基于多标签分类的引文全局功能识别研究*

刘兴帮[1,2]，陆伟[1,2]，孟睿[1,2]
(1. 武汉大学信息资源研究中心，武汉 430072；2. 信息检索与知识挖掘研究所，武汉 430072)

摘要：引文功能是科研工作者引用一篇文献的动机。其中，相比较于只考虑引文前后文语句的引文局部功能，引文全局功能关注的是参考文献在全文范围内的信息，是被引文献在施引文献中价值的综合体现，其自动识别研究对于引文推荐、引文索引、语义化引文网络构建等学术文本挖掘研究具有重要意义。文章根据“参考文献在施引文献中存在一处或多处具体引用”这一特点，将引文全局功能识别研究转化为多标签分类问题，并构建引文全局功能数据集，在此数据集之上进行引文全局功能自动识别实验，取得较好的效果。

关键词：引文全局功能；多标签分类；学术文本挖掘；引文分析
中图分类号：G353.4
DOI：10.3772/j.issn.1673-2286.2016.3.001

1 引言

引文功能描述了科研人员引用他人科研成果的目的和动机[1]。引文功能识别是学术文本挖掘（如重要引文识别、引文网络构建和科研成果评价指标构建等）的基础性工作之一，在文献计量、研究主题演化分析、科研趋势预测等多个领域具有重大应用价值。近年来，随着文献数量的快速增长以及文本技术的发展，引文功能的识别逐渐由传统的人工识别转向基于机器学习、数据挖掘等技术的自动识别。

根据引文作用范围的不同，引文功能的自动识别可以分为引文局部功能识别（local classification）和引文全局功能识别（global classification）[2]。局部功能指的是一篇文献在施引文献中的某一具体引用处体现功能；全局功能则是该参考文献在施引文献中体现的整体功能。如Simone Teufel[1]分别在第二章和第三章引用了Greg Myers[3]的内容，引文局部功能识别是分别识别Greg Myers的文章在第二章和第三章被应用所体现的功能；引文全局功能识别则是识别Greg Myers的文章在施引文献中的整体作用。

现有的引文全局功能识别研究主要是将全局功能识别问题转化成一个或者多个局部功能识别问题，随后将局部功能进行整合，选取权重最高的一种或数种作为全局功能。这种引文全局功能识别本质上是引文局部功能识别。本文认为这种引文全局功能识别方法存在以下不足：①一篇参考文献的各引文局部功能之间存在一定隐含联系，单独识别一篇参考文献的各引文局部功能，忽视了这种可以作为识别特征的隐含联系，会在一定程度影响引文功能识别的准确率。②一篇参考文献的各个引文局部功能均是施引文献和被引文献之间引用关系的体现，任何一个单一功能并不能准确将该参考文献的所有引文功能概括。

针对这些不足，笔者提出基于多标签分类的引文全局功能识别方法。这种多标签表示方式在更加准确及全面描述引文功能的基础上，能够为其他研究如引文网络构建、引文演化、引文推荐等提供更丰富而全面的线索。

本文的主要贡献如下：①在Dong、Schäfer等的引文局部功能标注数据[4]的基础上构建了引文全局功能数据集，构建的数据集能够为相关研究工作提供数据支持。②提出了一种基于多标签分类的引文全局功能识别方法，该方法相较于已有方法能够更加有效地识别引文全局功能。

文章后续组织如下：第二部分对相关研究进行调研；第三部分具体阐述基于多标签分类的引文全局功能识别方法，并对本研究所使用的引文分类体系、分类方法及特征方案进行介绍；第四部分给出实验思路和

* 本研究得到国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”（编号：71473183）和武汉大学与中国科学技术信息研究所合作项目“科学文献的语义功能识别与深度利用”资助。

实现结果，并对实验结果进行分析；本文最后分析了研究的不足，并提出下一步的工作思路。

2 相关研究

引文功能识别研究早期使用的方法主要是人工标注、统计等[5-6]，随着引文规模的增加及计算机技术的发展，引文功能的自动识别研究成为研究热点。

引文局部功能自动识别相关研究成果较多。早期的学者通过分析具体引用处的一些语言特征，手工构造一些规则进行引文功能的自动识别。1997年，Garzone[7]手工构建了包含35个类目的分类体系，并通过手工构建的语法解析规则及模板匹配规则在20篇文献上进行了引文功能识别实验，识别准确率为84%。1999年，Nanba与Okumura[8]改进了Garzone的分类体系，将引文句的功能分为“基于”“比较”以及“其他”，并通过构建特征词表的方式进行引文功能识别，在构建的数据集上识别准确率为83%。

由于手工构造的规则覆盖度有限，基于规则和模板的方法在大规模数据上识别准确度不高。学者逐渐开始采用机器学习方法识别引文功能：2006年，Teufel[1]首次实现了将机器学习分类应用于引文局部功能识别，构建了包含12个类别的分类体系，通过KNN算法在其自建的数据集上取得了57%的宏F值；2008年，Radoulov[9]将引文局部功能识别问题转化为词义消歧问题，构建了包含9个类别的分类体系，并使用朴素贝叶斯网络方法在其自建的数据集上取得了69.4%的宏F值识别效果；2011年，Dong与Schäfer[4]认识到手工标注的局限性，尝试使用半监督方法识别引文局部功能，他们构建了包含4个类别的分类体系，并采用ensemble分类器，在自建数据集上达到了66%宏F值的识别效果；2012年，Jochim[10]使用了一个基于二元选择的分类体系，并提出了若干语言结构特征，使用最大熵分类器进行引文功能识别，识别的宏F值为68.2%。2013年，Abu-Jbara[11]在Dong与Schäfer分类体系基础之上区分“批评”和“证实”，构建了包含6个类目的分类体系，使用支持向量机分类器，在其自建的数据集上达到58%宏F值。

引文全局功能的自动识别成果相对较少，笔者在进行文献调研时仅发现一篇：2013年，Xu[2]等人在进行引文全局功能分类时，将引文全局功能分为“functional”“perfunctory”以及“hard to tell”，随后引文全局功能识别时，通过投票的方式选取一处引文局部功能作为其全局功能。由于引文的各局部功能均可能体现该引文的全局功能，这种只选取一种局部功能作为代表功能的做法，在识别的全面性上稍显不足，并且其忽视了各局部功能之间的联系，而这种局部功能间的隐含关系是有助于提升整体识别效果的。本文针对引文全局功能识别研究的不足，通过多标签分类的方法，选取具体引用处上下文特征及参考文献整体特征，提出了一种新的引文全局功能识别方法。

3 引文全局功能识别方法描述

3.1 问题定义

本文将引文功能的自动识别问题视为分类问题加以解决。假设一篇文献的参考文献集合$R=\{r_1,r_2,\cdots r_i,\cdots,r_m\}$，对每一个参考文献$r_i$，其在该篇文献中的每一个被引用可以组成集合为$C_i=\{c_1,c_2,\cdots c_j,\cdots,c_k\}$。给定一个引文功能的分类标签集合$L=\{l_1,l_2,\cdots l_p,\cdots,l_n\}$，引文局部功能的识别是判断$r_i$的具体引用集合$C_i$中每个元素$c_j$的功能，即学习$c_j$向标签集合L的映射。而引文全局功能识别则是识别参考文献r_i的整体功能，由于r_i可能在文中存在多处具体引用，此时的识别问题即识别r_i的功能集合L_{ri}（其中$L_{ri}\subseteq L$）。针对这一研究问题，主要有两种解决方法：

（1）对r_i的被引用集合C_i中的每个元素进行引文局部功能识别，汇总得到功能集合L_{ri}。

（2）将问题转化为多标签分类问题，提取r_i的特征，识别出功能集合L_{ri}。

由于方法（1）是引文局部功能识别后汇总，而上文中已提到引文局部功能识别已有很多相关研究，本文主要进行方法（2）的研究，并在实验部分与方法（1）进行比较。接下来介绍本文的引文功能分类体系、多标签分类的具体方法及分类特征。

3.2 引文功能分类体系

引文功能分类体系的构建是引文功能标注以及引文功能自动化分类的基础。先前大量的引文功能分类研究都构建了自己的分类体系[1,4,10,12-18]，并进行了数据集的标注。这些分类体系在类别的数目和种类上存在较大差异。本文在比较了这些分类体系后，选择了Dong与Schäfer的分类体系[4]进行实验，具体见表1。

表1 Dong与Schäfer的引文功能分类体系

类别	描述
背景（Background）	被引文献在文章中全局性地描述了研究背景，或者是简略提供了相关研究和前沿研究成果的介绍
基本思想（Fundamental Idea）	被引文献激发或者引导了施引文献的研究工作
技术基础（Technical Basis）	施引文献中使用了被引工作中的重要工具、方法、数据或者其他资源
比较（Comparison）	被引文献激发或者引导了施引文献的研究工作

选择该分类体系的原因有以下几点：首先，该分类框架下有公开的引文局部功能标注数据集，方便本文在此基础上构建引文全局功能数据集。其次，该分类体系将价值较大的类别独立考虑，对于价值较小的类别（如“相关研究”“背景”等）不作细致区分。这种做法确保类别数较少，更适合自动化识别，同时也保证了识别出价值较大功能标签的能力。

3.3 多标签分类

多标签分类适用于分类样本具有多个标签的问题情景，广泛运用于基因功能识别、音乐情感分类以及图像语义标注等领域[19]。本文采用RAKEL（Random k-Labelsets）[20]这一多标签分类方法进行引文全局功能的识别。相较于其他多标签分类方法而言，RAKEL考虑到各分类标签之间的相互关联，还能弥补数据偏斜的问题。下面介绍用 RAKEL方法进行引文全局功能多标签识别的具体过程：

（1）随机从引文功能标签集合L中选取m个k元子集，组成集合L_m。

（2）对L_m的其中一个元素L_k，将其幂集（所有子集的集合）记为L'_k。利用单标签分类算法（如决策树等）训练一个分类器h_k：输入样本x，输出L'_k中的一个元素，即元素L'_k的一个子集。

（3）对L_m中的每个元素重复步骤（2），则得到m个分类器。

（4）对于一个新的引文全局功能分类样本x，将其置于步骤（3）中的m个分类器，则得到m个分类结果组成的集合Q_m。

a）对每个L中的标签l_i进行如下操作：设定$V_{li}=0$。将集合Q_m中l_i的个数记为S_{li}；对每个集合L_m的元素，若该元素包含l_i，则$V_{li}=V_{li}+1$；求得$t_i=S_{li}/V_{li}$。

b）设定一个阈值t。L中$t_i>t$的标签组成的集合即为该样本的引文全局功能集合。

3.4 分类特征

本文引文全局功能识别使用的分类特征主要包括以下五类：

（1）词汇特征

词汇特征一直是引文功能识别研究的核心特征之一。本文一方面借助词表（具体见表2），以“引文句中是否包含该词表中的词”作为词汇特征，另一方面，为了弥补词表涵盖率有限的问题，本文还选取1-Gram作为词汇特征。

表2 本文采用的词表特征

关键词特征	说明
代词	we, our, us, ours, this work, this study, this paper
比较词	compare, differ, contrast, comparison, equal, exceed, outperform, oppose, consistent with, signify, golden standard, than, unlike
基本思想	follow, same, similar to, motivate, inspire, idea, spirit
技术基础	provided by, taken from, extracted from, based on, use, apply, extend, measure, evaluate, modify, extract
时态词	early, previous, prior, recent, recently
数量词	many, some, most, several, number of, numerous, variety, range of
频率词	usually, often, common, commonly, typical, typically, traditional, traditionally
举例词	such as, example, for instance, e.g.

（2）句法特征

除了词本身可以体现引文功能之外，引文上下文中的一些特定的句式及语法结构也能体现出引文功能，比如“is better than”“but…”等可以表示两者的对比。

本文抽取的句法特征主要包括：文献[4]提出的七个用于捕捉特定词性序列的正则表达式、斯坦福句法分析工具[21]解析的引文句中存在的依存关系以及Jochim[10]提出的四个句法特征（见表3）。

（3）物理特征

表3 Jochim提出的四种句法特征

特征	说明
self-comp	比较级词汇的主语是否为第一人称代词
self-good	积极词的主语是否为第一人称代词
other-comp	比较级词汇的主语是否为被引文献
other-contrast	转折词的主句是否为被引文献

引文的物理特征指引文的分布特征及频次特征。不同位置的引用，在功能上可能会存在一定差异，比如“引言”“相关工作”的引用更可能是“背景”，而“实验”部分的引用更可能是“技术基础”或者“比较”。本文采用的分布特征包括：引文句所处的章节、引用标记在引用句中的位置（前部、中部、后部）。频次特征在识别“背景”这一功能标签时能起到良好效果，当一个引用句中存在大量引用时，往往是背景知识介绍，本文选取的频次特征包括：引文句中的引文是单个出现还是成组出现、引文句中引用的个数。

（4）整体特征

除引用句中存在的特征外，参考文献本身也含有一些有助于引文全局功能识别的特征，比如根据参考文献的作者信息判断该引用为自引，则很可能施引文献的成果是在该参考文献基础之上完成的。本文选取的参考文献级别特征包括是否为自引以及参考文献在施引文献中存在几处引用。

（5）其他特征

除了上述特征之外，本文还选取了两类特征进行补充。首先，本文认为作者引用的情感与引用的功能具有很强的关联。所以利用情感极性词表[22]，将引文句中的情感词抽取后与其情感倾向拼接起来组成情感特征。此外，引用“工具”和“数据集”时，引文的功能较为明确，所以本文采用Jochim[10]的命名实体识别的模型，并将“引文句是否包括工具”及“引文句是否包含数据集”纳入特征。

4 实验与结果分析

4.1 数据准备

本文的引文全局功能数据集的构建基于两个数据集合：一个是由Dong与Schäfer[4]在其分类体系（即本文采用的分类体系）基础之上构建的引文局部功能标注数据（本文称之为DFKI数据集）。另一个是由Schäfer与Weitz[23]构建的xml格式的ACL Anthology全文数据集（本文称之为paperXML数据集）。

DFKI数据集包括1768个引用句信息，每一行数据包括六个字段：第一个字段为引文句编号，由ACL_ID和一个序号构成；第二个字段为引文句内容，其中引文标记由”()”代替；第三个字段为引文句词性标注结果；第四到第六个字段是分别从功能、相关程度、情感倾向三个维度对引文句的标注结果。数据样例如图1，第四个字段即本文所需的该引文句的功能标注结果。

08-2002-8	The parser's outputs define a relation on word pairs()
DT NNS NNS VV DT NN IN NN NNS()	BackGround SRelated Neutral

图1 DFKI数据样例

由于DFKI数据集只包括具体引用处的信息，为了构建参考文献级别上的引文功能数据集，需要引入paperXML数据集，将引用句在文献中的具体位置找出，从而找到参考文献与引文句的对应关系。本文构建引文全局功能数据集的具体做法如下：首先，通过ACL_ID将DFKI数据集中的每个引文句a与paperXML数据集中的对应文献关联起来。其次，利用Jaccard系数[24]判断文献中每个句子与a的相似程度，选择相似程度最高的句子t，作为a对应的句子。随后，通过人工简单校验，笔者得到DFKI数据集中引文句在paperXML中的对应句，由于paperXML数据集包含引文信息，笔者得到引文句和参考文献的对应关系，将属于同一篇参考文献的引文句的功能进行简单合并（示例如图2），得到参考文献功能。最后，本文构建了包含1805个参考文献信息的引文全局功能数据集。其中每个参考文献都有其对应的功能标签及施引文献正文信息。

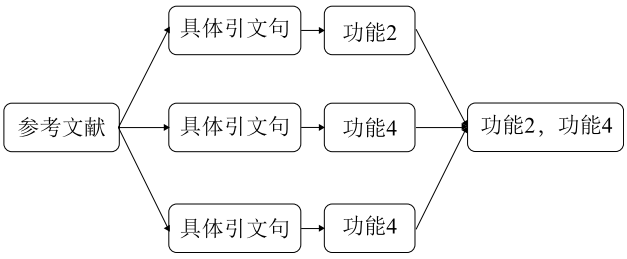

图2 引文句功能合并示例

4.2 数据处理

在多标签分类时，需要把多个具体引用处的特征合并到参考文献级别上。本文的做法是：将这些具体引用处的特征分为词类型特征（如N-Gram等词汇特征）、布尔型特征（如是否存在指定词表中的词、是否包含自引等特征）、数值型特征（如包含特定词的个数、引用在文中出现的位置等）。将词类型的特征汇总，得到的词汇集作为参考文献的词类型特征；将布尔型特征求“与”运算（如两处局部引用，一处包含某词，一处不包含，则参考文献的此特征取不包含）；将数值型特征求平均值得到参考文献的数值型特征。通过这样的方式，将具体引用处的特征转化到参考文献的特征上来。

4.3 实验设置

本文设置了三组实验：

（一）（基准实验）N-Gram特征+基于多标签分类的引文全局功能识别：将1-Gram、2-Gram、3-Gram作为特征运用RAKEL方法进行多标签分类。

（二）本文提出的特征+基于引文局部功能合并的引文全局功能识别：利用本文的引文功能分类特征，进行引文局部功能分类。随后将分类结果汇总，得到引文全局功能。

（三）本文提出的特征+基于多标签分类的引文全局功能识别：利用本文的引文功能分类特征，采用RAKEL方法进行多标签分类。

实验一与实验三的对比，探究本文的特征是否可以很好运用于引文全局功能分类。实验二与实验三对比，探究引文全局功能多标签分类方案与将引文局部功能分类后汇总的方案之间的效果差异。

文中引文局部功能分类的分类器以及引文全局功能分类的分类器均为线性支持向量机[25]，在10折交叉验证基础上计算召回率、准确率及F值作为主要实验结果。

4.4 实验结果的评价指标

本文采用准确率、召回率和宏F值评价实验效果，三个指标的定义如下：

$$Precision=\frac{|Setp\cap Set_L|}{Setp} \quad (1)$$

$$Recall=\frac{|Setp\cap Set_L|}{Set_L} \quad (2)$$

$$F\text{-}value=\frac{2*Precision*Recall}{Precision+Recall} \quad (3)$$

针对每一个功能类别，Set_P表示识别结果中包含该功能的参考文献集合，Set_L表示标注数据集中包含该功能的参考文献集合，$|Set_*|$表示一个集合中元素的个数，Precision表示准确率，即被正确识别出该功能的参考文献数目占被识别为该功能的参考文献总数的比例，Recall表示召回率，即被正确识别出该功能的参考文献数目占被标注为该功能的参考文献总数的比例，F-value表示准确率和召回率的调和平均数。

4.5 结果分析

本文三组实验的总体分类效果如表4所示。从表中可以看出，实验二与实验三在F值指标上，都比实验一有较大提升。其中，实验二的召回率最高，为69.96%，实验三的准确率最高，为83.41%。实验三的F值最高，达到71.52%。这说明本文提出的分类特征与分类方案在引文全局功能识别上起到了良好的实验效果，但两个引文全局功能识别方案的侧重点不同。综合而言，实验三的准确率及F值有着一定优势。

表4 三组实验准确率、召回率、F值指标的宏平均结果

	准确率	召回率	Macro-F
N-Gram特征+多标签分类	0.7808	0.5746	0.6475
本文特征+局部功能汇总	0.6578	0.6996	0.6753 (+4.29%)
本文特征+多标签分类	0.8341	0.6450	0.7152 (+10.46%)

实验二与实验三各类别的实验结果如表5所示。实验三在“基本思想”“比较”及“背景”三个类别的F值上比实验二分别相对提升18.68%、9.41%及1.86%；“技术基础”上存在一定劣势，低1.56%。“基本思想”和“比较”在分类框架中表示了施引文献和被引文献成果之间的关系，属于价值较大的标签。总体比较两个方案，实验三在识别价值更大功能标签的效果上更出众。

进一步分析实验三的识别效果。根据3.3节算法描述，RAKEL识别出的标签数目一定程度上由阈值决

表5 实验二与实验三不同类别实验结果对比

	实验二（本文特征+局部功能汇总）		
	准确率	召回率	F值
基本思想	0.5691	0.5833	0.5761
技术基础	0.7277	0.7204	0.7240
比较	0.4267	0.5926	0.4961
背景	0.9077	0.9022	0.9049
平均值	0.6578	0.6996	0.6753
	实验三（本文特征+多标签分类）		
	准确率	召回率	F值
基本思想	0.7988	0.5976	0.6837（+18.68%）
技术基础	0.8057	0.6389	0.7127（-1.56%）
比较	0.8274	0.4039	0.5428（+9.41%）
背景	0.9046	0.9394	0.9217（+1.86%）
平均值	0.8341	0.6450	0.7152（+5.91%）

定，实验三在不同阈值下的实验结果如图3所示。随着阈值的增加，识别出的标签数目减少，准确率上升，召回率下降。这表明，在阈值降低时模型返回更多的功能标签，更全面地识别引文全局功能；而阈值提高时模型返回引文的少量功能标签，识别精度更高。总体来看，本文模型的F值维持在一个较高的水平（0.6以上），最高可达0.7152。这反映了根据实际情况调节识别的引文功能标签数目时，本文模型均能表现出良好的识别效果。

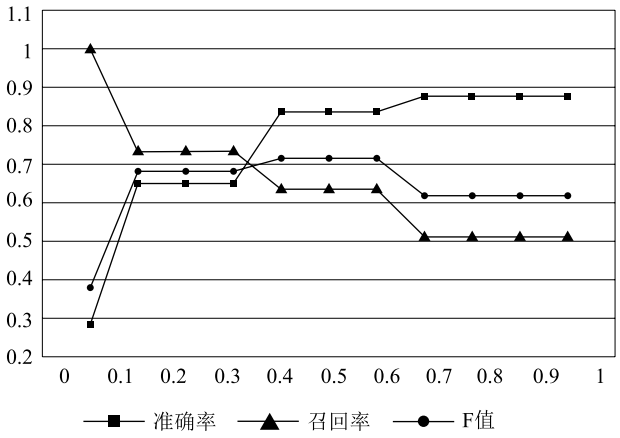

图3 实验三准确率、召回率及F值随阈值变化情况

4.6 实验分析

笔者对多标签方法识别效果优于局部功能识别汇总的原因进行了一定分析：

在引文局部功能识别中，“基本思想”等价值较大的功能标签一直是识别的重点及难题，因为“基本思想”引用常常和大量“背景”引用一样出现在文献的“引言”和“相关研究”中，单独对局部引用处特征进行分析，很难准确识别出该局部引用处的功能。但通过实验二和实验三对比可看出，多标签分类对“基本思想”这一功能识别准确率非常高，为79.88%。这是由于多标签分类可以综合考虑各引文局部特征，并将各局部功能之间关联加入到识别过程中。虽然一处引文局部功能难以判断“背景”和“基本思想”的差异，但由于“基本思想”往往伴随着其他章节的“比较”引用及“技术基础”引用等，通过综合一篇参考文献的各引文局部特征，可以准确地将“基本思想”这一功能识别出来。本文的实验论证了综合考虑一篇参考文献的各局部引用处的特征，可以更准确地识别出引文全局功能。

5 结语

本文将引文全局功能识别问题转化为多标签分类问题，利用具体引用处上下文特征及参考文献整体特征，在构建的引文全局功能数据集上进行实验。实验结果表明，通过挖掘引文局部功能之间的隐含联系，基于多标签分类的引文全局功能识别方法能够更加有效地识别引文全局功能。

本文的研究成果具有较大的应用价值：将引文全局功能融入引文推荐研究中，能够从语义层面优化推荐结果；多个功能标签有利于构建节点关系更加丰富的引文网络；通过引文全局功能的不同，区分科研文献引用价值，进而优化现有科研文献评价指标；除此之外，其他引文分析相关研究也能从中受益。

本文研究仍存在一定的不足，需要进一步探索。未来的研究思路主要包括：①进一步挖掘有助于引文全局功能识别的特征。②采用多种多标签分类方法，对比不同分类方法的分类效果，进一步提高引文全局功能的识别效果。③将识别成果运用于引文推荐等相关研究中，解决相关实际问题。

参考文献

[1] Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function [C]// Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational

Linguistics, 2006: 103-110.

[2] Xu H, Martin E, Mahidadia A. Using heterogeneous features for scientific citation classification [C]// Proceedings of the 13th conference of the Pacific Association for Computational Linguistics, 2013.

[3] Myers G. In this paper we report… - Speech acts and scientific facts [J]. Journal of Pragmatics, 1992, 17(4): 295-313.

[4] Dong C, Schäfer U. Ensemble-style Self-training on Citation Classification [C]// IJCNLP, 2011: 623-631.

[5] Oppenheim C, Renn S P. Highly cited old papers and the reasons why they continue to be cited [J]. Journal of the American Society for Information Science, 1978, 29(5): 225-231.

[6] Hanney S, Frame I, Grant J, et al. Using categorisations of citations when assessing the outcomes from health research [J]. Scientometrics, 2005, 65(3): 357-379.

[7] Garzone M A. Automated classification of citations using linguistic semantic grammars [D]. University of Western Ontario London, 1997.

[8] Nanba H, Kando N, Okumura M. Classification of research papers using citation links and citation types: Towards automatic review article generation [J]. Advances in Classification Research Online, 2011, 11(1): 117-134.

[9] Radoulov R. Exploring automatic citation classification [D]. University of Waterloo, 2008.

[10] Jochim C, Schütze H. Towards a generic and flexible citation classifier based on a faceted classification scheme [C]// Proceedings of the 2012 International Conference on Computational Linguistics, 2012: 1343-1358.

[11] Abu-Jbara A, Ezra J, Radev D R. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics [C]// HLT-NAACL, 2013: 596-606.

[12] Garfield E. Can citation indexing be automated [C]// Statistical association methods for mechanized documentation, symposium proceedings, 1965(1): 189-192.

[13] Lipetz B A. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators [J]. American Documentation, 1965, 16(2): 81-90.

[14] Herlach G. Citation patterns: mechanistically identifiable characteristics of citation links [D]. University of Chicago, Graduate Library School, 1973.

[15] Chubin D E, Moitra S D. Content analysis of references: adjunct or alternative to citation counting? [J]. Social studies of science, 1975, 5(4): 423-441.

[16] Oppenheim C, Renn S P. Highly cited old papers and the reasons why they continue to be cited [J]. Journal of the American Society for Information Science, 1978, 29(5): 225-231.

[17] Zhang G, Ding Y, Milojevic S. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content [J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1490-1503.

[18] 陆伟,孟睿,刘兴帮.面向引用关系的引文内容标注框架研究[J].中国图书馆学报,2014(6):93-104.

[19] Tsoumakas G, Zhang M L. Learning from multi-label data [EB/OL]. [2016-01-05]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.207.2982.

[20] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification [C]// Machine learning: ECML 2007, Springer Berlin Heidelberg, 2007: 406-417.

[21] The Stanford Parser: A Statistical Parser [EB/OL]. [2009-01-10]. http://nlp.stanford.edu/software/lex-parser.shtml.

[22] Wilson T, Hoffmann P, Somasundaran S, et al. OpinionFinder: A system for subjectivity analysis [C]// Proceedings of HLT/EMNLP on interactive demonstrations. Association for Computational Linguistics, 2005: 34-35.

[23] Schäfer U, Weitz B. Combining OCR outputs for logical document structure markup: technical background to the ACL 2012 contributed task [C]// Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. Association for Computational Linguistics, 2012: 104-109.

[24] Jaccard P. The distribution of the flora in the alpine zone [J]. New phytologist, 1912, 11(2): 37-50.

[25] Cortes C, Vapnik V. Support-vector networks [J]. Machine learning, 1995, 20(3): 273-297.

作者简介

刘兴帮，男，1991年生，武汉大学信息管理学院硕士研究生，研究方向：信息检索、数据挖掘等，E-mail：baronliu0710@foxmail.com。
陆伟，男，1974年生，武汉大学信息管理学院教授，副院长，研究方向：信息检索、知识管理、数据挖掘等，E-mail：reedwhu@gmail.com。
孟睿，男，1990年生，美国匹兹堡大学博士研究生，研究方向：信息检索、数据挖掘。

A Study of Global Citation Function Recognition Based on Multi-Label Classification

LIU XingBang[1,2], LU Wei[1,2], Meng Rui[1,2]
（Center for Studies of Information Resources, Wuhan University, Wuhan 430072）

Abstract: Citation function is defined as the author's reason for citing a given paper. Compared with the local citation function which only considers the few context sentences, global citation function takes the global information into account. Its automatic recognition research will exert important impact on academic text mining, e.g. citation recommendation, citation indexing, semantic citation network construction, etc. Based on the idea that “a cited article may appear once or multiple times in the citing article”, this paper proposes to deal with Global Citation Function problem in a way of Multi-Label Classification. A significant improvement over baseline method shows the effectiveness of multi-label classification method on this research problem.

Keywords: Global Citation Function; Multi-Label Classification; Academic Text Mining; Citation Analysis

（收稿日期：2016-03-19）

2016年全国知识组织与知识链接学术交流会征文通知

为了探讨网络环境下知识组织与知识链接的新理念、新思路、新方法，中国科学技术信息研究所、国家科技图书文献中心和中国科学技术情报学会拟于2016年9月召开第七届“全国知识组织与知识链接学术交流会”，由华中师范大学信息管理学院承办。特向国内图书情报界及相关领域的专家学者征文。具体会议时间、地点另行通知。

一、会议主题

知识组织

- 知识组织研究进展
- 大数据中的数据清洗、组织和分析
- 元数据整合
- 知识组织体系的构建和应用

知识链接

- 科研实体关系揭示
- 网络资源链接及其关联分析
- 面向项目研发产出的关联分析
- 数据关联挖掘和揭示

知识评价

- 科学计量与评价
- 引文分析、主题分析
- 专利分析与利用
- Web科技资源评价

知识服务

- 知识服务、知识管理研究进展
- 数字科研环境与开放共享服务
- 用户分析与个性化用户服务
- 知识图谱及可视化分析

欢迎广大图书馆学、情报学相关研究、教学与实践者，图书馆和信息机构的管理者以及相关信息技术人员踊跃投稿。优秀论文将发表在《数字图书馆论坛》上。

二、征文要求

1、文章要求观点明确、主题突出；来稿必须为未经发表的论文；稿件统一用A4纸排版，以电子邮件方式提供Word格式文档；正文字数应控制在4000~8000字。

2、来稿请提供：中英文题目、中英文作者及单位、中英文摘要和关键词、正文、参考文献。论文后请附作者简介，包括作者单位、联系电话、电子邮箱、通讯地址及邮政编码等。

3、截稿日期：2016年8月15日。

三、联系方式

联系人：赵莹莹，彭帆　　联系电话：010-58882061　　邮箱：KOLink@istic.ac.cn

会议网址：http://168.160.16.186/conference