Concepts and Lattices: An Investigation

Scott W. Cunningham and Jan H. Kwakkel

Faculty of Technology, Policy and Management, Delft University of Technology
Email: S.Cunningham@tudelft.nl

Abstract
The dominant paradigm of tech mining (Porter and Cunningham 2004) is based upon information retrieval. The paradigm is tabular, is primarily based on geometric operations, and is designed to find exemplars in concepts and in text. The second is based on artificial intelligence -- the paradigm is based on lattices, uses logical operations, and is designed for delineating sets of terms or concepts. The second paradigm is valuable, if under-exploited for tech mining applications. For ease of reference, but at some cost of oversimplifying the account of theories and methods, we refer below to the first paradigm as information retrieval (IR) and the second paradigm as artificial intelligence (AI). The purpose of this paper is to identify the paradigm, provide a tutorial, endorse the merits, and to demonstrate how it might be exploited to produce improved algorithms and software for tech mining applications. Our approach is to examine each of the following in turn: the representation of the data, the operations performed on the data, and the resulting outputs for tech mining or other purposes.

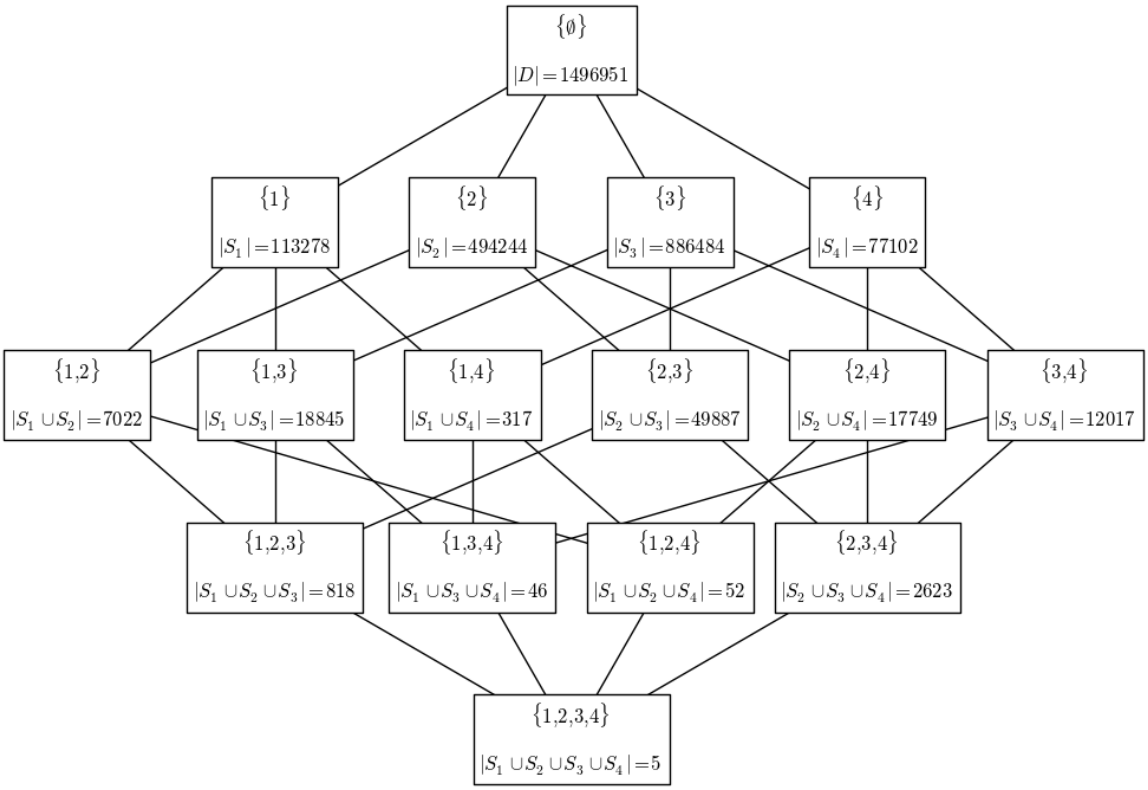

Figure 1. A Simple Galois Lattice on Nanotechnology Publication

The Representation of the Data
Tech mining is largely based on modern information retrieval (Salton and McGill 1986). The technique is based upon "vector spaces" of terms and documents. These are essentially matrices where the rows indicate documents, and the columns indicate the presence, extent or absence of terms. The space is used to index, position and retrieve relevant documents. The technique has been further refined to examine robust subspaces of the data (Deerwester,

Dumais et al. 1990), and to identify additional, authoritative sources of information (Kleinberg 1999). The AI approach enriches the data with conditional, higher-order relationships between terms and documents. The approach, based on earlier work in logic and sets, is known as a Galois lattice (Ganter, Stumme et al. 2005). The higher-order information represented in a lattice is often incorporated into Boolean queries, only to be discarded when adopting the vector space approach. A simple example of a Galois lattice, using nanotechnology data, is shown in figure 1.

Measures on the Data

Both the IR and AI paradigms require the use of measures to structure and order the data. A measure is simply a generalized means of quantifying observed relationships in the data. Four kinds of measure theories are commonly used for this purpose. These measure theories, further described below, are geometric, probabilistic, logical, and information measures. The IR paradigm has largely been concerned with the geometric and probabilistic measures and operations, while the AI paradigm entails the use of logical and information theoretic measures.

Geometric measure theories involve creating similarity or distance measures between terms or documents. The spaces created by embedded objects are then analyzed. Various coordinate systems for traversing these spaces exist, including Euclidean and spherical coordinates systems. *Probabilistic measure theories* entail measuring the state of uncertainty concerning the user, the query, and the corpus. The objective is then to formulate findings which are robust to the presence of noise or uncertainty. *Logical measure theories* are built upon order and set theory. The operators are shared with Boolean algebra and involve combining sets (through the join operator) or refining overlapping sets (through the meet operator). The objective of such measures is to build an ontology – a formal representation of knowledge which can be used to model a domain of knowledge and to support reasoning about the domain and its constituent parts. *Information theory* entails quantifying limits on the reliable communication, transmission and storage of data. Information theory uses the familiar measurement of a bit, a unit for the quantification of information content. Operations on the data entail finding taxonomies which communicate a lot of information while requiring a minimum amount of overhead for communication.

Analytic Outputs

The principal analytic output from the IR paradigm is an exemplar. Exemplars are groups of related terms or documents, presented to the user as summary, representative content for the whole set. Exemplars may be variously structured – lists, clusters, factors and networks are all common means of presentation. In contrast the AI paradigm is often more concerned with rules rather than exemplars. Rules help explain the differences between terms and concepts, defining how and why certain terms and documents are included in one set and not another. Rules are useful for accounting and retrieval. Even extensive Boolean queries can be presented to others as documentation, and can be independently verified against large databases of science, technology and innovation content.

This paper examined two paradigms for tech mining – the information retrieval and artificial intelligence paradigms. The two paradigms are contrasted in order to better determine the comparative merits and demerits of the two approaches (table 1). The contrast presented is somewhat artificial as there is a body of research which spans and combines the two paradigms. However the framework of table 1 provides design guidance for which elements from each approach should be adopted for new tech mining research and application.

Table 1. Two Paradigms for Tech Mining

	Information Retrieval	Artificial Intelligence
Representation	Vectors	Lattices
Measures	Geometric, Probabilistic	Logical, Information
Operations	Similarity/Distance	Composition/Decomposition
Outputs	Exemplars	Rules

We believe that lattices are a superior representation of terms and documents. While they can be derived from bit matrices, the lattice representation aids comprehension of complex relationships in the data and speeds computation. The most interesting measures on the lattice, we argue, are either probabilistic or information-based. The chief challenge in tech mining is information incompleteness – new sources of information are constantly emerging, and we must come up with timely and robust findings regardless of the perpetual lag between data and discovery. Information and probability assist in this challenge. Tech mining applications must rapidly position a field of research in a larger context, while permitting a close focus on individual research findings where appropriate. Operations of composition and decomposition are therefore far more important for decision-makers than querying for similarity. Finally, rules are more significant outputs than exemplars. Widespread availability of partial, overlapping sources of data mean that the field should pay much more attention to clear queries which can be widely applied across diverse data sets.

References

Deerwester, S., S. T. Dumais, et al. (1990). "Indexing by Latent Semantic Analysis." Journal of the American Society for Information Science **41**: 391-407.

Ganter, B., G. Stumme, et al., Eds. (2005). Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence, no. 3626. Berlin, Springer-Verlag.

Kleinberg, J. M. (1999). "Authoritative Sources in a Hyperlinked Environment." Journal of the ACM **46**(5): 604-632.

Porter, A. L. and S. W. Cunningham (2004). Tech Mining: Exploiting New Technologies for Competitive Advantage. Hoboken, N.J., Wiley.

Salton, G. and M. J. McGill (1986). Introduction to Modern Information Retrieval. New York, NY, McGraw-Hill, Inc.