2000

Holistic Information Retrieval Through Textual Data Mining

Richard Orwig
Washington State University Vancouver, orwig@wsu.edu

Mark Pendergast
Florida Gulf Coast University, mpenderg@fgcu.edu

Holistic Information Retrieval Through Textual Data Mining

Richard Orwig, Ph.D., Washington State University Vancouver, orwig@wsu.edu

Mark Pendergast, Ph.D., Florida Gulf Coast University, mpenderg@fgcu.edu

Abstract

Information retrieval can be likened to a mining process. Searchers drill through a document space using keywords to extract document subsets. These subsets must be reviewed to extract the topically relevant documents from the irrelevant. Searchers interactively learn from the relevance of the document subsets and re-submit more search arguments to perhaps narrow the search to obtain more relevant documents or broaden the search to improve the likelihood of recalling the highest percentage of relevant documents from the document space. Searchers may be aided in the search by using a document organization scheme used by human categorizers to organize the document space. Such schemes, such as the Library of Congress Classification System, tend to be rigid and dated. Documents are greatly increasing and number and organizational schemes such as the LCCS are not adapting well to the varying content of books and documents being added to the document space. What is needed is an automatic mapping tool that 1) takes the document space as it is, 2) creates a conceptual map of the space, and 3) clusters like documents and places them together on the map. This research (in progress) is an attempt to determine the value of the Kohonen Self-Organizing Map (SOM) (Kohonen, 1995) for use as an interactive textual data mining tool for categorization of large sets of documents. The SOM algorithm analyzed 339 *Management Information Systems Quarterly* abstracts from 1985 to 1997. The first analysis resulted in a map of two major regions -- Information and Systems. This demonstrated that the SOM was working correctly but produced a potentially uninteresting map. What may be more interesting is the next level of conceptual detail, i.e., the major conceptual areas of the *MISQ* document space below this high level of abstraction. To obtain this map, "management," "information," and "systems" was added to a stop-word list and the Kohonen algorithm was reapplied to obtain a mapping of the *MISQ* literature at this second level of detail below Management Information Systems. At both levels of abstraction, the 339 abstracts were partitioned among the conceptual regions. This suggests the possibility for an interactive tool that aids searchers in exploring large document spaces by using a "divide and conquer" approach of information retrieval whereby the tool clusters similar documents into topical regions of the map for exploratory browsing.

Introduction

Typical searches for information retrieval involve sending a search argument of key terms into a document space to retrieve a subset of those documents. The searcher judges the relevance of each returned document to the topic of the search. Even with more sophisticated Boolean searching, the process amounts to sending very specific terms into the document space in hopes of documents using exactly those terms in discussing relevant documents such that a match occurs and those documents are retrieved. The quality of the returned set of documents varies with respect to relevance of each of the documents to the topic of the search. The quality of the complete search session varies with respect to whether every relevant document in the document space is actually collected through multiple queries.

Categorization schemes may be created to organize the document space. The most common is the Library of Congress Classification System used by many libraries to organize book collections by subject area. These schemes are difficult to create and maintain as new topical areas come into being. Further, more and more documents are created daily as new information sources such as the internet support them. The need is for a tool that automatically organizes a non-organized document space into conceptual (topical) regions as well as clusters like documents into their appropriate conceptual regions.

This paper describes the application of the Kohonen Self-Organizing Map (SOM) to 339 *Management Information Systems Quarterly* abstracts from the years 1985 to 1997. The next section describes the SOM and how it was applied to the *MISQ* abstracts. The third section describes the results. Section Four contains conclusions and suggestions for further research work.

2. Kohonen Self-Organizing Map

Neural network classifiers can be organized into two major types based upon the method with which they "learn:" supervised or unsupervised (Lippmann, 1987). Supervised neural networks use a set of data with known attributes and categories to train the network. A back-propagation neural network is an example that uses a supervised learning technique. Once the network is trained with known data new data can be presented to it for classification. An unsupervised learning technique lets the data itself determine the attributes upon which it will organize. The Kohonen Self-Organizing Map (SOM)

repeatedly examines the input data, formulating a two-dimensional view of multi-dimensional data.

Input nodes to the Kohonen algorithm are represented by a vector where each vector value represents a feature of the specific input that it represents. In our case, each input was an *MISQ* abstract and each vector value represented a specific term. The first term of the vector represented the most frequent term occurring in all of the abstracts. The second term of the vector represented the second most frequent term. The top 217 most frequently occurring terms were used (those terms appearing in 15 abstracts or more). For example, given one abstract, if the most frequent term in all of the document space appeared in this abstract, the first vector value was assigned a '1.' If not, it was assigned '0.' If this abstract contained the second most frequently occurring term, the second position contained a '1' and '0' otherwise. Table 1 shows the first ten terms on the frequency list. "Information systems" was the term represented in the first vector value for each abstract.

No.	Freq.	Term
0	290	INFORMATION SYSTEMS
1	245	INFORMATION
2	217	SYSTEMS
3	126	INFORMATION TECHNOLOGY
4	119	SYSTEM
5	118	MANAGEMENT
6	111	RESULTS
7	110	TECHNOLOGY
8	104	DATA
9	100	DEVELOPMENT

Table 1 - First 10 Frequent Terms

The Kohonen neural network algorithm is a two-layer network. One layer is the input layer. In this case, there were 339 input nodes representing the 339 *MISQ* abstracts. The second layer is the output layer. In our case, we set up an output layer of 200 nodes on a 10-row by 20-column grid. Each of the 200 nodes was represented by a 217-term vector that was initialized to random values. The algorithm works by repeatedly going through the input layer, comparing each input vector value to every node value on the map. The node with the least Euclidean distance from the input is the "winning" node. The value of the winning node is adjusted by a learning factor times the difference between that node and the input node. In addition, each node in the neighborhood (say every node within 5 nodes of the winning node) is also adjusted. The neighborhood size is gradually reduced to ensure that the algorithm will finish. After approximately 10,000 iterations, the Kohonen map is considered trained by the input.

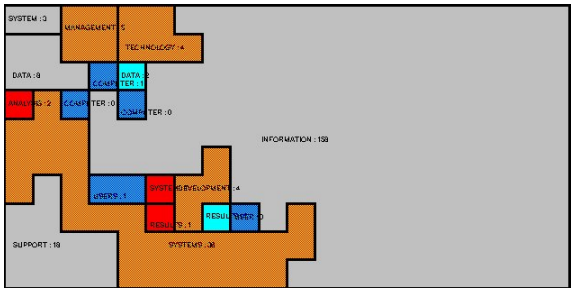
Once trained, the input nodes are run against the map one last time to assign each abstract to a node within a specific region.

The Kohonen software is freely available for research purposes at www.cis.hut.fi/~hynde/lvq/.

3. Results

Output of the initial application of the *MISQ* abstracts is represented in Figure 1:

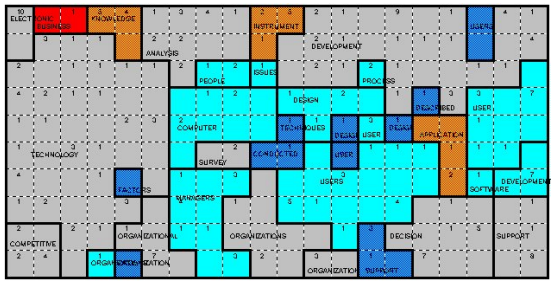
Figure 1 - Initial Kohonen Output



The SOM organized the abstracts into conceptual regions. Figure 1 shows the largest (most important) region is INFORMATION. Related conceptual regions are physically close together. Thus, while SYSTEMS is also a comparatively large area that is conceptually related to both INFORMATION as well as SUPPORT.

It would seem to make sense that the two most important topics pertaining to *MISQ* abstracts would be "information" and "systems." This level of abstraction is too high to be informative as to the conceptual regions of *MISQ*. If a searcher wanted to stay at this level and explore *MISQ* in greater detail, s/he could click on the regions and read the abstracts. However, it would be more interesting to treat INFORMATION and SYSTEMS as meaningless terms for this exploration and re-organize the document space. Since "management," "information," and "systems" are the main conceptual components of *MISQ* we added all three terms to the stop-word list and re-ran the algorithm. The new map is shown in Figure 2. In this case, however, we have shown the Node View of the map.

Figure 2 - Node View Kohonen Output with Updated List



The Node View shows the location of the abstracts contained within the region. Interestingly, ELECTRONIC and COMMERCE are co-located concepts in the upper-left portion of the map. The colors of each region are color-coded based upon the average relative age of the abstracts contained within the region. Red is more recent. Dark blue is oldest. Gray is average.

The red color of the COMMERCE region indicates the one abstract in that region is more recent. In fact, clicking on that node of the map will display the abstract of the El Sawy and Bowes December, 1997, article entitled "Redesigning the customer support process for the electronic economy: Insights from storage dimensions":

"Insights are provided for redesigning IT-enabled customer support processes to meet the demanding requirements of the emerging electronic economy in which fast response, shared knowledge creation, and internetworked technologies are the dynamic enablers of success. A description is provided of the implementation of the TechConnect support system at Storage Dimensions, a manufacturer of high-availability computer storage system products. TechConnect is a unique IT infrastructure for problem resolution that includes a customer support knowledge base whose structure is dynamically updated based on adapted learning through customer interactions. The impacts of TechConnect and its value in creating a learning organizations are assessed. Insights are then drawn for redesigning knowledge-creating customer support processes for of business conditions of the electronic economy. Copyright University of Minnesota, MIS Research Center Dec 1997"

4. Conclusions

Humans are notoriously inconsistent and slow in classifying and organizing information spaces. In our search for information, we have need of a tool that will automatically organize the information space where the organizing scheme of that space is created by the current attributes of the data itself. The Kohonen Self-Organizing Map is a tool that can help organize textual data (Upson, 1989) (Mikkulainen, 1993) (Lin, et al., 1991) (Honkela, 1996) (Chen, et al., 1998) (Orwig, et al., 1997). However, additional interface tools are needed to make the use of the tool more interactive. At least two directions are possible: terms manipulation and regional manipulation with drill down capabilities. Terms manipulation involves two further possibilities: Stop-word list manipulation and vector weighting. We should enable easy adding or removing terms from the stop-word list that are meaningless or uniquely meaningful, respectively, to the topic of the search as we added "management," "information," and "systems" in our second iteration. Vector weighting means that we need to experiment with altering the vector values of terms representing each document (abstract). Current vectors are either "1" or "0" depending upon whether the term for that position appears in the document. It may be possible to adjust the value of a given term in a document relative to other terms contained within the specific document using a function such as the cosine function (Lippmann, 1987). Finally, additional work is needed to allow user combination of regional areas that s/he may feel is actually one region (e.g., ELECTRONIC and COMMERCE in Figure 2).

References

Chen, H., Orwig, R., Titkova, O. and Nunamaker, J.F., Jr. "Information Visualization for Collaborative Computing," *IEEE Computer* (31:8), 1998, pp. 75-82.

Honkela, T. "Newsgroup exploration with WEBSOM method and browsing interface," Helsinki University of Technology, 1996 1996.

Kohonen, T. *Self-Organizing Maps*, Springer-Verlag, Berlin, 1995.

Lin, X., Doergel, D. and Marchionini, G. "A self-organizing semantic map for information retrieval," Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, New York, NY, 1991,

Lippmann, R.P. "An Introduction to computing with Neural Nets," *IEEE ASSP Magazine*), 1987, pp. 4-22.

Mikkulainen, R. Subsymbolic Natural Language Processing: An Integrated Modelof Scripts, Lexicon, and Memory, MIT Press, Cambridge, MA, 1993.

Orwig, R., Chen, H. and Nunamaker, J.F., Jr. "A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output," *Journal of the American Society of Information Science* (48:2), 1997,

Upson, C. "The Application Visualization System: A Computational Environment for Scientific Visualization," *IEEE Computer Graphics and Applications* :July), 1989, pp. 30-42.