how far a specific news article has been processed. Downloading sessions designate specific processing operations to specific computers in our network. The database contains all raw and processed news articles, extracted keywords, various special navigational links, etc. The news parsing service groups together all the various procedures involved in preparing the news articles for further processing, such as the removal of common words, stemming, and extraction of special fields like news title, author, and date. Finally, all the interaction between the user and the database is handled by the interface module.

3. STRUCTURED SOM FOR TEXT ARCHIVES

Various preprocessing operations on the downloaded news articles, such as stop-word removal, stemming, and feature selection, are performed to prepare the news articles for SOM training and labeling. These standard document parsing procedures still result in far too many dimensions (in the order of 10,000 unique terms). Fortunately, there is a fast and very effective way of drastically reducing (to just 500 or even less) the dimensions of the news document vectors using a method referred to as “Random Projection” [5].

3.1 Document Archives based on SOM

The SOM training algorithm has been widely studied and carefully analyzed and has been the subject of rigorous mathematical analysis leading to formal results about its convergence properties [3, 4]. By far, the most prominent use of SOMs in text document archiving and retrieval has been the WEBSOM [6]. The WEBSOM is capable of archiving up to 7 million text documents. Other uses of SOM in document and news classification as well as other related techniques in document clustering have been described in [7].

Our own SOM implementations using the popular Reuters 21,578 collection and the CNN news collection [1, 2, 6] show that the resultant trained SOMs offer a radically different way of browsing through a document collection. This is consistent with claims by WEBSOM proponents [6].

In a SOM-based document archive, documents that are similar to each other are grouped together in the way any document clustering method would do it. On top of that, and this is SOM’s main strength, document clusters are organized and laid out in a simple 2D grid in such a way that document clusters that are similar are positioned next to each other, while clusters that are less similar are positioned farther away. Given this underlying spatial organization of document clusters, a user just has to view the labels extracted for each node (cluster) in the map in order to zero in on potentially interesting clusters. Once an interesting document is found, the user just has to click on other documents in the same cluster as well as on documents in its vicinity (in the 2D grid). Similarly, the search for “related stories” (for automatic generation of links) for a selected story can be narrowed down to a small sub-region in the map.

3.2 Structured SOM

One drawback of SOMs, however, is that since training is unsupervised, the distribution of the resultant clusters can vary significantly from one SOM training to another. We have thus decided to deviate from the SOM methodology in that we now impose a *structure* on the resultant trained map. The “structure” is imposed on a SOM by pre-assigning each node to a given class label.

Although we allow the SOM algorithm to freely organize news articles according to which news articles are similar and which are not, we require that news articles of a certain class label (e.g. Singapore news) will only locate its clusters in a pre-designated region. Procedure 1 presents the revised SOM training algorithm.

Prior to training, pre-assign each node in the map to some class label.

1. *Randomly select a training vector* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ *with class label k*
2. *From among the nodes labeled with the same class label k, find the best matching unit,* $\boldsymbol{m}_c$, *using an appropriate distance metric (e.g. cosine of angle between x and reference vectors)*
3. *Modify the weights of all units i in the winning neighborhood using the following weight update rule:*

$$m_i(t + 1) = m_i(t) + \alpha(t)[x(t) - m_i(t)]$$

where ***t*** *denotes the cycle number,* $\boldsymbol{\alpha}$ *is the learning rate (note that the learning rate decreases and the neighborhood size shrinks as* ***t*** *increases)*

Procedure 1. Structured SOM training algorithm

4. CONCLUSION

We described the Bveritas system, a South-East Asian centric system that integrates and organizes news articles from English news websites based in Singapore, Malaysia, Philippines, and Thailand, plus news stories from CNN and Reuters International. Aside from the usual components of online news sites, the following are the novel features of the Bveritas system:

- automatic clustering of news documents into groups of related news articles (from different websites)
- automatic organization of news clusters in a 2D theme map, where the spatial proximity between clusters in the map reflects the similarity between the clusters
- automatic extraction of meaningful labels for each cluster of news articles
- automatic generation of links to related news articles.

These novel features of the system directly result from the use of a *structured SOM* as the underlying structure for the news archive. We are currently designing alternative ways of producing different structure maps for different user profiles.

5. REFERENCES

[1] Azcarraga, and T. Yap Jr. (2001). SOM-Based Methodology for Building Large Text Archives. *7th Intl Conference on Database Systems for Advanced Applications*, DASFAA01. Hong Kong. April 18-20.

[2] Azcarraga, and T. Yap Jr. (2001). "Comparing Keyword Extraction Techniques for WEBSOM Text Archives", *13th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI 2001), Dallas, Texas, USA, November 7-9.

[3] Clark, D. & Ravishankar, K. (1990) A Convergence Theorem for Grossberg Learning, *Neural Networks*, 3(1), 87-92.

[4] Kraaijveld, MA, Mao J, and Jain AK (1995), A non linear Projection Method Based on Kohonen’s Topology-Preserving Maps, *IEEE Trans on Neural Networks*, 6 (3), pp 548-559, May.

[5] Kohonen, T. (1988*) Self-Organization and Associative Memory*, Series in Information Sciences, Second Edition, Berlin, Springer-Verlag.

[6] Kohonen T. (1998). Self-Organization of Very Large Document Collections: State of the Art. *International Conference on Artificial Neural Networks,* ICANN98. Skovde, Sweden. September 2-4.

[7] Dittenbach M, D. Merkl, and A. Rauber (2000). Using Growing Hierarchical Self-Organizing Maps for Document Classification. ESANN2000. Bruges, Belgium. April 26-28.