情报学报 ISSN 1000-0135
第29卷第6期1138-1142，2010年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.29 No.6,1138-1142 December 2010

doi:10.3772/j.issn.1000-0135.2010.06.024

学术期刊论文审稿人的自动选择研究[1)]

陆 伟[1] 王 伟[1] 刘 丹[2]

（1. 武汉大学信息资源研究中心，武汉 430072；2. 北京大学信息管理系，北京 100871）

摘要 对于学术期刊论文审稿，通常采用人工阅读待审论文的方式确定审稿人。这种方式不仅耗时，而且带有一定的主观性。如何辅助期刊编辑快速准确地为论文选择审稿人，对于缩短审稿周期、提高论文质量，都具有重要的意义。本文借鉴现代信息检索的思想，分别运用概率模型中的BM25模型和统计语言模型中的Jelinek Mercer平滑模型对论文审稿人的自动选择问题进行了研究。实验结果表明，使用Jelinek Mercer平滑模型自动选择的审稿人能够较好的覆盖论文主题，具有较高的准确性。

关键词 论文审稿人 自动选择 BM25 Jelinek Mercer平滑

Research on Auto Reviewer Assignment for Journal Manuscripts

Lu Wei[1], Wang Wei[1] and Liu Dan[2]

(1. *Center for Studies of Information Resources, Wuhan University, Wuhan* 430072;
2. *Department of Information Management, Peking University, Beijing* 100871)

Abstract When assigning papers to reviewers, the editors often read the papers' information artificially. This approach is not only time consuming, but with a certain degree of subjectivity. How to quickly and accurately assist editors to select reviewers for papers is of great significance to shorten cycle and improve the quality of papers. In this paper, we introduce the idea of information retrieval into this problem, and use BM25 model and Jelinek Mercer smoothing model to study on auto reviewer assignment for papers. We create a review assignment test set and do experiments on it. The results show that reviewers auto-selected by Jelinek Mercer smoothing model can cover the papers' subjects with high accuracy.

Keywords reviewer, auto assignment, BM25, Jelinek Mercer smoothing

1 引 言

我国学术期刊一般实行三级审稿制度，即编辑初审、专家评审、主编或副主编决审[1]。在这三审中，专家评审是整个审稿程序中最关键的一环。为待审论文选择合适的审稿人，是期刊编辑日常工作中的一项重要内容，也是确保刊物学术质量的重要环节。

在传统的审稿人选择工作中，一般先由期刊编辑阅读待审论文的标题、摘要等信息，确定论文的主题，然后从编委或审稿人名录中选择与论文主题对应的审稿人，再将论文转交给审稿人进行评阅，并由审稿人提出修改意见。在这个过程中，一方面由于

收稿日期：2009年10月10日

作者简介：**陆伟**，男，1974年生，博士，教授，主要研究领域：信息检索与智能挖掘、数字图书馆、知识管理等。E-mail：reedwhu@gmail.com。**王伟**，女，1985年生，硕士研究生，主要研究领域：信息检索应用。**刘丹**，女，1988年生，博士研究生，主要研究领域：信息检索与文本处理。

1） 本文系教育部人文社会科学研究项目“专家专长智能识别与检索系统实现研究”和国家自然科学基金重点项目“基于生命周期的数字信息资源深度开发与管理机制研究”成果之一。

编辑自身专业知识的限制，容易在确定待审论文的主题时出现偏差，选择了不合适的审稿人，使所提的评审意见不到位，为稿件的取舍带来一定的困难；另一方面，当待审论文数量较多时，编辑要阅读大量的稿件，不仅耗时耗力，延长了论文的审阅周期，而且还容易由于个人惰性使待审论文主题的确定出现偏差。综合多方面考虑，实现学术论文审稿人的自动选择，不论是对减轻期刊编辑的工作量，还是对缩短审稿周期、提高论文质量，都具有重要的现实意义。

作为对学术期刊论文审稿人自动选择的初步探索，本文以图书情报类中文核心期刊的编委会成员从1999～2008年发表的论文的标题、摘要、关键词等信息为基础，构建了一个测试数据集；同时借鉴现代信息检索的思想，运用两种目前比较成熟的信息检索模型，通过在测试数据集上的具体实验，对期刊论文审稿人的自动选择问题进行了研究，并对两种模型应用于该问题的实验结果进行了分析和评价。

在下文章节2中将介绍国内外审稿人选择的研究现状；章节3将详细介绍本文的研究方法与模型；章节4将介绍本文的测试数据集和实验结果；章节5将介绍存在的问题及未来的研究方向。

2 国内外研究现状

2.1 国外研究现状

国外学者从20世纪90年代初已经开始了学术期刊论文或会议论文审稿人的自动选择研究，并取得了一定的成果。

文献[2]首先运用信息检索思想和隐语义索引(LSI)模型，对会议论文审稿人的自动选择进行了研究。在文献[3]中，作者通过在Web上寻找候选审稿人发表的论文的摘要，然后使用TF-IDF权重对候选审稿人排序，以确定最终人选。文献[4]将该问题看作一种基于文档的专家专长发现，在语言模型的基础上，构建了一种新的主题模型，用来发现审稿人专长，并与提交的论文进行自动匹配。文献[5]使用隐语义索引模型，从待审论文的标题和摘要中识别论文主题，以及从候选审稿人发表的文献的标题中识别其专长，然后运用一个专家系统自动实现论文审稿人的选择。在近期的研究中，文献[6]针对为具有多个主题的论文选择审稿人的情况，分别从审稿人和论文的角度提出了三种不同的研究策略，对这一问题进行了深入的研究，并建立了四个新的测评指标，用于评价每种策略的效果。

2.2 国内研究现状

国内学者对审稿人的选择研究多集中于审稿人选择制度、影响审稿人选择的因素、从哪里选取审稿人等方面，对应用现代信息技术进行审稿人自动选择的文献目前还很少。

文献[7]通过对审稿人选择工作的分析，阐述了审稿人选择的重要性和审稿人在审稿过程中面临的各种问题，提出了一种新的审稿判断标准。文献[8]针对科技论文学科相互渗透、专业交叉的特点，分析了合理选配审稿人应考虑的几个因素。文献[9]提出了利用网络学术数据库，如国家自然科学基金项目数据库、中国期刊网、万方数据库等，多渠道选取期刊审稿人。此外，国内学术期刊的审稿流程正在向网络化、信息化方向转变，一些期刊编辑部已经建立起在线投稿、审稿系统。但是稿件通过网络提交之后，仍然需要编辑人工阅读论文和选择审稿人，没有涉及论文与审稿人的自动化匹配问题。

与审稿人自动选择类似的是，香港城市大学与东北大学的学者合作[10]，运用一种混合知识模型对R&D项目评审人的自动选择进行了研究，并建立了一个原型系统用于实际问题，取得了良好的效果。

3 方法和模型

3.1 研究方法

当前，科学研究越来越呈现出一种多学科交叉融合的趋势，学术论文中大多数会涉及两个或两个以上不同的学科或主题。如果将一篇涉及多个主题的待审论文只提交给一位审稿人进行审阅，由于审稿人自身研究领域和知识的局限性，往往不能全面覆盖待审论文的研究主题，使所提的评审意见不到位，从而影响论文的取舍。因此，在实际的审稿中，一般会把一篇待审论文提交给两个或两个以上的审稿人进行审阅。

如果直接把待审论文的摘要作为查询式，从代表候选审稿人的文档集中检索出前 n 篇文档作为该待审论文的审稿人，可能导致“扎堆”现象，即检索出的前 n 个审稿人只能覆盖待审论文的一个主题，而使其他主题不能被覆盖。因此本文采用聚类的方法，先对每一篇待审论文摘要中的词进行聚集，得到 n 个词类；然后用得到的每一个词类作为一个查询式，从代表候选审稿人的文档集中检索出得分最高的一篇文档，将该篇文档所代表的候选审稿人作为

一个审稿人;最后将 n 次检索出的 n 个候选审稿人作为待审论文的最终审稿人。

笔者的具体做法是:①对每一篇待审论文的摘要进行中文分词,同时去掉无意义的噪声词,得到待聚类的词集合。笔者使用极易软件公司开发的中文分词组件作为处理待审论文摘要的工具,该分词组件使用带词典的正向最大匹配算法和词尾多重消歧技术,支持英文、数字、中文混合分词和中英文噪声词过滤,分词速度和精度良好。②以词集合中的每一对词在代表候选审稿人的文档集中的互信息为依据,使用 K-Means 聚类算法对词集合进行聚类,得到 n 组词类。K-Means 聚类作为一种经典的基于划分的聚类算法,因其理论可靠、算法简单、速度快等优点被广泛用于文本分类、数据挖掘等领域。使用词对在代表候选审稿人的文档集中的互信息作为聚类依据,是由于词与词之间的互信息可以较好地反映词之间的关联程度:两个词同时出现的次数越多,关联程度就越高,互信息也越大;反之关联程度小,互信息小。通过聚类,那些经常同时出现的关系密切的词会聚集在一起,而这些聚集在一起的词又能够一定程度地表征某一主题领域。这样用 n 组词类中的每一个词类作为查询式,从文档集中检索出的 n 个文档就有可能覆盖待审论文的多个主题。

本文的实验流程如图1所示。

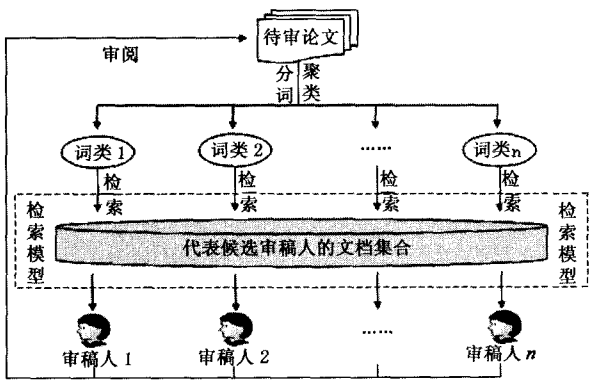

图1 论文与审稿人自动匹配的流程

3.2 使用的检索模型

笔者采用两种当前比较成熟的信息检索模型作为研究论文审稿人自动选择方法中的实验模型,分别是 BM25 模型和 Jelinek Mercer 平滑统计语言模型。本文以 $R=\{r_1,\cdots,r_N\}$ 表示审稿人研究领域的文档集合,r_i 表示描述审稿人 i 的研究领域的文档。

3.2.1 BM25 模型

BM25 模型是目前比较流行的信息检索模型之一,它作为一种典型的概率检索模型,是著名的检索实验系统 Okapi 检索模块的核心模型[11]。该模型的具体公式参见文献[11]。

依据该模型,在给定查询式 q 后,可以获得文档 r_i 的权重得分:

$$W(r_i,q,R)=\sum_j w_j(r_i,R) \tag{1}$$

其中,$w_j(r_i,R)$ 为第 j 个检索词在文档 r_i 中的权重。

3.2.2 Jelinek Mercer 平滑模型

统计语言模型的平滑算法有多种,在实际研究中,应用较多的主要有 Jelinek Mercer 平滑、Dirichlet 平滑和 Absolute Discounting 平滑三种。文献[12]通过实验证明,Dirichlet 平滑适合于长度较短的查询式,而 Jelinek Mercer 平滑应用于长度较长的查询式时效果更好。在本文的研究中,对待审论文摘要进行分词、聚类后,平均每一个词类包含的词数超过10个,这些词构成的检索式长度较长,因此,本文选择 Jelinek Mercer 平滑算法作为实验中统计语言模型的平滑方法。

Jelinek Mercer 平滑算法是一种典型的线性插值平滑方法,它将查询词在文档和文档集合中出现的概率进行加权处理,作为查询词概率的最大似然估计,其公式为:

$$p_\lambda(w_j\mid r_i)=(1-\lambda)p(w_j\mid r_i)+\lambda p(w_j\mid R) \tag{2}$$

其中,$p(w_j\mid r_i)$ 代表第 j 个检索词在文档 r_i 中的统计概率,$p(w_j\mid R)$ 代表第 j 个检索词在文档集 R 中的统计概率。

最终在查询式 q 下,文档 r_i 的得分为:

$$p(r_i,q,R)=\sum_j p_\lambda(w_j\mid r_i) \tag{3}$$

4 实验结果与评价分析

4.1 实验数据

由于没有现成可用的中文数据用来代表审稿人,因此本文选择了图书情报类19种中文核心期刊[13]的编委会成员作为候选审稿人,共192人。然后从"中国知网"上获取每个人从1999~2008年十

年间署名发表的所有学术论文的标题、关键词和摘要，简单的汇总在一个文档中，作为代表其研究领域的特征文档，从而本文得到了一个包含 192 个文档的数据集。另外本文选取《图书情报工作》2009 年第一期和第二期[14]中包含摘要信息的所有学术论文，共 63 篇作为待审论文，用于测试。

要评价方法的有效性，需要有一个评价标准。由于本文是为了研究待审论文与候选审稿人的自动匹配，每次检索只选出得分最高的审稿人，并且使待审论文主题能够被检索出的审稿人最大程度覆盖，因此需要确定每个候选审稿人的研究专长领域，以及待审论文的研究主题。笔者通过人工阅读候选审稿人 10 年间发表的论文的摘要和待审阅论文摘要的方式，确定 192 位候选审稿人的研究领域，以及待审论文的主题，以此作为判断标准，用于评估本文使用的方法的准确性。

4.2 评价指标

本文在每次检索时，只检索出得分最高的一篇文档，传统的查全率和查准率等评价指标不能反映选出的审稿人对待审论文主题的覆盖程度，因此需要使用新的评价指标体系。本文采用文献[6]中建立的评价指标体系中的两个评价指标，分别是 Coverage 和 Confidence。Coverage 用来衡量选出的审稿人对待审论文主题的覆盖程度；Confidence 用来衡量待审论文的每个主题被选择出来的审稿人覆盖的冗余度，即待审论文的每一个主题被尽可能多的审稿人覆盖。Coverage 和 Confidence 越大表示自动匹配方法越好。

这两项评价指标的计算公式分别为：

$$Coverage = \frac{n_r}{n_A} \tag{4}$$

$$Confidence = \frac{\sum_{i=1}^{n_r} \frac{n_{A'_i}}{n}}{n_r} \tag{5}$$

其中，n_A 表示每一篇待审论文的主题数目，$n_{A'_i}$ 表示选择为每一篇待审论文所选择的每一个审稿人所能覆盖的主题数目，n_r 代表为每一篇待审论文选择的 n 个审稿人所能覆盖的主题数目。

4.3 实验结果

本文在分词后对词集合进行聚类时，分别设定聚类的类目数为 2 类、3 类、4 类(类目数代表为每篇待审论文选择的审稿人数)，以比较选择不同审稿人数时两种方法对待审论文主题的覆盖率和冗余度。实验的结果如表 1 所示。

表 1 不同审稿人数目下两种模型的效果比较

模型 \ 指标	2 个审稿人/篇		3 个审稿人/篇		4 个审稿人/篇	
	Coverage	Confidence	Coverage	Confidence	Coverage	Confidence
BM25 模型	0.34	0.37	0.42	0.29	0.55	0.29
Jelinek Mercer 平滑模型	0.68	0.68	0.76	0.54	0.81	0.46

从表 1 可以看出，在本文实验所选择的 2～4 个审稿人范围内，论文主题覆盖率和冗余度方面，Jelinek Mercer 平滑模型都要明显优于 BM25 模型。BM25 模型一般只能覆盖待审论文 50% 左右的主题，而 Jelinek Mercer 平滑模型则能够覆盖待审论文主题的 70%～80%。此外，随着为每篇待审论文选择的审稿人数的增加，待审论文主题的覆盖率逐渐升高，表明为每篇待审论文选择较多的审稿人时可以覆盖更多的待审论文主题；而冗余度逐渐降低，则是由于聚类的类目数越多，聚类后词类之间的关联性就越小，从而使选择出的审稿人所能覆盖的论文主题之间存在更少的交叉。

在测试的待审论文集中，并不是所有的待审论文都具有相同的主题数。本文只考虑了具有两个以上主题的待审论文共 63 篇，其中 2 个主题的论文有 28 篇，3 个主题的论文 30 篇，4 个主题的论文有 5 篇。因此本文根据待审论文的主题数目把待审论文测试集分为三组，编号为 G2、G3、G4，用于详细考察 Jelinek Mercer 平滑模型对不同主题数目的论文的有效性。本文分别统计了不同审稿人数目下，每种待审论文主题数目的主题覆盖率和冗余度，结果如表 2 所示。

表 2 使用 Jelinek Mercer 平滑模型在不同审稿人数目下，不同论文主题数的覆盖率和冗余度比较

指标 \ 分组	2 个审稿人/篇			3 个审稿人/篇			4 个审稿人/篇		
	G2	G3	G4	G2	G3	G4	G2	G3	G4
Coverage	0.7	0.68	0.58	0.81	0.72	0.71	0.84	0.8	0.71
Confidence	0.68	0.68	0.71	0.54	0.53	0.55	0.48	0.45	0.45

在本文所选择的审稿人数范围内，随着待审论文主题数目的增加，对待审论文主题的覆盖率都有所下降，而冗余度的变化则不太明显。这表明待审

论文的主题数目越多,自动选择的审稿人覆盖所有的论文主题的难度也越大。

尽管Jelinek Mercer平滑模型还不能达到主题的完全覆盖,但是已经能够说明使用Jelinek Mercer平滑模型作为实现待审论文与审稿人自动匹配的方法具有一定的可行性。

5 结 论

本文借鉴信息检索的思想,分别采用概率模型中的BM25模型和统计语言模型中的Jelinek Mercer平滑模型,在构建的测试数据集上,对学术期刊论文审稿人的自动选择问题进行了实验研究,并对实验的结果进行了评价。评价结果显示,使用Jelinek Mercer平滑模型,在待审论文主题的覆盖率和冗余度方面,都比BM25模型的效果更好。因此可以使用Jelinek Mercer平滑模型作为实现待审论文审稿人自动选择的方法。

本文所描述的学术期刊论文审稿人的自动选择方法不仅可以用于学术期刊领域,还可以运用在与此相似的领域,如学术会议论文和博士学位论文审稿人选择匹配。

由于本文在实验中需要对测试论文的摘要进行中文分词和聚类,因此使用的中文分词算法和聚类算法会对构造的检索式有影响,从而影响最后的检索精确度。另外由于使用的数据集与测试论文集数据都比较少,所获得的结论还需要在更多的数据上验证其有效性。未来笔者将对不同的分词算法和聚类算法的有效性作进一步比较,并将构建更大规模的数据集进行广泛的测试和验证。此外,如何根据摘要信息的特征进行聚类数目的动态分配,以实现实验与实际应用的最大符合,也有待于进一步深入研究。

参 考 文 献

[1] 施才能. 选准审稿专家是确保审稿质量的关键[J]. 编辑学报, 1995, 7(4):198-199.

[2] Dumais S, Nielsen J. Automating the assignments of submitted manuscripts to reviewers[C]// Proceedings of SIGIR 1992, 1992:233-244.

[3] Basu C, Hirsh H, Cohen W, et al. Recommending papers by mining the web[C]// Proceedings of IJCAI Workshops on Learning about Users and Machine Learning for Information Filtering, 1999.

[4] Mimno D, McCallum A. Expertise modeling for matching papers with reviewers[C]// Proceedings of the 13th ACM SIGKDD, 2007:500-509.

[5] Ferilli S, Di Mauro N, Basile T, et al. Automatic Topics Identification for Reviewer Assignment[J]. Advances in Applied Artificial Intelligence, 2006:721-730.

[6] Maryam K, Zhai C X, Geneva B. Multi-Aspect Expertise Matching for Review Assignment[C]// CIKM'08: Proceeding of the 17th ACM conference on Information and knowledge management, 2008:1113-1122.

[7] 汪宏晨, 柳建乔. 基于新审稿判断标准的审稿人制度的研究[J]. 湖北民族学院学报(自然科学版), 2005, 23(4):408-410.

[8] 蔡玉琪. 合理选配审稿人是提高审稿质量的有效途径[J]. 农业图书情报学刊, 2007, 19(2):135-136.

[9] 黎贞崇, 唐莲英. 网络审稿人的选择及实现方法[J]. 广西广播电视大学学报, 2004, 15(1):68-70.

[10] Sun Y H, Ma J, Fan Z P, et al. A hybrid knowledge and model approach for reviewer assignment[C]. 40th Annual Hawaii International Conference on System Sciences, 2007: 817-824.

[11] Lu W, Robertson S E, Macfarlane A, et al. Window-based Enterprise Expert Search[C]// Proceedings of the 15th Text Retrieval Conference (TREC 2006), Gaithersburg, MD, USA, 2006.

[12] Zhai C X, Laferty J. A Study of Smoothing Method s for Language Models Applied to Ad Hoc Information Retrieval [C]. Proceedings of the ACM-SIGIR, 2001:334-342.

[13] 北大核心期刊要目总览(2009版): 图书情报类[EB/OL]. [2009-06-29]. http://162.105.140.111/tugongwei/info/detail.asp? lngID=459.

[14] 《图书情报工作》[EB/OL]. [2009-07-06]. http://www.lis.ac.cn/CN/article/showTenYearVolumnDetail.do? nian=2009.

(责任编辑 王建平)