Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

Extracting Information from Text Data Bases

Rudolf Albrecht[1]

Space Telescope European Coordinating Facility, European Southern Observatory, Karl Schwarzschild Str. 2, D-85748 Garching, Germany, Email: ralbrech@eso.org

Dieter Merkl

Institut für Softwaretechnik, Technische Universität Wien, Resselgasse 3/188, A-1040 Vienna, Austria, Email: dieter@ifs.tuwien.ac.at

Abstract. The problem of computer assisted literature search is often misstated as the attempt to find, in a repository of electronic literature, a paper which, according to title and keywords, contains the desired information. This assumes that the original author was correct and complete in assigning keywords, and that the search criteria are free of preconceived expectations on the context in which the information might be found. Crossing the boundaries of disciplinary subfields is very difficult in this manner.

To avoid these problems it is necessary to categorize and classify the documents in the text data base using the full text. This will identify papers of interest, and of unexpected cross-relevance. We show the results of experiments using unsupervised classification based on neural networks.

1. Electronic Publishing and Literature Data Bases

In astronomy, several years of major journals are now accessible through the Internet. Within a very short time we will have the current body of astronomical knowledge available for computer processing.

Using appropriate tools and networks we can consider the electronically available astronomical literature to be one huge text data base, consisting typically of technical/scientific articles, written in scientific English and using well defined terminology. In addition to improved access and timely availability electronic publications have the advantage of being searchable. Organizations like the NASA Astrophysics Data System (ADS; `http://adswww.harvard.edu/`) specialize in such services, freeing the user from having to read an increasingly enormous amount of material in order to find the desired information.

[1]Astrophysics Division, Space Science Department, European Space Agency

2. Text Data Mining Using Neural Networks

Artificial neural networks are well suited for tasks that are characterized by noise, poorly understood intrinsic structure, and changing characteristics, each of which are present when dealing with text. Learning techniques are favorable in such an environment compared to algorithmic and knowledge-based approaches.

Neural networks consist of one or more layers of processing elements (neurons). The "program" is stored in a distributed fashion within the "weights" of the connections between processing units. Input patterns presented to the neural network will propagate selectively through it, depending on the different weights of the different connections. Neural networks are inherently parallel computational models. It is possible to "train" the network by changing the weight of the connections such that certain inputs generate certain outputs.

The self-organizing map (Kohonen 1995) is an unsupervised neural network for ordering high-dimensionality statistical data in such a way that similar input items will be grouped close to each other. The utilization of self-organizing maps for text data mining already has found appreciation in information retrieval research (cf. Kohonen et al. 1996; Lagus et al. 1996; Merkl 1997a, 1997b, 1998).

Each of the units i of the self-organizing map is assigned an $n-$dimensional weight vector m_i, $m_i \in \mathcal{R}^n$. The weight vectors have the same dimension as the input patterns (the document representation in our application). Each training iteration t starts with the random selection of one input pattern $x(t)$, which is presented to the self-organizing map and each unit determines its activation. The unit with the lowest activation is referred to as the *winner*, c, of the training iteration, i.e., $m_c(t) = min_i\|x(t) - m_i(t)\|$. Finally, the weight vector of the *winner* as well as the weight vectors of selected units in the vicinity of the *winner* are adapted. This adaptation is implemented as a gradual reduction of the difference between input pattern and weight vector, i.e., $m_i(t+1) = m_i(t) \cdot \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)]$. Geometrically speaking, the weight vectors of the adapted units are moved a bit towards the input pattern. The amount of weight vector movement is guided by a so-called learning rate, α, decreasing in time. The number of units that are affected by adaptation is determined by a so-called neighborhood function, h_{ci}. This number of units also decreases in time.

As the Euclidean distance between those vectors decreases and the weight vectors become more similar to the input pattern, the respective unit is more likely to win at future presentations of this input pattern. The consequence of adapting not only the *winner* alone but also a number of units in the neighborhood of the *winner* leads to a spatial clustering of similar input patters in neighboring parts of the self-organizing map. Thus, similarities between input patterns that are present in the n-dimensional input space are mirrored within the two-dimensional output space of the self-organizing map.

Note that there is no attempt to "understand" the contents of the documents. The goal, at least initially, is "document routing" or "information filtering": identifying those documents which are, or should be, of interest to the user given a particular information profile.

3. Classification of Documents

The general problem of classification is to place N data points into M bins, where $1 < M << N$. The width of the bins is not necessarily M/N and the difference between bins should be meaningful.
There are two different basic classification strategies:

1. Try to reproduce a classification which is already being used, has been shown to be useful, and can be improved by adding new data.
2. Try to discover classes which are intrinsic to the data.

In the case of text mining the desired output pattern is generally not known. The conventional classification system as used in libraries is not applicable, because it is too coarse, and because even documents classified in dissimilar categories (e.g., cosmology and asteroids) can be of relevance to each other because they contain common topics (e.g., spectral analysis)

A possible case with a predefined output pattern is the “user profile”: based on a user’s previous document retrieval behavior the network can be trained to identify documents which fit the pattern of the user.

4. Results and Plans

ApJ Letters articles from November and December 1997 were merged into an input data set of 128 documents. Full-text indexing of all documents was performed. Words that appear in less than 13 documents or more than 116 documents were excluded. 1451 words (terms) remained. The terms were weighed according to $tf * idf$ (term frequency times inverse document frequency) (Salton & Buckley 1988). The scheme ensures that terms that occur frequently within a document but rarely within the whole collection are assigned high weights, providing best discrimination between documents. 10^4 iterations were performed to train the network.

The output consists of an 8 by 8 element map. Documents contained in a bin are considered similar, similarity decreases with distance between bins. The largest bin contains 10 papers, seven of them on spectroscopy of galaxies, which also was the common denominator with the three outliers (solar, brown dwarf, and one paper on a cool degenerate star); another common denominator is “lensing”, which occurred as gravitational lensing in seven papers, and in one more as part of the name of an author, an obvious shortcoming.

Other prominently populated bins contain papers on brown dwarfs (4 out of 4) and solar mass ejections (4 out of 4). However, the potentially most interesting bins are the ones which contain papers with very dissimilar topics, for instance a bin populated by six papers with all different topics (e.g., the Sun, Seyffert galaxies, SN187a). Closer inspection shows, however, that all papers deal with the physics of hot gas in a magnetic environment. The “Subject Heading” keywords of the papers would not have helped: the only common keyword is “ISM” (Interstellar Medium), and it occurs only in three papers. Our system is essentially telling the Seyfert galaxy researcher to read, among others, a paper on solar mass ejections. This is something which Seyfert specialists do not ordinarily do, but which, as it turns out, might be of enormous cross-specialty relevance.

At this point the freedom of the system to determine classification criteria is still too unrestricted: there are too many non-domain specific terms. For example, we find bins with dissimilar papers describing similar techniques for observing or data analysis. While this is still acceptable we also find bins with purely linguistic classification criteria, focusing on terms like possible, probable, theoretical, and systematic. Although these are valid terms they should not be taken into account.

The obvious next step is to perform lexical pre-processing on the input data set. All terms will be compared with a set of valid terms, and only these will be retained. The valid terms can be derived from a training set of representative input data (i.e., a private dictionary), or the Astronomy Thesaurus (`http://msowww.anu.edu.au/library/thesaurus/`) can be used. To further increase the efficiency of the classification all synonyms should be eliminated (like magnitude and brightness). Super-terms can be established and all sub-terms can be replaced by them. The effect will be to turn the input papers into sets of keywords which characterize (but not describe) the papers.

It is obvious that these first experiments are only a small step towards the ideal result. It is equally obvious, however, that strategies such as the one we describe represent the only hope of coping with the enormous amount of published literature and the ever increasing fractionality of the field.

References

Kohonen, T. 1995, Self-organizing Maps, (Berlin: Springer-Verlag)

Kohonen, T., Kaski, S., Lagus, K., & Honkela, T. 1996, in Proc. of the Intl. Conf. on Artificial Neural Networks, (Bochum), 269

Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. 1996, in Proc. of the Intl. Conf. on Knowledge Discovery and Data Mining, (Portland), 238

Merkl, D. 1997a, in Proc. of the European Symp. on Principles of Data Mining and Knowledge Discovery, (Trondheim), 101

———, 1997b, in Proc. of the Intl. ACM SIGIR Conference on R&D in Information Retrieval, (Philadelphia), 186

———, 1998, in A Handbook of Natural Language Processing - Techniques and Applications for the Processing of Language as Text, ed. R. Dale, H. Moisl, & H. Somers (New York: Marcel Dekker), in press

Salton, G. & Buckley, C. 1988, Information Processing & Management 24, 513