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

Co-occurrence Evidence for Subject Vocabulary Reconciliation in ADS Databases

Jonghoon Lee, David S. Dubin

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Michael J. Kurtz

Harvard-Smithsonian Center for Astrophysics

Abstract. The NASA Astrophysics Data System gives astronomers all over the world access to over a million abstracts in the areas of astronomy and astrophysics, instrumentation, physics and geophysics. A mixture of indexing vocabularies has limited ADS searchers' ability to conduct precise subject searches, but prospects for a single consistent vocabulary of descriptors are promising. We report results of an ongoing project to reconcile the heterogeneous indexing (keywords, index terms, and subject headings) applied to the existing ADS records. Evidence for term mappings are sought in the consistent assignment of different descriptors to the same documents. We describe the apprehension of this evidence through the use a spreading activation network. Design and deployment of an experimental interface to assess these mappings is now under way.

1. Introduction

This paper reports results in an ongoing project to reconcile heterogeneous indexing vocabularies in the NASA Astrophysics Data System (ADS; Eichhorn et al. 1998). Over 15,000 astronomers use ADS each month to retrieve abstracts and full text articles in the areas of astronomy and astrophysics, instrumentation, physics and geophysics. The ADS databases support a number of different search methods, including access by title, author, and astronomical object name. Users can search on a combined index of the abstract and keyword fields, but cannot (through the main web interface) limit their searches to a single controlled vocabulary of subject descriptors. The merging of the abstract and keyword indexes was a deliberate decision by the ADS administrators, because documents in ADS are indexed with several different descriptor vocabularies. Some documents have been indexed by professional indexers using terms from the NASA Thesaurus. Others have had keywords suggested by authors and approved by journal editors.

Controlled indexing vocabularies overcome the variation in authors' natural language by providing standardized labels for concepts. They also permit searchers to limit their queries to the most important ideas or topics in a document. The mixture of different descriptor vocabularies in ADS defeats

the standardization goal, and the merging of the abstract and keyword indexes limits the search precision function of the subject indexing. Descriptors representing identical concepts (or closest counterpart) can stand in several different relationships to each other. For example, counterpart terms may be synonyms (e.g., **Andromeda Galaxy** and **Galaxies: Individual Messier Number: M31**). A descriptor in one vocabulary may be a pre-coordinated combination of terms in the other vocabulary (e.g., **Microwave Background Radiation** vs. **Background Radiation** and **Microwaves**).

An ongoing project at the University of Illinois investigates sources of evidence to support the automatic and/or computer-assisted reconciliation of the heterogeneous indexing in ADS. Two sources of evidence have been investigated: lexical resemblance between descriptors and consistent assignment of descriptors from different vocabularies to the same documents. Use of lexical resemblance evidence is discussed in an earlier paper (Dubin 1998).

2. A Spreading Activation Model for Co-Assignment Evidence

Our source of evidence is a subset of the ADS database, in which each document has been indexed with two or more different vocabularies. The consistent assignment of two or more terms from different vocabularies to the same documents suggests some kind of semantic relationship or connection among the terms. Not every connection represents a context-free pairing of synonymous terms: sometimes a pair of coordinate index terms assigned from one vocabulary imply the assignment of a pre-coordinated descriptor from another vocabulary. We therefore developed a model that accommodates such context-sensitivity.

We developed a spreading activation model, similar to those employed for modeling human associative memory (Collins & Loftus 1975). In our model, the network is composed of three layers: an input term layer, document layer, and output term layer. The activation of terms in the input layer is spread through the network to the connected documents and from there to the output terms. As a result, this model produces a list of terms with their activation levels representing the degree of relatedness to the input term(s) (Lee 1998).

Calculation of the activation level employs a simple weighting rule. The weight assigned to the link between a document and a term is determined by the number of connections and the direction of activation. We use the conservation of activation principle which guarantees that the sum of input activation equals to that of output activation. The activation received by a document from input terms is computed as follows:

$$D_j = \sum_{i=1}^{l} S_i w_{ij} \quad T_k = \sum_{j=1}^{m} D_j w_{jk}$$

In these formulas, S_i is the activation of an input term i; D_j is the activation of a document j; T_k is the activation of an output term k; w_{ij} is a weight between an input term i and a document j; and w_{jk} is a weight between a document j and an output term k.

3. Evaluation

We applied the spreading activation model to investigate connections between two different subject vocabularies employed in ADS: Astrophysical Journal keywords (ApJ) and terms assigned to documents from the NASA Thesaurus (STI)[1]. These two vocabularies, one used by professional indexers and the other by authors, differ in degree of specificity and level of pre-coordination (the ApJ terms include pre-coordinated descriptors).

Experimental materials included two sets of documents, one indexed by 10,200 ApJ terms, and the other by 3,335 STI terms. The ApJ set contained 39,366 documents, and the STI set 22,139 documents. Among these, 14,956 documents were identified as co-indexed. The merged network representation included 4,120 STI term nodes in one term layer, 14,956 document nodes in the middle layer, and 2,305 ApJ term nodes in the other term layer. Two spreading activation models were constructed according to the direction of activation (from source term to target term): ApJ $\rightarrow$ STI and STI $\rightarrow$ ApJ.

4. Results

Each term from the source vocabulary was used as an input to the network. The output of the spreading activation process was the list of activated terms from the target vocabulary along with their activation level. For example, when the STI term (**Andromeda Galaxy**) is given as an input node, 171 ApJ terms were identified as activated terms in the output layer. One of the problems here is that the output usually includes a large number of terms most of which seem to be barely related (very low activation level like 0.01). Therefore, a cutoff criterion is needed to differentiate more related terms from less related ones.

We adapted the so-called "Mexican-Hat" function for our cutoff criterion. The Mexican-Hat is the second derivative of the Gaussian curve, and has been successfully used for vision processing, especially edge detection (Charniak & McDermott 1987). The set of output activation levels was convolved with the second derivative of the Gaussian in order to find the point where the slope of the activation value distribution drops most dramatically. Only terms above this cutoff point were selected as mapping terms. The average number of terms above the cutoff was 1.9 out of 135 activated terms for STI $\rightarrow$ ApJ, and 1.6 out of 354 activated terms for ApJ $\rightarrow$ STI.

Term mappings identified by the spreading activation model include a variety of term-to-term relationships. They ranges from simple spelling variants to complicated semantic factoring. We were encouraged to observe that the model identified many of the same connections uncovered in our lexical resemblance study, although the current model employs no lexical evidence. Examples of relationships apprehended with the network included:

1. Exact match **(Astrometry $\rightarrow$ Astrometry)**
2. Spelling variants **(Galactic Clusters $\rightarrow$ Galaxies: Clustering)**
3. Ordering **(Clusters: Globular $\rightarrow$ Globular Clusters)**

---

[1]Indexers at NASA's Scientific and Technical Information (STI) Group assigned the descriptors.

4. Pre-coordination **(Cosmic Rays: Abundances → Abundance, Cosmic Rays)**
5. Term omission **(Catalogs → Astronomical Catalogs)**
6. Class relation **(Galaxies: Individual Messier Number M33/M81 → Spiral Galaxies)**

5. Discussions and Future Studies

These preliminary results suggest our model can successfully reveal the complex nature of term relationships between two astronomical subject vocabularies. For example, we note the asymmetric characteristics of term mapping. The difference between STI → ApJ mapping and ApJ → STI mapping is observed for many pair of terms. This bi-directional relationship is well revealed since the spreading activation model is sensitive to the direction of term mapping. Current experiments include a generalization of the term-term mapping approach described in this paper: a many-to-many mapping (in which all the terms for a document are activated simultaneously) affords apprehension of the context-sensitive relationships described earlier.

We are evaluating the usefulness of the spreading activation output as evidence for vocabulary merging. We plan a user study in which expert astronomers evaluate the connections suggested by the model. We will follow up with a user evaluation test, assessing the impact of mergings on search precision. We're also evaluating the scalability of the model with larger databases.

We are developing a visualization tool in order to help users understand the complex nature of term relationship. It will show term mapping structure among participating vocabularies using a graphical interface. A directed graph method will be used to represent the directional information of term relationship.

References

Charniak, E. & McDermott, D. V. 1987 Introduction to Artificial Intelligence, (Reading: Addison-Wesley), 101

Collins, A. M. & Loftus, E. F. 1975, Psych. Rev., 82, 407

Dubin, D. S. 1998, in ASP 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 ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 378

Lee, J. 1998, A Theory of Spreading Activation for Vocabulary Merging, (Grad. Sch. of Library and Information Sci., Univ. of Illinois, unpublished report), (Champaign: Univ. of Illinois)