AMETHYST: A System for Mining and Exploring Topical Hierarchies of Heterogeneous Data

Marina Danilevsky, Chi Wang, Fangbo Tao, Son Nguyen, Gong Chen, Nihit Desai, Lidan Wang, Jiawei Han
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
{danilev1,chiwang1,ftao2,nguyenb1,gchen10,nhdesai2,lidan,hanj}@illinois.edu

ABSTRACT

In this demo we present AMETHYST, a system for exploring and analyzing a topical hierarchy constructed from a heterogeneous information network (HIN). HINs, composed of multiple types of entities and links are very common in the real world. Many have a text component, and thus can benefit from a high quality hierarchical organization of the topics in the network dataset. By organizing the topics into a hierarchy, AMETHYST helps understand search results in the context of an ontology, and explain entity relatedness at different granularities. The automatically constructed topical hierarchy reflects a domain-specific ontology, interacts with multiple types of linked entities, and can be tailored for both free text and OLAP queries.

Categories and Subject Descriptors

I.7 [**Computing Methodologies**]: Document and Text Processing; H.2.8 [**Database Applications**]: Data Mining

Keywords

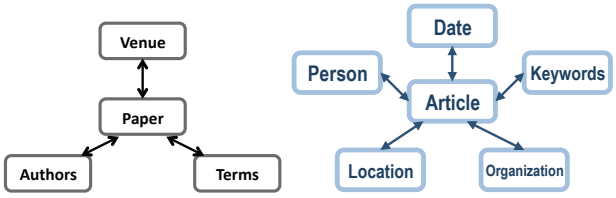
Topic Modeling, Network Analysis, Heterogeneous Network, Entity Mining

1. INTRODUCTION

Heterogeneous Information Networks (HINs), composed of multiple types of entities and links are very common in the real world, and have become increasingly more important to analyze and understand [3]. Many of them have a text component, inviting many interesting recent applications of topic mining techniques [2, 5, 4, 1]. One such important application is constructing a high quality hierarchical organization of the concepts in a dataset at different levels of granularity [6], which heralds a more nuanced understanding of topics and subtopics in the dataset as a whole.

In order to apply the automatically constructed hierarchy to browsing, search, and summarization tasks on the HIN dataset, we present AMETHYST (Analyzing, Mining, and Exploring a Topical HierarchY SysTem). Compared with other systems that support topical analysis and multi-dimensional mining tasks, such as Arnetminer [5], Topic Cube [7] and Microsoft Academic Search[1], AMETHYST has the following key features:

- The topics are organized in a hierarchy where more general topics are parent of more specific topics. The hierarchy helps understand search results in the context of an ontology, and explain entity relatedness at different granularities.
- The topical hierarchy is automatically constructed from the data and therefore reflects the domain-specific ontology and accommodates user preference.
- The topical hierarchy interacts with multiple types of linked entities, and can be tailored for both free text and OLAP queries.



(a) Bibliographic network. (b) News articles network.

Figure 1: Schemas of two HINs with text components.

There are many examples of HIN's with text components, such as a bibliographic network, a collection of user reviews. For example, Figure 1a shows the schema of the DBLP publication network consisting of a paper and its related entities: authors, publication venue, publication year, and some key terms. Figure 1b shows the schema of another HIN, constructed from news articles, consisting of an article and its related entities: people, locations, and organizations mentioned in the article, the publication date, and keywords.

Throughout this paper we will use a publication network dataset consisting of papers published in DBLP to illustrate the functionality of our system. However, AMETHYST can work with any HIN which can be used to successfully build a topical hierarchy. Furthermore, if multiple topical hierarchies are generated for the same dataset, exploring them with our system can provide a deeper understanding of the implications of the different ways of organizing information.

2. SYSTEM FRAMEWORK

Figure 2 illustrates the offline (red) and online (blue) architecture components of AMETHYST:

[1] academic.research.microsoft.com

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

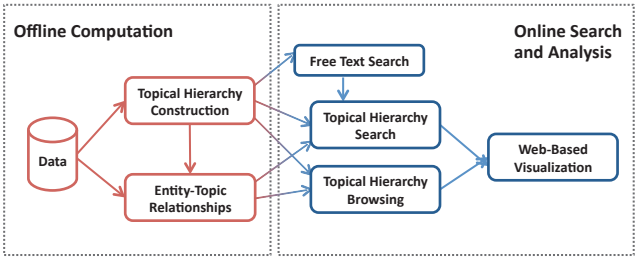


Figure 2: System framework of AMETHYST.

• **Topical Hierarchy Construction:** Construct a topical hierarchy given an HIN dataset, representing the topics present in the dataset, with each topic represented by a ranked list of mixed-length topical phrases.

• **Entity-Topic Relationships:** Precompute the relationship between every entity present in the HIN and every topic in the hierarchy, to facilitate browsing and searching.

• **Topical Hierarchy Browsing:** Users can choose an available dataset, and browse the topical hierarchy via a web interface. For each entity type, the most relevant entities are listed for each topic in the hierarchy.

• **Topical Hierarchy Search:** Users can perform free-text queries, and/or explore just the slice of the hierarchy that represents the topics of an OLAP query, with related phrases and entities re-ranked accordingly for each topic.

3. OFFLINE COMPUTATION

For a given dataset, we first precompute the topical hierarchy (or possibly several topical hierarchies), as well as the entity-topic relationships between the entities in the dataset and the topics in the hierarchy.

3.1 Topical Hierarchy Construction

The input to the construction of a topical hierarchy is an HIN with a text attribute, as described in Section 1, and the output is a hierarchical tree of topics. The basic unit of a topical hierarchy is a phrase. The output is a tree of topics, where each topic is represented by a ranked list of mixed-length topical phrases, such that a child topic is a subset of its parent topic. Every non-root topic t in a topical hierarchy is represented by a ranked list of phrases $\{\mathcal{P}^t, r^t(\mathcal{P}^t)\}$, where $\mathcal{P}^t$ is the set of phrases for topic t, and $r^t(P^t)$ is the ranking score for the phrases in topic t. For every non-leaf topic t in the tree, its children C^t are its subtopics. For example, in a topical hierarchy based on the DBLP dataset, the topic of query processing and optimization may be described by the phrases {'query processing', 'query optimization',...}, while its parent topic of general problems in databases may be described by {'query processing', 'database systems', 'concurrency control',...}. A phrase can appear in multiple topics, though it will have a different ranking score in each topic (e.g. 'query processing' in the above example).

Topical phrases that would be regarded as high quality by human users are likely to vary in length. Unlike existing phrase extraction and ranking methods which are term-centric, our approach is phrase-centric and is able to naturally compare *mixed length* phrases with each other. Regardless of length, a phrase is ranked highly within a topic if it has good coverage, is discriminative, has high phraseness (it is a true phrase, e.g. 'active learning' in a topic about machine learning, and not simple a combination of frequent unigrams, e.g. 'learning classification'), and is complete (e.g. 'vector machines' is incomplete, since it is nearly always a subset of the longer phrase 'support vector machines').

3.1.1 Topical Frequency

The process of constructing the topical hierarchy results in every phrase in the hierarchy having a *topical frequency* value for every topic in the hierarchy.

DEFINITION 1 (TOPICAL FREQUENCY). *The topical frequency $f_t(P)$ of a phrase is the count of the number of times the phrase is attributed to topic t. For the root node o, $f_o(P) = f(P)$. For each topic node in the hierarchy, with subtopics C^t, $f_t(P) = \sum_{z \in C^t} f_z(P)$, i.e., the topical frequency is equal to the sum of the sub-topical frequencies.*

Table 1 illustrates an example of estimated topical frequency of phrases for a computer science topic that has 4 subtopics. For instance, the phrase 'support vector machines' is estimated to belong entirely to the machine learning (ML) topic with high frequency, while 'social networks' is fairly evenly distributed among three of the topics. Each phrase's topical frequency is recursively estimated for subtopics, in order to perform hierarchical topic construction.

Phrase	ML	DB	DM	IR	Total
support vector machines	85	0	0	0	85
query processing	0	212	27	12	251
world wide web	0	7	1	26	34
social networks	39	1	31	33	104

Table 1: Example of estimating phrase topical frequencies

AMETHYST also uses the phrase topical frequency values to define the relationships between entities in the original HIN and topics in the hierarchy.

3.2 Entity-Topic Relationships

In order to browse the heterogeneous entities associated with each topic node, as well as run user queries, we calculate the entity-topic relationships offline. We use the link data from the original network, as well as the topical frequency of every phrase present in the hierarchy. We also keep a persistent index of the document-phrase relationships for the dataset. We estimate the topical score of each document based on the topical frequencies of the phrases contained in the document. A document containing phrases ranked highly in topic t will therefore have a high topical score in that topic. We then estimate the topical score of each entity based on the topical scores of the documents which link to it in the HIN. This offline computation of the entity-topic relationships is then used in the online tasks of browsing and searching.

4. ONLINE ANALYSIS, MINING, AND EXPLORATION

We first present the web interface of AMETHYST and demonstrate how a user might explore a topical hierarchy constructed from an HIN dataset. We then describe the different search functionalities.

4.1 Exploring a Topical Hierarchy

Figure 3 demonstrates the interface of AMETHYST for browsing. The main menu is in the top right corner, allowing

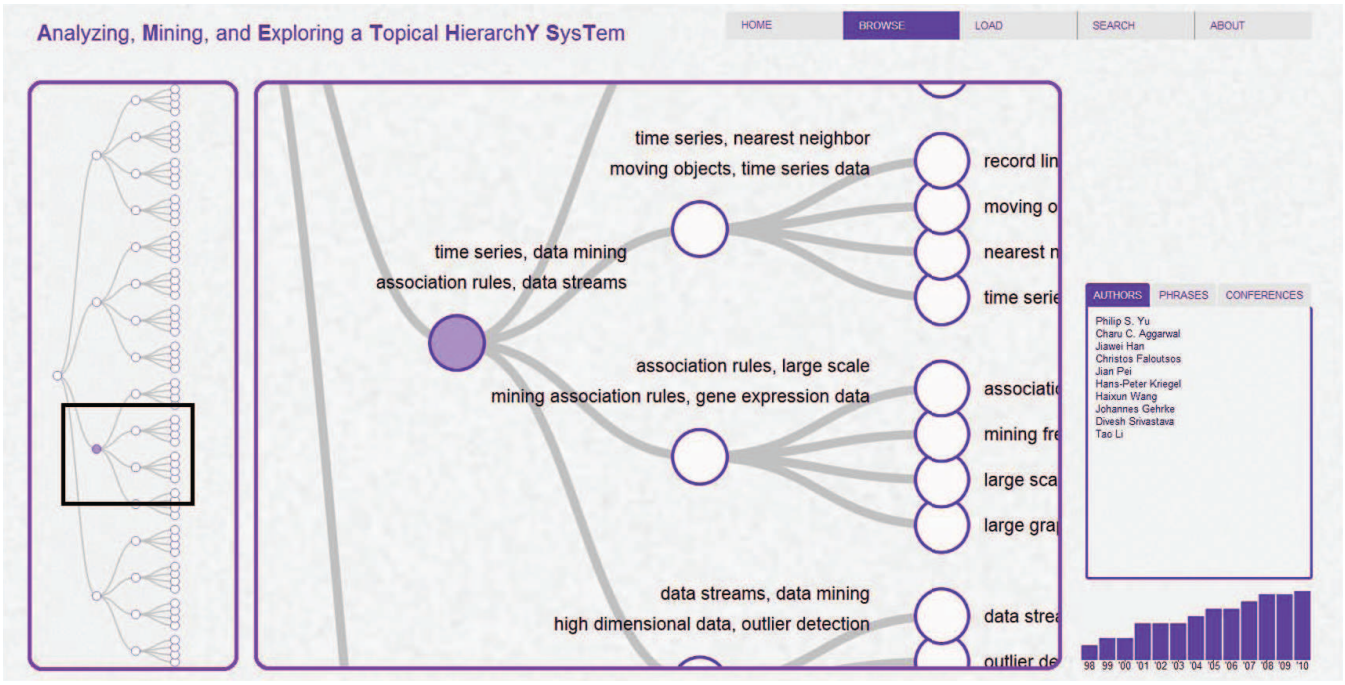


Figure 3: Browsing the DBLP topical hierarchy. The user has selected a node that seems to be generally about data mining, and the Authors tab of the Topic Detail View is displaying a number of well-known data mining authors. The Topic Temporal view most likely reflects the ever-increasing number of publications in the DBLP dataset as whole, including in the area of data mining.

users to choose a dataset to load, browse the resulting topical hierarchy, or search the topical hierarchy (search details are discussed in Section 4.2) To illustrate the interface, we assume that the user has loaded the aforementioned DBLP dataset, and is exploring the topical hierarchy.

• **Structure View:** leftmost panel, visualizing the structure of the entire topical hierarchy. Clicking a node zooms in on that area of the topical hierarchy. In Figure 3, the user has selected the third child of the root node.

• **Zoom View:** central panel, which zooms in on the selected node (highlighted). Each node represents a topic in the hierarchy, and displays its top ranked topical phrases. The user may click other nodes or use the mouse to move around in this view. The rectangle in the Structure View will shift accordingly, as a constant reference for the user's current location in the hierarchy.

• **Topic Detail View:** the rightmost panel, below the main menu, which provides detailed information about the selected topic. The 'Phrases' tab shows a significantly longer list of the topical phrases. Other entity types present in the HIN dataset are also represented by tabs (authors and conferences, in the case of the DBLP dataset). Each tab shows a ranked list of the most relevant entities to the selected topic. In Figure 3, the user has selected a topic that is generally about data mining, and so the Authors tab of the Topic Detail View displays well-known data mining authors.

• **Topic Temporal View:** the histogram below the Topic Detail View, showing the relative temporal distribution of documents which are associated with this topic (if the currently loaded HIN dataset has a temporal attribute). In Figure 3, the temporal distribution is generally growing, most likely because the number of publications in the DBLP dataset increases with each year. In Section 3.2 we described how we calculate the topical score of each document offline. Therefore, the histogram value for a given topic t and a given time period y can be estimated from the topical scores of all documents, $s_t(d)$ where $Time(d) = y$.

4.2 Searching a Topical Hierarchy

The other online components of AMETHYST enable the user to perform different search tasks. By clicking on the Search option in the menu bar, users can choose to perform a topical search using entities or free text as input, using an input area that appears beneath the menu bar.

4.2.1 Topical Search

A user may be interested in focusing on a subset of the topical hierarchy. For instance, which topics discovered from the DBLP dataset reflect the work of a particular author, or a group of authors? Or, which news topics have cropped up in a particular location? Our topical search functionality makes it easy to slice and dice the topical hierarchy via an OLAP implementation.

Given a query consisting of a specific set of criteria (e.g. an author, a conference, or a year range for the DBLP dataset), we do not construct a new topical hierarchy. Rather, we filter the existing hierarchy so that only topics relevant to the query are highlighted, and we re-rank the phrases and entities within each topic. For a given query we can filter the original HIN and select only those central document objects whose entities satisfy the criteria given in the query (e.g., all papers published in a queried year). For each topic, we are now able to use the precomputed relationships described in Section 3.2 to re-rank its phrases and entities based on the strength of their links with the selected documents.

For example, the user may be interested in the topics that the author Jiawei Han has worked in since the year 2000. Figure 4 demonstrates the result of the user issuing this

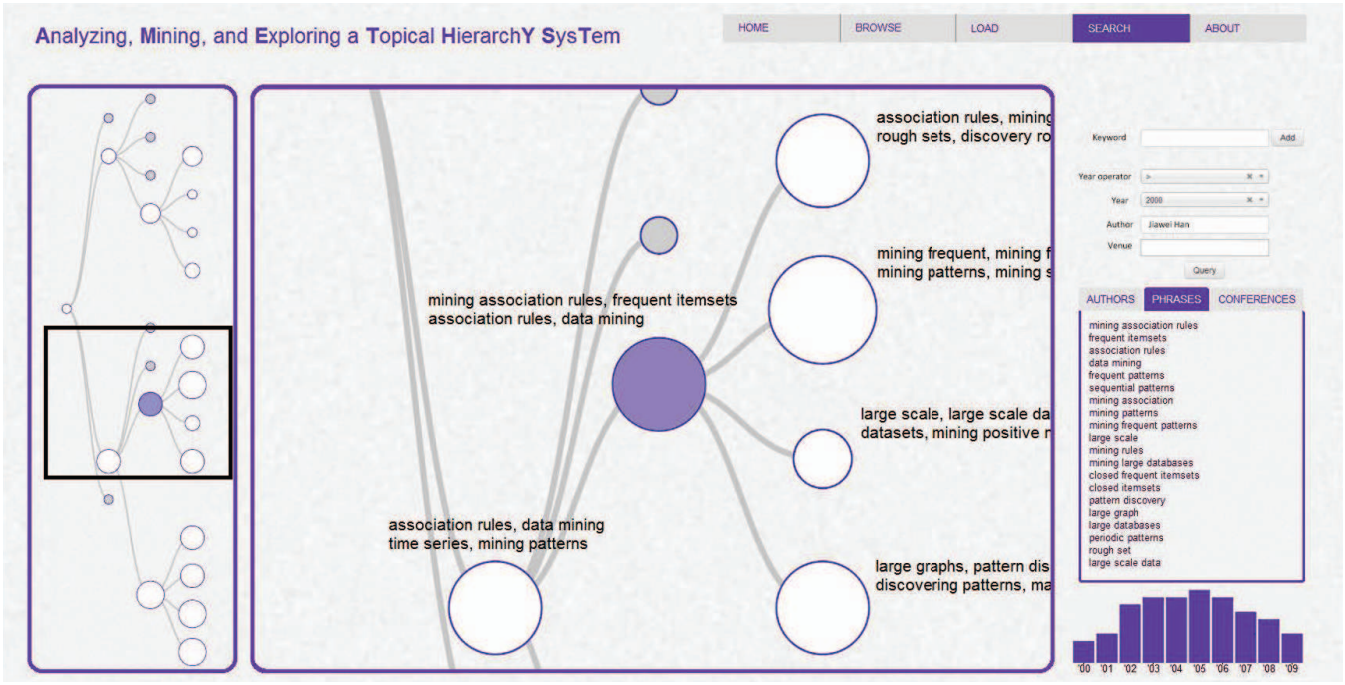


Figure 4: Topical hierarchy query on the author 'Jiawei Han' publishing since the year 2000. The highlighted node is on the general topic of association rule mining, as can be more clearly seen in the Phrases tab of the Topic Detail view. The Topic Temporal View reflects the fact that Dr. Han's number of publications on this topic was highest in the mid-2000s.

query. The query itself is visible in the input area below the menu bar. The **Structure View** now reflects the importance of various topics from the original hierarchy, relative to the query. The phrases within each topic, as well as the entities connected to each topic are reordered. Notice how in the example result, the selected node in the **Zoom View** shows only those phrases that are closely related to association rule mining to be ranked highly (compare with the top-ranked phrases for this same node visible in Figure 3 which mention gene expression data, a phrase that is not very relevant to the query, and which has therefore been downranked in the topic.) The **Topic Detail View** shows that the top ranked phrases for this topic continue to be fairly closely related to association rule mining. Finally, the **Topic Temporal View** reflects the fact that Dr. Han's publication rate on this topic peaked in the mid-2000s.

4.2.2 *Free Text Search*

A user may be interested in examining the contexts in which a phrase, or a set of phrases appears in the topical hierarchy. The topical hierarchy search interface shown in Figure 4 allows the user to query using free text. All topical hierarchy phrases which exist in the free text query are identified. The topics containing these phrases are then visualized, accentuating the topics in which the phrases are highly ranked, and thus providing the user with the topical context of their query. The free text search may also be combined with an OLAP query, as indicated in Figure 2.

5. CONCLUSION

Our system can work with a variety of datasets such as DBLP, news articles, and other data collections with similar schema. Other applications of mining HINs with text components may also be explored with AMETHYST, such as related entity search, expert finding, and ontologies which incorporate user guidance.

6. ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation grants IIS–0905215, CNS-0931975, and IIS-1017362; U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA); and IIS-1017362, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. Chi Wang was supported by a Microsoft Research PhD Fellowship. Marina Danilevsky was supported by a National Science Foundation Graduate Research Fellowship grant NSF DGE 07-15088.

7. REFERENCES

[1] X. Chen, M. Zhou, and L. Carin. The contextual focused topic model. In *KDD*, 2012.
[2] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*, 2011.
[3] J. Han. Mining heterogeneous information networks: the next frontier. In *KDD*, 2012.
[4] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008.
[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.
[6] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. A phrase mining framework for recursive construction of a topical hierarchy. In *KDD*, 2013.
[7] D. Zhang, C. Zhai, and J. Han. Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM*, 2009.