Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

Knowledge Discovery in Literature Data Bases

Rudolf Albrecht[1]

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

Dieter Merkl

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

Abstract. The concept of knowledge discovery as defined through "establishing previously unknown and unsuspected relations of features in a data base" is, *cum grano salis*, relatively easy to implement for data bases containing numerical data. Increasingly we find at our disposal data bases containing scientific literature. Computer assisted detection of unknown relations of features in such data bases would be extremely valuable and would lead to new scientific insights. However, the current representation of scientific knowledge in such data bases is not conducive to computer processing. Any correlation of features still has to be done by the human reader, a process which is plagued by ineffectiveness and incompleteness.

On the other hand we note that considerable progress is being made in an area where reading all available material is totally prohibitive: the World Wide Web. Robots and Web crawlers mine the Web continuously and construct data bases which allow the identification of pages of interest in near real time.

An obvious step is to categorize and classify the documents in the text data base. This can be used to identify papers worth reading, or which are of unexpected cross-relevance. We show the results of first experiments using unsupervised classification based on neural networks.

1. The Astronomical Research Process

The research process has only recently been defined in epistemological terms: the model developed by Sir Karl Popper (1972) comes closest to what most natural scientists do when they "do science".

Briefly, the research process starts with the input of signals, either through sensory perception, or through measuring devices which register signals which

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

are either too faint or not suited for our senses. In science we know this step as data acquisition.

The next step is the transformation of the input data into meaningful values, quite often literally the "data reduction" from a jumble of instrument dependent individual measurements to a much smaller, coherent and consistent set of parameters.

By injecting concepts into the collection of parameters we construct models. Concepts range from very simple, such as a linear correlation, to the very complex, like evaporating black holes. The injection of concepts happens spontaneously and associatively, it is a result of the evolution of our brain (Albrecht 1988).

Models come in two flavors, hypotheses and theories, the difference being that a hypothesis is an as-of-yet unsubstantiated and incomplete theory. Given the fact that no theory is ever complete it is more correct to say that all models are hypotheses. This is in agreement with the historical observation that even "wrong" models served well as good hypotheses in a heuristic sense.

Good models allow to make predictions as to future observations. They also allow to add to our pool of concepts by abstraction and generalization. If a model conflicts with observations we have to discard it. Since we can never be certain that any model will forever withstand the test of future observations Popper concludes that in science we can never demonstrably attain the "truth".

Asking the question where in this process the most progress has been made historically in astronomy we tend to think that it has been in the first step: the introduction of ever more powerful telescopes and detectors, and the opening of more spectral windows has allowed to quite literally include observations of the whole universe into the building of models.

We contend, however, that the most progress has been made in the application of concepts: the scientific revolution (i.e. paradigm change) during the period of enlightenment removed concepts like that of the supernatural, of magic and of the subjective from our model building tools, which indeed provided us with the very basis of what we today call scientific thinking.

2. Representation of Models

Models found through the above process are described by scientists using a combination of natural language (with exactly defined semantic content of crucial elements usually called technical terms) and mathematical representation. In other words, a scientific publication, and, more generally, the scientific library, constitute a knowledge base, right now encoded in the idiosyncratic literary style of different authors with different cultural and language backgrounds.

In astronomy we have converged on one main representation language which we call scientific English, the quality of which, however, differs considerably between scientists, severely limiting their ability to convey, as an author, or to internalize, as a reader, a scientific model. It is thus desirable to define a meta language for conveying scientific information, which is both human readable and computer processable.

For the past 20 years essentially all important astronomical publications have been published in English. While this is a disadvantage for the non-native

English speakers it is an enormous advantage for the science of astronomy. In almost no other science are all active scientists able to communicate with each other so easily. All activities which have the potential of a deviation from this situation must therefore be forcefully resisted.

However, even with all-English publications human-to-human knowledge transfer is suboptimal. The sheer volume of the published material makes it impossible to read more than what pertains to ones immediate area of interest. This has, over the recent past, led to an ever increasing fractionalization of the natural sciences in narrow areas of specialization. It is highly likely that a large body of new knowledge is "hidden" (i.e. implied) in the existing literature, but cannot be extracted because serendipitous reading is exceeding the capacity of humans to absorb information. This applies even stronger to interdisciplinary reading, like astronomy and physics.

The obvious way to solve this problem is to use the information processing capability of computers. However, the current publication procedures are not well suited to this: the use of a natural language makes computer-assisted processing of published knowledge impossible. A potential solution would be to use a meta language. First steps towards a meta language have been taken: there is a reference dictionary[2] (Lortet et al. 1994) and a thesaurus[3] (Shobbrook & Shobbrook 1993) which cover the field of astronomy.

If we extrapolate this concept we arrive at the following vision: 1) "publishing" will not be done in the form of papers, but as additions or modifications to a global knowledge base; 2) Consistency checking, novelty assurance, truth maintenance, etc., is immediately and easily possible, eliminating refereeing; 3) the knowledge base, or segments of it, can be mapped into different natural languages, (even languages which the contributors do not speak) and at different levels (such as textbook, or popular description).

HOWEVER: This is unlikely to happen soon, if at all, and even if it did, it would not include the already existing literature. This means that we have to start with the processing of literature published in scientific English.

3. Electronic Publishing and Literature Data Bases

Astronomy has pioneered electronic publishing. Several years of major journals are now accessible through the Internet. Even for literature which is not published electronically, the abstracts are available on abstract data bases. It is safe to predict that within a very short time we will have the current body of astronomical knowledge available for computer processing.

Using appropriate tools plus the already existing network connections we can consider the electronically available astronomical literature to be one huge text data base. This text data base consists typically of technical/scientific articles of several pages in length, written in scientific English and using well defined terminology.

---

[2]http://vizier.u-strasbg.fr/cgi-bin/Dic

[3]http://msowww.anu.edu.au/library/thesaurus/

In addition to improved access and timely availability electronic publications have the advantage of being searchable. There are organizations like the NASA Astrophysics Data System (ADS, http://adswww.harvard.edu/) which specialize in such services. The aim is to free the user from having to read an increasingly enormous amount of material in order to find the desired information.

Advanced search services are becoming available because of a medium in which reading through all available material is totally prohibitive: The World Wide Web. Such search services are convenient and useful. However, as of today they are still mainly text-string oriented and not context oriented.

4. Text Mining

An immediate goal of text data mining is to construct synopses of the material: summaries of the topics which are covered by the documents in the text data base according to criteria defined by the scientist. Another goal is to identify salient points: concise lists of different topics, if possible in order of importance, adjustable in depth.

An important goal is taxonomy: determination of the topics in the documents which are (or should be) of interest to the scientist. This is to be followed by classification: the grouping of documents containing different topics, either as defined by the scientist, or as defined by the information content. The most valuable help for the scientist consists of the identification of dependencies of the different topics on each other, especially of unexpected relationships.

The long term goal has to be to break up the body of astronomical literature into processable pieces of information. In analogy to knowledge discovery in a numerical data base we can then do knowledge discovery in this data base which contains concepts, models, and hypotheses. We can aim for the discovery of implied, previously unknown, and potentially useful knowledge from such a data base. Alternatively, candidate hypotheses can be injected into such a data base with the purpose of either supporting or disproving the hypothesis.

Having the contents of this data base represented in a meta language would obviously facilitate this process enormously. However, some advances should be possible even on the basis of just text in scientific English. It is evident that the capability to do this would immediately lead to enormous advances in scientific productivity.

5. Experiments: Text Data Mining using Neural Networks

The field of neural networks in a wide variety of applications has attracted renewed interest, which is at least partly due to increased computing power available at reasonable prices and the development of a broad spectrum of highly effective learning rules. Artificial neural networks are well suited for areas that are characterized by noise, poorly understood intrinsic structure, and changing characteristics. Each of which are present when dealing with natural language text as the data collection. Learning techniques are favorable in such an environment compared to algorithmic and knowledge-based approaches. The obvious reason is that learning techniques are capable of self-adjustment to changing

conditions whereas algorithmic and knowledge-based approaches require manual, thus expensive, adaptation.

In general, neural networks consist of one or more layers of processing elements (neurons) which are connected to each other. The "program" (or "knowledge") of a neural network 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. However, they are normally implemented as software on a conventional computer.

It is difficult to "program" the network. It is, however, possible to "train" the network by changing the weight of the connections such that certain inputs generate certain outputs.

Neural networks can be used for supervised and unsupervised classification. Supervised classification requires previous knowledge of the classes, represented by a training set. Unsupervised classification allows the neural network to uncover the intrinsic structure of the text data base. We feel that especially unsupervised neural networks are well-suited for text data mining because particularly non-obvious associations between documents are of interest. Contrary to that, supervised neural networks may only be used to represent predefined associations. These models are thus less effective in uncovering non-obvious associations.

The self-organizing map (Kohonen 1982, Kohonen 1995) is an unsupervised neural network architecture for ordering high-dimensionality statistical data in such a way that similar input items will be grouped close to each other. As a consequence, the training results of self-organizing maps provide a convenient starting point for interactive exploration of document spaces. 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, Lin et al. 1991, Merkl 1997a, Merkl 1997b, Merkl 1998).

The self-organizing map consists of a number of neural processing elements, i.e. units. Each of these units i is assigned an n-dimensional weight vector m_i, $m_i \in \Re^n$. It is important to note that the weight vectors have the same dimension as the input patterns (the document representation in our application). The training process of self-organizing maps may be described in terms of input pattern presentation and weight vector adaptation. Each training iteration t starts with the random selection of one input pattern $x(t)$. This input pattern is presented to the self-organizing map and each unit determines its activation. Usually, the Euclidean distance between weight vector and input pattern is used to calculate a unit's 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.

This movement has as a consequence, that the Euclidean distance between those vectors decreases and thus, 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. The training process of the self-organizing map describes a topology preserving mapping from a high-dimensional input space onto a two-dimensional output space where patterns that are similar in terms of the input space are mapped to geographically close locations in the output space.

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.

6. 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. Noisy input data cause problems at bin boundaries. The difference between bins should be meaningful and preferable reflect the physics behind the data.

Conventional classification in astronomy is often limited by the human classifier: physiology and dimensionality. This implies that human based classification does not always best reflect the physics of the data.

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. The number of classes and the dimensionality can be chosen (or iteratively determined) to meet the above requirements.

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).

The requirement for a document classification system is to group (clump, cluster, i.e. classify) SIMILAR documents. Self-organizing maps use the learning ability of neural networks to achieve this. A possible case with a predefined output pattern is the "user profile": based on previous retrieval behavior the network can be trained to identify documents which fit the pattern of the user. The criterion for similarity is basically the frequency of the occurrence of terms and combinations of terms.

7. Results

Articles in ApJ Letters November and December 1997 were merged into an input data set consisting of 129 input documents, and 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, Salton 1989). Such a weighting scheme ensures that terms that occur frequently within a document but rarely within the whole document collection are assigned high weights. These terms are generally credited for providing best discrimination between documents. Ten thousand iterations were performed to train the network (each document was presented about 80 times).

The output consists of an 8 by 8 element map. Documents contained in a bin are considered similar, with similarity decreasing with the distance between bins. With only 129 input documents and 64 bins some of the bins are rather sparsely populated. However, the results were quite satisfactory: the largest bin contains 10 papers, seven of them dealing with spectroscopic work on galaxies, which also was the common denominator with the three outliers (one solar, one 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 which has to be addressed.

Other prominently populated bins contain, for instance, 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. The most prominent one is a bin populated by six papers, all of which have different topics (e.g. the Sun, Seyfert galaxies, SN1987a). 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.

It is obvious that these first experiments fall short of the ambitious goals outlined in the first sections of this paper. 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

Albrecht, R., 1988, On the Interrelation Between Technology and Evolution, In: Frontiers and Space Conquest, J. Schneider & M. Leger-Orine (eds.) Kluwer Academic Publishers, 221

Kohonen, T., 1982, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43, 59–69

Kohonen, T., 1995, Self-organizing maps, Berlin: Springer-Verlag.

Kohonen, T., Kaski, S., Lagus, K., & Honkela, T., 1996, Very large two-level SOM for the browsing of newsgroups, In: Proceedings of the Int'l Conference on Artificial Neural Networks, Bochum, Germany, 269–274

Lagus, K., Honkela, T., Kaski, S., & Kohonen, T., 1996, Self-organizing maps of document collections – A new approach to interactive exploration, In: Proceedings of the Int'l Conference on Knowledge Discovery and Data Mining, Portland, OR, 238–243

Lin, X., Soergel, D., & Marchionini, G., 1991, A self-organizing semantic map for information retrieval, In: Proceedings of the Int'l ACM SIGIR Conference on R&D in Information Retrieval, Chicago, IL, 262–269

Lortet, M.-C., Borde S., & Ochsenbein F., 1994, Second Reference Dictionary of the Nomenclature of Celestial Objects, Astron. Astrophys., Suppl. Ser. 107, 193

Merkl, D., 1997a, Exploration of document collections with self-organizing maps – A novel approach to similarity representation, In: Proceedings of the European Symposium on Principles of Data Mining and Knowledge Discovery, Trondheim, Norway, 101–111.

Merkl, D., 1997b, Exploration of text collections with hierarchical feature maps, In: Proceedings of the Int'l ACM SIGIR Conference on R&D in Information Retrieval, Philadelphia, PA, 186–195.

Merkl, D., 1998, Text Data Mining, In: A Handbook of Natural Language Processing – Techniques and Applications for the Processing of Language as Text, R. Dale, H. Moisl, & H. Somers (eds.), New York: Marcel Dekker, in Press

Popper, K., 1972, The Logic of Scientific Discovery, London: Hutchinson

Salton, G., 1989, Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer, Reading, MA: Addison-Wesley

Salton, G., & Buckley, C., 1988, Term weighting approaches in automatic text retrieval, Information Processing & Management 24(5), 513–523.

Shobbrook, R. M. & Shobbrook, R. R., 1993, The Astronomy Thesaurus, Version 1.1, Epping, New South Wales, Australia: Anglo-Australian Observatory for the International Astronomical Union