Taking the load off the conference chairs: towards a digital paper-routing assistant

David Yarowsky and Radu Florian
Computer Science Department and Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland 21218
{yarowsky,rflorian}@cs.jhu.edu

Abstract

This paper describes and extensively evaluates a system for the automatic routing of submitted papers to reviewers and area committees, without the need for any human annotation from the reviewers or the program chair. Routing is based on a profile of previous writings obtainable on-line for the reviewer pool, a generally stable and reusable resource that requires no manual adaptation for new submission streams. The paper explores a wide set of variations and extensions on the core model, and achieves system accuracy approaching that of several human judges on the same task.

1 Introduction and Problem Statement

Routing submitted papers, abstracts or grant proposals to qualified reviewers is a central task of the academic enterprise, and a remarkably difficult one. Typically it is conducted under significant time pressure in a conference reviewing cycle. As the number of submissions and size of the reviewer pool grows, it becomes increasingly difficult for a conference chair to be familiar with the different expertise of all members of the program committee. It is also difficult for one person to master the subtleties of fine subject area distinctions as topic diversity in a conference becomes large. For these reasons, conferences such as ACL (the Association for Computational Linguistics) often use a hierarchical program committee structure, where submitted papers are first routed to area committees, and then more specialized area chairs have the task of assigning papers to individual reviewers in the committee. However, in a diverse and multidisciplinary field such as natural language processing, it is often difficult to define clear cut committee descriptions and the program chair still must be cognizant of the detailed expertise of the area committee members in order to route atypical or multidisciplinary papers to committees with the most appropriate pool of reviewers. The low inter-rater consistency results shown in Table 12 indicate that humans find even area committee routing to be a difficult task.

The following paper focuses on a range of automated solutions to this task of routing papers to their most appropriate area committee. It presents extensive empirical investigation and evaluation of a wide range of issues related to this task.

Previous published research into the problem of automatic routing of conference paper submissions is surprisingly limited. Approaches to this task can be essentially broken down into four major strategies:

The first strategy is keyword based. Authors are required to specify a list of topic/subtopic areas for their papers (often from a prespecified term list), and reviewers then complete a survey of their relative level of expertise on this list of topics/subtopics. This approach is followed by AAAI conference reviewing. It suffers from the problem that authors often have a difficult time selecting keywords to adequately describe their work. It works best in conferences that are very broad, and is least effective in more focused workshops where routing distinctions in subject area and paradigm are more subtle.

The second strategy is to build a statistical profile of reviewers' expertise by eliciting relevance judgments on a set of abstract data. AAAI also requires its reviewers to rank (bid on) submitted abstracts, and there is currently unpublished work exploring the application of supervised routing to the ranked reviewer bids on AAAI submitted abstracts (Hirsh, personal communication). In groundbreaking work, Dumais and Nielsen (1992) developed a system for the routing of Hypertext'91 abstracts using latent semantic indexing (Deerwester et al., 1990), trained from available text sources including a small set of reviewer-submitted abstracts, on-line books and ACM articles as a source for the term-by-document matrix used in their singular value decompositions. Reviewers manually ranked their interest in all submitted abstracts, and best performance was achieved when reviewers were assigned twice their target number of abstracts and asked to choose their preferred half.

One problem with the modeling of reviewer rankings/bids is that these may be based more on what the reviewer finds interesting rather than what

Paper 185

word sense disambiguation with an ensemble of naive Bayesian classifiers

score	committee	Rough committee characterization
0.377	com4	statistical NLP (focus on sense tagging)
0.365	com3	statistical NLP (focus on MT, statistical parsing)
0.258	com6	generation and systems
0.242	com5	lexicons, some non-statistical sense tagging
0.228	com2	syntax/parsing (mostly non-statistical)
0.224	com1	discourse/dialog

Table 1: Committee routing system output

he/she is most qualified to review. Even with instructions, there may be a natural human tendency to bid higher on exciting/interesting abstracts in a more distant area and possibly bid lower on weak/uninteresting papers in the reviewer's core expertise. In addition, AAAI reviewers also report this as a long and tedious abstract ranking process, that shifts the burden of labor for paper routing onto the reviewers rather than the program chairs.

A third option is to learn the reviewer/committee profiles by having the program chair assign a portion of the submissions to reviewers and/or committees and then attempt to model these assignments in order to compute the assignment of remaining submissions to the reviewer pool based on these models. We evaluate such a strategy below.

A fourth option, the focus of this paper, is to create a statistical profile of reviewers' expertise by modeling the collection of their previously published papers and other writings or statements of research interest extracted automatically from what is available on the web. One advantage of this approach is that frees the reviewer from a laborious abstract ranking/bidding process. Another is that profiles based on a large collection of the reviewer's own writings is perhaps a better model of areas of demonstrated expertise rather than simply the papers a reviewer finds most "interesting". And, finally, such profiles based on collected writings tend to be relatively stable and reusable from conference-to-conference (reviewers often serve on many program committees) and may optionally be updated when a reviewer's representative publications grow of change significantly. The effort in creating such publications-based profiles need not be repeated as the pool of submitted abstracts change. This approach is remarkably cost effective and the empirical results below indicate that it can achieve performance competitive with human paper routers.

2 Task Description

The primary task investigated in this paper is the routing of full-length submitted conference papers to one of 6 area committees for ACL'99, with the committees ranked in order of appropriateness in Table 1 (actual output of the system on sample paper #185). The 6 committees are best defined by their members (listed with their committee numbers in Figures 2), but they are *very* roughly characterized in Table 1.

Score	Reviewer	Committee
0.540	ng	com4
0.426	bruce	com4
0.420	roth	com4
0.414	golding	com4
0.369	wiebe	com1
0.368	resnik	com3
0.351	daelemans	com4
0.344	shin	com3
0.337	lee	com4
0.327	hang	com5

Table 2: Reviewer routing system output (for paper 185, above)

A secondary task is to provide a proposed ordered list of appropriate reviewers, as shown in Table 2. Note that this list can be filtered to include just the first choice committee, or can include the most appropriate reviewers independent of committee structure.

2.1 The Data

The evaluation data used in these experiments consisted of full-length articles submitted to the general session of ACL'99. Thematic session submissions were ignored because the reviewing committee was preselected by the author in these cases. The ACL'99 call for papers included a statement requesting voluntary submission of electronic versions of their papers for a paper routing experiment. Of the 180 general session authors, 51% (92) participated in the study through electronic submission.

As noted above, electronic copies of representative papers were also solicited from members of the general session area program committees. Participants had the option of including a numeric ranking (1 to 10) indicating the representativeness of the papers with respect to their areas of expertise, but few chose to do so. In the numerous cases where none or insufficient numbers of papers were received from reviewers, their self-selected sample of previous publications were augmented by large numbers of downloaded reviewer papers from cmp-lg (xxx.lanl.gov/cmp-lg), their own home pages and the www.cora.jprc.com[1] archive.

Papers were received and processed from 5 acceptable formats: latex, postscript, plain text, portable document format (pdf) and html, all of which were converted to a marked-up plain text normal format. Distinct regions of the papers (title, abstract, main body, bibliography) were identified and extracted, when possible, in support of differential region weighting.

2.2 Evaluation Methodology

The primary "gold standard" for evaluation consisted of the committee numbers actually assigned to each paper by the ACL'99 program chair performing the committee routing. These judgments were obtained prior to his seeing the results of the automatic routing experiments. Because the program chair considered other factors including potential conflicts of interest in assigning papers, this is not a perfect annotation of the most appropriate committee based strictly on mass of reviewer expertise. Three other judges (2 NLP faculty members and one 3rd year NLP grad student) also routed those papers voluntarily submitted from the authors for the routing experiments, with their names, addresses and institutions stripped. Greatest committee appropriateness based on topic and reviewer expertise was the sole criterion for these paper assignments. A second evaluation gold standard was obtained from the weighted consensus of the 4 reviewers (described in Section 7 below).

The 92 submitted papers were divided into two equal halves: a primary test set on which all major results were evaluated, and a secondary devtest set, via which some global parameters were estimated and the one instance of supervised training took place.

Several evaluation measures were used to reflect system performance. The first is exact match classification *accuracy* (the percentage of the papers on which the gold standard and system agreed exactly on the committee assignment). Because the system returns a full preferred rank order of the 6 committees for all papers, a second natural performance measure is the *average position* of the truth (gold standard committee selection) in this rank list. This measure gives an assessment of how many committees the human judge would have to consider, on average, before it found the correct classification; smaller is better. Because in many cases there are two equally viable committee contenders, a third measure *One-of-best-2* indicates the percentage of cases where the gold standard classification is in the top two choices ranked by the system. In many cases, the whole histogram is given, indicating the position of the gold standard classification in the system's committee ranking.

3 Routing Methodologies

There are numerous methods described in the information retrieval literature for article routing. Assuming that there are n classes and a set of m articles, the *article routing* task attributes each of the m articles to one of the n classes. It is clear that our task fits well in this paradigm; each paper has to go to one committee. The two major approaches tested in this model are the standard Salton-style *vector space* model (Salton and McGill, 1983) and the *Naive Bayes* classifier (Mosteller and Wallace, 1964). [2] These and several permutations and extensions are detailed and evaluated below.

3.1 Vector Routing Model

Unless we specify otherwise, we shall assume that the vocabulary is selected by removing a set of common (stop) words from the text. Both the submitted papers and the reviewer papers are represented in the space $[0;\infty)^{|\mathcal{V}|}$, as vectors D_i: $D_{ij} = c_{ij} \cdot w_j$, where c_{ij} is the count (the number of occurrences) of the jth word in document D_i and w_j is an "importance" weight associated with the jth word. One typical weighting function is IDF (Inverse Document Frequency): $w_j = \log\left(\delta + \frac{N}{\mathrm{docf}_j}\right)$, where N is the total number of documents and docf_j is the *document frequency* of the jth word (the number of documents the word appears in). One can measure the similarity between 2 documents by using the *cosine similarity* between their vector representations:

$$\text{cosine\_similarity}(D_i, D_j) = \frac{\sum_{k=1}^{|\mathcal{V}|} D_{ik} \cdot D_{jk}}{\sqrt{\sum_{k=1}^{|\mathcal{V}|} D_{ik}^2}\sqrt{\sum_{k=1}^{|\mathcal{V}|} D_{jk}^2}} = \left\langle \frac{D_i}{\|D_i\|_2}, \frac{D_j}{\|D_j\|_2} \right\rangle$$

the dot product of the normalized[3] vectors (see (Salton and McGill, 1983)). This measure of similarity yields values close to 1 for similar vectors and close to 0 for dissimilar ones.

[1] This is a web search engine specialized in searching Computer Science related papers (see (McCallum et al., 1999)).

[2] Routing using these and other models is a central task in information retrieval, discussed in depth in (Hull, 1994),(Lewis and Gale, 1994), (Larkey and Croft, 1996) and (Voorhees and Harman, 1998) and many other articles.

[3] $\|\cdot\|_2$ being the Euclidean norm.

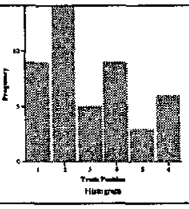
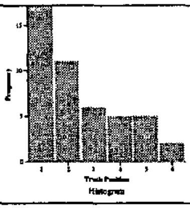
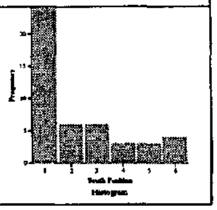
	Baseline	**Keywords Only**	**Keywords, Title, Abstract Only**
Accuracy:	19.6%	36.9%	52.2%
Average position:	3.02	2.48	2.28
One-of-best-2:	50.0%	60.9%	65.2%
Histogram:			

Table 3: Baseline performance measures

The main algorithm proceeds as follows:

1. For the ith reviewer ($i = 1, \ldots$), compute a centroid R_i - a vector presumably associated with the main research interests of the reviewer:

$$R_{ij} = \sum_{P \in \mathcal{P}_i} r(P) \cdot c_j(P) \cdot w_j \qquad (1)$$

   where $\mathcal{P}_i$ is the pool of papers for ith reviewer, $r(P)$ is the weight/relevance of paper P and $c_j(P)$ is the word count of jth word in paper P (a given word might weight differently in different regions - see region weighting below).

2. For each committee, compute its centroid as the sum of the composing reviewers' centroids:

$$C_{kj} = \sum_{R_i \in \mathcal{C}_k} R_{ij} \qquad (2)$$

   where $\mathcal{C}_k$ is the pool of reviewers for committee k.

3. For each paper, rank all the committees based on the cosine similarity between the paper's vector and the committee centroids - the one that ranks highest is chosen as the classification of the paper:

$$\text{classification}(P_l) = \underset{k=1\ldots6}{\text{argmax}}(\text{cosine\_similarity}(P_l, C_k)) \qquad (3)$$

Table 3 gives results for several basic baseline models. Section 2.2 describes these measures.

Clearly different regions of a paper have different importance in determining its semantic context. We automatically separate the text into title, abstract, keywords, body and bibliography regions and investigate different weighting parameters for these regions.

The results for full text and region weighting are given in Table 5. Consensus evaluation is described in Section 7. A confusion matrix[4] showing region weighting results is given in Table 4. Note that the primary confusion is between the difficult to distinguish committees 3 and 4.

Truth **Router**	com$_1$	com$_2$	com$_3$	com$_4$	com$_5$	com$_6$
com$_1$	**1**	0	0	0	0	2
com$_2$	1	**9**	1	0	0	3
com$_3$	0	0	**6**	2	0	1
com$_4$	0	0	4	**3**	0	2
com$_5$	0	0	1	0	**5**	0
com$_6$	0	1	1	0	1	**2**

Table 4: Confusion matrix for the full text, region weighting case

The remainder of this section describes the modifications made to this model, the results we obtained, conclusions and explanations of the results.

3.1.1 Weighting Paper Sources Differently

As noted before, the reviewers' papers were obtained from different sources, with potentially different relative indicativeness of a reviewer's expertise. A variety of relative weighting parameters for these sources were explored on the devtest. None yielded a significant improvement over the equally weighted model.

3.1.2 Term Selection and Weighting

Experiments were conducted to test the efficacy of two variants of IDF (based on the concepts of 1 document per reviewer and 1 document per committee), entropy-based term weighting, use of stemming, and

[4] A cell (i, j) in the confusion matrix shows how many times committee i was chosen where committee j was the true assignment. It is an indication of the nature of the misclassification observed, not merely its absolute number.

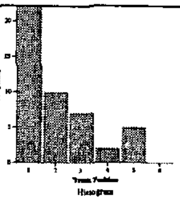
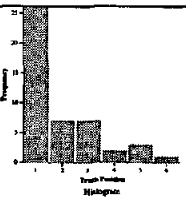
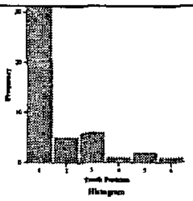
	Full Text, Equal Weighting	Full Text, Region Weighting	Full Text, (RW) Consensus Evaluation
Accuracy	47.8%	56.5%	67.4%
Average position	2.09	1.96	1.72
One-of-best-2	69.6%	71.7%	78.3%
Histogram:			

Table 5: Performance on Full Text Routing

Title	Abstract	Body	Bibliography	Topic Area
30	30	1	1	10

Table 6: Word Weights Based on Region

vocabulary selection based on statistically significant cross-class frequency variation. No variation outperformed the region weighting model shown in Table 5.

3.2 Naive Bayes Classifier

The naive Bayes model makes an independence assumption relative to the words in a text. It chooses the committee C_j that maximizes the probability $P(C_j|P_i)$; formally

$$\begin{aligned}\mathrm{argmax}_j\, P(C_j|P_i) &= \mathrm{argmax}_j \frac{P(C_j)\cdot P(P_i|C_j)}{P(P_i)} \\ &= \mathrm{argmax}_j\, P(C_j) \prod_{w_k \in P_i} P(w_k|C_j) \\ &= \mathrm{argmax}_j \left(\log(P(C_j)) + \sum_{w_k \in P_i} \log(P(w_k|C_j))\right)\end{aligned}$$

and, furthermore, if one assumes equal *a priori* probability on the committees ($P(C_j) = ct$), then one looks for

$$\underset{j}{\mathrm{argmax}} \left(\sum_{w_k \in P_i} \log(P(w_k|C_j)) \right)$$

where the words w_k are the target words in the article P_i (usually all the non-stopwords).

One of the issues that need to be addressed when considering naive Bayes approaches is smoothing. One cannot afford to have null probabilities, as they would just nullify the results. The smoothing method used in this approach is the simple additive smoothing method, that adjusts the maximum likelihood estimates as follows:

$$P(w_k|C_j) = \frac{\delta + C(w_k, C_j)}{\delta \cdot |\mathcal{V}| + N(C_j)}$$

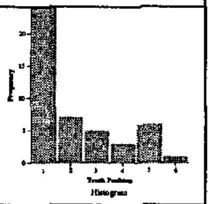
Accuracy:	52.2%
Average position:	2.20
One-of-best-2:	67.4%
Histogram:	

Table 7: Results for the Naive Bayes Classifier

where $N(C_j) = \sum_k C(w_k, C_j)$ and $\mathcal{V}$ is the whole vocabulary. This is a very simple strategy, but we believe that it works relatively well for unigrams. Results are shown in Table 7; it underperforms the region weighted vector-based model with similar parameters.

To check whether unseen words are a problem in our case,we varied the parameter δ. Since the results were almost the same for δ values varying from 0.01 to 1, we conclude that more sophisticated smoothing methods (e.g. Good-Turing, Knesser-Ney) would not have made a difference, either.

3.3 Voting

As an alternative approach to the top-down hierarchical routing strategy, we investigated the initial direct assignment of papers to reviewers, and then allowed the top k reviewers vote for his or her own committee. Although optimal performance here was slightly lower than for the reference system (46.5%, 2.22), the gold standard is based on the primacy of human committee assignments and have no guarantee that the committee has an adequate number of well qualified reviewers. Without the ability

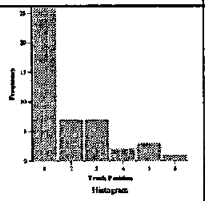
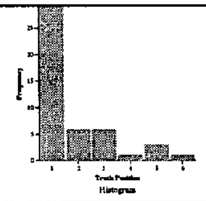
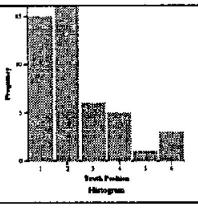
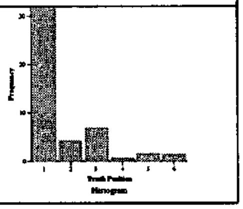
	Sim_{text} only	Sim_{text} and Sim_{bib}	Sim_{bib} only	Consensus evaluation
	$\beta = 0$	$\beta = 0.76$	$\beta = 1$	$\beta = 0.76$ consensus
Accuracy:	56.5%	63.0%	39.1%	69.0%
Average position:	1.96	1.85	2.35	1.65
One-of-best-2:	71.7%	73.9%	56.5%	78.3%
Histogram:				

Table 8: Performance of routing based on bibliographic similarity

for cross-committee reviewing, a committee with 3 moderately-well qualified reviewers would probably be preferable to a committee with only a single qualified reviewer but with extremely strong expertise.

3.4 Routing based on (transitive) bibliographic similarity

Appropriate reviewers for a paper can often be determined through analysis of the paper's bibliography. Clearly direct citation of a potential reviewer is partial evidence of that person's suitability to review the paper. This relation is also somewhat transitive, as the authors who cite or are cited by an author directly cited in the paper also have increased likelihood of being relevant reviewers.

The goal of this section is to identify transitively related authors via chains of the bibliographic relations $Cites(author_i, author_j)$ and $Coauthor(author_i, author_j)$. To estimate these relations, we automatically extracted and normalized bibliographic citations from a large body of on-line texts including all of the reviewer-submitted papers. Via transitive use of this extensive citation data, reviewer-paper similarity could be estimated even when there was no direct mention of the reviewer in the text to be routed.

To formalize this approach, let us assume that there exists an indexed set of authors $\mathcal{A}=\{a_1, \dots, a_{n_a}\}$. The reviewers are part of this set; let $\mathcal{R} = \{r_1 \dots r_{n_r}\}$ denote the set of reviewers. We also dispose of a set of papers submitted by reviewers, $\mathcal{P} = \{p_1, \dots, p_{n_p}\}$. Using the set $\mathcal{P}$ we compute 2 matrices: *Cites* and *Coauthor*:

$$Cites_{ij} = \frac{\sum_{p \in \mathcal{P}} N_p(a_i,a_j)}{\sum_{k=1}^{n_a} \sum_{p \in \mathcal{P}} N_p(a_i,a_k)} \qquad Coauthor_{ij} = \frac{Nc(a_i,a_j)}{\sum_{k=1}^{n_a} Nc(a_i,a_k)}$$

where $N_p(a_i, a_j)$ is the number of times a_j was cited in the paper p if a_i is an author of p, 0 otherwise, and $Nc(a_i, a_j)$ is the number of papers in which a_i and a_j were coauthors identified either from the head of a paper $p \in \mathcal{P}$, or a bibliographic citation extracted from p. The relation *Cited_by* can be captured by the transposition of the citation matrix $Cites^T$.

Distance d	1	2	3	4
Accuracy	30.3%	30.4%	32.6%	34.5%
Average Position	2.28	2.22	2.17	2.17
One-of-best-2	65.2%	71.7%	73.9%	71.7%

Table 9: Performance comparison at different levels of parameter d, $\lambda = 0.8$ and $\beta = 1$, evaluated on devtest data

A symmetric similarity matrix combining these base relations is defined as:

$$Sim^1 = \lambda \tfrac{1}{2}\left(Cites + Cites^T\right) + (1-\lambda)\tfrac{1}{2}\left(Coauthor + Coauthor^T\right) \qquad (4)$$

where λ is a weighting factor between the contributing sources of similarity. The index 1 (Sim^1) denotes "direct" (non-transitive) bibliographic similarity. We enforced that $Sim^1_{ii} = 1$ for all authors i.

The submitted articles, $P_1, \dots, P_{n_P}$ were routed to committees based on similarities between the authors cited in the paper and the reviewers forming a committee:

$$Sim(P_l, C_k) = \frac{1}{|C_k|} \sum_{r_i \in C_k} Sim(P_l, r_i) \qquad (5)$$

where

$$Sim(P_l, r_i) = \sum_{j=1}^{n_a} \frac{C(P_l, a_j)}{\sum_{t=1}^{n_a} C(P_l, a_t)} \cdot Sim^1(a_j, r_i)$$

$C(P_l, a_j)$ being the number of times author a_j was cited in paper P_l. A paper is routed to the committee that maximizes the paper/committee similarity given in (5). Tuning the parameter λ on the training set yielded $\lambda = 0.8$.

The similarity relation computed in formula (4) is very sparse, as a large number of values are 0. To compute a more robust similarity, one can consider the transitive closure of the graph defined by Sim^1. The weights in the resulting graphs are:

$$Sim^{\infty}(i,j) = \sum_{\substack{i=i_1,\ldots,i_n=j \\ i_l \neq i_p}} C(i_1 \ldots i_n) \qquad (6)$$

where $Sim^{\infty}(i,j)$ is the similarity between the i^{th} and j^{th} author. The similarity along one path could be any function of the weights of the composing links. The one we considered is:

$$C(i_1 \ldots i_n) = \prod_{k=1}^{n-1} Sim^1(i_k, i_{k+1})$$

Computing the values in (6) proves to be computationally expensive, and it appears that extending the transitive similarity relationship indefinitely may become counterproductive. Therefore, we limited the length of the paths involved in computing the formula (6):

$$Sim^d(i,j) = \sum_{n=1}^{d} \sum_{\substack{i=i_1,\ldots,i_n=j \\ i_l \neq i_p}} C(i_1 \ldots i_n)$$

Let us observe that $Sim^{\infty}(i,j) = \lim_{d\to\infty} Sim^d(i,j)$, hence the name. In Table 9, one can observe that the routing performance increases as d increases up through a transitive distance of 3, with mixed results beyond that point.

Section 3.4 has, until now, described a routing similarity based only on transitive bibliographic citation and co-authorship (Sim_{bib}). However, routing a paper solely on this basis is not optimal as it ignores similarity between the the terms in the full text (Sim_{text}), as described in Section 3.1 using region weighting. We combined these two measures through interpolation:

$$\begin{aligned} Sim(P_l, r_i) = \beta \cdot Sim_{\text{bib}}(P_l, r_i) + \\ (1-\beta)\, Sim_{\text{text}}(P_l, r_i) \end{aligned} \qquad (7)$$

On the training set, a value of $\beta = 0.76$ was found to maximize performance, for $d = 3$ and the previously fixed $\lambda = 0.8$.

The full evaluation of the transitive bibliographic similarity measure are given in Table 8. Performance using exclusively Sim_{bib} ($\beta = 1$) is considerably lower (39.1%) than the previous best text-based similarity (Sim_{text}) performance of 56.5% exact match accuracy. However, combining the two evidence sources yields a substantially higher routing accuracy of 63.0%. This result is also observed when evaluating on the consensus gold standard described in Section 7, where combined model accuracy of 69.0% exceeds the Sim_{text} only accuracy of 67.4%. As shall be shown, for both evaluation standards the combined system accuracy rivals that of several human judges.

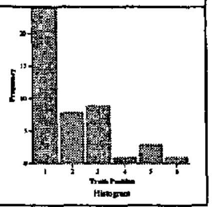
Accuracy:	52.2%
Average position:	2.00
One-of-best-2:	69.6%
Histogram:	

Table 10: Author-based paper routing

3.4.1 Routing based exclusively on the paper's author

Prior to now, we have ignored a submitted paper's author(s) when making the routing decision. However, ACL'99 reviewing was not blind and an interesting question is what is routing performance when classification is based *exclusively* on the authors' identity. Using *only* $Sim_{bib}(author, reviewer_j)$ for the paper's author(s), exact match accuracy completely ignoring the submitted paper (52.5%) approaches that of the accuracy using only the submitted text (56.5%), as shown in Table 10. This suggests that an author's identity alone is largely sufficient for routing the paper to the committee most appropriate for evaluating her or his work.

4 Supervised Learning

The algorithms presented so far are unsupervised; the only use for labeled data in the devtest was for global parameter optimization. This is a strength of the approach presented here, because it can be used successfully without any human annotation. In this section, we tested the efficacy of training supervised models based on initial program chair annotation of a portion of the submitted papers. Models of the types of papers initially assigned to each committee can help select further papers appropriate for that committee. Using the vector model, we can define the centroid D_{ij} of papers initially routed to a given committee as in (2), where $D_{ij} = \sum_{P_k \in C_i} c(w_j, P_k)$ and $c(w_j, P_k)$ is the count associated with paper P_k and the jth word. Rather than use these models in isolation, we combine them with the previously described reviewer centroids for each committee C_{ij} into $C'_{ij} = C_{ij} + \lambda \cdot D_{ij}$, where the parameter λ was optimized in the devtest to be 3. The results are presented in Table 11, and outperform the simple unsupervised model 60.9% to 56.5%, given initial program chair annotation of 1/2 of the data (the devtest set).

The updates to the base centroids were made offline in our method; however, this is not required;

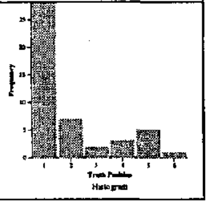
Accuracy:	60.9%
Average position:	1.98
One-of-best-2:	76.1%
Histogram:	

Table 11: Adaptation to the primary judge partial annotation of the data

once the decision is made (a new paper is routed), the "true" label can be used to update the corresponding centroid. There are numerous methods that could be borrowed from AI and IR to implement this strategy, including Active Learning (Lewis and Gale, 1994). Such online adaptation can maximally leverage program chair feedback and minimize the need for initial tagged training data.

5 Automatic Area Committee Generation

In a hierarchical routing system, clearly the composition of the committees is crucial. Suboptimal results are achieved if the 3 most appropriate reviewers for a paper are spread out over different committees.

As an experiment to see if the committee organization could possibly be improved, we investigated empirically committee structures using several clustering strategies.

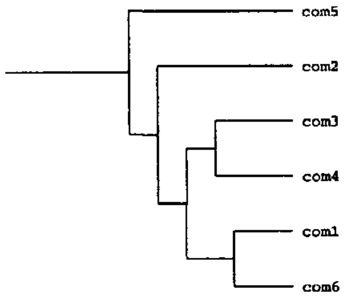
In the first test, we generated a hierarchical agglomerative cluster of the entire reviewer set based on the pairwise cosine similarity between their publication vectors, using maximal linkage clustering (Duda and Hart, 1973; Jain and Dubes, 1988). The results are given in Figure 2a, showing the full tree and extracted cluster list. The numbers in brackets indicate the actual committee assignment of the reviewers; basic inspection will indicate that the derived clusters correspond closely to existing committee compositions (although this information was completely ignored in the clustering process). Analysis of the substructure in the tree shows a natural sub-clustering by research subfocus (e.g. ((isabelle (knight (fung wu))) somers)). Inspection will also show that people with close research focus are spread out among 3 or more different committees, raising some doubts about the optimality of any committee-based routing process.

In another experiment, we tested the extent to which committees could more productively be reformed by beginning with the initial committee centroids and redistributing the reviewers using K-means clustering. We used a modified version of it to obtain reviewer groups that are balanced in size similar to the original committees. This was done by limiting the class size to the the maximum number of reviewers in an original class; the starting point of the algorithm was based on the original committees. The resulting clusters are shown in Figure 2b. The basic initial committee composition is preserved, with some outliers reassigned.

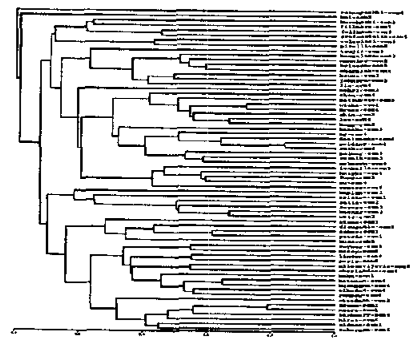
A third experiment was conducted to see if committees could be reconstructed to better match the committee assignment of papers as proposed by the program chair. Specifically, we "reversed" the routing problem by computing committee centroids based on the set of submitted papers assigned to the committee by the program chair, and then routed the reviewers to the committees as if each reviewer was an abstract. In this case, we did not impose any restriction on committee size. The results are shown in Figure 2c. One can still see the original committees in the new organization; the fact that the third committee is large (21 reviewers, almost one third of the whole population) can be probably explained by the fact that the papers routed to committee 3 were interdisciplinary, therefore they had a lot in common with many reviewers.

Another meaningful measure for clustering is the $Sim_{bib}(author_i, author_j)$ based on transitive bibliographic citation and co-authorship (Section 3.4). Figure 2d shows the results of applying maximum linkage agglomerative clustering to this similarity measure. This also shows some correlation with the manually chosen committees.

Finally, it is readily noted by the human judges that certain committees (such as 3 and 4) were quite similar and difficult to distinguish. We can use agglomerative hierarchical clustering of our committee profile centroids to achieve some measure of relative committee distance. The following tree confirms human intuition regarding committee similarity:



One application of this tree and associated distances is to weight the cost of committee misassignments by the severity of the error. The majority of the system errors noted in Table 4 are between (3,4) and (1,6), which this empirical clustering would indicate are relatively low cost mistakes.



- ratnaparkhi[4]
- bel[5]
- kurohashi[2] fillmore[5] fellbaum[5] grefenstette[5] calsolari[5] pirelli[5]
- tsujii[3] bangalore[2] manning[2] briscoe[3] charniak[4] bouma[2] johnson[2] lin[4]
- mohri[3] chen[4] matsumoto[3] wiebe[1] bruce[4] shin[3] lee[4] hang[5]
- tanaka[3] ng[4] daelemans[4] golding[4] roth[4] palmer[5] resnik[3] mckeown[6]
- isabelle[3] knight[3] fung[3] wu[4] somers[6]
- kaplan[2] cristea[1] satta[2] rogers[2] becker[2] weir[2]
- stone[1] dieugenio[1] asher[1] poesio[1] busa[5]
- rayner[3] mccoy[6] linden[6] paris[6] milosavljevic[6] oberlander[6] hahn[1] bateman[6] busemann[6] elhadad[6] rambow[6]
- steedman[2] green[1] moore[1] carberry[6] hirst[6] sidner[1] zukerman[6]

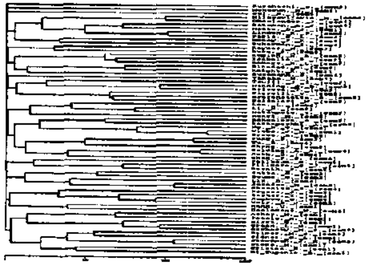
Figure 2a: Committees obtained by average-linkage agglomerative clustering of reviewer papers

- **kcom1** asher[1] busa[5] calzolari[5] cristea[1] dieugenio[1] hahn[1] paris[6] poesio[1] sidner[1] stone[1] wiebe[1]
- **kcom2** bangalore[2] becker[2] bouma[2] johnson[2] lin[4] manning[2] mohri[3] rogers[2] satta[2] tsujii[3] weir[2]
- **kcom3** briscoe[3] charniak[4] fung[3] isabelle[3] knight[3] matsumoto[3] mckeown[6] rayner[3] resnik[3] shin[3] wu[4]
- **kcom4** bruce[4] chen[4] daelemans[4] golding[4] lee[4] ng[4] ratnaparkhi[4] roth[4]
- **kcom5** bel[5] fellbaum[5] fillmore[5] grefenstette[5] hang[5] kaplan[2] kurohashi[2] palmer[5] pirelli[5] somers[6] tanaka[3]
- **kcom6** bateman[6] busemann[6] carberry[6] elhadad[6] green[1] hirst[6] linden[6] mccoy[6] milosavljevic[6] moore[1] oberlander[6] rambow[6] steedman[2] zukerman[6]

Figure 2b: Committees obtained by k-means reclustering of initial committees

- **rcom1** asher[1] cristea[1] dieugenio[1] green[1] moore[1] poesio[1] sidner[1] steedman[2] pirelli[5] carberry[6] hirst[6]
- **rcom2** hahn[1] becker[2] bouma[2] johnson[2] manning[2] rogers[2] satta[2] weir[2] briscoe[3] tsujii[3] lin[4]
- **rcom3** wiebe[1] bangalore[2] isabelle[3] knight[3] matsumoto[3] mohri[3] rayner[3] shin[3] bruce[4] chen[4] daelemans[4] golding[4] lee[4] ratnaparkhi[4]
- **rcom4** fung[3] resnik[3] tanaka[3] charniak[4] ng[4] hang[5] mckeown[6]
- **rcom5** kurohashi[2] fellbaum[5] fillmore[5] grefenstette[5]
- **rcom6** kaplan[2] bel[5] busa[5] bateman[6] busemann[6] elhadad[6] linden[6] milosavljevic[6] oberlander[6] paris[6] rambow[6] zukerman[6]

Figure 2c: Committees obtained by reverse routing reviewers to the centroids of assigned papers



- kurohashi[2] mckeown[6] bel[5]
- milosavljevic[6] oberlander[6] bateman[6] moore[1] linden[6] paris[6] mccoy[6] zukerman[6]
- rayner[3] busa[5] kaplan[2] isabelle[3] somers[6] elhadad[6] knight[3] busemann[6]
- tanaka[3] tsujii[3] manning[2] briscoe[3] pirrelli[5] grenfenstette[5] calzolari[5]
- lee[4] li[5] matsumoto[3] mohri[3] bangalore[2] ratnaparkhi[4] ng[4] wiebe[1] bruce[4] fung[3] wu[4] chang[3] golding[4] roth[4] chen[4]
- fellbaum[5] fillmore[5] resnik[2] satta[2] bouma[2] lin[4] johnson[2] charniak[4] rogers[2] weir[2] hahn[1] daelemans[4]
- asher[1] poesio[1] hirst[6] green[1] carberry[6] palmer[5] sidner[1] steedman[2] stone[1] cristea[1] di eugenio[1]

Figure 2d: Reviewer clusters based on agglomerative clustering using bibliographical similarity

6 System Usage and Confidence Measures for Routing

The routing algorithms presented here have two natural modes of application. The system's committee recommendations can be used either for post-hoc routing error identification (as a sanity check) or for pre-hoc initial automatic assignment with human verification[5] . The latter strategy requires some measure of system confidence for optimal application. Such a measure would help a human judge minimize the time spent in performing the task. If the system is very confident, one might even decide to accept the decision without careful review. On the other hand, in cases where the system is not confident, full attention is required.

Based on the ranked output of the system, we searched for feature transformations whose output can be used in determining confidence intervals. A reasonable one is $\delta = \frac{x_1 - x_2}{x_1}$ where the x_1 and x_2 are the scores associated with the first and second choices of the system. A plot of the averaged accuracy of this operator is depicted in Figure 1 (the value interval was divided in 10 equal and partially overlapping bins and average accuracy was computed on each one of them). The graph on the right shows the accuracy in the case where ranking the gold standard as the system's second committee choice is not considered an error.

One conclusion that can be drawn from the plots is that one can be relatively confident in the system classification if the value of δ is above the 0.25 threshold, while $\delta < 0.1$ tends to indicate lowest expected accuracy and greatest need for careful human inspection. Such confidence measures may also be

[5] The former strategy was actually employed in ACL'99 reviewing.

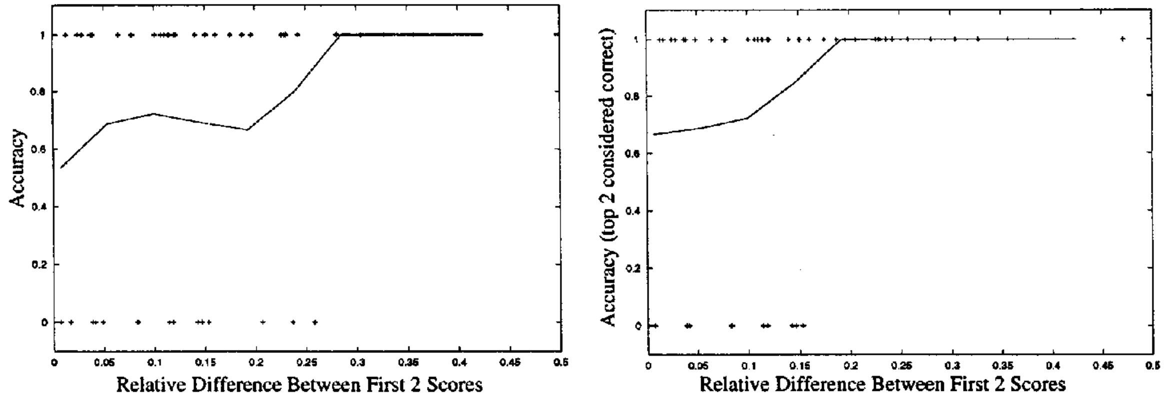


Figure 1: Measures of confidence

used in posthoc correction of human assignments to rank the most likely human errors for re-inspection.

7 Human Performance and Consensus Generation

Area committee routing is a difficult task for humans. Table 12 shows the relatively low inter-judge agreement rates for the 4 judges mentioned in Section 2.2 when annotating the 46-word primary test set. Judge 1 (the program chair) had a slightly different objective function for routing (including the avoidance of conflicts of interest and perhaps some committee size balancing), explaining slightly lower agreement rates than that between the two faculty members (Judges 2 and 3) who had the same task description of finding the most appropriate committee without constraint. Judge 4 was a knowledgeable but less experienced 3rd year graduate student, and his lower performance relative to his colleagues may have been due to more limited familiarity with the reviewers and their expertise.

In order to improve the quality of the gold standard, a consensus standard was generated by taking the majority vote of Judges 1-3. In case of a tie, the program chair was used as the definitive assignment. In nearly 80% of the data, the consensus was identical to the program chair's assignment.

Table 13 illustrates the performance of Judge 4 and the reference Systems (Section 3.1 and 3.4) for both the Judge 1 and Consensus gold standards. Both Judge 4 and the System agreed substantially more with the consensus than the Judge 1 standard, providing some evidence for the relative merit of the consensus standard. The most interesting result, however, is that the system performed better than the graduate student Judge 4 for both standards (although generally lower than the performance of the more experienced faculty members). This suggests that system performance, by virtue of its inherently much greater familiarity with the publications and hence the expertise of the reviewers, more than compensates for its rather limited skills at generalization and inference. This would suggest that the proposed algorithm may be as effective (or even more effective than) human paper routers except for the most knowledgeable human judges.

	Judge$_1$	**Judge$_2$**	**Judge$_3$**	**Judge$_4$**
Judge$_1$	**100**	60.9	65.2	45.6
Judge$_2$	60.9	**100**	73.9	47.8
Judge$_3$	65.2	73.9	**100**	52.2
Judge$_4$	45.6	47.8	52.2	**100**

Table 12: Human judge agreement

The final observation is that in cases where there is high agreement among the human judges, system routing accuracy is also very high. Table 14 divides the data by thresholds of minimum agreement between the Judges 1-3, as the primary partitioning principle using the Section 3.1 system without the Sim_{bib} extension. Given a certain level of agreement (e.g. all 3 judges agree), it's also useful to consider whether the 4th Judge agreed or not with that consensus. By giving the less-experienced 4th Judge an effective 1/2 vote, further refinement in the granularity of consensus can be obtained without effectiving the primacy of the votes of Judges 1-3. In the 57% of the data where only the first 3 judges agree, system accuracy exceeds 80%. In the most confidently classified 35% of the data where all 4 judges agree, system accuracy approaches 88% and in 100% of these cases the consensus committee was one of the system's top two choices. These results strongly suggest that in the clear-cut cases where humans consistently agree on a classification, system performance is very reliable too. The large bulk of system "errors" are in cases where humans tend to disagree as well.

	Judge$_1$	Judge$_2$	Judge$_3$	Judge$_4$	System 3.1	System 3.4
Judge$_1$ assignment = Truth	100	60.9	65.2	45.6	56.5	63.0
Consensus = Truth	78.3	82.6	82.6	56.5	67.4	69.0

Table 13: Human judge and system agreement with 2 goldstandards

Minimum Agreement	% of data	System Accuracy	Average Position	One-of-2-best
1	100	67.4%	1.72	78.3%
1.5	98	68.8%	1.64	80.0%
2	87	75.0%	1.38	87.5%
2.5	72	78.8%	1.30	90.9%
3	57	80.8%	1.23	96.2%
3.5	35	87.5%	1.12	100.0%

Table 14: Routing results given levels of minimum human agreement on committee assignment

8 Conclusions

This paper has presented and extensively evaluated a class of algorithms for automatic routing of submitted papers to reviewers and area committees, without the need for any human annotation from the reviewers or the program chair. Routing is based on a profile of previous writings obtainable on-line for the reviewer pool, a generally stable and reusable resource that requires no manual adaptation for new submission streams. The paper explored a wide set of variations and extensions on the core model, and system accuracy approaches or exceeds that of human judges on the same task. This research demonstrates that such automated paper routing techniques may have merit for paper routing for future conferences, especially those with relatively large and diverse program committees where it is difficult for one person to be familiar with the full range of expertise of all committee members.

9 Acknowledgements

Many thanks are owed to Ken Church and Robert Dale, ACL'99 Program Committee co-Chairs, the ACL'99 area chairs, and the participating reviewers and authors for their essential support of this project. It is hoped that insights gained in this feasibility study will be useful for future program chairs deciding if, where and how to utilize automatic routing strategies in support of their committee and reviewer assignment tasks. This research was partially supported by the National Science Foundation grant IRI-9618874.

References

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harsham. 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. John Wiley.

Susan Dumais and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of SIGIR '92*, pages 233–244, Copenhagen, Denmark.

Haym Hirsh. Personal communication.

D. Hull. 1994. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of SIGIR '94*, pages 282–291, New York.

Anil K. Jain and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice Hall.

L. Larkey and W.B. Croft. 1996. Combining classifiers in text categorization. In *Proceedings of SIGIR'96*.

D. Lewis and W. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin.

Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 1999. Building domain-specific search engines with machine learning techniques. In *Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace*.

F. Mosteller and D. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Massachusetts.

G. Salton and M. McGill. 1983. *An Introduction to Modern Information Retrieval*. New York, McGraw-Hill.

E. Voorhees and D. Harman. 1998. Overview of the 6th text retrieval conference (trec-6). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication, 500-240.