Mining for Proposal Reviewers: Lessons Learned at the National Science Foundation

Seth Hettich
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 9 94043
sjh@ics.uci.edu

Michael J. Pazzani
Rutgers University
CoRE Building, Rm 706
96 Frelinghuysen Rd
Piscataway, NJ 08854-8018
Pazzani @ rutgers.edu

ABSTRACT
In this paper, we discuss a prototype application deployed at the U.S. National Science Foundation for assisting program directors in identifying reviewers for proposals. The application helps program directors sort proposals into panels and find reviewers for proposals. To accomplish these tasks, it extracts information from the full text of proposals both to learn about the topics of proposals and the expertise of reviewers. We discuss a variety of alternatives that were explored, the solution that was implemented, and the experience in using the solution within the workflow of NSF.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data Mining

General Terms
Algorithms, Human Factors,

Emerging applications, technology, and issues

Keywords
Keyword extraction, similarity functions, clustering, information retrieval.

1. INTRODUCTION
The National Science Foundation receives over 40,000 proposals a year. Each proposal is reviewed by several external reviewers. It is critical to the mission of the agency and the integrity of the review process that every proposal is reviewed by researchers with the expertise necessary to comment on the merit of the proposal. If there is not a good match between the topic of a proposal and the expertise of the reviewers, then it is possible that a project is funded that will not advance the progress of science or that a very promising proposal is declined. We explore the problem of using data mining technology to assist program directors in the review of proposals. Care is taken to match the technology to the existing workflow of the agency and to use technology to offer suggestions to program directors who ultimately make all decisions. Although this paper reports on reviewing proposals, we argue that the lessons and technology would also apply to the reviewing of papers submitted to conferences and journals.

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

Many proposals are reviewed in panels, i.e., a group of typically 8-15 reviewers who meet to discuss a set of 20-40 related proposals, with each panelist typically reviewing 6-10 proposals. Most proposals are submitted in response to a particular solicitation (e.g., "Information Technology Research") or to a specific program (e.g., "Human Computer Interaction"). Individual program directors, or for larger solicitations teams of program officers, perform a number of tasks to insure that proposals are reviewed equitably. These tasks include:

1. Divide the proposals into "clusters" of 20-40 related proposals to create panels.
2. Finding reviewers:
   - Identify potential external reviewers to invite for each panel.
   - Assign panelists as reviewers of proposals.
   - If there is not adequate expertise on a panel to review a proposal, obtain "ad hoc" reviews from people with that expertise who are not on a panel.

In addition to this lengthy process, reviewers must not have a conflict of interest with proposals they are reviewing (e.g., they may not be from the same department as the proposal's author), and a diverse group of panelists (both scientifically and demographically) is desirable to insure that multiple perspectives are represented in the review process. Furthermore, due to scheduling or workload conflicts, not every invited reviewer accepts the invitation, requiring an iterative process of inviting a batch of reviewers and then inviting others to fill in gaps after the initial reviewers respond to the invitation.

A particular consideration at NSF is that many proposals are multidisciplinary, e.g., mining genome data. To determine if such a proposal is meritorious, it is important to consult some experts with backgrounds in data mining (to insure that the methods proposed are likely to work) and in the biological sciences (to insure that the problem addressed is an important open problem). If all reviewers have expertise in one area, it's possible that an important problem would be addressed by a technique that isn't

very promising or that very promising technology would be applied to a problem that is already solved.

2. Exploring Potential Solutions

Over the past decade, vendors have proposed various text mining technologies to NSF to help with the reviewing process. The most common technology proposed is automated text clustering to help organize proposals into panels. A variety of alternative approaches (e.g., hierarchical [1] or k-means [2]) have been explored. While these present interesting views of proposal submission data, they do not produce results that fit easily into the workflow of NSF or that have gained universal acceptance by program officers who organize panels and assign reviewers. Automated clustering approaches suffer from a number of flaws that have reduced their utility in dividing proposals into panels.

1. The size of clusters. Most clustering algorithms produce clusters of quite different size. Often, there are a few very large clusters and a larger number of very small clusters. In contrast, NSF panels are often approximately the same size due to logistical constraints ranging from the size of rooms to the number of proposals that can be discussed per day.

2. The stability of clusters. Dividing proposals into panels often occurs incrementally. Although most solicitations have deadlines, some proposals that come in before the deadline are misrouted and then found a few weeks later. Occasionally, due to severe weather or natural disasters, a deadline is extended for some regions of the country. Many clustering algorithms if rerun on a slightly expanded data set produce drastically different results. Some algorithms are stochastic in nature and produce different clusters when rerun on the same data (e.g., see [3]). It is difficult to convince program officers with different backgrounds and expertise that a computer system has found an ideal organization of a group of proposals if that organization changes drastically.

3. Lack of alignment with organizational structure of NSF. The clusters produced by clustering algorithms rarely correspond to the scientific and organizational structure of NSF. Each panel has a program officer (or occasionally a team of 2-3 program officers) with specific expertise. When clusters are created automatically without regard to the organization and program officers' expertise, some clusters do not correspond to established scientific fields and no program director wants to be responsible for reviewing proposals that don't fall within their general area of expertise.

4. Lack of alignment with the goals of the solicitation. For example, some solicitations focus on broadening participation in the scientific workforce, and it is useful to group proposals into panels that address issues such as increasing the participation of women and others that focus on increasing the participation of underrepresented ethnic groups. These panels have heterogeneous scientific content. Other solicitations focus on advancing the frontier of science and might divide proposals into panels by scientific subfield. Within a scientific panel proposals might have a heterogeneous broader impact such as increasing the participation of underrepresented groups or creating results of interest to undustry.

In general, the problem with fully automated text clustering solutions is that they don't leave room for human input of preferences or constraints. There has been some research that addresses issues raised. For example, the simplest k-means clustering algorithm is incremental and would allow for the late additions to the existing clusters. However, the results of k-means are not stable so it results in different partitioning of the same data on different runs. Several investigators (e.g., [4] and [5]) have looked at adding constraints to the clustering process so that constraints are approximately the same size. However, none of these address the lack of alignment with the organization structure and workflow. In Section 3, we discuss an approach to "cluster checking" in which algorithms related to text clustering and classification are used to suggest improvements to clusters produced by people and new proposals are added to existing panels.

NSF has also explored and experimented with technology for assigning reviewers to proposals. One approach is to create a database of reviewers with keywords indicating user expertise. These databases are populated by users filling out a form with their expertise. Experience within NSF on prototypes of reviewer databases have found mixed results. Common problems include:

1. It is difficult for a scientific community to agree upon a taxonomy of keywords. One need only examine the ACM Computing Classification Scheme at http://www.acm.org/class/1998 to gain an appreciation for the difficulty. While this classification is adequate for a coarse sorting of papers into topic areas, the topic areas tend to be too coarse to be of much use in bringing expertise into the reviewing process. For example, the most fine-grained term representing the topic area of this conference is "Data Mining." If this were used as the basis for assigning reviewers, then a system that uses a keyword-based approach would believe that anyone publishing in this conference would be considered equally qualified to review a proposal or paper on any topic in the conference. The Data Mining field has become sufficiently specialized that one can be an expert in one area (such as association rules) and not have detailed expertise in other areas (such as text classification) and an ideal reviewer for a proposal in one area may not be qualified for another area.

2. It is difficult to maintain such a keyword database over time. New topics arise in rapidly growing fields requiring the taxonomy and database to be updated frequently. This is particularly important for a funding agency that has the goal of funding work at the frontier of science rather than

concentrating on incremental work in mature fields.

3. If unrestricted text is allowed as descriptions for expertise, it is rare that potential reviewers, program directors, and proposal authors all select the same free text terms. Numerous studies of information retrieval systems have found low agreement among individuals assigning keywords to content (e.g., [8]).

4. There is not high compliance with requests of users to enter information into the database. Many researchers are too busy to fill out forms or hesitant to "volunteer" for reviewing. While agreeing to review proposals is a service to the funding agency, being asked to review proposals is as welcome to some as other forms of service such as serving on jury duty.

5. The interface for submitting proposals to NSF, Fastlane, does not allow keywords to be entered describing the proposals. While this could be added to the interface, doing so would require consensus that this will facilitate proposal handling and this has not been demonstrated convincingly.

Due to the limitations of keyword-based database systems, when they are used within NSF, they are limited to suggesting a pool of candidates for a panel on a given topic. While Computer and Information Science and Engineering at NSF has experimented with a keyword system (e.g., in the 2001 ITR competition), it was not used in subsequent years.

Finally, NSF has experimented with systems that allow panelists to indicate preferences for reviewing proposals within a panel. In such systems, panelists indicate their preference for reviewing a proposal on a numeric scale. Many conferences also use similar systems such as Cyberchair [9]. In Cyberchair, a constraint satisfaction algorithm assigns people proposals they are most interested in. These systems only address part of the reviewer assignment problem. They do not assist with identifying panelists but only assigning proposals to panelists once they have been identified. There has been an issue with compliance on these systems as well, i.e., not every panelist promptly enters preferences data and a single person not replying can delay the assignments for all others. In addition, it isn't clear what the preference scores mean or how much thought goes into the assignments. While the intent is to judge how well qualified a reviewer is to review a proposal, we have observed many panelists having a strong preference for proposals by well known researchers and fewer having a preference for proposals by less established researchers. While NSF typically asks for preferences on 20-30 proposals, some conferences ask for preference data on 200-300 papers. The second author admits that when presented with 300 papers in Cyberchair, not as much time is spent reviewing the abstracts of the last batch of papers as the first to determine preferences. Finally, there is also a problem with multidisciplinary proposals if people from one discipline have a preference for a paper. It can occur that all computer scientists and no biologists give high preference scores to a bioinformatics proposal, in which case a preference-based system will result in one aspect of the proposal not being reviewed.

3. Revaide

We have deployed a prototype system, Revaide, within NSF that addresses the problems with previous fully autonomous systems. The philosophy behind the system is to assist program directors and not replace their judgment with a black box system. One key design criteria is that Revaide offers suggestions that may be accepted or declined individually. In this section, we introduce Revaide, its tasks and solution, and evaluate the utility of using Revaide. We introduce a measure to evaluate how well the expertise of a group of reviewers is suited for a proposal. Following the discussion of the key components of Revaide in this section, we will report on the experiences using the algorithm.

3.1 Representing Proposals

Proposals are submitted to NSF in PDF form. Revaide converts the proposals to ASCII and represents proposals in the standard TF-IDF vector space [10] as term vectors in the space of all words in the document collection. The entire proposal is used including the references and resume of the investigator. One simple use of Revaide is to annotate spreadsheets of proposals with the 20 terms with highest TF-IDF weights. These keywords are often more informative to program directors than the title to determine what a proposal is about. While early versions of Revaide used stemming [11] to convert words to root forms, we found that stemming reduced the human comprehensibility of the resulting term vector representation. Experience showed that using stemming did not increase the quality of the suggestions made by Revaide. Therefore, we no longer use stemming.

One other enhancement also increased the comprehensibility of the resulting term representation. We augmented the stoplist of items that should not be used as keywords. While most stoplists include common words such as articles and prepositions, we augmented the stoplist to include words that appeared in proposals that were not descriptive of the proposal content, including the e-mail addresses of PIs and the name and city of the university. These words frequently occur within a few proposals and not in many others giving them high TF-IDF weights, but they confused program directors when used as keywords and degraded the quality of Revaide's suggestions.

An example will illustrate the representation used by Revaide for one proposal. The terms with the highest weights and their weights were `image: 0.031, judgments: 0.028, feedback: 0.027, relevance: 0.026, multimodal: 0.020, retrieval: 0.019,` and `preference: 0.017`. To preserve the privacy of the submitter, we cannot provide the title or abstract, but we find that the automatically extracted keywords do indeed provide a compact representation that makes sense to program directors and provides a basis to assist reviewers.

3.2 Representing Reviewer Expertise

Revaide represents the expertise of a reviewer with the TF-IDF representation of the proposals they have submitted to NSF in the past. While it would be possible to use published papers of authors downloaded from Citeseer [12] or Google Scholar as measures of expertise, there are advantages in using NSF proposals in a practical system deployed at NSF. First, all proposals are similar in style and length. These conditions are

ideal for keyword extraction with TF-IDF. Second, the proposals have a variety of meta-data that is useful in other aspects of the process. This meta-data includes the PI's name, e-mail address and other contact information, and an NSF ID for the PI's university. This meta-data simplifies contacting the PI and checking for conflicts of interest between proposals and reviewers. Third, NSF has a strong preference for using people with PH.D. degrees as reviewers, and one can't distinguish new graduate students from professors on published papers. By using people who have submitted to NSF as a reviewer pool, this problem is avoided since those eligible to apply to NSF are eligible to review. Finally, using proposals also avoids the problem of disambiguating people with common names. Finally, it automatically creates a large pool of potential reviewers. A disadvantage of this approach is that it does include people who do not submit to NSF, such as researchers from industry or from outside the US. Of course, program directors may identify such people through usual means, such as checking the editorial board of journals and program committees of conferences.

In practice, we restrict Revaide's pool of reviewers to those authors of proposals that have been judged as "fundable" by the review process to insure that the reviewers were thought by their peers to have expertise in the area. We also leave out proposals with more than one author so that it is clear who has the expertise in a proposal. When more than one past proposal is available for a given author, all of the proposals are combined by adding and then re-normalizing the term vectors to form a model of the expertise. The example proposal representation in the previous section would also serve as the expertise representation of the author that submitted the proposal.

3.3 Cluster Checking

The first task we consider is assisting groups of program directors to form panels. The most help is needed in large competitions where 500-1500 proposals may be submitted at a time. NSF's system produces a spreadsheet that includes columns containing information such as the author's name, institution, the title of the proposal and links to the abstract and the PDF of the entire proposal. Teams of program directors manually sort these proposals first into general areas and then into panels of 20-30 proposals. Due to the short time and large number of proposals, it is possible that a proposal could be put into a panel with only a loose relationship to the majority of the proposals. Due to the distributed nature of the work, it is also possible that no one claims responsibility for a proposal.

As described earlier, attempts to use automated clustering failed at this task when program directors didn't accept the results of the clustering system. Instead of automatically clustering, Revaide checks the clusters produced by program directors for coherence and suggests improvements. In addition, Revaide suggests panels for "orphan" proposals that are not assigned to a panel. Furthermore, before program directors form panels, the spreadsheet they use is augmented first with the terms that have the highest TF-IDF weights[1] of each proposal.

The first step in cluster checking is to form a representation of the important terms of the cluster. In Revaide, this is done by finding the centroid [10] of the proposals that are in each cluster, essentially creating a term vector for each cluster that is the "average" of the term vectors of the proposals. Next, the cosine similarity [10] is found between each proposal's term vector and each cluster's term vector. REVAIDE produces a summary of the important terms in each cluster. These terms are chosen based on a weighted TF/IDF score. The example below illustrates such a summary. In addition to the TF-IDF weight of each term[2], Revaide also prints out the number of proposals in the cluster that contain each term.

```
The top 20 terms of panel ROB are: robot: 0.267(in
24/28) sensor: 0.203 (in 28/28) vehicl: 0.144 (in
22/28) imag: 0.114 (in 22/28) motion: 0.107 (in
22/28) intellig: 0.104076 (in 25/28) mobil: 0.102
(in 23/28) agent: 0.094 (in 18/28) autom: 0.091
(in 25/28) movement: 0.078 (in 17/28) action:
0.077 (in 23/28) sens: 0.068554 (in 26/28)
autonom: 0.068 (in 25/28) self: 0.068 (in 21/28)
assembl: 0.064 (in 18/28)
```

If the most similar cluster to a proposal is not the cluster to which a proposal has been assigned, that is a sign that a proposal is potentially in the wrong cluster. Such discrepancies are pointed out to the program director with a suggestion to move the proposal to another panel. Below, the output of cluster checking is shown omitting any identifying information from the output.

```
The top 20 terms of panel CIP-SC are: sensor:
0.355 (in 31/32) vehicl: 0.2493 (in 22/32)
wireless: 0.178 (in 29/32) monitor: 0.157 (in
32/32) node: 0.147 (in 27/32) transport: 0.136 (in
29/32) devic: 0.132 (in 30/32) signal: 0.129 (in
30/32) traffic: 0.129 (in 22/32) grid: 0.119 (in
21/32) event: 0.116937 (in 32/32) energi: 0.107
(in 29/32) transmiss: 0.105 (in 25/32) protocol:
0.103 (in 27/32) flow: 0.103 (in 26/32) layer:
0.100317 (in 25/32) mobil: 0.100 (in 26/32) rout:
0.096 (in 23/32) agent: 0.092 (in 17/32) safeti:
0.091 (in 25/32)

    Panel DSP is a better match for proposal
NSF04XXXX1 than cluster CIP-SC.
```

In our experience, Revaide recommends a better panel for approximately 5% of the proposals. We have received comments from program directors that include, "Thanks, I don't know how I overlooked that," in response to Revaide's cluster checking. Often, Revaide finds a better panel that is a matter of emphasis within a proposal, e.g., determining that a proposal will make a contribution to computer vision for astronomical applications as opposed to making a contribution to astronomy using existing computer vision techniques.

A special case of the cluster checking is when a proposal has not been put into any panel. This can occur if no member of the distributed team of program directors has identified that a proposal falls within the scope of the panel. In this case, the panel that is most similar to the proposal is found, together with the next three, as determined by cosine similarity between the orphan

[1] Although the weights are not included, the terms are ordered by weight.

[2] This example shows an earlier version of Revaide that used stemming [9], perhaps also illustrating why we turn stemming off in later versions.

proposal vector and the centroids of the panels. The output below illustrates this process.

```
Top terms of NSF04XXXX2 are: sensor, wireless,
hierarch, node, channel, energi, signal, rout,
alloc, poor, radio, path.

Cluster WON2 is the best match for NSF04XXXX2
Alternate panels for orphan: Cluster WON3, Cluster
WON, Cluster CIP-SC
```

This algorithm for assigning an orphan proposal to a panel is related to Rocchio's algorithm for text classification [13]. The MailCAT system [14] used the idea of displaying a few possible folders for filing e-mail messages analogous to the way that Revaide finds a few possible panels. In both cases, the idea is to cope with the reality that text classification is not 100% accurate while providing benefit by focusing a person on a few possibilities out of the many that are available.

3.4 Proposal Classification

Revaide has the capability of performing text classification. The algorithm for recommending a panel for orphan proposals is one use of text classification. This section describes another use: performing an initial assignment of proposals to program directors. Recall that teams of program directors sort through proposals to identify the major area before further subdividing into panels. Revaide can use a text classification algorithm to perform this initial sort. In this case, the training data is the previous year's proposals and the class is the name of the program officer who organized the review panel the previous year. That is, the goal of the text classification is to find the person who will assume initial ownership of this year's proposals based upon their responsibilities in the prior year[3]. The initial program director either places a proposal into a panel they will organize or passes it to another program officer who is a better match for the proposal.

In a study using cross validation of the 2004 proposals submitted to Information and Intelligent Systems, the classification accuracy was 80.9%. This clearly is not good enough for a fully automated system. However, it provides tremendous benefits within the existing workflow. For example, rather than having 10 people each sort through 1000 proposals to find proposals of interest, each person is initially assigned approximately 100 by the text classification algorithm. Each program director then reviews those 100 proposals and on average needs to find a better program director for 20 proposals. This has greatly reduced the amount of effort required to identify the best program officer for each proposal.

Revaide assists with each step of the panel formation process, first by recommending an initial program officer. Once the final program officer is decided upon for each proposal[4], the proposals are manually subdivided into panels and the panels are checked for coherence. A proposal might be "orphaned" if it was initially misrouted or delayed or if no program officer claimed responsibility in the initial sort. It is then assigned to a program director in the panel checking stage. In the next section, we discuss assisting in the assignment of reviewers to proposals.

[3] Because many program officers are rotators who spend a short time at NSF, the initial assignment may be based upon the program officer's predecessor's proposals.

[4] This overview slightly simplifies the process. Two program directors may decide to hold a joint panel, e.g., at the intersection of databases and artificial intelligence.

3.5 Assigning Reviewers

The most straightforward way to choose N reviewers for a proposal would simply be to select the N authors of the previous proposals that are the most similar to the new proposal to be reviewed. This is the approach that has been used in some past efforts at automatic reviewer assignments (e.g., [15]). This approach does a fair job but has some important drawbacks. The main problem occurs when a proposal has more than one topic (a fairly common occurrence) and one topic dominates the match with other proposals. This leads to a set of reviewers that all have the same expertise, often leaving other topics in the target document uncovered. For example, consider a document about data mining using Gaussian mixture models to predict outcomes in a medical context. Ideally you would want a mix of reviewer expertise for this document: general data mining, the specific technique being used, as well as the field it is being applied to. Simply selecting reviewers by document similarity would tend to select reviewers who matched most closely to the primary topic of the paper (as determined by the TF-IDF weighting process) possibly failing to select any reviewers at all for an important secondary topic of the document.

To solve this problem, we approach the task slightly differently. Instead of finding the N closest matches for the target proposal, we look for the set of N proposals that together best match the target document. We define a measure that indicates the degree of the overlap between the terms in a proposal vector and a set of expertise vectors.

We represent a proposal as a normalized weighted vector of terms:

$$\vec{P} = \langle p_1, ..., p_n \rangle .$$

Similarly, we represent a reviewer's expertise as a normalized vector:

$$\vec{E} = \langle e_1, ..., e_n \rangle .$$

Where p_i is the weight of term i in a proposal and r_i is the weight of term i in a reviewer's expertise vector. We define a residual term vector to represent the relevant terms in the proposal that are not in the expertise of the reviewer. The weight of each of the residual term vectors is the difference between the weight in the proposal and expertise vector with a minimum of 0.

$$\vec{R} = \langle \max(0, p_1 - e_1), ..., \max(0, p_n - e_n) \rangle .$$

More generally, there is typically more than one reviewer and we define the residual term vector when there are k reviewers to be

$$\vec{R} = \left\langle \max(0, p_1 - \varepsilon \sum_i^k e_{1,i}), ..., \max(0, p_n - \varepsilon \sum_i^k e_{n,i}) \right\rangle$$

where ε controls the amount of overlap in expertise desired in the reviewers. If ε is 1, then it is sufficient to have one reviewer whose expertise about a term equals the importance of that term to the proposal. If ε is 0.5, then two reviewers should have expertise on every term in the proposal.

To compare alternative sets of reviewers and alternative approaches for finding reviewers we define a measure called Sum of Residual Term Weight (SRTW) to be:

$$SRTM = \sum_i \max(0, p_i - \varepsilon \sum_j^k e_{i,j})$$

We define the goal of assigning reviewers to be finding a set of reviewers that reduces the sum of residual terms to be 0 and the one set of reviewers is better suited to review a proposal than a number if that set of reviewers has a lower SRTM.

We have implemented a hill-climbing search algorithm to find a set of reviewers for each proposal. We start by finding the "best" reviewer and then iteratively select another reviewer until *N* are found. At each step, the reviewer that minimizes SRTM is selected. This iterative process will reduce the residual term weight. The residual term weight with no reviewers is 1.0 (since we work with normalized vectors). As each reviewer is selected, the term weights are adjusted according to the expertise of the reviewer. By subtracting the expertise vector from the document vector, the sum of residual term weights in the document vector will decrease.

Table 1 shows a trace of how the residual term weights are reduced by selecting reviewers. The row shows the most important terms in the term vector of a proposal and the remaining table shows the residual term vector after subtracting each expertise vector (with ε =0.5). A proposal on relevance feedback for image retrieval is to be reviewed. The first reviewer selected is an expert on image retrieval. Once that contribution has been accounted for, we see terms such as "image" have a lower term weight, reducing their impact on finding the next reviewer. The second reviewer has greater experience in image relevance judgments and these terms are reduced in weight. The process repeats until the desired number of reviewers are found.

An important aspect of this algorithm is that it can easily be started from a partial solution. This turns out to be a very useful property when considering the context in which the system is used. By allowing program directors to provide a partial solution that will then guide the system towards its final solution, we allow the experts to use Revaide as a tool to assist them to complete their jobs rather than using it to completely replace their judgments.

Another benefit of SRTM is that it may be used to determine whether a proposal has reviewers with adequate expertise. When there is no reviewer with expertise on an aspect of the proposal, the value of SRTM for that proposal would be higher than others. This might occur if the pool of reviewers is too small or if the proposal is on a topic that had not received submissions in the past. One way to find a reviewer in this case is to use the terms with the highest residual weights as query to a specialized search engine such as Google Scholar. Figure 1 illustrates the results of Google Scholar using the three terms with the highest residual weights from table 1. Although Google Scholar is not integrated with the entire workflow of Revaide (e.g., it doesn't identify the e-mail address and affiliation of the authors), it still provides a useful way of recommending reviewers.

As we have described assigning reviewers and SRTM so far, the goal is to find a set of reviewers for a single proposal. However, at NSF panels, reviewers typically review several proposals in a panel. Revaide can easily be used to recommend panelists for a set of proposals. Recall that in cluster checking, Revaide creates a term vector for each panel that is the centroids of the proposals in the panel. This cluster term vector represents the terms that are most important to the proposals in the panel. To invite panelists, Revaide simply finds the panelists whose expertise best reduces the SRTM of the centroid of the panel. In this case, rather than assigning four reviewers to a proposal, 12 reviewers might be selected for a panel of 24 proposals. A lower value of ε is used when selecting reviewers for a panel. For example, a value of 0.2 will bias Revaide toward finding 5 reviewers with expertise in the major areas. In reality not everyone who is invited to review actually agrees to. Therefore, we typically ask 20 with the expectation of getting a 50% yield. Once many reviewers have accepted, Revaide can be run again using the confirmed reviewers as a starting point and finding reviewers to complement their expertise.

Proposal	image	0.031	judgments	0.028	feedback	0.027	relevance	0.026	multimodal	0.020
After Reviewer 1	judgments	0.280	feedback	0.023	relevance	0.022	image	0.020	multimodal	0.020
After Reviewer 2	feedback	0.023	image	0.020	multimodal	0.020	preference	0.016	judgments	0.015
After Reviewer 3	feedback	0.020	multimodal	0.019	preference	0.016	judgments	0.015	solicit	0.011

Table 1. A trace of the residual term vectors after assigning reviewers.

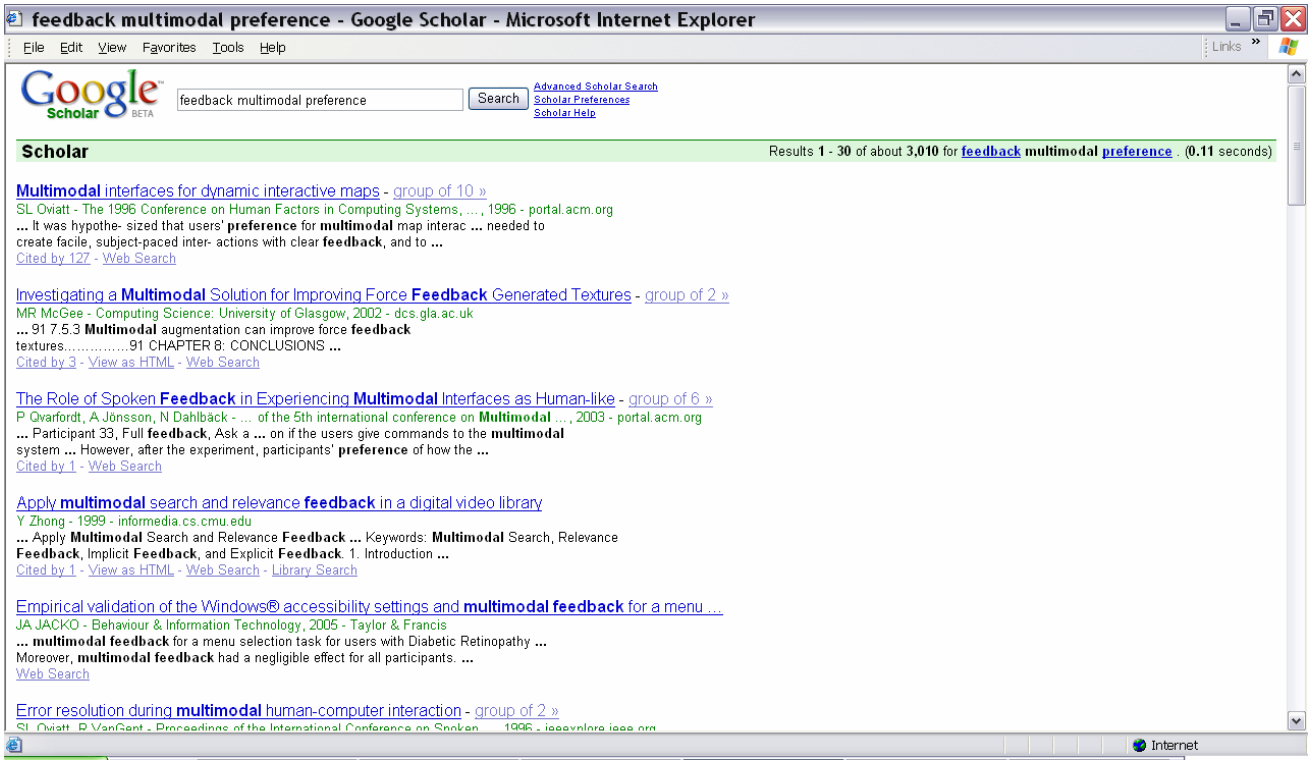

Figure 1. Using the terms with the highest residual weights as a query to Google Scholar.

3.6 Integration with NSF Systems

Over the course of three years, Revaide has transitioned from a set of utilities run remotely at UC Irvine to a prototype deployed within the CISE directorate at NSF. In the first year, Revaide was used only in parallel with existing systems and only had the ability to find the N most similar to the previous proposals. This showed the promise of the technique, but the utility was diminished by the loose coupling with NSF's systems and workflow. For example, the first version of Revaide could find the closest proposals but didn't have the meta-data to automatically associate a title, author and the author's contact info with the proposal. These were later manually added to spreadsheets. At this point, Revaide also could not perform conflict of interest checking and would recommend a reviewer from the same institution as the proposal's author, a violation of NSF's policies. Furthermore, it would even recommend that an author review a proposal by that author based on the author's prior proposal.

The inability to perform conflict of interest checking also lead to a serendipitous finding: Revaide could be used to spot probable revisions of prior year's proposals. A new proposal was typically much more similar to a prior version of that same proposal than any other previous proposal. In a small number of instances, we found a proposal that was "too similar" to a previously submitted funded proposal, a clear violation of NSF policy. In other cases, we found too much similarity to a proposal currently under review at another part of NSF, another violation of NSF policy. Revaide merely alerted program directors to these possible violations. Program directors decided whether there was a probable violation, which in the case of resubmission of funded work was then investigated by NSF's Inspector General. While not emphasized in this paper, Revaide still retains these capabilities.

In the second year, Revaide was re-engineered to accept meta-data on proposals so that it can do conflict of interest checking and produce output that includes names and contact info of potential reviewers. Revaide also has access to the previous summary reviews rankings and funding decisions on proposals so only those whose expertise has been validated by the peer-review process were considered as potential reviewers. Revaide also helped NSF achieve its diversity goals by including some demographic data on reviewers. If a proposal or panel did not include female reviewers, reviewers from underrepresented groups, or reviewers from EPSCOR states (i.e, states that do receive much federal research funding), additional reviewers were recommended from these groups, insuring that proposals are not just reviewed by an "old boys club" and that a diverse group of investigators has the benefits of participating in funding decisions.

At this point, Revaide was also changed from using cosine similarity for selecting reviewers to using the residual term weight approach described earlier. This was done in response to the problem of cosine similarity on interdisciplinary proposals leading to recommending proposals only from a single discipline. However, Revaide was still used remotely from California when the results and data were in Arlington, VA. Delays caused by computation to converted proposals from PDF to ASCII and index proposals, transferring gigabytes of data, minor errors in the meta-

data[5] and the difference in time zone typically resulted in a two- or three-day turnaround in running the system. Nonetheless, the system illustrated its utility by finding proposals that were obviously assigned to the wrong panel and suggesting qualified reviewers that were overlooked by program officers.

In the third year, encouraged by the results of the second year, NSF purchased the appropriate computer equipment and ran Revaide in house. Furthermore, this enabled tighter integration with NSF's databases, e.g., proposal meta-data was accepted in the exact format produced by NSF's systems rather than requiring an intermediate step of manually reformatting the data. Furthermore, processes were put into place to accurately record and maintain the data used by Revaide. This reduced the time required to get results from Revaide from a few days to a few hours. Plans are now being evaluated to have a contractor fully integrate Revaide with NSF's internal systems and build a web interface to Revaide. In the next section, we summarize the experiences in the third year of using Revaide.

4. Evaluation and Lessons Learned

In this section, we report on two experiments that empirically evaluate the utility of the residual term weight approach in assigning reviewers. We also report on the lessons we have learned in deploying Revaide in the government context.

4.1 Selecting Reviewers for Proposals

We consider selecting reviewers independently for proposals. In particular, for each proposal submitted to the 2004 Information Technology Research program in the division of Information and Intelligent Systems, a total of approximately 1,500, we compare finding the three closest reviewers as determined by cosine similarity to the three that best reduce SRTM. In each case the pool of reviewers is the people who submitted proposals to the division in the prior three years. The average sum of residual term weights (with $\varepsilon = 0.5$) decreases from 0.636 for the three closest to 0.569 for Revaide's approach. Note that this average does not tell the entire story. For more than five percent of the proposals, perhaps the most interdisciplinary proposals, there was a difference of greater than 0.15 in the sum of residual terms, demonstrating the importance of finding a set of reviewers with complementary expertise. Of course, it may seem like a tautology to show that a system that attempts to minimize SRTM has a lower SRTM. However, this also puts a number behind the intuition that similarity alone isn't sufficient for finding reviewers for interdisciplinary proposals.

4.2 Selecting Panelists

Here we consider alternative strategies for selecting panelists for two panels of proposals submitted to the 2005 Universal Access solicitation. For each panel, we compare using six randomly selected people funded in the prior year as reviewers (analogous to the common conference practice of inviting a program committee before papers are submitted or the NIH practice of having a standing panel), the six reviewers closest to the centroid of the proposals in the panel, and the six reviewers that best reduces the sum of residual term weights from the centroid of the panel (with $\varepsilon = 0.5$). Once the panelists are selected, then four panelists are assigned to proposals by Revaide using SRTM with $\varepsilon = 0.5$. The mean residual term weight under these conditions is shown in Table 2. It is apparent from this figure that both approaches that examine the proposals to select panelists have a benefit over picking panelists who are experts in the general subject area (Standing Panel). Furthermore, selecting panelists with complementary expertise (SRTM) has an advantage of selecting panelists whose expertise is most similar to the central theme of the proposals (Similarity).

Standing Panel	Similarity	SRTM
0.783	0.662	0.521

Table 2. Sum of residual term weights with three alternative approaches to selecting panelists.

4.3 Experiences and Lessons Learned

In the third year of Reviade's development, it was relied upon heavily in the Information and Intelligent Systems (IIS) with the evaluation of a competition that received slightly over 1000 proposals and several other competitions with 200-500 proposals. It was also used in competitions in the Computer and Communication Foundations Division and a Computer and Information Science and Engineering interdisciplinary competition. In IIS, Revaide was relied upon to initially dispatch proposals to program officers, to check panels for coherence, to find panels for orphan proposals, and to recommend reviewers for most panels. Some summary results and lessons learned included:

- Revaide greatly reduced the time required to form panels. In one competition, this was essentially completed in two weeks compared to approximately six weeks for a smaller competition that didn't use Revaide.
- Revaide increased the pool of reviewers beyond those normally called upon by program officers. While some members of the community had been called upon repeatedly, others with similar expertise had been overlooked. In many cases, people who had not reviewed before agreed to review nearly immediately when asked, while those frequently called upon are more reluctant to serve another time.
- Revaide greatly reduced the amount of time to find reviewers for panels. One program officer reported it took a week rather than a month to finalize two panels.
- One program officer after using Revaide asked panelists to select which proposals they were most interested in reviewing. Frequently, the most desired proposals by the panelists were indeed the proposals that led to the reviewer's invitation.
- In one case, a program officer thought the reviewers suggested had expertise that wasn't relevant to the proposal. However, after the

[5] For example, an unexpected carriage return in a proposal title resulted in an ill-formed tab separated file.

program director read the proposal and not just the abstract, it was found that the proposal did indeed touch on all the topics for which the reviewers were selected.

We believe there are several factors responsible for the success of Revaide:

1. Rapid turnaround is quite important in getting the system accepted. Even a day's delay at a critical time cannot be tolerated. This implies a close integration between the existing databases and processes and the reviewer recommendation system.
2. The system was put within the existing workflow of the organization. Other alternatives explored, such as automated clustering, redefined the roles of people in the organization.
3. The system is not a black box that produces a solution but rather provides a basis for its recommendations in terms of automatically derived keywords. For example, the keywords for an AI panel were `logic, reasoning, inference, planning, action, reinforcement, game, variables, agent, classifiers, planners, inhabitant, decision, graph, motifs, probabilistic, propositional, and rule.` Similarly, a confusion matrices for proposal assignment convinced people that the solution was much better than chance but not omniscient.
4. Each recommendation was subject to validation and could be ignored independently of others. Furthermore, the system was designed to supplement the capabilities of program officers and serves as "another set of eyes" to focus program officers' attention on potential improvements. This also means that imperfect technology (e.g., a classifier with 80% accuracy) can still be beneficial in an organization that has higher standards.

5. Related Work

Revaide addresses the challenge of assigning reviewers (cf [16]. The main technical contribution of Revaide is the use of the sum of residual term weights measure in reviewer assignment. In implementation, we used a well established but simple document model: TF-IDF weights on words. The residual term weight approach is independent of the document model and could just as easily be used with hand-selected keywords, LSI terms (e.g., [17]), or author and topic models (e.g., [18] and [19]). We did indeed consider using LSI in Revaide but have decided against it because LSI doesn't produce terms that are easily understood by people and can easily be used as queries for a text search engine. If we had access to only abstracts, LSI might prove particularly useful, but in longer documents such as full proposals, the benefits of LSI are less dramatic and not worth the lack of comprehensibility in this application.

Our goal with residual term weight is to represent the terms in a proposal left uncovered by a partial set of reviewers. One approach to this problem is Maximal Marginal Relevance (MMR) [20]. MMR provides a way to adjust the ranking (or re-rank) the retrieved results of a query to produce a diverse set of documents. MMR is based on comparing retrieved documents to each other in order to select a diverse group. In contrast, SRTW is a more focused measure that seeks to achieve diversity to satisy the goal of covering terms in a source document.

6. Future Work

NSF is evaluating plans to more closely integrate Revaide into its data infrastructure and workflow. Revaide would then be able to directly access NSF databases rather than going through intermediate files. We plan on conducting further research on the general topic of reviewer assignment. In particular, we are exploring approaches that will balance reviewer assignments across reviewers on a panel. We believe such an approach will need to consider the residual term weights, the number of proposals assigned to a reviewer, and the distance between a proposal and a reviewer's expertise (because in our experience reviewers have a strong aversion to reviewing proposals outside their expertise).

7. Conclusions

We have described Revaide, an emerging application deployed at NSF as prototype. While much of Revaide relies upon existing technology for representing documents, Revaide makes two contributions to the practice of text mining. First, we have defined a new measure of similarity suited for insuring that expertise is found for all aspects of a proposal to be reviewed. Second, we have shown that text mining technology can be deployed to augment rather than replace human judgment.

8. ACKNOWLEDGMENTS

Thanks to all at NSF who were instrumental in the development and deployment of Revaide. Feedback of early users helped in the design of later versions. Many also helped navigate the approval.

9. REFERENCES

[1] Willett, P. (1998). Recent Trends in Hierarchic Document Clustering: A Critical Review, Information Processing and Management, 24(5), 577-597.

[2] Larsen , B. and Chinatsu A. (1999). Fast and effective text mining using linear-time document clustering, *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 16-22.

[3] Hopcroft, J., Khan, O., Kulis, B. & Selman, B. Tracking evolving communities in large linked networks.*Proc. Natl Acad. Sci. USA* 101(Suppl.1), 5249-5253

[4] Bradley, P., Bennett, P and. Demiriz., A. (2000) *Constrained k-means clustering*. Technical report, MSR-TR-2000-6 5Microsoft Research.

[5] Banerjee , A. &G hosh,, J. (2002). *Frequency Sensitive Competitive Learning for Clustering on High-dimensional*

Hypersphere, International Joint Conference on Neural Networks (IJCNN), pp. 1590-95.

[6] Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. (1987): The vocabulary problem in human-system communication. Commun. ACM 30. 964-971

[7] Ding, W., and Marchionini, G. *A Study on Video Browsing Strategies.* Technical Report UMIACS-TR-97-40, University of Maryland, College Park, MD, 1997.

[8] Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. (1987): The vocabulary problem in human-system communication. Communications of the ACM 30. 964-971

[9] van de Stadt, R. (2000). CyberChair, an Online Submission and Reviewing System or: A Program Chair's Best Friend, WWW9.

[10] Salton, G., & McGill, MJ (1983). Introduction to modern information retrieval. NY: McGraw-Hill

[11] Porter, M.F., (1980), An algorithm for suffix stripping, Program, 14(3) :130-137

[12] Giles, C. Bollacker, K., Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. Third ACM Conference on Digital Libraries, pp. 89-98, 1998.

[13] Rocchio, J. (1971) *Relevance feedback in information retrieval,* in . The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall Inc., pg 313-323.

[14] Segal, R and Kephart, J (1999).. MailCat: An Intelligent Assistant for Organizing E-Mail. In *Proceedings of the Third International Conference on Autonomous Agents*.

[15] Basu, C., Hirsh, H., Cohen, W., and Nevill-Manning, C., (1999). *Recommending Papers by Mining the Web*, Proc. IJCAI Workshops on Learning About Users and Machine Learning for Information Filtering, IJCAI 99, Stockholm, Sweden.

[16] Geller, J. and Scherl, R., 1997 *Challenge: Technology for Automated Reviewer Selection*, IJCAI 1997 55-61

[17] Dumais, S., Nielsen, J. (1992, *Automating the Assignment of Submitted Manuscripts to Reviewers* Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval: 233-244

[18] Steyvers, M., Smyth, P., Griffiths, T. (2004) *Probabilistic Author-Topic Models for Information Discovery* KDD'04, Seattle, Washington USA.

[19] Mann, G., Mimno, D. and McCallum, A (in press). *Bibliometric Impact Measures Leveraging Topic Analysis*. Joint Conference on Digital Libraries (JCDL).

[20] Carbonell, J. and Goldstein, J (1998). *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*, SIGIR'98, Melbourne Australia