一种基于本体的稿件-审阅人相关度度量方法[†]

肖刘明镜[1]，周志[2]，邹小军[2]，胡俊峰[2,*]

（1. 北京大学 信息科学技术学院,北京 100871;

2. 北京大学 计算语言学教育部重点实验室,北京 100871）

摘要：随着稿件数量的不断增长，审阅人指派越来越成为国内外各会议期刊编委和基金委员会的一项费时费力的工作，计算机辅助审阅人指派研究应运而生。而稿件-审阅人相关度度量则是其中的一个重点研究问题。本文在关键词相似度的基础上，设计了一种基于本体的稿件-审阅人相关度度量方法。该方法涵盖了关键词提取、关键词分布相似度计算、本体自动构建和基于网络流的稿件-审阅人相关度计算等部分。每一部分均由计算机自动完成，减轻了人工指派的压力。初步实验表明，本文方法优于基于关键词字串的相关度计算方法。

关键词：审阅人指派；相似度计算；本体；信息检索

An ontology based manuscript-reviewer correlation measurement method

XIAO Liumingjing[1], ZHOU Zhi[2], ZOU Xiaojun[2], HU Junfeng[2]

(1. School of Electronics Engineering & Computer Science, Peking University, Beijing, 100871, China;

2. Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing, 100871, China)

Abstract: With the growing amount of manuscripts, reviewer assignment is increasingly becoming a laborious task for conference organizers, journal editors and foundation committees, which stimulates the growth of automatic reviewer assignment researches. Similarity measurement between manuscripts and reviewers is a key research issue in reviewer assignment problem. Based on keyword similarity, this paper designed an ontology based manuscript-reviewer correlation measurement method. This method covers keywords extraction, keyword distributional similarity computation, automatic ontology construction and network flow based manuscript-reviewer correlation measurement. Each part does not need manual interventions, which saves human efforts in reviewer assignment. Preliminary experiments show that this method outperforms string similarity based method.

Keywords: Reviewer assignment; Similarity computation; Ontology; Information retrieval

1 引言

在学术界，将候选稿件准确高效地投递到合适的审阅人手中是一项很重要的工作，它关系到论文评审、基金申请等工作的公正性、合理性。近年来，稿件规模数量呈现大量增长且有继续增长的趋势。截止到 2012 年 3 月国家自然科学基金（The National Natural Science Foundation of China, NSFC）申请截稿时，共收到各类项目申请 170792 项，比 2011 年同期增加 23089 项[1]。每份申请书会被指派给 4-5 名同行审阅专家，如此大量的申请书指派任务已经成为一个非常繁重的工作。

审阅人自动指派需要考虑一系列问题：1. 稿件与审阅人的专业领域、研究兴趣点要相合，尽可能提高稿件与审阅人之间的匹配程度；2. 每份稿件要有一定数量一定资历的审阅

[†]**基金项目：**国家自然科学基金资助项目（M1321005）

[*]**作者简介：**胡俊峰（1967—），通信作者，男，副教授，主要研究方向为自然语言处理，知识发现；肖刘明镜（1992—），男，本科生，主要研究方向为自然语言处理。

[1] 数据来源：http://www.nsfc.gov.cn/nsfc/cen/xmzn/2013xmzn/index.html

人审阅来保证公平公正；3. 每位审阅人审阅的稿件数量不能太多，需要均衡审阅人的工作负载；4. 稿件作者与审阅人之间的回避关系，比如曾经是共同作者，或者在同一个研究工作机构工作。本文主要关注第一点，即稿件与审阅人之间的匹配程度度量的优化。

审阅人自动指派问题已经受到学术界越来越多的关注，尤其是稿件-审阅人相关度的计算问题，国内外学者在这一问题上进行过大量的研究。最早研究这一问题的学者是 Dumais 和 Nielsen，他们提出了一种基于潜在语义索引（Latent Semantic Indexing， LSI）的稿件-审阅人相关度计算方法，即将稿件和审阅人投射到 k 维潜在语义空间中，使用向量夹角余弦值计算稿件-审阅人相关度[1]。Biswas 和 Hasan 等人采用向量空间模型（Vector Space Model, VSM）来计算稿件-审阅人的匹配程度[2]。为了改进算法效率，他们还尝试用自动提取关键词来替代用所有词语表示文本向量，以及基于 ACM 计算机分类系统（Computing Classification System，CCS）构建的领域本体来计算稿件与审阅人的相关度。马建等利用国家自然科学基金的资助项目建立了一个研究本体，并基于该本体进行自然科学基金申请书的指派[3]。Yarowsky 和 Florian 等人综合了审阅人的论文发表信息，采用朴素贝叶斯方法计算稿件与领域委员会的相关度，他们还尝试对审阅人所发表论文进行层次化聚类，自动构建领域委员会的层次结构[4]。Watanabe 等人引入关键词协作网络来计算稿件-审阅人相关度，并提出了话题新奇程度度量来优化指派效果[5]。

本文提出了一种基于本体的稿件-审阅人相关度计算方法，该方法先通过关键词在本体树上的距离来度量关键词之间的语义相似度，再构建网络流模型计算稿件-审阅人相关度，以此来优化审阅人指派的准确率。该方法一个有所改进的地方是不仅仅使用了稿件的文本内容，还通过引入本体作为外部知识，补充了小文本语料信息不足的缺陷。另一个改进的地方是通过计算机从文本语料中自动学习本体，用于优化关键词匹配程度的计算，而不是通过人工建立小规模本体用于辅助稿件指派。该方法的主要优势在于不仅仅基于关键词字串相似度进行稿件-审阅人匹配，，来发现关键词之间的近义关系。本文后续各节组织如下，第二节简述本体及本文所采用的面向文本的本体学习技术，第三节详述我们提出的稿件-审阅人相关度度量方法，第四节通过实验比较本文方法和基于关键词字串的相关度计算方法，第五节总结全文并提出展望。

2 本体与本体学习

在计算机科学与信息科学领域，本体是指一种“形式化的，对于共享概念体系的明确而又详细的说明”[6]。本体是一种特殊类型的术语集，有明显的结构化特点，它表达了特定领域中对象类型或概念及其属性和相互关系[7]。作为现实世界中的概念知识相关关系形式化表达，本体目前的应用领域包括信息架构、信息学、软件工程、人工智能、语义网等。本体大多描述的都是个体、类、属性以及关系。领域本体是针对某个特定领域，或者现实世界的一部分建模，表达概念在适用于该领域中的特殊含义。Erol[8]给出了本体的形式化定义：

一个本体结构是一个五元组，可表示为$O:=(C,R,\leq_C,\sigma,\leq_R)$。其中，$C$表示概念集，其元素称为概念标识符 c（concept identifiers）；R表示关系集，其元素称为关系标识符 r（relation identifiers），例如$c_1 \times c_2 \ldots \times c_n$，表示$c_1$，$c_2$，…，$c_n$之间存在 n 元关系；$\leq_C$表示概念集C的某个偏序，称为概念层次（concept hierarchy）；σ表示特殊关系$R \rightarrow C^+$；$\leq_R$表示关系集R的某个偏序，称为关系层次（relation hierarchy）。当$r_1 \leq_R r_2$，意味着$|\sigma(r_1)| = |\sigma(r_2)|$。

目前很多的本体都是手工构建的，这个过程需要耗费大量的人力、物力和财力，不仅开发周期长，而且极易出错。例如 Cyc[9]和 WordNet[10]等系统需要大量的专家人工为本体输入知识，然后其它系统和应用才能运用这些知识进行推理或者获取新的知识。鉴于手工方式费时费力，实

现本体的自动或者半自动构建的技术—本体学习（ontology learning）[11]技术应运而生。何劭达等[12]在分布相似度的基础上，设计了一种基于 HITS[13]的面向中文文本的本体学习框架。该框架可以根据文本语料自动学习得到该语料范围内的本体层次结构，其中第底层为术语，以上各层均为概念，概念用一个同义词的集合来表示。实验表明[12]该算法在概念发现和概念层次聚类上均优于 Google 基于 RNN 和 k-means 的算法。由于学习得到的本体包含的是该语料范围内的领域知识，因而它能够辅助我们衡量该领域下关键词之间的语义距离，而不仅仅是关键词之间的字串距离，并且我们可以随时增加或改变用于本体学习的语料，以适应时代的变化，因此本文借助该本体学习框架生成领域本体并借以改进稿件-审阅人相关度度量。

3 稿件-审阅人相关度度量

3．1 稿件及审阅人建模

我们采用周志等提出的基于带权复杂网络的关键词提取算法[14]，对每一份稿件自动提取关键词列表。这种算法提取出来的关键词是按重要程度排序的，排列越靠前，表示该关键词在列表中越重要。实验表明[14]，该算法优于经典的 TF-IDF[15]关键词提取算法，因此本文选用该算法提取的关键词列表来表征稿件的内容信息。同时我们也综合了作者在稿件中自己给定的关键词，直接插入到自动提取的关键词列表的前面并对列表中的关键词去重。表 1 给出了我们对一篇稿件进行关键词提取的结果，关键词前面的序号表示它在列表中的排名。

表 1 对一篇稿件进行关键词提取的结果

序号	关键词						
1	图像压缩算法	8	低复杂度	15	JPEG 算法	22	像素
2	全色调 BTC	9	计算复杂度	16	量化	23	立足点
3	图像重构	10	编码效率	17	隐含	24	去除
4	空间信息	11	BTC 算法	18	视觉	25	三维虚拟现实
5	图像质量	12	远程监控	19	图像处理		
6	半色调	13	信号处理	20	信息系统		
7	二值信息	14	脉冲噪声	21	视频压缩算法		

在审阅人建模上，我们采用前述方法对审阅人过去发表过的论文摘要也自动提取了关键词列表，同时还考虑了审阅人在已发表论文中自己给定的关键词，综合起来作为论文摘要的关键词列表。我们统计每一篇摘要中得到的关键词列表中各词的出现频率，按频率降序排列起来作为该审阅人的关键词列表，用来表征该审阅人。

3．2 词语分布相似度计算与领域本体构建

我们采用了何邵达等提出的计算词语分布相似度的方法[12]，并应用于专业文本语料的领域本体学习中。该方法在词汇上下文相似度的基础上，通过词性模板引入了词性上下文相似度，二者结合起来表示词语分布相似度。该方法支持在较小的专业文本语料库中学习领域本体。具体工作流程如下：

1) 通过上下文共现计算词语两两间的点互信息（PMI）（本文以窗长为 5 计算上下文共现），把 PMI 向量作为词汇上下文特征向量，采用其夹角余弦值作为词语间的词汇上下文相似度。
2) 对文本做词性标注，对每个词语统计词性模板频率（本文以窗长为 3 划定词性模

板）作为词性模板向量，采用其夹角余弦值作为词语间的词性上下文相似度。最终计算词汇上下文相似度与词性上下文相似度的乘积作为词语间的分布相似度。词语分布相似度定义如下：

$$Sim_{w_iw_j} = Sim_{w_iw_j}{}^{lex} \cdot Sim_{w_iw_j}{}^{pos} \tag{1}$$

$$Sim_{w_iw_j}{}^{lex} = \frac{PMI_{w_i} \cdot PMI_{w_j}}{|PMI_{w_i}| \cdot |PMI_{w_j}|} \tag{2}$$

$$Sim_{w_iw_j}{}^{pos} = \frac{PAT_{w_i} \cdot PAT_{w_j}}{|PAT_{w_i}| \cdot |PAT_{w_j}|} \tag{3}$$

$$PMI_{w_iw_j} = \log \frac{Freq_{w_iw_j} \cdot \sum_w Freq_w}{Freq_{w_i} \cdot Freq_{w_j}} \tag{4}$$

其中，$Sim_{w_iw_j}$表示词语w_i和w_j的分布相似度，$Sim_{w_iw_j}{}^{lex}$表示词汇上下文相似度，$Sim_{w_iw_j}{}^{pos}$表示词性上下文相似度，PMI_{w_i}表示词汇上下文特征向量，PAT_{w_i}表示词性模板向量，$Freq_{w_i}$表示w_i的词频，$PMI_{w_iw_j}$表示w_i和w_j的共现频率。

本文中我们均采用了中科院分词软件 ICTCLAS[16]进行分词和词性标注。我们在收集的相关专业领域的文档集中用该方法计算得到该领域下的词语分布相似度矩阵。基于该矩阵，我们采用何邵达等提出的基于 HITS 的本体学习方法自动生成专业领域本体，实验表明[12]该算法在概念发现和概念层次聚类上均优于 Google 基于 RNN 和 k-means 的算法。领域本体使用树状的 XML 格式存储，通过树形结构描述了概念及概念间的从属关系。我们采用网上收集的信息科学领域科技文献摘要，来生成该领域的本体[2]。在本文中，我们还增加了 NSFC 提供的部分数据资源作为语料生成电子学与信息系统领域本体[3]。

3．3 稿件-审阅人相关度计算

对于稿件的关键词w和审阅人的关键词w'，我们通过关键词的字串相似度（本文采用两个字符串的最长公共子序列长度来计算）以及关键词的语义相似度（本文采用关键词在本体树上的距离来计算）综合计算两个关键词之间的匹配程度：

$$Match_{ww'} = \gamma \cdot Match_{ww'}{}^{sem} + (1-\gamma) \cdot Match_{ww'}{}^{str} \tag{5}$$

$$Match_{ww'}{}^{sem} = \frac{1}{\left(\frac{Dis_{ww'}{}^{tree}}{2}\right)^2 + 1} \tag{6}$$

$$Match_{ww'}{}^{str} = \frac{LCS_{ww'}{}^2}{Len_w \cdot Len_{w'}} \tag{7}$$

其中，$Match_{ww'}$表示稿件的关键词w和审阅人的关键词w'的匹配程度，$Match_{ww'}{}^{sem}$表

[2] 数据链接：http://www.klcl.pku.edu.cn/clr/ontology/ontology_InformationScience.rar
[3] 由于使用授权限制无法提供公开下载

示w和w'的语义相似度，$Match_{ww'}{}^{str}$表示w和w'的字串相似度，$LCS_{ww'}$表示 w和w'的最长公共子序列长度，Len_w和$Len_{w'}$分别表示w和w'的字符串长度，$Dis_{ww'}{}^{tree}$表示w和w'在本体树上的距离，γ为调节字串相似度和语义相似度比重的参数。

我们引入网络流算法[17]来计算稿件与审阅人之间的相关度。我们构建如图 1 所示的网络 $G=(V, E)$，其中 $V=M\cup R\cup\{s, t\}$ 为顶点集, M 表示稿件的关键词集合，R 表示审阅人的关键词集合，s 和 t 分别表示网络的源点和汇点， E 为边集。E 中每条边(u,v)均有容量属性 $Cap_{u,v}$。边集 E 由以下规则确定：

1) 对 M 中的每个关键词 m，令源点 s 到 m 的边容量 $Cap_{s,m}$=1；

2) 对 R 中的每个关键词 r，令 r 到汇点 t 的边容量 $Cap_{r,t}$=1；

3) 对 M 中的每个关键词 m，R 中的每个关键词 r，令 $Cap_{m,r}=Match_{m,r}$（若 $Cap_{m,r}$=0，则(m,r)不属于边集 E）。

对上述网络从源点 s 向汇点 t 求解最大流。$M\cup R$ 中每个顶点满足入流流量等于出流流量，E 中每条边满足流量不超过容量的限制。我们将整个网络的最大流的流量作为稿件与审阅人的相关度。最大流流量越大，表明稿件与审阅人的匹配程度越高。采用网络流模型的优势在于，既通过关键词之间的匹配程度作为边容量来增加流量，又通过关键词与源点或汇点间的边容量限制来防止单个关键词过度增大整个网络的流量。

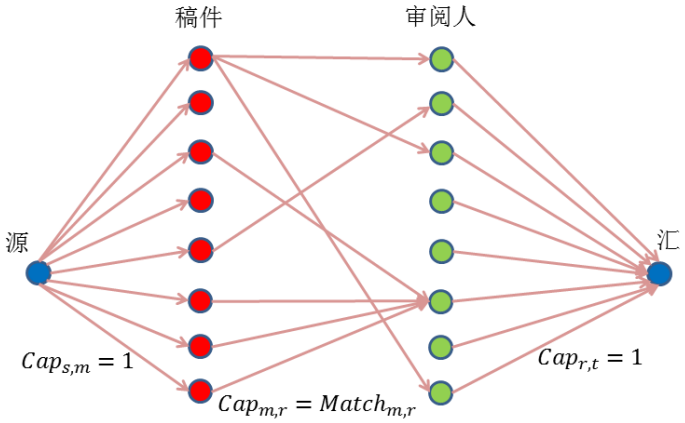

图 1 稿件-审阅人相关度计算的网络流模型

4 实验

我们收集了 NSFC 已经结题的一些项目申请书作为评测数据。NSFC 设立了八个主要的学部，分别用 A 到 H 表示，每个学部自顶向下建立了共三级学科分类体系，每一级学科由一个唯一的学科分类代码来表示。表 2 给出了信息科学学部的学科分类体系的部分结构，其中“F01 电子学与信息系统”是信息科学学部的一个一级学科代码，其下包括“F0101 信息理论与信息系统”、“F0102 通信理论与系统”等二级学科代码，再往下一层均为三级学科代码。

申请人在提交基金申请书时，被要求填写两个最合适的学科分类代码。而在实际的申请书指派工作中，所有评审专家也会被要求填写五个以内的最熟悉的学科分类代码。学科代码在申请书-审阅人指派工作中扮演了很关键的作用。在一些情况下，申请书和审阅人被

按学科分类代码直接分成组，然后由专业人员直接进行指派。因此，如果我们能把申请书正确的投送到合理的学科分类代码下，就意味着初步实现了对审阅人的智能指派。

表 2 信息科学学部学科分类体系的部分结构

F01 电子学与信息系统			
F0101	信息理论与信息系统	F0102	通信理论与系统
F010101	信息论	F010201	网络通信理论与技术
F010102	信源编码与信道编码	F010202	无线通信理论与技术
F010103	通信网络与通信系统安全	F010203	空天与水下通信
F010104	网络管理与服务	F010204	多媒体通信理论与技术
F010105	信息系统建模与仿真	F010205	光、量子通信理论与系统
F010106	认知无线电	F010206	计算机通信理论与系统

考虑到实际的申请书指派结果具有保密要求，不适宜用作评测数据，本文的研究就把申请书对学科分类代码的指派模拟为对审阅人的指派来完成对本文算法的有效性测试。学科分类代码采用类似前述审阅人建模的方式，从填报该学科分类代码的申请书的摘要文本集合中产生关键词列表。首先默认两个用户自填学科分类代码作为申请书-审阅人指派的标准答案。用本文中基于本体的稿件-审阅人相关度度量方法为申请书指派相关度最高的两个学科分类代码，并与用户自填代码进行比较。如果本文方法在指派学科分类代码上的准确率有所提升，那么有理由认为我们的方法在审阅人指派问题上相比基线方法指派的效果也会更佳。

为了验证本方法的有效性，我们设计了如下实验。我们选用三种不同的稿件-审阅人相关度计算方法来比较指派的准确率。方法一是只考虑关键词的字串相似度，即令公式(5)中的γ等于零，并且仅采用作者在稿件中自己给定的关键词作为关键词列表。方法二相对于方法一引入了 3.1 节中的方法来对稿件进行建模，即综合了作者自己给定的关键词和关键词提取算法自动提取出来的关键词列表。方法三是采用本文提出的基于本体的稿件-审阅人相关度计算方法。实验中选取的参数为γ=0.6，每篇稿件选取前 10 个关键词，每个学科分类代码选取前 30 个关键词。

我们采用 NSFC“F01 电子学与信息系统”领域已结题的项目申请书作为评测的数据集，实验结果如表 3 所示。其中指派一个代码的实验是指派相关度最高的一个学科分类代码，只要自动指派的代码包含在用户自填的两个代码中，就算作正确指派。指派两个代码的实验是指派相关度最高的两个学科分类代码，如果自动指派的两个代码与用户自填代码完全相同，才算作正确指派，如果只有一个代码包含在用户自填代码中，则只算作 50%的正确率。实验结果如表 1 所示，可以看出，基于本文方法的学科分类代码指派在准确率上相对于其他两种方法有所提升。在指派一个代码的实验中，本文方法比方法一和方法二的准确率分别提高了 14.80%和 2.00%，在指派两个代码的实验中，本文方法比方法一和方法二的准确率分别提高了 15.79%和 3.16%。方法二之所以优于方法一，主要是因为方法二在稿件建模上引入了自动提取的关键词，表征稿件内容信息更为准确丰富；本文方法通过领域本体挖掘了关键词间的近义关系，因此相比于方法一和方法二中基于字串相似度的指派更为准确。

表 3 三种方法进行学科分类代码指派的比较

所用方法	指派一个代码			指派两个代码		
	总数	正确数	准确率(%)	总数	正确数	准确率(%)
方法一	1000	633	63.30	1679	924	55.03
方法二	1000	761	76.10	1679	1136	67.66
本文方法	**1000**	**781**	**78.10**	**1679**	**1189**	**70.82**

5 结论与展望

本文设计了一种基于本体的稿件-审阅人相关度度量方法，该方法结合了关键词字串相似度与基于领域本体计算的语义相似度，并通过构建网络流模型计算稿件-审阅人的相关度。初步实验表明，我们的方法在指派准确率上要优于基于关键词字串相似度的方法。

下一步工作我们将从以下几个方面展开，一是优化网络边容量的设计，本文目前的设计存在较多经验的因素，例如参数和函数选取等。今后可以考虑通过机器学习的方法来优化参数的设定。二是考虑引入话题模型，综合语义相似度和文本话题相似度来优化稿件-审阅人相关度的度量。

参考文献：

[1] Dumais S T, Nielsen J. Automating the assignment of submitted manuscripts to reviewers[C]//Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1992: 233-244.

[2] Biswas H K, Hasan M. Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment[C]//Information and Communication Technology, 2007. ICICT'07. International Conference on. IEEE, 2007: 82-86.

[3] Ma J, Xu W, Sun Y, et al. An ontology-based text-mining method to cluster proposals for research project selection[J]. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2012, 42(3): 784-790.

[4] Yarowsky D, Florian R. Taking the load off the conference chairs: towards a digital paper-routing assistant[J]. 1999.

[5] Watanabe S, Ito T, Ozono T, et al. A paper recommendation mechanism for the research support system papits[C]//Data Engineering Issues in E-Commerce, 2005. Proceedings. International Workshop on. IEEE, 2005: 71-80.

[6] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge acquisition, 1993, 5(2): 199-220.

[7] Fensel D. Ontologies[M]. Springer Berlin Heidelberg, 2001.

[8] Bozsak E, Ehrig M, Handschuh S, et al. KAON—towards a large scale Semantic Web[M]//E-Commerce and Web Technologies. Springer Berlin Heidelberg, 2002: 304-313.

[9] Lenat D B, Guha R V. Building large knowledge-based systems; representation and inference in the Cyc project[M]. Addison-Wesley Longman Publishing Co., Inc., 1989.

[10] Snasel V, Moravec P, Pokorny J. WordNet ontology based model for web retrieval[C]//Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in. IEEE, 2005: 220-225.

[11] Sanderson M, Croft B. Deriving concept hierarchies from text[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 206-213.

[12] He S, Zou X, Xiao L, et al. Construction of Diachronic Ontologies from People's Daily of Fifty Years[C]//Proceedings of the 9th edition of the Language Resources and Evaluation Conference. 2014.

[13] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632.

[14] Zhou Z, Zou X, Lv X, et al. Research on Weighted Complex Network Based Keywords Extraction[M] //Chinese Lexical Semantics. Springer Berlin Heidelberg, 2013: 442-452.

[15] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[16] Zhang H P, Liu Q, Cheng X Q, et al. Chinese lexical analysis using hierarchical hidden markov model[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 63-70.

[17] Ahuja R K, Magnanti T L, Orlin J B. Network flows: theory, algorithms, and applications[J]. 1993.

作者：肖刘明镜

通讯地址：北京市海淀区颐和园路 5 号，北京大学信息学院计算语言所

邮编：100871

手机：15201318039

E-mail 地址：xlmj531@163.com

作者：胡俊峰

通讯地址：北京市海淀区，北京大学信息学院计算语言所

邮编：100871

手机：13311322031

E-mail 地址：hujf@pku.edu.cn