Unsupervised Classification and Visualization of Unstructured Text for the Support of Interdisciplinary Collaboration

Lisa J. Miller
ljmiller@hawaii.edu

Rich Gazan
gazan@hawaii.edu

Susanne Still
sstill@hawaii.edu

Department of Information and Computer Sciences
University of Hawaii at Mānoa
Honolulu, HI 96822-2327

ABSTRACT

We present a computer supported tool for cooperative work in interdisciplinary fields, which we tested within the area of astrobiology. Our document classification and visualization system is fully automated and data driven, based on unsupervised learning algorithms and network visualization tools. A new feature selection algorithm was created to aid this process that indicates which words should be used for mutual information-based clustering. Our system can extract information about collaborations from unstructured databases with no meta-data and reveals structure that can aid the planning of collaborative research. We analyzed publications produced by researchers from NASA's Astrobiology Institute. We presented this analysis as a cultural probe and recorded reactions from researchers that indicated that our method can help scientists from different disciplines to work together. We have made an interactive version of our visualization and analysis available as a website for long-term use.

Author Keywords

interdisciplinary science; document analysis; unsupervised learning; feature selection

ACM Classification Keywords

H.5.3 Group and Organization Interfaces:

General Terms

Algorithms; Measurement

INTRODUCTION

The field of Computer Supported Cooperative Work (CSCW) can contribute useful tools to interdisciplinary scientific collaborations. In practice, however, the use of CSCW as a bridge to more productive collaborations has been modest and situational. Challenges involved in making CSCW applicable to aiding interdisciplinary cooperation have been identified in the CSCW literature, which we discuss below.

We present here a computer supported tool, designed for scientists funded by the NASA Astrobiology Institute (NAI), to facilitate cooperative work in this highly interdisciplinary field. We have analyzed a corpus of documents published under the current NAI funding period, looking for commonalities by means of automated cluster analysis. This allowed us to identify clusters of related documents, to extract topics that characterize these clusters, and to evaluate how these relate to the Science Objectives listed in the NAI planning guidelines [**?**].

The cluster analysis furthermore reveals a meta-structure, which we visualized using a force-directed graph. This visualization provides an opportunity for scientists to look at the large body of unstructured document data in a structured way, and thereby identify how their work fits into the entire body of work produced under NAI funding. One explicit goal of the funding agency is to "carry out, support and catalyze collaborative, interdisciplinary research". Our data analysis and visualization tool allows for a time efficient way of assessing where cooperation between teams and/or interdisciplinary research occurs within the funded work. Thus, it may be useful in helping to plan in which areas new collaborations could be pursued. Furthermore, the visualization aids both scientists and the funding agency in identifying potentially relevant areas that are not yet being actively addressed within the funded body of work.

CSCW and Interdisciplinarity

There is a long history of CSCW research in the domain of interdisciplinary science collaborations [**?**, **?**, **?**]. Distributed interdisciplinary collaborations have been studied from a CSCW perspective at several levels of analysis, including organizational forms [**?**], interfaces [**?**], and human infrastructure [**?**].

One study conducted retrospective interviews of principal investigators (PIs) and Co-PIs of distributed interdisciplinary teams [**?**]. Those teams explicitly stated dissemination and integration across fields as their goals. Integration was defined as "the extent to which a research team combines its distinct expertise and work into a unified whole" [**?**]. Integrated teams required both administrative support and internal com-

CSCW'14, February 15–19, 2014, Baltimore, Maryland, USA.

ACM 978-1-4503-2540-0/14/02...$15.00.
http://dx.doi.org/10.1145/2531602.2531666

mitment, and the authors call for CSCW tools that create opportunities for cross-team discussion and feedback [**?**].

Within an interdisciplinary collaboration, individual researchers will use tools that work for them, and are less likely to use groupware or similar common tools [**?**]. However, scientific research at the boundary between disciplines can result in the negotiation of new roles [**?**] and lead to the creation of boundary objects [**?**], artifacts across which diverse collaborators can communicate and negotiate shared meaning.

Cultural probes [**?**], defined as "designed objects, physical packets containing open-ended, provocative, and oblique tasks to support early participant engagement with the design process" [**?**], are artifacts designed to engage a group or community and elicit interactions. Introducing cultural probes can generate both boundary objects [**?**] and more lightweight, transient, boundary-negotiating objects [**?**], which in turn, can elicit cross-boundary communication [**?**].

The results of our analysis were presented to scientists as a cultural probe and they remain available to them as an interactive tool on our website. This work is part of the ongoing development of the Astrobiology Integrative Research Framework (AIRFrame) [**?**], a system [**?**, **?**, **?**, **?**] funded by the NASA Astrobiology Institute (NAI) that uses document analysis techniques to allow astrobiology researchers from diverse fields to identify the subset of publications relevant to their work, but which may have appeared in journals specialized for a different audience, as well as relevant concepts and researchers from outside their discipline. The goal of AIRFrame is to foster understanding across areas, and thereby catalyze interdisciplinary collaboration.

Astrobiology and the NAI

The field of astrobiology is concerned with questions of life in the universe. Researchers in areas as diverse as astronomy, geology, biology, chemistry, and oceanography address questions about habitable environments, prebiotic chemistry, water in the universe, extremophiles, and the physical limits of life, to name a few examples. Understanding more about the evolution and distribution of life may also help to understand more about the future of life on Earth.

NASA is the main funding agency for astrobiology. The NASA Astrobiology Institute (NAI) is a virtual, distributed research organization consisting of 14 teams across the US and approximately 840 researchers, hospitable for CSCW. Researchers work both in their areas of expertise and on interdisciplinary research projects. NAI teams share science results internally and with other teams in weekly meetings; principal investigators at each site have monthly meetings to share results; and each team is responsible for producing an annual report of its publications, presentations, outreach, and other activities. Critically, both publications and annual reports are the primary means by which researchers and teams are evaluated.

The NAI requires researchers to articulate in what way their work is interdisciplinary. However, little or no support structure is in place for researchers to find potential interdisciplinary collaborations or to assess how their work fits into the body of research being conducted. The AIRFrame system addresses these issues. AIRFrame analysis results are shared with researchers in team meetings and seminars to elicit feedback as input to iterative development [**?**].

CSCW for the NAI

A crucial challenge in providing CSCW tools to scientists is that they naturally have a preference for more traditional means of communication. However, established methods of transmitting scientific information such as journal publications can be used to create tangible tools and assessment instruments [**?**, **?**, **?**]. Scientists can employ these tools to identify areas of potentially productive crossover in large distributed collaborations, and to provide evidence of their interdisciplinary impact.

We chose to analyze the articles produced by the NAI teams without causing disruption to ongoing research. Instead of using expert help to construct a structured database that fit into a given set of criteria, we use unsupervised machine learning algorithms to extract interpretable structure directly from document texts by means of cluster analysis. We visualize the results of the cluster analysis to further aid ease of interpretation. This automatization is useful as it saves valuable expert time. Not imposing pre-conceived criteria may also be useful, as that allows for the possibility of discovering, for example, which Science Objectives appear together in document clusters, revealing the extent to which the clusters correspond to stated NAI funding guidelines.

The collected documents are publications reported by the 14 NAI-funded research teams in three years of annual reports. Some of these were available to us as full text, for others we only had access to abstracts. NAI does not have a structured citation database, hence no bibliographic metadata was used in the automated analysis. We used unsupervised learning (cluster analysis) based on the distributions of words over documents, as estimated from the data. In the Methods Section, we detail data collection, discuss the challenges that have to be overcome to perform meaningful cluster analysis of document data, and discuss how we addressed them using existing algorithms. We saw fit to develop a new preprocessing method which provides an automated indication of how many words should be used in the analysis. This method has not been published elsewhere and is also described in the Methods Section.

Results of the cluster analysis are discussed in the Results Section, together with their visualization and the resulting meta-structure that was observed. These were presented to the University of Hawaii NAI team. Both the automated document analysis and the visualization of the results met with great success, and were deemed useful by the team. We have created a website to allow continued and more detailed access to these results. We report on the presentation of our work and reactions of our audience in the Presentation and User Reaction Section.

METHODS

Astrobiology Document Dataset

Often, only the abstract text is used in cluster analysis, but prior research with the Textpresso semantic search engine [**?**] suggests that abstracts of scientific journal articles are not enough to discover important connections among documents. Abstracts are rich with keywords, but they traditionally do not include more subtle evidence of linkage such as common research methods, equipment, and shared references.

We manually collected 1,346 publications, compiled from the 2009, 2010, and 2011 NAI annual reports. We attempted to retrieve the full text of all articles but if it was unavailable the article abstract was used. This resulted in a corpus of 724 long multi-page articles including references, 106 long abstracts or short articles with references, 484 short abstracts, and 32 very short database posting records or citations. Of these, 832 were in pdf format, the rest were retrieved as plain text. Pdf documents were converted to plain text using the Unix utility *pdftotext.*

The *NASA Astrobiology Roadmap* document [**?**] is the planning guide for the NAI research represented in the dataset. It outlines 18 high-priority Science Objectives ranging from strict astronomy topics (e.g., *Indirect and direct astronomical observations of extrasolar planets*) to interdisciplinary topics (e.g., *Biosignatures to be sought in Solar System materials*) to strongly biological themes (e.g., *Co-evolution of microbial communities*). This document was included in the document data set and was used to analyze and evaluate the output of cluster analysis (detailed in the Results Section).

Cluster Analysis

By grouping the documents into a number of different clusters we wish to extract useful structure from this diverse and unstructured collection for which manual analysis by an expert is too time consuming. We are particularly interested in the topics that emerge within clusters when as few assumptions as possible are made about the data.

Clustering can be viewed as a form of summarizing data in a meaningful way and a large number of different clustering methods have been developed, e.g., [**?**, **?**, **?**, **?**]. There are a few serious challenges involved in cluster analysis [**?**]. Determining what level of detail should be used in the summary, or equivalently choosing the appropriate number of clusters, is considered the hardest problem in clustering [**?**]. Choosing an appropriate clustering criterion such that similar objects end up in one cluster is another fundamental issue, involving both selection of a measure of similarity and identification of relevant features in the data [**?**, **?**].

Many popular text clustering methods are based on text modeling algorithms developed within the area of Information Retrieval for querying large databases [**?**, **?**, **?**, **?**]. These methods, such as Latent Dirichlet Allocation (LDA) [**?**] and Latent Semantic Analysis (LSA) [**?**], use statistical modeling methods to reduce document representation from a vector of word counts to a more compact vector of model features. A simple similarity measure, such as the cosine distance or average linking, is then used to group the reduced vectors into clusters [**?**, **?**, **?**]. LSA identifies a linear subspace in a matrix of document-word vectors which captures variance between the documents [**?**, **?**]. LDA is a probabilistic expansion of LSA where latent multinomial variables (referred to as topics in this literature) are modeled as discrete distributions over words, while documents are modeled as discrete distributions over "topics".

Statistical document models require assumptions be made about the structure of the data and what part of it is relevant for clustering. For example, with LDA one must use *ad hoc* methods to estimate the parameters which fix the shape of the assumed prior distributions of words over topics and topics over documents. Additionally, the number of topics must be estimated. Model-based, parametric clustering methods have the drawback that incorrect assumptions or poor selection of parameters can lead to poor performance or skew results toward finding certain types of clusters [**?**, **?**, **?**].

Alternatively, one can define a notion of relevance directly, using information theory [**?**]. This results in a framework, the Information Bottleneck (IB) method, within which a clustering algorithm and a notion of similarity can be derived mathematically [**?**], rather than needing *ad hoc* specification.

Applied to document clustering, the IB method can be interpreted as lossy compression of documents into clusters in such a way that the information about the words is maximally retained. Thus, the words are defined as the relevant variable, W. The mutual information between clusters, C, and documents, D, $I[C;D]$ is minimized, while the mutual information between clusters and words, $I[C;W]$, is maximized. The parameter β controls a trade-off between the amount of compression and keeping relevant information, that is, information about words.

$$\min_{p(c|d)} \left(I[C;D] - \beta I[C;W]\right) \tag{1}$$

Minimizing this equation yields the optimal cluster assignments, $p(c|d)$, the probability of cluster c given document d.

Even prior to the development of the Information Bottleneck framework, information theoretic methods were shown to be preferrable to parametric methods in data analysis for science studies [**?**]. For more than ten years, the IB framework and related methods, have been viewed as the state of the art for clustering highly multidimensional data, such as text [**?**]. It is difficult to compare outcomes of different clustering methods because of diverse definitions of relevance and similarity, and the relatively ambiguous meaning of correctness in grouping unlabeled data. However, tests using labeled benchmark document datasets have shown that IB-based methods produce as good or better clustering precision with regard to the labeling than parametric models such as LDA [**?**, **?**]. Additionally, with the IB framework, perturbation theory allows one to work out the difficult problem of how many clusters can be used maximally without over-fitting the data [**?**].

Document Preprocessing

Before documents can be clustered, preprocessing must be done to extract valid text and format it for input into the clustering algorithm. Most commonly, each document is initially represented as a single *feature vector* whose dimensions or *features* are normalized word counts [**?**, **?**, **?**, **?**]. Removing or combining words to reduce the dimensionality of the feature space while maintaining the most important elements is known as *feature selection*. Important elements are those features which define subsets in the data relevant to the intended goal of a particular clustering project [**?**, **?**, **?**].

In different clustering methods, feature selection may be an integral part of the clustering process or it can be just another step in the preprocessing before clustering occurs. With text modeling, such as LDA, most of the processing to differentiate between clusters is done within the feature selection step itself and it is considered part of the clustering process. With other methods, such as the IB, a complex clustering algorithm performs most of the work to fit the data into clusters and feature selection is part of preprocessing, used only to eliminate extraneous elements.

The System for the Mechanical Analysis and Retrieval of Text (SMART) project worked to evaluate and develop feature selection methods for more than 30 years [**?**]. We employ two of those methods which remain standard practice [**?**]: *stop word* removal which removes a list of extremely common function words; and *stemming* which removes the endings of words, transforming related words into identical word stems.

For all of the documents, a standard list of stop words were removed and all remaining words were stemmed using the Porter Stemming Algorithm [**?**]. We then applied our new feature selection method to determine which and how many of those words to keep for clustering.

New Feature Selection Method

A feature selection method particularly popular and methodologically consistent in IB-based clustering is to build a list of words, w, ranked by $I[w]$, their individual contribution to the mutual information between all words and documents in the data, $I[W;D]$, as in Equations 2 and 3 [**?**, **?**, **?**].

$$I[w] = \sum_{d \in D} p(w|d)p(d)\log_2\left(\frac{p(w|d)}{p(w)}\right) \quad (2)$$

$$I[W;D] = \sum_{w \in W} I[w]. \quad (3)$$

Where $p(w|d)$ is the probability of word w given document d, $p(w)$ is the probability of w over the whole dataset, and $p(d) = \frac{1}{N_D}$ is the probability of a particular document, with N_D being the number of documents .

However, with this method, the number of words to keep is not indicated, a cutoff threshold must be decided upon. We developed a new method, introduced here, to address this issue.

Algorithm 1 Greedy Corrected $I[W;D]$ Word Ranker

```
for all words w in the data do
    Calculate and save I[w], Equation 2.
end for
while all words not ranked do
    for all not yet ranked words, w do
        Calculate C[w] from Equation 4.
        Calculate corrected information contribution:
        I_corr[w] = I[w] − C[w].
        Keep track of the maximum I_corr[w].
    end for
    Add word with largest I_corr[w] to ranked list
    Add N_Dw for that word to N_Dr
    Add N_w of that word to N_r
    Add I_corr[w] to I_sum
end while
return The ranked word list.
```

Based on ideas taken from [**?**, **?**, **?**, **?**], our new method for choosing words attempts to correct the error made when estimating information from finite samples. The idea is to retain only those words that give an increase to the corrected mutual information between words and documents. Words that do not increase this value are seen as adding noise to the data due to finite sampling effects.

A greedy algorithm is used due to the combinatorial nature of the problem. Complete pseudo-code is given in Algorithm 1 and Equation 4, where N_{Wr} is the number of ranked words, N_{Dr} is the number of documents containing any ranked word (a document is only counted once, the first time any word it contains is ranked), N_r is the total number of occurrences of all ranked words in the data, N_w is the number of occurrences of a single word w in the data, and N_{Dw} is number of documents a word w occurs in not already counted in N_{Dr}.

$$C[w] = \frac{(N_{Dr} + N_{Dw} - 1) * N_{Wr}}{2\log(2)(N_r + N_w)} \quad (4)$$

The correction term changes as each word is added to the ranked list. Therefore, it must be recomputed for each as-yet unranked word in each round of ranking. In each round, the word with the largest corrected contribution, $I_{corr}[w]$, is added to the ranked list. After all are ranked, all words that increase I_{sum}, the cumulative $I_{corr}[w]$, are retained for clustering.

For this dataset, the cumulative per-word contributions of the top-ranking 10,000 words out of a total of 48,421 are shown in Figure 1. Our method indicated 1,943 words to be kept as can be seen by the red square at the maximum of the inset in the figure. Those words were then used for all clustering procedures.

Visualization of Cluster Analysis Results

The use of node-link plots in association with text clustering to visualize structure in scientific research has a long history [**?**, **?**, **?**, **?**]. Commonly, metadata or modeled text are

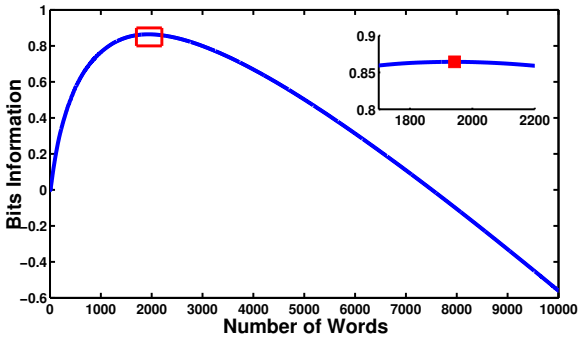

Figure 1. Cumulative corrected per-word contribution to mutual information between documents and words. 10,000 highest contributing words. Words ranked by largest to smallest contribution. Inset shows the section within the red box of the larger plot, highlighting the point at 1,943 words after which contributions become negative. All words giving a positive contribution were kept for clustering.

represented on a network graph framework, graph clustering algorithms are used to simplify and group the representation, and layout algorithms are employed to spatially manipulate the graph for easier visual comprehension [**?**, **?**, **?**, **?**].

In our method, we use the IB to cluster mixed-length documents into several different numbers of clusters. Then, we use the Gephi graph visualization program [**?**] to visualize the combined multiple clustering results on one force-directed node-link plot. The graph layout was done using Gephi's Force Atlas algorithm and Gephi's partition functionality was used to color the document nodes. All other data analysis and processing was done outside of Gephi.

An exponentially large number of edges would have resulted if we had represented each cluster as a fully connected subplot, where all document nodes in a cluster are connected to each other by an edge, as is common. Inspired by network topology for large-scale networks, we reduced the number of edges by representing the clusters as star networks with each document represented by a small node connected to a central cluster node, often referred to as a hub [**?**]. Each cluster node is sized by the number of documents assigned to that cluster, and each document node has an edge connecting it to the cluster it was assigned to. This is visualized simultaneously for several clustering results with different numbers of clusters.

Our method of reducing edges is related to graph reduction techniques such as node and edge aggregation into meta-nodes or aggregating edges into a single meta-edge [**?**, **?**], but different in that we aggregate edges into the hub. The original document nodes are retained and are connected to each other through the hub. This allows us to preserve both the overall structure of the entire dataset and the ability to visualize properties of individual document nodes, such as coloring by discipline.

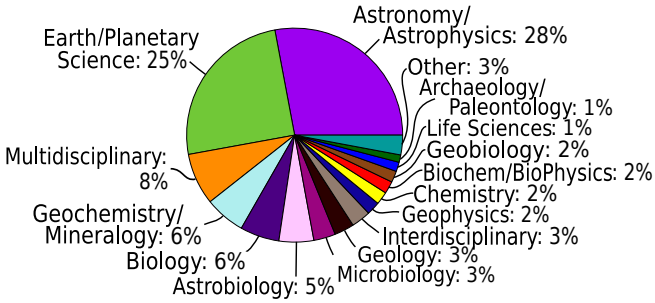

Figure 2. Breakdown of astrobiology document dataset by discipline. The multidisciplinary category includes articles in journals such as *Science, Nature* and *PNAS*. The interdisciplinary category contains articles in journals which specifically state that they publish interdisciplinary studies, such as *The Journal of Cosmology*.

RESULTS OF CLUSTER ANALYSIS

For a given number of clusters, k, the IB clustering algorithm returns a representation of each cluster, which is a list of words sorted by their probability of occurring in that cluster. The algorithm also returns a list of assignments of documents to each cluster.

We looked at the 50 most probable words in each cluster and inspected the documents assigned to each. The documents are listed by projects in the annual reports to the NAI, those projects include a subset of Science Objectives from the Astrobiology Roadmap. From this, we obtained both projects and related Science Objectives for each document. We identified the most prominent Science Objectives in each cluster. This evaluation allowed us to identify topics that appeared in each cluster.

We hand-labeled the documents with 22 disciplines listed on their publishers' websites, the distribution of these is shown in Figure 2. Over half of the documents appeared in Astronomy/Astrophysics (AAP) or Earth and Planetary Sciences (EPS) journals and conferences. Approximately 8% were published in multidisciplinary journals such as *Science* and *Nature*, while the rest come from 19 disciplines that each make up 6% or less of the total. The "Other" category in contains the least represented disciplines: Marine Science, Chemical Physics, Aerospace/Astronautics, Physics, Meteorology/Climate, and Information Science. Discipline labels were used only in our evaluation and visualization, not as input to the clustering algorithm.

We examined the best cluster analysis results and chose to evaluate $k = 8$, 12, and 16 clusters in detail to show a progression in relationships amongst the documents at different levels of compression. Less than 8 clusters compressed the data too much. Clusters were very large and detailed topics were not present. More than 16 clusters resulted in very small clusters with few documents in them. These were no longer a useful summary of the data. Within the chosen range of 8, 12, and 16 clusters, the differences between clustering results were enough to discern obvious changes in the cluster memberships without adding too much complexity for evaluation.

In our results, most clusters are a mix of long full-text documents and shorter abstracts all related to the same topic.

However, some small clusters containing only abstracts were made. These indicate a limitation of clustering short documents. Despite having similiar topics, the sparsity of words in some shorter abstracts make their representations very different from longer documents with much richer word content, so much so, that they can be forced into separate clusters.

In the following sections, clusters are identified with a number-letter combination, such as 8a, where the number indicates k, the number of clusters made, and the letter is an arbitrary name for the cluster. Cluster topics are listed in italics following the cluster identifier.

Eight Clusters

Classifying the dataset into 8 clusters results in 6 with topics directly related to combinations of Science Objectives. Three clusters contain documents from very mixed disciplines, with no single discipline making up more than 30% of the total: 8a - *beginning of life and biosignatures*; 8c - *prebiotic materials in ice and water*; and 8d - *genetics and biomolecules*. Two clusters contain a mix of primarily AAP and EPS documents: 8f - *detecting planetary habitability* and 8h - *planet formation and planetary qualities*. Cluster 8b - *observations and properties of stars and planetary systems* contains only AAP documents.

Cluster 8g - *methods of research, equipment and instruments, field expeditions, and planning for future missions* is about methods and planning, rather than NAI Science Objectives. This cluster contains mixed disciplines with 46% of them from EPS. Predictably, it includes the Astrobiology Roadmap document. The smallest cluster, 8e - *observations of extrasolar planets* is a collection of AAP abstracts from a single NAI team.

Twelve Clusters

Seven of the 12 clusters contain topics closely related to Science Objectives, quite similar to clusters 8a-c, 8f, and 8h, but split slightly more along disciplinary lines. Out of those 7 only 2 contain very mixed disciplines (no individual discipline makes up more than 30% of the total): 12j - *origins of organic matter, early Earth ecosystems, microbes, and biosignatures* and 12l - *microbes, RNA, and biomolecules*. EPS documents make up more than 30% of clusters 12g - *habitable planet environments, early Earth and Mars*; 12h - *planet, asteroid, and meteorite surface features*; and 12k - *microbial fossils and evolution of minerals*. AAP makes up slightly more than 30% of cluster 12a - *ice chemistry, icy worlds, prebiotic materials, and chemical complexity* and more than 90% of cluster 12c - *stars, planetary systems and exoplanet detection*.

Both 12b - *general discussions of astrobiology, exploration missions, and planning* and 12i - *environment and instrument testing* are directly related to cluster 8g. 12b is a cluster of general topics including the Astrobiology Roadmap, while 12i contains only 32 EPS abstracts.

Additionally, there are 2 small clusters of under 40 abstracts each which have the same topic: *extrasolar and habitable planets*, 12d, which is all AAP articles, and 12f with mixed disciplines. Finally, there is a large cluster containing mostly EPS documents, 12e - *MESSENGER mission to Mercury*. Interestingly, the MESSENGER mission to Mercury is not listed in the Astrobiology Roadmap.

Sixteen Clusters

When 16 clusters are used, the topics of 6 do not change significantly from 12 clusters. Listing their identities: 12a - 16f; 12e - 16e; 12h - 16i; 12j - 16g; 12k - 16c; 12g - 16o. We now discuss the remaining 10.

There is more than one biology-based cluster: 16d - *evolution, genetics and RNA* which contains the Roadmap document and 16p - *microbes, bacteria and extremophiles*. The previously discussed AAP clusters, 8a and 12c, split into 16n - *stars, planetary systems and observations of extrasolar planets* and 16h - *planet formation and planet qualities*. Documents contained in clusters 12b and 12i recombine and split into 3 mixed-discipline clusters: 16l - *instruments, analysis methods, analog environments, and field studies*; 16m - *Earth as model for habitable extrasolar planets*; and an all abstract cluster, 16k - *bio-signatures in meteorite materials and asteroids*.

Three small clusters of abstracts each contain less than 30 documents: 16j - *extrasolar planet detection*; and 16a and 16b which are too small, containing only very short documents, such that we could not identify meaningful topics. This is an indication that using more than 16 clusters would result in a summary that contains meaningless detail.

Visualization

To visualize the meta-structure resulting from cluster evaluation over different numbers of clusters, we plotted all clustering results in one graph, with node size based on degree and document nodes colored by discipline (Figure 3).

The layout was made using the Force Atlas algorithm provided with Gephi [**?**]. This algorithm causes all nodes to repel each other with a force based on size while edges draw connected nodes together. Therefore, document nodes were drawn toward the hubs representing their assigned clusters. The plot expanded along directions where there are less connecting edges (documents and cluster topics are less closely related) and contracted towards areas with similar topics where many connections occur. The resulting structure reveals cluster topic relatedness; how different disciplines are spread over the result space; and how documents are reassigned when the number of clusters is changed.

The graph shows that there is no clear hierarchical structure. Clusters do not split up evenly when going from 8 to 16 clusters. Instead, re-assignments are complex as indicated by the large number of isolated documents in the graph. Some documents, however, do stay together and we discuss this behavior in detail later in this section. We now discuss the highlighted areas in Figure 3.

The area inside the dashed line marked with I in Figure 3, contains clusters with star and exoplanet detection topics

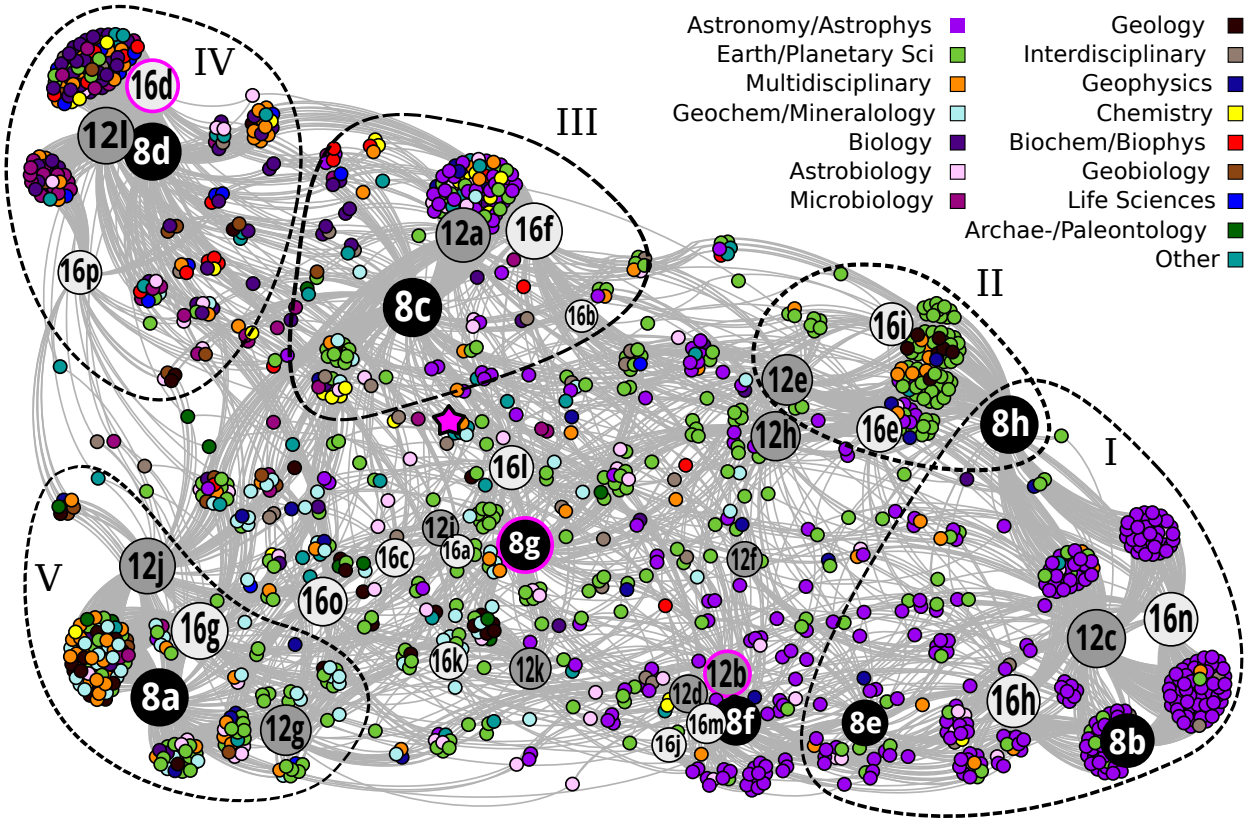

Figure 3. Force directed plot of clusters and documents for k = 8, 12, and 16 clusters, showing clumping behavior and distribution of disciplines. Cluster hubs, colored gray or black, are labeled with k plus a letter to aid evaluation. Document nodes are colored by discipline as indicated. Areas within dashed lines are approximate locations of topics discussed in the text. The dark pink star node is the NASA Astrobiology Roadmap document, its assigned clusters are outlined in the same pink.

and their associated documents. These are almost exclusively AAP publications as indicated by the majority of purple nodes. When assigned to different numbers of clusters, documents either stay together, such as in cluster 12c, or they split up into subtopics, such as in 16h and 16n. This is the most prominent place in the graph where an almost hierarchical structure can be found. Many documents in area I never cluster with other documents outside of this area. They remain very close together because there are few edges to pull them away. This causes the area to be somewhat segregated from the rest of the graph.

Within area II we find a predominantly green group of document nodes, which are the EPS documents related to the MESSENGER Mercury mission that make up cluster 12e. A large portion of those documents is also assigned to cluster 8h, together with large groups of AAP documents found in area I. Therefore, there are short paths connecting documents in these two groups pulling the document nodes and their related clusters closer together, giving a visual clue about the bi-disciplinary content of cluster 8h.

Area III contains an heterogeneous selection of disciplines (indicated by document node colors), but most of the documents share similar topics, namely the study of prebiotic materials (particularly ice and water) in lab experiments, on Earth, in the Solar System, and outside the Solar System.

Inside area IV are many dark purple biology, magenta microbiology, and brown geology document nodes. These are the documents clustered into genetics, microbes, and biomolecule-related topics. The large group of documents near the cluster 16d node stay together when different numbers of clusters are used. They are also assigned to clusters 12l and 8d.

Clusters with topics related to biosignatures and the beginning of life are located in area V with cluster nodes 8a, 12g, 12j, and 16g. These are multidisciplinary, indicated by the amount of differently colored document nodes. They are also not as separated from the bulk of the data as the biology-based clusters (in area IV) due to the fairly large amount of connections they share with the ice-related work contained in area III, and the more loosely connected documents related to mission planning, analog environments, and instrument testing located near the center of the plot and cluster 8g.

The *NASA Astrobiology Roadmap* document contains descriptions of all of the 18 Science Objectives. It is by definition a summary of the intended outcomes of research produced by the NAI Teams. In Figure 3, the *Roadmap's* location is marked by a node shaped as a pink star with a thick black border. The clusters it was assigned to, 8g, 12b, and 16d, are highlighted by pink borders. This document is located close to the center of the plot, as should be expected, and the clus-

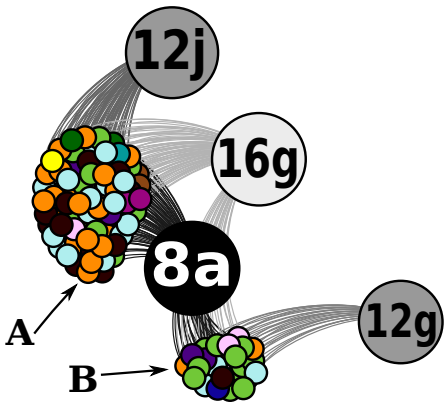

Figure 4. Detail from area V from Figure 3 showing how documents always assigned together to the same clusters form clumps of nodes. As explained in the text, the documents contained in clumps marked "A" and "B" are all assigned to clusters 8a and 16g, but are split at k=12 into 12g and 12j. Document nodes colored by discipline as in previous figures.

ters it was assigned to do not consistently represent the same topic or the same disciplines. This illustrates an advantage of looking at more than one clustering solution on the plot. Examining only one solution would make it appear that a single cluster topic (where the *Roadmap* is assigned) represents the research guidelines much more strongly than any of the others.

Plotting multiple clustering solutions also reveals how some document nodes are always assigned together to the same cluster, for all values of k considered. These documents form clumps in the force directed plot. However, one clump of document nodes rarely makes up an entire cluster itself and examination of these clumps reveals further insights into the data.

As an example, we focus on area V from Figure 3. A detail is provided in Figure 4, with two clumps of documents labeled **A** and **B**. In-depth examination of all documents within **A** and **B**, respectively, including texts, authors, and citations, reveals a high degree of relatedness within each clump.

The 84 documents in **A** are predominantly geochemistry (light blue), multidisciplinary (orange), and geology (dark brown). Most of these articles originate from projects tagged with Science Objectives pertaining to the early Earth and the search for biosignatures in materials of extraterrestrial origin (such as meteorite samples.) All clusters that the documents in **A** were assigned to have topics related to the beginning of life and biosignatures.

The 26 documents in **B** contain a larger percentage of EPS articles (green). The majority of these articles come from projects tagged with Mars exploration and general planetary habitability-related Science Objectives.

Interestingly, the documents in **B** cluster together with those in **A** in both 8a and 16g, but they separate into clusters 12g and 12j at the intermediate compression level of 12 clusters.

The authorship and citations within both clumps are very tightly coupled. We find that active collaborations between many authors and connections via direct citation are present. Some authors are represented in both **A** and **B**. These authors have done work on the formation of both the Earth and Mars. Their papers about Earth are in **A**, and their papers about Mars are in **B**. Several of the Mars papers by these authors are cross-institutional collaborations with other authors also represented in **B**.

From this examination it can be seen that our method extracts a large amount of relevant detail from the unstructured database we had at hand. Documents that cluster together share similar topics and some share authorship and citations. Documents that clump together over a range of k, are indicative of active collaboration and citation amongst the authors, thus providing, for example, an excellent starting point for CSCW researchers looking for existing collaborations in an unstructured database.

PRESENTATION AND USER REACTION

Our analysis results were presented to 17 NAI researchers from the University of Hawaii (roughly $\frac{2}{3}$ of the entire University of Hawaii Team), during a planning session. The primary goals of the session were to examine (i) the team's collaborative efforts and research output, and (ii) how both could be compared with other NAI teams. The visualizations generated in the present study served as cultural probes during the session, providing researchers in diverse areas a data-driven representation of the larger context of their work and a basis to explore potential connections across disciplines. We noted 4 main reactions:

1. The scientists thought that the model-free nature of the clustering process, which included no preconceptions about the disciplinary membership of any paper, was appropriate to the nature of astrobiology research. It contributed to the perceived objectivity and trustworthiness of the results.

2. The Hawaii team thought that their primary distinguishing strengths were in the areas of astronomy and earth and planetary sciences. However, the visualization suggested that those strengths are relatively common across NAI. One of the strongest interdisciplinary clusters, and one of the rarest across NAI, is in ice chemistry and related areas (seen in area III of Figure 3). The researchers had not been aware of this and realized this was an area of strength unique to Hawaii that should be emphasized in forthcoming reports to NASA. This led to a discussion about whether a physical chemist specializing in ice chemistry might have a productive collaboration with an astronomer specializing in comets. Those researchers and several others used the visualization to form themselves into breakout groups, map out the next steps of their research, and discuss areas of potential crossover.

3. The researchers were not surprised to see documents in the biological sciences cluster together, somewhat removed from those in astrobiology's other constituent fields. They

discussed the relative benefit of having some teams specialize in biological aspects of astrobiology as opposed to trying to force interdisciplinary connections in every area. They argued that the Hawaii team's focus on water, a precursor and substrate for life, might help bridge biological and physical sciences.

4. The scientists expressed interest in examining documents from outside their home discipline that consistently clustered within their own areas, suggesting that this clustering process might also be used as an interdisciplinary document discovery and recommendation system.

As a result of the feedback from the NAI scientists, we have built a website with a fully interactive version of our analysis visualization. This is located at: `http://airframe.ics.hawaii.edu/visualizations/`. The site allows a much richer experience with our data and analysis than is possible to represent here. The entire graph, similar to Figure 3, can be zoomed and panned. Mousing-over nodes pops up document bibliographic information or cluster topic. Clicking a node highlights related clusters and documents, hiding all others. Users can choose to see different numbers of clusters, from 2 - 16, or combinations of those. They may also choose to color the document nodes by discipline or by NAI team. We anticipate that the site will be useful not only for astrobiology researchers, but for NASA administrators and interested members of the public as well.

CONCLUSION AND OUTLOOK

Interdisciplinary research is, by definition, open to contributions from diverse fields, and it is useful to have a mechanism for information to migrate across domains. In this paper we used computer supported methods, powered by unsupervised learning (clustering) and network visualization, to aid cooperative work by helping interdisciplinary scientists work together.

We have created a document classification and analysis tool to aid interdisciplinary research in the field of astrobiology. To that end we created a new algorithm for document preprocessing, resulting in an indication of how many words to use for the mutual-information based clustering method we applied. We have presented the resulting document analysis to scientists at a planning session, as a cultural probe, and have recorded their reactions. We have created a website providing a web-based version of our analysis to allow for its continued use.

Our classification and visualization method is automated and uses unsupervised machine learning (clustering) methods that make minimal assumptions about the structure of the data. It thus saves valuable expert time.

Analysis of the astrobiology documents shows that our method reveals useful structure in an unstructured database without bibiographic meta-data. We were able to identify: (i) which documents are topically related, (ii) how a given document fits into the entire body of work, and (iii) where collaborations take place. This is useful as it provides a time efficient visual guidance that can help researchers to identify both publications and other researchers that are relevant to their own interests. It can thus aid in strengthening existing collaborations and the formation of new ones. It can also aid in guiding searches and in building reading lists.

Most importantly, our method met with success when it was presented to a subset of astrobiology researchers. Our document analysis contained some expected results, but also contributed new insights. It elicited interest and discussions. This demonstrates that our document clustering and visualization method can be a valuable collaborative tool.

The utility of our method should become even more obvious in application areas in which the number of publications is much larger than what we analyzed here. The size of our database ($< 10^3$ documents) implies that probably most PI's are aware of a large fraction of most of the published articles. However, in areas where there are one or even two orders of magnitude more documents, our automated and systematic approach may provide a valuable starting point to guide search and discovery of connections between documents, thereby saving time and making an otherwise infeasible task more tractable. Computing speed limitations are being addressed in active research by others [**?**] and should be less of an issue in the future.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science.