Multi-Aspect Expertise Matching for Review Assignment

Maryam Karimzadehgan  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801  
mkarimz2@uiuc.edu

ChengXiang Zhai  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801  
czhai@cs.uiuc.edu

Geneva Belford  
Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801  
belford@cs.uiuc.edu

ABSTRACT

Review assignment is a common task that many people such as conference organizers, journal editors, and grant administrators would have to do routinely. As a computational problem, it involves matching a set of candidate reviewers with a paper or proposal to be reviewed. A common deficiency of all existing work on solving this problem is that they do not consider the multiple aspects of topics or expertise and all match the entire document to be reviewed with the overall expertise of a reviewer. As a result, if a document contains multiple subtopics, which often happens, existing methods would not attempt to assign reviewers to cover all the subtopics; instead, it is quite possible that all the assigned reviewers would cover the major subtopic quite well, but not covering any other subtopic. In this paper, we study how to model multiple aspects of expertise and assign reviewers so that they together can cover all subtopics in the document well. We propose three general strategies for solving this problem and propose new evaluation measures for this task. We also create a multi-aspect review assignment test set using ACM SIGIR publications. Experiment results on this data set show that the proposed methods are effective for assigning reviewers to cover all topical aspects of a document.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms
Algorithms, Experimentation

Keywords
Topic models, Expert retrieval, Review assignment, Evaluation Metrics

1. INTRODUCTION

Review assignment is a common task that many people such as conference organizers, journal editors, and grant administrators would have to do routinely. The task usually has to do with assigning a certain number of experts to review a research paper or a grant proposal and judge its quality. Since the assignment must be based on both the content of the paper or proposal and knowledge about the expertise of all candidate reviewers, it is not an easy task, and it is quite challenging to optimize all the assignments. Currently such assignments are done manually by editors or conference organizers. Some conference systems support bidding on papers by reviewers, shifting some of the assignment task to the reviewers. But it is still mostly a time-consuming manual process. Another deficiency of manual assignment is that the assignment may be biased. For example, in bidding, a popular paper may attract more bids than a non-popular paper. Also, the matching of expertise may be more accurate for reviewers whose expertise is largely known to an editor or a conference program chair, but less accurate for reviewers whom the assigning person is not very familiar with. For these reasons, it is highly desirable to develop techniques to automatically do this task. Indeed, recently, a prototype system for reviewing proposals has been developed at National Science Foundation (NSF), which has been shown to improve productivity [9].

As a research problem, automatic review assignment appears to have been first studied by Dumais and Nielsen [6], and later studied by a few other researchers [2, 9, 13]. (For a detailed review of the previous work, please refer to Section 6.) In most existing works, the problem is cast as a retrieval problem in which the query is a paper or proposal to be reviewed and an expert is represented as a text document (e.g., based on his/her publications). Then, top N documents are taken as assignments.

One common deficiency of all the existing works is that a paper is matched as a whole unit without considering multiple aspects of a paper. As a result, if a document contains multiple subtopics, existing methods would not attempt to assign reviewers to cover all the subtopics; instead, it is quite possible that all the assigned reviewers would cover the major subtopic quite well, but not covering any other subtopic. Since a paper or proposal often contains multiple subtopics, such simplistic view of matching often leads to non-optimal assignments. For example, if a paper is about applying machine learning techniques to develop new retrieval models for...

1 Aspect or subtopic will be used exchangeably throughout the paper.
Web search, ideally it should be reviewed by reviewers with expertise to cover three subtopics, i.e., “machine learning,” “retrieval models,” and “Web search.” However the existing methods may assign three reviewers that are all experts on machine learning but do not know well about Web search.

In this paper, we study how to match reviewers with a paper based on matching of multiple aspects of expertise so that the assigned reviewers would not only have the required expertise to review a paper but also can cover all the aspects of a paper in a complementary manner. This is a challenging task because we have no knowledge in advance about the aspects covered in a paper and it is also unclear how we should select reviewers to maximize their coverage of all the aspects.

We propose three general strategies for multi-aspect review assignment: the first one is Redundancy Removal in which we model both relevance and redundancy of reviewers’ expertise and attempt to achieve broad coverage through avoiding assigning reviewers with similar expertise to other assigned reviewers. In other words, we hope to increase the aspect coverage indirectly through eliminating the redundancy among reviewers. The second method is Reviewer Aspect Modeling, in which we improve the aspect coverage more directly through modeling the possible aspects in reviewers’ profiles with mixture language models such as PLSA [10]. With the learned expertise/topic aspects, we can represent both a reviewer and a paper or proposal to be reviewed in the space of these aspects and match them at the level of aspect representation. The third method is Paper Aspect Modeling in which we find aspects in a submitted paper. Our general idea is to partition the words in the paper into the same number of subsets as the number of reviewers we would like to assign, and we can then use each subset as a query to retrieve a reviewer. We further propose two different ways to implement this general idea, leading to two specific methods: The first method is to partition the words through word clustering based on the mutual information similarity between words. Each cluster of words can then be used to retrieve a reviewer. The second method is to segment the paper based on similarity between sentences and group adjacent sentences that are similar to each other together as a segment. Each segment can then be used to retrieve one reviewer.

Evaluating the review assignment task as we defined (i.e., matching multiple aspects) is challenging in two ways: First, it is unclear how to define appropriate metrics to measure assignment performance, and second, no existing test set is available. To solve these two problems, we propose four evaluation measures for this task and create a multi-aspect review assignment test set using ACM SIGIR publications from years 1971–2007. Experiment results show that all the proposed methods are effective for increasing the coverage of subtopics in review assignment and outperform a state of the art baseline retrieval method. The reviewer aspect modeling method using PLSA is shown to outperform both redundancy removal and paper aspect modeling method.

In summary, our contributions are as follows:

- We conduct the first study of review assignment based on matching multiple aspects of expertise.
- We propose three general strategies and specific instantiations of them for multi-aspect review assignment, so that the assigned reviewers would not only have the required expertise to review a paper but also can cover all the aspects of a paper in a complementary manner.
- We have created the first gold standard data which can help other researchers working on the same area.
- We have defined new evaluations measures to evaluate this new task.

All the proposed methods are general, thus they can be applied to all review assignment tasks, including related expertise matching tasks such as forming a thesis committee for a Ph.D. student (matching reviewers with a thesis) and forming a panel on a topic (matching candidate panelists with a topic description).

The rest of the paper is organized as follows. We first define the problem of multi-aspect review assignment in Section 2. We then propose several different methods for multi-aspect review assignment task in Section 3. In Section 4, we describe our experimental design and propose evaluation measures. In Section 5, we present our evaluation results. We discuss related work in Section 6 and conclude in Section 7.

2. MULTI-ASPECT REVIEW ASSIGNMENT

In many ways, the review assignment problem can be viewed as an expert retrieval problem in which the paper is a query and the reviewer is a document, and we want to retrieve k experts to review a paper. There are two basic tasks:

1. Obtain text representation for a reviewer
2. Match reviewers with the paper.

We assume that a reviewer can be represented by his/her research area and focus on studying how to perform the second subtask. In particular, we want to study how to match experts with papers based on multiple aspects of expertise. For us, thus, the goal is to retrieve k experts that are not only relevant to the paper but also are able to work together to cover all the aspects of a paper. There are several challenges in doing review assignment based on multiple aspects:

1. Ranking of experts must consider dependency between experts, which means we must go beyond the independent relevance assumption typically made in many information retrieval models.
2. The notion of multiple aspects is implicit, making it difficult to maximize the aspect coverage.
3. Evaluation of the task is also challenging as the traditional precision and recall measures no longer apply well.

3. METHODS FOR MULTI-ASPECT REVIEW ASSIGNMENT

In this section, we propose three general strategies for solving the multi-aspect review assignment problem: (1) Redundancy Removal, (2) Reviewer Aspect Modeling, and (3) Paper Aspect Modeling. Throughout this section, we will use $R = \{r_1, ..., r_N\}$ to denote the set of text documents representing the expertise of N reviewers, respectively, and $Q$ to denote the text of the paper to be reviewed.
3.1 Redundancy Removal

The idea of the redundancy removal approach is to reduce the redundancy among the assigned reviewers so as to indirectly maximize the coverage of topic aspects by the assigned reviewers. Specifically, when picking reviewers, we would consider both the relevance value and novelty value (or equivalently redundancy) of a reviewer relative to the other reviewers assigned to the same paper.

One way to implement this strategy is to use a greedy algorithm to gradually select a reviewer for a paper. In each step, given a set of previously picked reviewers, the next best reviewer is the one that is both relevant to the query topic and different from previously picked reviewers. This strategy is often called maximal marginal relevance (MMR) ranking [4]. With this strategy, after we've already selected some reviewers, we may potentially achieve better results.

3.2 Reviewer Aspect Modeling

Instead of attempting to indirectly cover multiple topic aspects of a paper to be reviewed through eliminating the redundancy among reviewers (e.g., MMR), an intuitively better strategy may be to directly cover topic aspects of the paper through explicitly modeling subtopics in the publications of reviewers with mixture language models. Indeed, even if we avoid redundancy, there is no guarantee that a new reviewer would cover the uncovered aspects, as the reviewer can be different from the already selected reviewers because of covering some non-relevant topics. Thus if we can somehow directly model the potential topic aspects and try to match a paper with reviewers based on their topic-aspect representation, we may potentially achieve better results.

To implement this idea, we assume that there is a space of $K$ topic aspects, each characterized by a unigram language model. Let $\tau = (\tau_1, ..., \tau_K)$ be a vector of topics. $\tau_i$ is a unigram language model and $p(w|\tau_i)$ is the probability of word $w$ according to topic $\tau_i$. Consider that we have a set of $m$ reviewer expertise documents (each representing a reviewer). We can then learn these $K$ topic aspects from this set of reviewer expertise documents using a topic model such as PLSA [10], which is described below.

3.2.1 Modeling Multiple Aspects (PLSA)

Let $R = \{r_1, ..., r_N\}$ be the set of reviewer expertise documents and $r_i$ be the document for reviewer $i$. The log-likelihood of the whole collection according to PLSA is:

$$\log p(R|\tau) = \sum_{i=1}^{N} \sum_{w \in V} c(w, r_i) \log \left( \sum_{a=1}^{K} p(a|\theta_i) p(w|\tau_a) \right)$$

where $V$ is the set of all the words in our vocabulary, $c(w, r_i)$ is the count of word $w$ in $r_i$, and $p(a|\theta_i)$ is the probability of selecting topic aspect $\tau_a$ for document $r_i$. Intuitively, $p(a|\theta_i)$ encodes the coverage of different topic aspects in the expertise document of reviewer $i$. $p(w|\tau_a)$ is the probability of word $w$ according to topic $\tau_a$.

We may use the EM algorithm to compute the maximum likelihood estimate of all the parameters including $\tau$ and $\theta_i$.

Once we obtain $\theta_i$ which is a distribution over all the possible topic aspects $p(a|\theta_i)$ ($a = 1, ..., K$), we can think of it as a new way to represent our document $r_i$ in terms of topic aspects. With a similar model to the one presented above, we can also estimate the subtopic coverage in our query paper $Q$ with the main difference being that we have $\tau$ already known so we only need to estimate $\theta_Q$ which would give us a distribution over subtopics for our query paper, i.e., $p(a|\theta_Q)$, where $a = 1, ..., K$.

We now describe how we match a paper with reviewers based on this new subtopic-based representation.

3.2.2 Matching based on Multiple Aspects

In order to match a paper with a reviewer based on their topic representations, we compute the following Kullback-Leibler (KL) divergence value:

$$D(\theta_Q||\theta_{r_i}^{k})$$

The KL-divergence between two distributions, $p$ and $q$, is defined as [5]:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$\theta_{r_i}^{k}$ is a weighted average of $\{\theta_a\}_{a=1}^{k}$ defined as follows:

$$p(a|\theta_{r_i}^{k}) = \frac{\sigma}{k-1} \sum_{a=1}^{k} p(a|\theta_a) + (1 - \sigma) p(a|\theta_k)$$

where $\sigma \in (0, 1]$ indicates how much redundancy we would like to model.

The intuition behind this is that $\theta_Q$ gives us a measure of which topic aspect is relevant; for example, a high $p(a|\theta_Q)$ indicates that topic aspect $a$ is relatively more relevant (than other aspects). As a result, if $\theta_Q$ assigns a high probability to some aspects, then we would like to cover these relevant aspects more than other less relevant ones. Thus, the best $r_k$ is the one that works together with $r_1, ..., r_{k-1}$ to achieve a coverage distribution most similar to the topic coverage given by the query. So, the parameter $\sigma$ controls how much we want to rely on the previously picked reviewers $r_1, ..., r_{k-1}$ to cover all topic aspects of the query. When $\sigma = 0$, the best $r_k$ is the one that covers all relevant aspects by itself and if
\( \sigma > 0 \), part of the coverage would be explained by previously seen reviewers, and the best \( r_k \) would be the best that covers those aspects which are not covered yet. Another parameter of this model is the number of aspects we want to have, which will be empirically set.

3.3 Paper Aspect Modeling

Instead of learning the subtopics/aspects from reviewer expertise documents, we can also use an alternative strategy of modeling aspects based on the paper to be reviewed. The general idea is the following: If we want to assign \( n \) reviewers, we can simply partition the words in the paper into \( n \) groups and use each to retrieve an expert. Let \( q_1, ..., q_n \) be queries based on the \( n \) partition groups, and \( r_j \) denote the text representation of the expertise of the \( j \)-th reviewer, then for each \( q_i \) we use the following language modeling approach (i.e., query likelihood) with Dirichlet prior smoothing [18] to retrieve the first top reviewer.

\[
p(q_i \mid r_j) = \prod_{w \in q_i} \frac{c(w, r_j) + \mu p(w|R)}{N_{r_j} + \mu}
\]

where \( N_{r_j} \) is the number of words in the reviewer representation \( r_j \), \( p(w|R) \) is a background language model estimated using the entire set of reviewer documents \( R \), and \( \mu \) is a smoothing parameter to be empirically set. We then merge the reviewers from all the clusters together. In case two clusters nominated the same reviewer, we resolve the conflict by allowing the reviewer to represent whichever cluster gives it a higher score.

We present two different ways to partition the words in a paper based on Mutual Information-based Clustering and Simple Discourse Analysis, respectively.

3.3.1 Mutual Information-based Clustering (MIC)

In this method, we use K-Means [12] algorithm to cluster words in the paper into \( k \) clusters (\( k = 3 \) in our experiments). We can then construct a query language model for each cluster and use it to retrieve a top-reviewer as described above.

Our idea is to cluster the words in a query paper into dissimilar categories, i.e., the resulting intracluster similarity is high but intercluster similarity is low. The similarity measure between words in the query is computed based on the Mutual Information [5] between each pair of words. Mutual Information measures the tendency of two words co-occurring in a collection. This similarity measure is the basis of the clustering algorithm; the similarity between word \( w \) and a cluster is computed as the mean of the similarity between word \( w \) and all the words in the cluster. We then, retrieve one reviewer for each cluster. In this case, we can ensure that reviewers with diverse expertise are retrieved (due to dissimilarity between clusters), so that these retrieved reviewers can work together to cover all topics of the query.

3.3.2 Simple Discourse Analysis (SDA)

Another way for finding aspects in the paper is to segment the paper into \( k \) parts. While in clustering each cluster may correspond to words scattering in the whole paper, in this simple discourse analysis approach, each segment corresponds to words occurring physically together in the paper. In general, any text segmentation method can be applied.

In our experiments, we used the following simple approach, which can be regarded as a version of the TextTiling approach proposed in [8]. Specifically, for partitioning a query into \( n \) parts, we first calculate the similarity between every two adjacent sentences based on Cosine Similarity of the corresponding TF-IDF weight vectors. We then iteratively segment the query abstract (or paper) at the boundaries where the similarity between two adjacent sentences is the lowest. Once we obtain \( n \) segments, similarly to the case of mutual information clustering, we can retrieve one reviewer for each segment according to the above language modeling approach.

4. EXPERIMENT DESIGN

In this section, we present our test data and measures used for evaluating our methods.

4.1 Data Set

Since there is not any data set available for review assignment task, we need to create our own data set. The data set in ”Enterprise Track” in TREC cannot be used for our task since we need a collection for matching reviewers with papers based on multiple aspects not only for retrieving experts/reviewers. So, to construct such a dataset, we have crawled abstract papers of ACM SIGIR proceedings from years 1971-2006 from the ACM digital library\(^2\). We have only considered abstract papers in conferences such as World Wide Web, SIGIR and CIKM in order to consider only IR researches for evaluating purposes. Then, we consider authors of these papers as prospective reviewers. For modeling reviewers’ expertise, we create a profile for each author in our pool by concatenation of all papers written by that specific author. If a paper has more than one author, we replicate that paper, one for each author. From the aforementioned dataset we further considered the authors having published more than three papers in these years so the number of prospective reviewers in our collection is reduced to 189.

We used the SIGIR 2007 papers to simulate papers that are to be reviewed. There are 73 papers with at least two aspects. We experimented with both abstracts and full papers.

Evaluating of such a system is very challenging. Since the actual assignments of reviewers to papers are confidential, we cannot use such data to evaluate the effectiveness of our methods. Even if we had such information, it would not necessarily indicate the best matching. As a result, we have to create a gold standard for evaluating our approaches. To achieve this purpose, an information retrieval expert identified 25 major subtopics based on the topic areas in the Call for Papers of ACM SIGIR in most recent five years and session titles in the recent ACM SIGIR conferences. The expert then reads the abstracts of all the 73 test papers and all the 189 expertise profiles, and assign relevant expertise/topic aspects to each paper and each reviewer. This serves as a gold standard to evaluate our methods. This data set is publicly available at http://timan.cs.uiuc.edu/data/review.html to facilitate further study of the multiple-aspect review assignment problem.

\(^2\)http://www.acm.org/dl
4.2 Evaluation Measures

While we can conveniently cast the multi-aspect review assignment problem as a retrieval problem, the traditional relevance-based precision and recall measures cannot be directly applied to measure performance because they are unable to reflect the coverage of multiple aspects in assigned reviewers.

Intuitively it is desirable to include reviewers covering many topic aspects so that they can collectively cover all aspects in the query. So, one way to account for this goal is to measure the number of different distinct topic aspects that are covered by \( n \) assigned reviewers as a function of aspects in the query. This measure tells us whether we can cover all aspects of the query. Consider a query with \( n_A \) topic aspects \( A_1, \ldots, A_{n_A} \) and let \( n_r \) denote the number of distinct topic aspects that these \( n \) assigned reviewers can cover. Coverage can be defined as the percentage of topic aspects covered by these \( n \) reviewers:

\[
\text{Coverage} \equiv \frac{n_r}{n_A} \quad (1)
\]

In addition to maximizing the coverage of topic aspects, we also would like to maximize the number of reviewers that can cover each aspect. Thus given the same level of coverage, we would prefer an assignment where each aspect is covered by as many reviewers as possible. Intuitively, this is related to the overall confidence of the assigned reviewers in reviewing each topic aspect of the “query paper.” We thus define a Confidence measure to capture the redundancy of reviewers in covering each aspect.

Formally, let \( A_1', \ldots, A_{n'} \) be the \( n_r \) distinct aspects covered by \( n \) retrieved reviewers and \( n_{A_i'} \) be the number of reviewers that can cover aspect \( A_i' \). The Confidence measure is defined as:

\[
\text{Confidence} \equiv \sum_{i=1}^{n_{A_i'}} \frac{n_{A_i'}}{n_r} \quad (2)
\]

Note that we normalize the confidence values with the number of covered topic aspects \( n_{A_i} \) such that a high confidence value may be obtained by intentionally covering fewer aspects. This is why confidence alone would not be so meaningful and it should be combined with the coverage. In this sense, the relation between coverage and confidence is similar to that between precision and recall. A perfect assignment should have both high coverage and high confidence.

One way to combine coverage and confidence is to normalize confidence over all aspects rather than just the covered aspects to compute an average confidence. Since a missed aspect would decrease the average confidence, it serves as a combination of coverage and confidence. Using the notations introduced earlier, the Average Confidence measure is defined as follows:

\[
\text{Average Confidence} \equiv \sum_{i=1}^{n_{A}} \frac{n_{A_i}}{n_{A}} \quad (3)
\]

We normalize all three measures by considering their corresponding optimal values that can be gained from our gold standard data set. Since measures are different; we describe algorithms for finding optimal values specifically for each measure. In general, finding the exact optimal values for these measures is NP-hard [16], so we opt to use greedy algorithms to compute an approximate value.

For finding the optimal Coverage, we use an MMR-based approach when selecting the next wide-coverage reviewer; i.e., we first pick the reviewer that covers the most number of aspects of the paper, then we take away the covered aspects and pick the next reviewer to cover as many of the remaining aspects as possible. For finding the optimal Confidence, we first pick the aspect, e.g. \( A_1 \), of the paper that most of the reviewers in the collection can cover it, we then select other reviewers that can cover aspect \( A_1 \) but do not cover any of the remaining aspects of the paper (to minimize \( n_r \)). For finding optimal Average Confidence, we select each reviewer to work independently to cover most aspects of the paper. In all cases, we follow some previous work [11, 16] and normalize each value with its corresponding optimal value to make the values more comparable across different query papers.

We have also considered the weighted harmonic mean of Coverage and Confidence measures called balanced \( F_{\text{score}} \):

\[
F_{\text{score}} = \frac{2 \cdot \text{Coverage} \cdot \text{Confidence}}{\text{Coverage} + \text{Confidence}} \quad (4)
\]

Ideally, we would like that the assigned reviewers to maximize both Coverage and Confidence but they are usually contradictory. Since our focus of study is to examine the effectiveness of multi-aspect matching, we use Coverage as the primary measure to evaluate different methods, though we report all other measures as well.

4.3 Baseline Method

All the proposed algorithms are implemented using the Lemur toolkit\(^3\). Since our work is the first study of review assignment based on multiple-aspect matching, there is no natural baseline to be compared with. One main hypothesis we would like to test is whether assigning reviewers based on aspect matching can achieve better aspect coverage than not considering aspect matching. In order to test this hypothesis, we compare all the aspect-based methods with a baseline method that doesn’t do aspect matching. We use a state of the art retrieval method (i.e., language modeling with Dirichlet smoothing [18]) as our Baseline approach. Specifically, for each author \( a \) in our collection, we construct one big document \( r_j \) as a profile for author by concatenation of all documents in our collection written by the \( j \)-th reviewer. Reviewers are ranked based on the probability of the query paper \( q \) given the reviewer document \( r_j \), which is computed as:

\[
p(q|r_j) = \prod_{w \in q} \frac{c(w,r_j) + \mu(p(w|R))}{N_{r_j} + \mu}
\]

where the variables have the same meaning as in the query likelihood formula used for paper aspect modeling (see Section 3.3.)

5. EXPERIMENT RESULTS

In order to evaluate the effectiveness of our methods, we compare all well-tuned methods. The parameter for BaseLINE, SDA and MIC is Dirichlet parameter whereas for MMR, the novelty parameter matters and for PLSA, the number of aspects and novelty parameters are important. We show sensitivity to these parameters in section 5.2. The number of reviewers assigned to each paper in our experiments is three.

\(^3\)http://www.lemurproject.org/
Table 1: Comparison of All Methods based on Abstract Queries and Full Paper Queries with at least Two Aspects

<table>
<thead>
<tr>
<th>Measures</th>
<th>Coverage</th>
<th>Confidence</th>
<th>AverageConfidence</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>Full</td>
<td>Abstract</td>
<td>Full</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.75</td>
<td>0.72</td>
<td>0.57 0.54</td>
<td>0.46</td>
</tr>
<tr>
<td>MMR</td>
<td>0.83</td>
<td>0.77</td>
<td>0.56 0.54</td>
<td>0.5</td>
</tr>
<tr>
<td>PLSA</td>
<td>0.87</td>
<td>0.86</td>
<td>0.58 0.57</td>
<td>0.53</td>
</tr>
<tr>
<td>MIC</td>
<td>0.8</td>
<td>0.8</td>
<td>0.48 0.51</td>
<td>0.42</td>
</tr>
<tr>
<td>SDA</td>
<td>0.8</td>
<td>0.78</td>
<td>0.52 0.56</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 2: Results for Abstract Queries Shown in Different Groups with Different Numbers of Aspects

<table>
<thead>
<tr>
<th>Methods</th>
<th>Coverage</th>
<th>Confidence</th>
<th>AverageConfidence</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>Full</td>
<td>Abstract</td>
<td>Full</td>
</tr>
<tr>
<td>Bins</td>
<td>B2</td>
<td>B3</td>
<td>B4</td>
<td>B2</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.73</td>
<td>0.78</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>MMR</td>
<td>0.8</td>
<td>0.85</td>
<td>0.87</td>
<td>0.61</td>
</tr>
<tr>
<td>PLSA</td>
<td>0.9</td>
<td>0.87</td>
<td>0.9</td>
<td>0.61</td>
</tr>
<tr>
<td>SDA</td>
<td>0.81</td>
<td>0.8</td>
<td>0.63</td>
<td>0.57</td>
</tr>
<tr>
<td>MIC</td>
<td>0.87</td>
<td>0.73</td>
<td>0.87</td>
<td>0.49</td>
</tr>
</tbody>
</table>

5.1 Basic Results

In this section, we show our evaluation results based on the four evaluation measures. Results are presented in Table 1 for all the methods considering both abstract and whole papers with at least two aspects (73 papers in total). For MMR method, we retrieved top 20 prospective reviewers from well-tuned Baseline method and do re-ranking based on this set. From the table, we can make the following observations:

Multi-aspect matching vs. Baseline: All the multi-aspect matching approaches outperform the Baseline method in terms of coverage. In most cases, the improvement is statistically significant as shown in Table 4, where we show the Wilcoxon Signed-Rank test for both abstract and full papers. This indicates that the proposed multi-aspect matching methods are effective in capturing and maximizing the coverage of multiple topic aspects by the assigned reviewers.

Table 4: Wilcoxon Signed-Rank Test Results (P-values) for Coverage Measure and all Queries (Abstracts and Full Papers)

<table>
<thead>
<tr>
<th>Measures</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>PLSA Baseline</td>
<td>.001</td>
</tr>
<tr>
<td>MMR Baseline</td>
<td>.005</td>
</tr>
<tr>
<td>MIC Baseline</td>
<td>.104</td>
</tr>
<tr>
<td>SDA Baseline</td>
<td>.036</td>
</tr>
<tr>
<td>PLSA MMR</td>
<td>.056</td>
</tr>
<tr>
<td>PLSA SDA</td>
<td>.012</td>
</tr>
<tr>
<td>PLSA MIC</td>
<td>.036</td>
</tr>
</tbody>
</table>

Coverage-Confidence tradeoff: Although all the multi-aspect matching methods outperform the Baseline in coverage, some of them have lower confidence than the Baseline; as a result, they do not all outperform the Baseline in terms of average confidence and F score, which combine coverage with confidence. This is not surprising and simply indicates that when we attempt to increase coverage, we tend to risk on lowering the confidence; such a tradeoff is similar to the familiar recall-precision tradeoff.

PLSA vs. others: Despite the effect of coverage-confidence tradeoff, the PLSA method stands out as a method that outperforms the Baseline by all measures; indeed, it also outperforms all other proposed methods by all measures. Some of the improvements in coverage are statistically significant as shown in Table 4.

Abstract vs. Full paper: While a full paper is far more informative than an abstract, our results show that abstracts may be sufficiently informative for multi-aspect review assignment. Indeed, for both MMR and PLSA, the performance on abstracts is consistently better than on full papers. MMR and Baseline are particularly hurt by using full papers as queries. Since both of them rely on ranking reviewers based on a single query language model constructed based on a full paper, this suggests that the query language model built based on a full paper may be noisier than one based on an abstract, which presumably is cleaner. Another possible reason is that our expertise representation for reviewers is built based on abstracts, so more detailed representation obtained using a full paper may not be so useful. In the future, we should further look into this issue by experimenting with using full papers to represent expertise.

In contrast, however, both paper aspect modeling approaches (i.e., MIC and SDA) apparently have benefited from full papers; they all perform better on full papers than on abstracts by all measures except in one case (coverage of SDA). This indicates that the more detailed information offered by the full paper does help us better capturing the multiple topic aspects in a paper than in the case of an abstract as we should expect.

Reviewer vs. Paper aspect modeling: The fact that PLSA outperforms MIC and SDA indicates that reviewer aspect modeling is a better strategy than paper aspect modeling. However, as discussed earlier, both paper aspect modeling approaches can gain more through using full papers that when we attempt to increase coverage, we tend to risk on lowering the confidence; such a tradeoff is similar to the familiar recall-precision tradeoff.

At the time of this study, we were not able to obtain full papers for these candidate reviewers.
Table 3: Results for Full Queries Shown in Different Groups with Different Numbers of Aspects

<table>
<thead>
<tr>
<th>Methods</th>
<th>Coverage</th>
<th>Confidence</th>
<th>Average Confidence</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.78</td>
<td>0.68</td>
<td>0.62</td>
<td>0.56</td>
</tr>
<tr>
<td>MMR</td>
<td>0.83</td>
<td>0.74</td>
<td>0.63</td>
<td>0.55</td>
</tr>
<tr>
<td>PLSA</td>
<td>0.87</td>
<td>0.84</td>
<td>0.87</td>
<td>0.61</td>
</tr>
<tr>
<td>MIC</td>
<td>0.81</td>
<td>0.78</td>
<td>0.87</td>
<td>0.5</td>
</tr>
<tr>
<td>SDA</td>
<td>0.8</td>
<td>0.76</td>
<td>0.79</td>
<td>0.56</td>
</tr>
</tbody>
</table>

5.2 Detailed Analysis

5.2.1 Different Numbers of Aspects

To further understand how these algorithms behave for papers with different number of topic aspects, we divide our queries into bins where each bin has a different number of aspects in the query. We have queries with at most 4 aspects, as a result, we have 4 bins. In bin 1 (one aspect), we have 12 queries and in bin 2, we have 32 queries, in bin 3, 35 queries and in bin 4, we have 6 queries. With such decomposition, we can obtain a detailed view of the performance of each method for both abstract queries and full papers. The results are shown in Table 2 and Table 3, for abstracts and full papers, respectively. These results provide a decomposed view of those in Table 1.

In general, we see that in most cases, the performance on papers with more distinct aspects is worse than that on those with fewer distinct papers. This is expected as when a paper has more distinct aspects to cover, it is harder to cover all aspects and harder to have all reviewers to cover every aspect. However, most of the observations we made earlier are confirmed in these decomposed results; in particular, we again see that PLSA outperforms all other methods in all but a few cases.

5.2.2 Parameter Sensitivity Analysis

The analysis made so far is based on optimal results obtained from parameter tuning. We now take a closer look at some of the parameters involved in all the methods and examine how they affect the performance and understand how they should be set.

The first parameter we study is the Dirichlet prior smoothing parameter $\mu$; it is relevant to the Baseline method and the two paper aspect modeling methods, MIC and SDA since they involve ranking experts using the KL-divergence retrieval model with unigram language models. (It also affects MMR, but the influence is through the Baseline method.) In general, we find that the performance is not very sensitive to this parameter when it is set to a relatively large value (e.g., larger than 4,000); a similar pattern of sensitivity was also found in using Dirichlet prior smoothing for regular ad hoc retrieval [18]. In Figure 1, we plot the sensitivity curves of coverage to this parameter for the Baseline, MIC, and SDA on both abstracts and full papers. We see that for all the
parameter values, MIC and SDA outperform the Baseline in coverage, again confirming that MIC and SDA are effective in modeling and matching multiple aspects.

The sensitivity curves of other measures are similar with a general trend of becoming more stable as the value of the smoothing parameter increases. The optimal setting appears to be between 4,000 and 5,000.

The MMR method has an additional parameter (i.e., $\nu$) controlling the degree of redundancy removal. In Figure 2, we plot the sensitivity curves of all the four measures to this parameter in both cases of abstracts and full papers. We see that as long as $\nu$ is set to a relatively large value ($\nu \geq 0.3$ for abstracts and $\nu \geq 0.5$ for full papers), the performance is stable and close to optimal.

PLSA has two parameters, the novelty parameter $\sigma$ and the number of aspects $K$. In Figure 3, we show the sensitivity curves of coverage to the novelty parameter for several different numbers of aspects (10, 20, 50, and 70). We see that using too few aspects (10) clearly hurts the performance, but the difference between 20, 50 and 70 aspects is relatively small, and they are all clearly better than 10 aspects. While the performance is fluctuating at different values of the novelty parameter $\sigma$, the magnitude of difference is not very big when a relatively large number of aspects is used ($K \geq 20$). In such cases, all the coverage values are higher than the Baseline values (0.75 for abstracts and 0.72 for full papers). The sensitivity curves for the other three measures (Confidence, Average Confidence, and F score) are all similar.

To see more clearly how the number of aspects affects performance in PLSA, we further plot sensitivity curves of all the four measures to this parameter in Figure 4, where we see that as long as a relatively large number of aspects ($K \geq 20$) is used, the performance is stable and close to optimal.

6. RELATED WORK

Review assignment task has a long history. Dumais and Nielsen [6] use Latent Semantic Indexing (LSI) trained on the abstracts provided by prospective reviewers. The recent work in [9], is a recommendation system for recommending panels of reviewers for NSF grant applications. They use
TF-IDF weighting measure for finding the match between reviewers and papers. Similarly, the work in [2] uses web to find abstracts of papers written by reviewers and then uses TF-IDF weighting to rank reviewers for a given paper. In a recent work on matching papers with reviewers [13], authors propose a novel topic model, Author-Persona-Topic (APT). They use statistical topic model to discover expertise. APT model is based on the idea that authors frequently write about different topic area combinations; so they learn the topical components as well as divide each author’s papers into several "personas". Each persona clusters papers with similar topical combinations. These works are similar to our work in the sense that they all try to match a paper/proposal to reviewers, but the main difference between our work and theirs is that none of the existing works consider matching reviewers with a paper based on matching of multiple aspects of expertise.

Expertise Modeling has received a lot of interests recently. There is also an Enterprise Track in TREC which is devoted to "Expert Finding" task. Expert finding is about finding people rather than documents and the goal is to retrieve a ranked list of experts with expertise on a given topic. There has been an extensive work on Enterprise track of TREC. For example, probabilistic models and language models are presented in [1, 14, 7] to solve the problem of expert finding. Our work is related to this body of work on expert finding, but the problem set up is different; as an expert finding problem, our problem is more challenging because we attempt to match expertise based on multiple aspects.

Authors in [4] argue for the value of diversity or relevant novelty in the ranked results of a query, and they propose the maximum marginal relevance (MMR) ranking function that incorporates such diversity in a ranked result set. In a recent work on a subtopic retrieval [16], authors assume that a query may have different subtopic interpretations, and reorder the results so that the top includes some results from each subtopic. Our work is quite related to this work in the sense that our problem and the subtopic retrieval problem share a lot of similarities and a similar strategy can be potentially applicable to both problems. Indeed, one of our proposed strategies (i.e., redundancy removal) and some of the evaluation methods are essentially similar to those proposed in [16]. However, despite the similarity, there are some important differences between our problem and the subtopic retrieval problem: (1) Our queries are long abstracts, which are quite different from keyword queries, and they make it appropriate to consider modeling multiple aspects on the query side, which we exploit to propose some new methods. (2) While in subtopic, the goal is high recall and ranking all the documents, our goal is to retrieve a fixed number of reviewers. As a result, we need different evaluation measures; in particular, the aspect precision measure proposed in [16] is completely inappropriate for our task.

Our work is also related to the study of topic modeling or aspect modeling. In [3, 10], the authors introduce generative probabilistic models for collections of text corpora. They model the collection of data as a finite mixture over an underlying set of topics. Each topic is a finite mixture over topic probabilities and each document is a finite mixture over a set of topics. In our problem, the reviewers’ expertise is modeled as a finite mixture over the set of topics. Author-Topic model is also proposed in [15]. They propose a new unsupervised learning algorithm for extracting topics of documents and the interests of authors. Their generative model extends the previous work by using probabilistic topics to author modeling by allowing the mixture weights for different topics which are to be determined by the authors of papers. When we directly model subtopics, we use PLSA [10] to generate subtopics in our corpus. Our work differs from this line of work in that we go beyond topic modeling to address the issue of matching expertise and maximizing the coverage of expertise based on topic modeling.

7. CONCLUSIONS AND FUTURE WORK

Review assignment is an important yet labor-intensive task. Automatic assignment of reviews is appealing for multiple reasons, including reducing the labor of the assigners and potentially improving the quality of assignments. A common limitation of existing approaches to automatic review assignment is that they all match a paper as a whole without modeling aspects of topics or expertise, resulting in unbalanced assignments. In this paper, we studied how to match reviewers with a paper based on matching of multi-
ple aspects of expertise so that the assigned reviewers would not only have the required expertise to review a paper but can also cover all the aspects of a paper in a complementary manner. We proposed three general strategies for solving this problem, including removing redundancy, modeling aspects based on reviewer expertise, and modeling aspects based on the paper to be reviewed, and proposed four specific methods to implement these strategies, including MMR, PLSA, MIC, and SDA. We also defined four measures (i.e., Coverage, Confidence, Average Confidence and F score) to evaluate methods for multi-aspect review assignment. We further created the first test collection for evaluating such a task based on ACM SIGIR publications.

We systematically tested all the methods with this data set. Experiment results show that all the proposed methods are effective for increasing the coverage of topic aspects in review assignment and outperform a state of the art baseline retrieval method. In particular, the PLSA multi-aspect matching method, which is a reviewer aspect modeling method, performs the best. Our results also show that while reviewer-aspect modeling outperforms paper-aspect modeling, the latter gains more as we move from using abstracts to using full papers for assignment, suggesting potential benefit of combining these two complementary strategies, which should be a very interesting future research direction.

Due to the lack of resources for evaluation, our conclusions are inevitably preliminary, thus another important future research direction is to further evaluate these methods with more data sets ideally created with multiple human assessors.

All the proposed methods are general, thus they can be applied to all review assignment tasks, including related expertise matching tasks such as forming a thesis committee for a Ph.D. student (matching reviewers with a thesis) and forming a panel on a topic (matching candidate panelists with a topic description). In all these other applications, it is also desirable to match experts based on multiple aspects. We plan to explore all these applications in the future. For our experiments, we retrieve the fixed number of reviewers, e.g., three; another interesting future research direction would be to vary this number and see how it affects the performance.

Finally, our problem setup has assumed that the assignment of reviewers to each paper is independently done, but in a real scenario of assigning papers for a conference, we generally have constraints such as review quota of a reviewer. Thus another interesting future research direction is to consider the larger problem of assigning a set of reviewers to a set of papers with review quota constraints. We can further extend our approaches to multi-aspect expertise matching to solve such a problem through constrained optimization.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their useful comments. This material is based in part upon work supported by the National Science Foundation under award numbers IIS-0347933, IIS-0713571, and IIS-0713581, and work supported by NIH/NLM grant 1 R01 LM009153-01.

9. REFERENCES