*Library and Information Services in Astronomy IV*
*July 2-5, 2002, Prague, Czech Republic*
*B. Corbin, E. Bryson, and M. Wolf (eds)*

# Vocabulary Mapping in the NASA ADS: Prospects for Practical Subject Access

Jonghoon Lee & David Dubin

*Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*
*j-lee43@uiuc.edu & ddubin@uiuc.edu*

**Abstract.** The popular NASA Astrophysics Data System includes bibliographic records indexed with terms from a variety of semi-compatible descriptor languages. These include coordinate index terms taken from the *NASA Thesaurus* and *Astrophysical Journal* subject headings, among others. We have worked to develop a system that takes as input the NASA terms assigned by professional indexers, and translates them into *ApJ* headings. Our system maps sets of descriptors, rather than individual descriptors, since two or more coordinate index terms may translate to a single pre-coordinated subject heading. We began our study with lexical resemblance as the main source of evidence and later developed a connected system that exploits patterns of consistent co-assignment in a subset of the ADS collection that is indexed using both *ApJ* headings and NASA terms. Our most recent efforts have been aimed at improving the network's performance via supervised learning. In this paper we present the results of our most recent formal evaluation studies and an examination of some specific documents drawn from a set we've mapped using the network.

## 1. The Heterogeneous Indexing Problem

In an ongoing project at the University of Illinois, we have investigated methods to support the automatic and/or computer-assisted reconciliation of heterogeneous indexing in the NASA Astrophysics Data System (ADS). ADS provides astronomers worldwide with access to over a million abstracts and full text articles in the fields of astronomy and astrophysics, instrumentation, physics and geophysics (Eichhorn et al. 1998). A mixture of controlled indexing vocabularies has limited ADS searchers' ability to conduct precise subject searches, and our investigations have focused on two sources of evidence for resolving the inconsistencies: lexical resemblance between descriptors and consistent assignment of descriptors from different vocabularies to the same documents (Dubin 1998; Lee 1998; Lee, Dubin, & Kurtz 1999).

### 1.1. Vocabulary Reconciliation

Indexing a document is a highly demanding task, and it is hard to elicit explicitly the set of formal rules for indexing. Accordingly, it is not feasible to make an automatic indexing system which is rule-based. Vocabulary merging is also difficult and requires the full understanding of the indexing schemes of all participating indexing vocabularies and complex relationships among terms employed. Hence, it is very difficult to extract the explicit rules for vocabulary merging. A symbolic or rule-based approach is very limiting for this type of problem. Alternatively, a connected approach can be applied to the problem of indexing and vocabulary merging since it tries to find input/output pattern relationships without the need to find the rules for indexing and term mapping.

In an earlier paper, we have described a spreading activation model, similar to those employed for modeling human associative memory (Lee 1998). In the model the network is described as a feed-forward three-layer network which is constructed out of evidence of document co-assignment pattern. A subset of documents, each indexed by two different indexing vocabularies, is selected from the database and then used as the evidence for merging different vocabularies. By identifying term descriptors from both vocabularies for each document, it is possible to relate terms from one vocabulary (source) to terms from the other (target). A node in the input layer corresponds to each term descriptor in the source vocabulary from which we want to find the mapping term(s) in the target vocabulary. A node in the middle layer corresponds to each document in the collection. A node in the output layer corresponds to each term descriptor in the target vocabulary.

The link between a term and a document has the connection strength between 0 and 1 if the document is indexed by the term. Initially, the weight assigned to the link is determined by the number of nodes connected from a sending node. The link gets the weight of $\frac{1}{N_i}$, where $N_i$ is the number of links from the node in one layer to the nodes in the subsequent layer. The assumption underlying this weighting scheme is that the activation of a sending node will be spread out across all the connected nodes. Specifically, when a term descriptor is employed to index many documents in the collection, the link weights between the input term node and the document nodes become small. This weighting scheme for the input and middle layers corresponds to the measure of Inverse Document Frequency (IDF). However, the link between a document node and an output term node will not be affected by the number of documents indexed by the terms from the output layer, but by the number of term descriptors applied for each document: so-called Inverse Term Frequency (ITF). Depending on the direction of spreading activation, link weights are determined by two different measures, IDF for input-to-middle links and ITF for middle-to-output links.

The network produces a set of activated output nodes as a result of spreading activation process. The activation of the nodes in the input term layer is spread through the network to the connected document nodes and from there to the output term nodes. The output of the network is a ranked list of output terms with their activation levels indicating the degree of relatedness to the input term(s).

We have used the network for context-dependent mapping of descriptor sets rather than basing a static term-to-term mapping on fixed associations. That is to say, all the terms assigned together to a particular document are mapped as a group. A document is assumed to be represented not by separate terms but by a set of terms which comprises a specific context. We have found that the network performs better for the mapping from specific to general term descriptors (STI → ApJ) than the mapping from general to specific (ApJ → STI). In addition, the STI → ApJ network showed more robustness against the removal of any single document node under testing than the ApJ → STI network (Lee & Dubin 1999).

## 2. Back-Propagation Learning for Vocabulary Merging

In the case of vocabulary mapping problem, supervised learning techniques can be used to find term relationships between indexing vocabularies if the training data of input-output patterns exist. Since we are testing the set of documents co-indexed by two different indexing systems, target output is known *a priori* for each input as a correct answer. The training exemplars consist of the set of co-indexed documents. For each document, source index terms are identified as an input pattern and target index terms as an output pattern. What is to be learned by the network are the input-output pattern mappings, that are source-target subject index term relationships.

Supervised learning occurs in two steps, via the back-propagation model (Rumelhart, Hinton, & Williams 1986). First, a set of terms is given to the corresponding nodes in the input term layer. The activations of the input term nodes propagate forward to the middle layer, thereby activating all the linked document nodes. Then the activation spreads to the output term layer along the connections. A set of term nodes activated in the output layer is identified: the product of feed-forward activation in the network is the list of activated nodes with their activation level.

The activation received by a document node is calculated by summing up the activation coming from all the connected input term nodes. The errors in the output layer are computed, and propagated backwards from there to the middle to the input layer. The actual output of a node is compared to the desired output of that node. The discrepancy between what is computed from the network and what is desired is the error measure for each node. The error ($\delta_k$) for an output term node $u_k$ is determined by the difference between the actual output ($a_k$) from the network and the correct output ($t_k$) for the given input $u_i$.

$$\delta_k = (t_k - a_k)a_k(1 - a_k)$$

For each document under testing, a set of subject descriptors is identified for both source and target indexing vocabularies. The input to the network is the set of subject descriptors from the source vocabulary for the given document, and the desired output is the set of subject descriptors (target terms) from the target vocabulary for the same document. The connection weight between the output term node $u_k$ and a linked document node $u_j$ is then updated to reduce this error. The errors are propagated backwards from the nodes in the output term layer to those in the document layer. The input-to-middle link weight is updated in the same way. The weight change is proportional to the estimated error in a document node and the activation of its incoming input term node.

## 3. Experiment

We have integrated the learning algorithm into the spreading activation network model to test whether the term relationships among heterogeneous indexing vocabularies can be learned based on the co-occurrence data of subject indexing. The network is trained on a training data set and tested against the new test data set. The learning performance is measured by precision and recall of the network prediction from STI terms to *ApJ* terms.

### 3.1. Data Set

The data set includes two different subject indexing vocabularies employed in ADS: *Astrophysical Journal* subject headings (*ApJ*) and index terms applied by NASA's Scientific and Technical Information group (STI). These two vocabularies differ in many respects. *ApJ* headings are assigned to the documents by authors while STI terms are by professional indexers. *ApJ* headings include pre-coordinated descriptors while STI terms come from the *NASA Thesaurus.* Another distinction between these two lies in their scope and degree of specificity in indexing: the NASA Thesaurus has many levels of broader and narrower terms, and professional indexers are trained to apply more specific terms wherever possible.

The test collection is composed of two sets of documents. one set has 39,366 documents, and indexed by *ApJ* headings. Another set has 22,139 documents, and indexed by STI terms. Out of these two sets of documents, the set of 14,956 documents was found to be co-indexed by both indexing vocabularies. Basic statistics of this test collection are presented in Table 1.

### 3.2. Network Representation

A three-layer feed-forward network is constructed out of a set of 14,956 documents with two different indexing systems applied. The input layer corresponds to the source vocabulary of STI (NASA thesaurus terms), the output layer to the target vocabulary of *ApJ* (subject keyword list for *Astrophysical Journal*), and the middle (or hidden) layer to the document set. Each term in the vocabulary is represented as a node in the input or output layer, and each document is represented as a node in the middle layer. As a result, the network will consist of 4,120 nodes in the input layer, 2,305 nodes in the output layer, and 14,956 nodes in the middle layer if the whole data set is used. The actual network constructed in this study includes 13,460 document nodes out of the training data set. The set of nodes in the input and output layer is determined by the randomly selected document nodes.

| | **STI** | **ApJ** |
|---|---:|---:|
| Number of documents | 14,956 | 14,956 |
| Number of descriptors | 4,120 | 2,305 |
| Average number of postings | 34 | 23 |
| Average number of descriptors per doc. | 9.6 | 3.5 |

Table 1. Statistics of the test collection

### 3.3. Procedure

The back-propagation network is implemented in four steps in the study. First, the indexing data are randomly divided into the training and test data sets. The training set is composed of 90% of the data and the test set includes the remaining 10%. Secondly, the training data set is used to construct the nodes in each layer in the network. Then, the weight matrices are initialized before training. Finally, the network is trained repeatedly with the training data set until it satisfies the stopping rule. For training, documents from the training data set are ordered randomly and used as training examples.

### 3.4. Formal Evaluation

During the training process, the network is evaluated against the test data set after each epoch of training. For term selection, we applied the so-called "Mexican-Hat" function which has been successfully used for detecting edges in vision processing (Charniak & McDermott 1987). The output of the spreading activation in the network is a large number of activated terms, some with very low activation levels. A cutoff point is determined by the distribution of term activation levels for each input. The ranked array of activation levels is convolved with the "Mexican Hat" curve, and the cutoff point is found where the slope of the activation value distribution reaches its steepest decline. Only terms above this cutoff point are selected as mapping terms for a given input and used as network output for evaluation.

We evaluate both the ranking and the sensitivity of the cutoff with conventional recall and precision measures. A perfect mapping is defined as one that exactly predicts assignments by human indexers. Our current efforts focus on improving the network's performance as measured by both precision and recall. The precision ratio is defined by the percentage of those above the cutoff that are correct while the recall ratio by the percentage of correct that are above the cutoff. A learning curve is obtained for both precision and recall measures in a series of training. Each iteration consists of 13,460 training documents, and the learned network is evaluated against 1,496 testing documents. The presentation order of documents changes with each iteration.

## 4. Quantitative Results

Before the training, the mapping performance of the network without learning is measured as 50% of average precision and 35% of average recall. After one iteration of training, the mapping performance increases to 58% and 40% respectively. Three iterations of training make the learning of the network stabilize at around 60% average precision and 42% average recall. These results indicate that supervised learning can improve the performance of term mappings. The network is constructed from and trained by the training data set which is based on the document co-assignment pattern. Each learning exemplar consists of two indexing indices: a document, STI terms, and *ApJ* terms. The performance of term mapping from STI to *ApJ* is largely dependent on the evidence of term relationships each learning exemplar brings to bear. The improvement in the mapping performance implies the network can learn term relationships out of document co-assignment pattern.

## 5. Qualitative Evaluation

Recall and precision studies give some sense of the network's success in predicting author and editor-assigned subject headings. But the measures tell us very little about what kinds of terms the network is likely to guess correctly or how its failures are likely to affect subject access in ADS if the model were put to practical use. We have therefore reviewed the results of mapping selected documents that are not part of our co-indexed collection. Although the target subject headings for these documents aren't part of the ADS bibliographic record, they do appear in the paper and scanned versions of the articles. With a few extra manual steps we can test the network's success in predicting those terms.

Each of the examples below contrasts the terms assigned to a document by a NASA indexer with the subject headings appearing in the journal and those assigned by the network. Thus the first column represents input to the network, the second our standard of success, and the third column is the output from the system. Some effort has been made to put corresponding or similar terms on the same line, but that is merely illustrative: actual output from the network is ranked based on activation level.

| **NASA** | **ApJ** | **Network** |
|---|---|---|
| galactic clusters | galaxies: clustering | galaxies: clustering |
| quasars | quasars | quasars |
| charge coupled devices | | |
| color-magnitude diagram | | |
| emission spectra | | |
| red shift | | |

This first example illustrates our definition of perfect success: starting with the six NASA terms, the network successfully predicts precisely the two terms that appear on the journal article. It is only such exact matches that contribute to increasing the recall and precision measures.

| **NASA** | **ApJ** | **Network** |
|---|---|---|
| accretion disks | accretion, accretion disks | |
| pre-main sequence stars | stars: pre-main-sequence | stars: pre-main-sequence |
| stellar mass ejection | stars: mass loss | stars: mass loss |
| stellar winds | | |
| absorption spectra | | |
| carbon monoxide | | |
| computational astrophysics | | |
| line spectra | | |

This second example is far more representative of the results we've seen. Precision is high: both terms predicted by the network are correct. None of the terms assigned by the NASA indexer have evoked a term that didn't appear on the article. But the missing heading could easily have been mapped on the basis of lexical resemblance.

| **NASA** | **ApJ** | **Network** |
|---|---|---|
| active galactic nuclei | galaxies: active | |
| | galaxies: nuclei | galaxies: nuclei |
| black holes (astronomy) | black hole physics | |
| quasars | quasars: general | quasars |
| interacting galaxies | galaxies: interactions | galaxies: interactions |
| cosmology | | cosmology |
| radio jets (astronomy) | radio continuum: galaxies | galaxies: jets |
| radio astronomy | | |
| luminosity | | |

The third example demonstrates that headings deemed incorrect in our recall and precision studies may be lexically or semantically very close to a correct heading. In this example, we would count "galaxies: nuclei" as a match, but "quasars" as a miss, since the actual heading assigned was "quasars: general."

| **NASA** | **ApJ** | **Network** |
|---|---|---|
| accretion disks | accretion, accretion disks | |
| magnetohydrodynamic stability | MHD | hydromagnetics |
| | instabilities | |
| x ray binaries | x-rays: bursts | x-rays: binaries |
| pulsar magnetospheres | | |
| astronomical models | | |
| computational astrophysics | | |
| stellar evolution | | |

Finally, the last example shows how what we've deemed an incorrect assignment can be understood from the input to the network. The system failed to assign either "x-rays: bursts" or "MHD" to this document. But one can easily recognize which input terms have yielded "hydromagnetics" and "x-rays: binaries."

## 6. Discussion

Predicting the exact headings assigned by authors and editors is not a realistic standard for success, considering the very different criteria that professional indexers bring to the task of subject analysis. The NASA STI indexers not only assigned more descriptors than authors and editors, but they also chose more specific terms. But, unrealistic though the standard may be, it means that our recall and precision measures are fairly conservative.

Our goal is not only to map index terms to another descriptor language: we wish to do so in a way that is as consistent as possible with the way the same headings are applied manually to other documents in the database. This is where the current strategy scores over static term-to-term mapping, either via the network model or using lexical resemblance evidence. A straight term-to-term mapping would produce far more subject headings than is typically applied to an article.

Nevertheless, we are not satisfied with the low levels of recall, and the fact that unmapped terms often bear a strong lexical resemblance to STI terms in the input. Due to the way our system is designed, strong evidence for one or two headings will trigger the threshold and close the door on other potential matches. Clearly we need a more liberal matching standard, even at the cost of some precision. Our next step will be to look for ways to either adjust the cutoff threshold or integrate other sources of evidence (e.g., lexical similarity).

There are also several issues of interest related to the learning of term relationships. First, learning occurs incrementally by updating the connection weights in order to adapt to the learning environment. The learning curve showing the rate of learning can provide useful information about the learning process. The comparison of human learning and machine learning may give some insights to the validity of cognitive/computational models of human information processing.

Second, the learning should be robust enough to generalize to the new data set which was not given to the network as learning exemplars. The success of a learning system is greatly dependent on this generalization of learning. If the network is well trained for the training instances, the knowledge is assumed to be represented in a distributed way across the network. In that case, the mappings of input to output are largely determined by patterns across the entire network and thus less prone to error. The network should be more robust if the mapping is drawing on evidence over many nodes, not just a few.

Third, the learning performance is very data-dependent, greatly influenced by the representativeness of the training instances. What is learned by the system are the regularities emerging from patterns in the data. Therefore, it is very critical to the success of the system to provide a representative data set which reveals the systematic input-to-output relationships.

## References

Charniak, E. & McDermott, D. V. 1987, Introduction to Artificial Intelligence (Reading, MA: Addison-Wesley)

Dubin, D. S. 1998, in A.S.P. Conf. Ser. Vol. 153, Library and Information Services in Astronomy III, ed. U. Grothkopf, H. Andernach, S. Stevens-Rayburn, & M. Gomez (San Francisco: ASP), 77

Eichhorn, G., Accomazzi, A., Grant, C. S., Kurtz, M. J., & Murray, S. S. 1998, in A.S.P. Conf. Ser. Vol. 145, Astronomical Data Analysis Software and Systems VIII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 378

Lee, J. 1998, A Spreading Activation Model for Vocabulary Merging, Tech. Rep. UIUCLIS--1999/4+IRG, University of Illinois, Graduate School of LIS, Urbana-Champaign, IL

Lee, J. & Dubin, D. 1999, in Proc. 1999 ACM SIGIR, ed. M. Hearst, F. Gey, & R. Tong, Association for Computing Machinery (New York: ACM), 198

Lee, J., Dubin, D., & Kurtz, M. J. 1999, in A.S.P. Conf. Ser. Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. D. Mehringer, R. Plante, & D. Roberts (San Francisco: ASP), 287

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533