一种面向权威度和多样性的自动学术调研框架

韩 晓[1),2)] 郭嘉丰[1)] 杜 攀[1)] 程学旗[1)]

[1)](中国科学院计算技术研究所网络数据科学与工程研究中心 北京 100190)

[2)](中国科学院大学 北京 100190)

摘 要 对某个领域或问题进行学术调研是科研工作的基本需求,然而随着越来越多的科研人员投身研究,大量的学术成果不断涌现,信息过载使得快速有效的调研工作变得越发困难.文中旨在提出一种自动学术调研框架,基于用户给定的关键词查询推荐最值得调研的论文及作者,以辅助科研人员高效完成调研任务.面向某个领域或问题最值得调研的论文和作者,需要具备显著的权威度且能覆盖该领域或问题的不同方面.因此,文中提出了一种面向权威度和多样性的两阶段排序模型:首先引入了 MutualRank 模型,同时考虑论文及作者信息以更好地建模他们的权威度;接着利用 PDRank 模型融合权威度和差异性两个因素对论文和作者排序,最终得到权威度高、覆盖面广的调研结果.通过实验作者证明了 MutualRank 对于权威度的学习效果优于传统的 PageRank,同时基于两阶段排序模型得到的调研结果也优于已有的基准方法.

关键词 MutualRank;PDRank;自动学术调研;多样性排序;社会计算;社交网络

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2015.00365

An Automatic Literature Survey Framework by Exploring Prestige and Diversity

HAN Xiao[1),2)] GUO Jia-Feng[1)] DU Pan[1)] CHENG Xue-Qi[1)]

[1)](*Research Center of Web Data Science and Engineering, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract Literature survey of domains or topics is the foundation of scientific research. Along with more and more researchers devoting themselves to their work, plenty of academic achievements come out continuously, which brings more difficulties to effective and efficient surveys. This paper aims at developing an automatic literature survey framework to help researchers survey effectively. This framework recommends papers and authors which are most worthwhile surveyed based on the keywords given by the user. These recommended papers and authors must be prestigious and cover different aspects of the domain or problem. This paper proposes a two-phase ranking model by simultaneously exploring prestige and diversity. Firstly we introduce MutualRank to learn the prestige of the papers as well as the authors by leveraging the two heterogeneous types of information. We then rank the authors and papers by using PDRank model which combines the prestige and diversity. Finally, we provide users with recommended survey results with high prestige and diversity. Experiments show that MutualRank is better than PageRank on modeling prestige, and the survey results based on two-phase ranking model is superior to the existing baseline methods.

Keywords MutualRank; PDRank; automatic literature survey; diversity ranking; social computing; social networks

---

收稿日期:2013-05-17;最终修改稿收到日期:2014-12-28.本课题得到国家自然科学基金(2013CB329601,61100175)、国家"二四二"信息安全计划(2011F45,2012G129)、国家科技支撑计划(2012BAH39B04,2012BAH39B02)和欧盟第七研发框架计划(FP7-PIRSES-318939)资助.**韩 晓**,女,1988 年生,硕士研究生,主要研究方向为信息检索、文本挖掘.E-mail:ifshall.han@gmail.com.**郭嘉丰**,男,1980 年生,博士,副研究员,主要研究方向为网络搜索和挖掘、用户数据挖掘、机器学习以及社交网络.**杜 攀**,男,1981 年生,博士,助理研究员,主要研究方向为信息检索、文本挖掘、机器学习.**程学旗**,男,1971 年生,博士,研究员,博士生导师,主要研究领域为网络科学、互联网搜索与挖掘、对等网络、信息安全以及分布式系统.

1 引 言

从事科研工作的研究人员、老师和学生都常会面临学术调研的任务,即为了了解某个领域或问题,需要阅读该领域或问题相关的代表性学术成果,掌握和追踪主要学者的工作.目前互联网上承载了大量的学术成果、学者信息,学术调研也主要基于互联网展开.通常的做法是,首先基于所需要调研的领域或问题构造查询关键词,通过搜索引擎或学术服务系统获取初始的一些论文,再通过有选择的延伸阅读,来发现和识别该领域或问题相关的代表性成果与学者.显然,学术调研是一项复杂而困难的工作,需要人们投入大量的精力,然而随着越来越多的科研人员投身研究,大量的学术成果不断涌现,信息过载使得快速有效的调研工作变得越发困难.因此,如何基于用户给定的关键词查询,通过自动挖掘与学习的方法,来推荐最值得调研的论文及作者,以辅助科研人员高效完成调研任务,成为了一个迫切又具有挑战性的任务.基于这种需求,自动学术调研系统应运而生.在这里,自动学术调研的任务是指给定一个查询,要求返回与查询相关的权威度高、覆盖面广的论文及作者.

在现有的学术系统中,很多都提供了基于关键词的学术搜索功能,比如 ACM Digital Library、Google Scholar、Academic Search 等等.给定用户的查询关键词,系统会返回与该查询相关的文章或作者.但是通过检索的方式并不能很好地满足用户调研的需求,这是因为检索主要考虑的是结果的相关性,一方面他们对权威度的建模通常还比较简单,另一方面检索结果也不能很好地解决多样性,即返回的文章或作者有大量的冗余和重复信息,无法满足调研的对象能够覆盖这个领域或问题的各个方面的需求.而已有的一些考虑多样性的排序模型,又往往对学术调研所需的权威度考虑不足.

因此,本文提出了一种面向权威度和多样性的两阶段排序模型.首先,我们提出了 MutualRank 模型来对权威度进行建模,该模型同时考虑论文及作者两类信息来学习他们的权威度.这是因为我们观察到,论文及作者的权威度之间存在着紧密的联系,权威度相对高的论文的作者权威度可能更高,同样的,权威度相对高的作者发表的论文也可能权威度更高,因此,这两种异质对象的权威度之间存在着相互增强的关系.接着,我们利用 PDRank 模型融合权威度和差异性两个因素对论文和作者进行排序,在这里为了建模论文或作者的差异性,我们提出了基于引用关系的论文和作者的差异度指标.最终,通过上述两阶段模型,我们可以得到权威度高、覆盖面广的调研结果推荐给用户.我们基于天玑社会化学术引擎采集到的 300 多万论文及 150 多万的作者数据进行了实验,通过实验我们证明了 MutualRank 对于权威度的学习效果优于传统的 PageRank[1],同时基于我们两阶段排序模型得到的调研结果也优于已有的基准方法.

本文第 2 节总结相关工作;第 3 节介绍面向权威度和多样性的两阶段排序模型;第 4 节给出实验结果及分析;第 5 节进行总结.

2 相关工作

自动学术调研的任务是给定一个查询,要求返回与查询相关的权威度高、覆盖面广的论文及作者,因此,该任务主要需要考虑排序结果的重要性以及多样性两个方面.

在关注重要性的排序模型中,影响最为广泛的就是基于随机游走模型的 PageRank[1-3] 模型.该模型由 Brin 和 Page 于 1998 年发表在 WWW 上,它的基本假设是,用户对网页的浏览过程可以通过随机游走模型来模拟,随机游走收敛时,用户在各个网页上的停留的稳态概率能够反映出该网页的重要程度.另一个关注重要性的排序模型是由 Kleinberg 于 1998 年提出的 HITS[3-4] 模型.该模型对网页的排序打分分为两个部分,一个是网页的 Authority 值,用指向该网页的链入度来表示;另一个是网页的 Hub 值,用该网页的链出度来表示.一个 Hub 值高的网页会同时指向多个 Authority 值高的网页,而一个 Authority 值高的网页同时会被多个 Hub 值高的网页所指向.因此,这两个值可以通过在网络上的互增强过程迭代计算求解.

为解决结果多样性的问题,最早可以追溯到由卡内基梅隆大学的 Carbonell 和 Goldstein[5] 在 1998 年提出来的方法 MMR.该方法考虑了相关性和多样性两个排序准则,使用新数据对象与查询对象的相似度及新对象与已经选取的数据对象的负值做线性加权和,通过迭代地选取具有最大边际相关度的对象,最终会得到一个同时具有查询相关性和内容

多样性的对象集合. 威斯康辛大学的 Zhu 等人[6-7]于 2007 年提出了 Grasshoper 的方法，该方法借助随机游走中的吸收态的影响，来实现排序中的多样性. 密歇根大学的 Mei 等人[8-9]于 2010 年提出了基于节点增强的随机游走模型的多样性排序方法 DivRank，该模型是一个基于节点增强的随机游走模型，模型的基本思想借助“富者更富”机制，使得网络中的代表性节点获得访问的概率远远超过其邻居节点，从而使得网络中的中心节点的排序得分被强烈放大，得到的结果最终既满足了重要性又满足了多样性. 现在有些工作中也有使用基于聚类方法来得到多样性的挖掘结果，如 NetClus[10-12]等. 然而，这些解决重要性和多样性问题的模型都没有考虑异质网络中的互增强对重要性的影响.

对于类似于学术网络的异质网络，Nie 等人[13]提出了 PopRank 模型来解决异质网络上的对象排序问题. 该模型考虑了异质对象重要性的相互增强，加入了异质对象间的转移概率来求得对象的 PopRank 值. 同样是对于异质网络，为解决文档摘要及关键字抽取问题，Wan 等人[14]加入词和句之间的转移概率，同时迭代计算词和句的 Rank 值，最终同时达到稳态便能得到词和句的 Rank 值. 但是现有的基于异质网络的工作多是求得重要性，而没有结合多样性的需求.

对于自动学术调研任务，现有的关注异质网络重要性的模型并没有考虑结果的多样性，而解决多样性问题的模型又没有考虑异质网络的互增强对重要性建模带来的影响. 我们受到以上相关工作的启发，结合自动学术调研这一实际任务，提出了一种面向权威度和多样性的两阶段排序模型来解决自动学术调研的问题.

3　面向权威度和多样性的两阶段排序模型

自动学术调研的任务是给定一个查询，要求返回与查询相关的权威度高、覆盖面广的论文及作者. 在这里，我们首先对问题进行详细的描述，接着引入 MutualRank 模型来对学术网络中的两类对象权威度进行建模和排序，然后，我们提出 PDRank 模型融合权威度和差异性两个因素对论文和作者进行排序，最后我们总结面向权威度及多样性的自动学术调研框架.

3.1　问题描述

自动学术调研任务关注于异质学术网络，该网络中存在作者及论文两种类型的对象，其中任意两个对象之间都可能存在着一定的关系. 在这里，我们形式化的定义一个学术网络 $G=\langle V,E,W\rangle$，其中 V 代表对象的集合，包含了论文 P 和作者 A 的集合，E 代表了对象之间的边的集合，任意的对象之间都可能存在着边（论文间引用关系、作者间引用关系及论文作者间的写作关系），W 代表对象边的权重，权重的定义如下：

$$w(x_i,x_j)=\begin{cases}1, & \text{如果 } x_i\in P(A),x_j\in A(P) \text{ 或者 } x_i\in P,x_j\in P\\ n, & \text{如果 } x_i\in A,x_j\in A\\ 0, & \text{其他}\end{cases}.$$

因此，学术网络也可以看成由 3 个子网络构成，分别是论文关系子网络、作者关系子网络和写作关系子网络. 在这个网络中，我们要求得同时具有权威度及多样性的论文及作者.

3.2　基于 MutualRank 的权威度排序

自动学术调研，如我们在引言中所说，用户主要需要识别和发现与所需调研的领域或问题相关的代表性学术成果，掌握和追踪主要学者的工作. 因此，学术论文和作者是学术调研的基本对象. 在对学术论文和作者进行权威度建模时，由于论文之间、作者之间都可以构成一个关系图，那么一种传统的作法就是分别基于论文关系图和作者关系图，利用 PageRank 等随机行走算法，来建模权威度. 然而我们发现，学术论文和作者这两种异质对象的权威度之间其实有着紧密的联系，需要融合起来考虑. 在这里我们的工作基于以下两点假设：

假设 1.　一篇论文如果被其他论文引用的次数越多，则它的权威度可能越高；一位作者如果被其他作者引用的次数越多，则他的权威度也可能越高.

假设 2.　权威度越高的作者发表的论文可能权威度也越高；权威度越高的论文的作者也可能权威度越高.

假设 1 与 PageRank 的假设类似，主要考虑了学术网络中同质对象之间的引用关系对学术对象权威度的影响. 假设 2 则与 HITS 的假设类似，它考虑了学术网络中异质对象的写作与被写作的关系对学术对象权威度的影响. 从随机游走模型的角度看，上述两个假设可以理解为：当用户对某个学术问题感兴趣，并阅读一篇相关论文时，他可能游走到该论文

所引用的另一篇论文去做更深入的阅读,也可能转去关注该论文的作者;同样,当用户关注与话题相关的某一位作者时,可能同时关注该作者引用过的另一位作者,也可能转去阅读该作者的某一篇论文.

基于以上两种假设,我们提出了 MutualRank 来对学术网络中的论文及作者的权威度进行建模.具体做法是,我们首先构建了一个异质关联的学术网络,其包括 3 个相互联系的子网络,即论文引用子网络、作者引用子网络和论文-作者间的写作关系子网络,我们在这样的异质网络上同时对论文和作者两个子网络进行迭代的随机行走,这两个随机行走通过论文-作者间的写作关系互相影响、互相加强.利用这样的迭代计算,我们可以同时建模论文和作者的权威度.

借鉴于 PageRank 算法,在 MutualRank 算法中,每个对象被访问的概率由三部分组成,一部分是随机跳转被选中的概率,另一部分是从指向它的同质对象顺着链接关系浏览的概率,还有一部分是指向它的异质对象顺着链接关系浏览的概率.

具体而言,我们假设网络中论文的总数为 M,作者的总数为 N.对于论文引用网络,任意两个节点 p_i 到 p_j 的转移概率是:

$$P(p_i,p_j)=\begin{cases} d\cdot\dfrac{1}{M}+(1-d)\cdot\dfrac{w(p_i,p_j)}{\sum\limits_{k=1}^{M}w(p_i,p_k)}, & \text{如果}\sum\limits_{k=1}^{M}w(p_i,p_k)>0 \\ \dfrac{1}{M}, & \text{否则} \end{cases} \tag{1}$$

其中,M 代表论文的总数量,$w(p_i,p_j)$ 代表论文 p_i 是否引用论文 p_j,$\sum\limits_{k=1}^{M}w(p_i,p_k)$ 表示在论文 p_i 的引用出度,d 代表阻尼系数.当论文 p_i 包含出链接时,节点 p_i 到节点 p_j 的转移概率由随机游走的概率及随机跳转的概率两部分组成.当论文 p_i 不包含出链接时,将随机地跳转到任意的论文中.

对于作者引用网络,任意两个节点 a_i 到 a_j 的转移概率是

$$P(a_i,a_j)=\begin{cases} d\cdot\dfrac{1}{N}+(1-d)\cdot\dfrac{w(a_i,a_j)}{\sum\limits_{k=1}^{N}w(a_i,a_k)}, & \text{如果}\sum\limits_{k=1}^{N}w(a_i,a_k)>0 \\ \dfrac{1}{N}, & \text{否则} \end{cases} \tag{2}$$

其中,N 代表作者的总数量,$w(a_i,a_j)$ 表示作者 a_i 引用作者 a_j 的次数,$\sum\limits_{k=1}^{N}w(a_i,a_k)$ 表示在作者 a_i 的引用出度,d 代表阻尼系数.当作者 a_i 包含出链接时,节点 a_i 到节点 a_j 的转移概率由随机游走的概率以及随机跳转的概率两部分组成.当作者 a_i 不包含出链接时,将随机地跳转到任意的作者中.

在式(1)、(2)中,我们根据论文引用关系及作者引用关系定义了论文之间的转移概率及作者之间的转移概率,这与 PageRank 的计算方法一致.接下来,我们将根据论文及作者间的关系来定义异质对象之间的转移概率.

对于论文写作子网络,当论文 p_i 是由作者 a_j 撰写时,则存在 p_i 到 a_j 的随机游走概率.任意两个节点 p_i 到 a_j 的转移概率是

$$P(p_i,a_j)=d\cdot\frac{1}{N}+(1-d)\cdot\frac{w(p_i,a_j)}{\sum\limits_{k=1}^{N}w(p_i,a_k)} \tag{3}$$

其中,N 代表作者的总数量,$w(p_i,a_j)$ 表示作者 a_j 是否是论文 p_i 的作者,$\sum\limits_{k=1}^{N}w(p_i,a_k)$ 代表论文 p_i 的作者数量,d 代表阻尼系数.节点 p_i 到节点 a_j 的转移概率由随机游走的概率以及随机跳转的概率两部分组成.

同样,当论文 p_j 是由作者 a_i 撰写时,则存在 a_i 到 p_j 的随机游走概率.任意两个节点 a_i 到 p_j 的转移概率是

$$P(a_i,p_j)=d\cdot\frac{1}{M}+(1-d)\cdot\frac{w(a_i,p_j)}{\sum\limits_{k=1}^{M}w(a_i,p_k)} \tag{4}$$

其中,M 代表论文的总数量,$w(a_i,p_j)$ 表示作者 a_i 是否是论文 p_j 的作者,$\sum\limits_{k=1}^{M}w(a_i,p_k)$ 代表作者 a_i 发表的论文总数,d 代表阻尼系数.节点 a_i 到节点 p_j 的转移概率由随机游走的概率以及随机跳转的概率两部分组成.

在式(3)、(4)中,我们定义了异质对象之间的转移概率.正是由于我们定义了异质对象间的转移概率,使得学术网络中的论文与作者的所有关系信息得到了充分的利用.

我们通过同时迭代来计算论文和作者的 MutualRank 值:

$$\pi^{(t+1)}(p_i)=(1-\alpha)\sum_{j=1}^{M}P(p_j,p_i)\pi^{(t)}(p_j)+\alpha\sum_{j=1}^{N}P(a_j,p_i)\pi^{(t)}(a_j) \tag{5}$$

$$\pi^{(t+1)}(a_i)=(1-\beta)\sum_{j=1}^{N}P(a_j,a_i)\pi^{(t)}(a_j)+\beta\sum_{j=1}^{M}P(p_j,a_i)\pi^{(t)}(p_j) \quad (6)$$

在迭代的过程中：

$$\sum_{i=1}^{M}\pi^{(t)}(p_i)=1,\ \sum_{i=1}^{N}\pi^{(t)}(a_i)=1 \quad (7)$$

经过足够长的时间 T 后，$\pi^{(T)}(p_i)$ 和 $\pi^{(T)}(a_i)$ 会趋于稳定，也就是我们要求得的 MutualRank 值. 该结果既考虑了同质对象之间的引用关系，也考虑了异质对象的相互增强，最终我们充分利用了论文和作者这两类异质的对象，来同时得到它们的权威度.

3.3　基于 PDRank 的多样性排序

在自动学术调研任务中，人们希望能够了解需调研的领域或问题的各个方面，以便有个较为全面的认识. 因此，除了要考虑论文和作者的权威度之外，一个重要的需求是希望得到的调研结果的覆盖面广，即调研结果的多样性是一个重要的指标.

MMR 方法同时考虑相关性和多样性两个排序准则，使用新数据对象与查询对象的相似度及新对象与已经选取的数据对象的相似度的负值做线性加权和，这个加权和被称为最大边际相关度，通过迭代地选取具有最大边际相关度的对象，最终会得到一个与查询相关性高且内容多样性丰富的对象集合.

受到 MMR 的启发，在这里，我们提出了 PDRank 框架来融合前面权威度和差异性两个因素：

$$PDRank \overset{\text{def}}{=} \arg\max_{D_i\in R}[\lambda Sim(D_i,Q)]+(1-\lambda)Diff(D_i,S)] \quad (8)$$

上式中，C 代表所有文档集合，S 代表已选择的文档，R 代表未选择的文档，则 $R=C-S$. $Sim(D_i,Q)$ 代表了当前文档 D_i 与查询 Q 的相似度，$Diff(D_i,S)$ 代表了当前文档与已选择的文档集之间的差异. 参数 λ 的调整可以使最终的结果在相关度和多样性中得到平衡，当 $\lambda=1$ 时，则系统只考虑结果的相关度而不考虑结果的多样性，当 $\lambda=0$ 时，则系统只考虑结果的多样性而不考虑相关度.

在这里，我们定义 $Sim(D_i,Q)$ 为 MutualRank，这一项代表了论文在调研的问题或领域中的重要程度. 学术调研任务中的多样性可以理解为调研结果的覆盖面，调研结果的引文覆盖该领域论文全集的范围越广，则越具备多样性特征. 因此，我们可以定义 $Diff(D_i,S)$ 为选择文章 D_i 后，新引入文章的比例，即

$$Diff(D_i,S)=\frac{Ref(D_i)-(Ref(D_i)\bigcap Ref(S))}{|C|} \quad (9)$$

上式中，C 代表论文全集，$ref(D_i)$ 表示文档 D_i 的引文集合，$ref(S)$ 表示已选择文档集 S 的所有引文集合，因此，$Diff(D_i,S)$ 可以衡量选择当前文档可以新引入的覆盖比例.

3.4　自动学术调研框架

在前两节中，我们分别介绍了基于 MutualRank 的权威度排序方法和基于 PDRank 的多样性排序方法. 在我们提出的自动学术调研框架中，用户给定一个查询关键词后，框架的工作如下：

(1) 首先我们检索出相关的论文集合，并构造出论文引用关系、作者引用关系及论文-作者写作关系.

(2) 在关系网络上使用 MutualRank 算法，计算出每篇论文及每位作者基于互增强的权威度.

(3) 使用 PDRank 算法，计算出同时满足权威度及多样性的论文及作者，得到调研结果.

经过以上 3 个步骤，我们最终会得到同时满足权威度和多样性的论文及作者.

4　实验及结果分析

4.1　数据集

本文使用了天玑社会化学术引擎的数据，包含计算机科学领域的论文 300 多万篇及作者 150 多万位. 实验随机选取了 200 个论文中出现的关键词(Keywords)作为查询，每个查询大约含有相关的论文 500～1000 篇，大约含有相关作者 1000～2000 位.

4.2　评估方法

在学术调研中，我们首先检索出与查询关键字相关的所有论文集合 C_P 及作者集合 C_A，接着可以使用某种排序模型求得排序最高的 K 篇论文 R_P 和 K 位作者 R_A. 我们从两方面评估自动学术调研的效果，即结果的权威度和覆盖面.

论文的权威度定义为，在与查询相关的论文集合 C_P 中，求解结果集合 R_P 中的每篇文章平均被 C_P 引用的次数，定义如下：

$$prestige(R_P)=\frac{\sum_{p\in R_P}\sum_{p'\in C_P}I[w(p,p')>0]}{|R_P|} \quad (10)$$

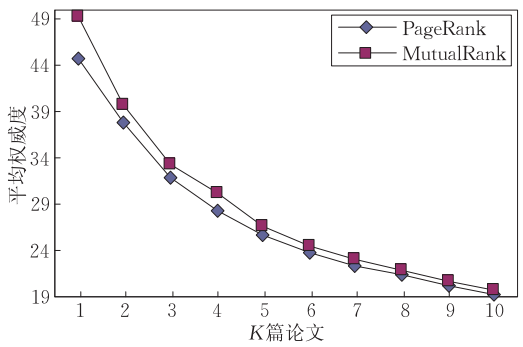
通常，作者的权威度可以使用作者的 H 指

数[15]来衡量，一名科研人员的 H 指数是指他至多有 H 篇论文分别被引用了至少 H 次. 在自动学术调研任务中，我们使用 R_A 中的每位作者在 C_P 中的平均 H 指数来衡量结果 R_A 的权威度.

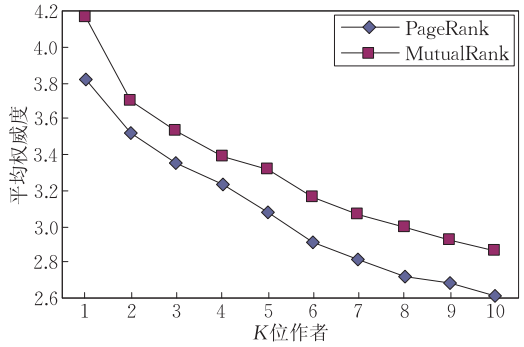
论文的覆盖度表示在与查询相关的论文集合 C_P 中，引用过结果集合 R_P 的比例，同样，作者的覆盖度表示与查询相关的作者集合 C_A 中，引用过结果集合 R_A 的比例.

4.3 对权威度建模的评估

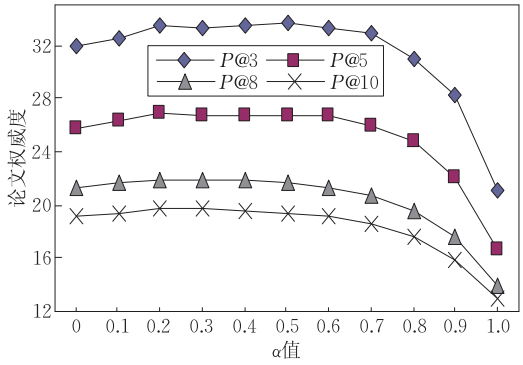
图 1 中的(a)和(b)分别展示了两种方法对论文和作者权威度排序的性能，从结果中我们可以看到，利用 MutualRank 计算得到的论文权威度及作者权威度均优于 PageRank 的结果. 这验证了我们前面提出的假设 1 和假设 2，利用论文和作者权威度的相互增强关系计算出的权威度优于传统的仅考虑单一数据对象的权威度.



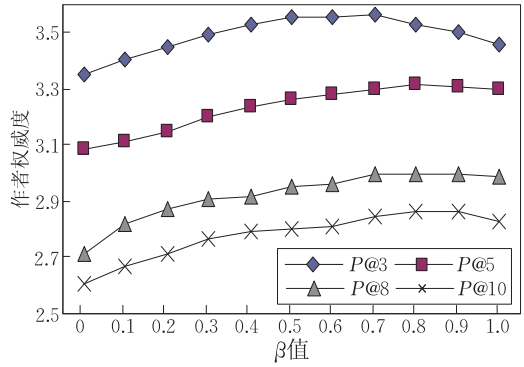
(a) Paper的权威度对比



(b) Author的权威度对比



(c) Paper权威度随α的变化



(d) Author权威度随β的变化

图 1 PageRank 和 MutualRank 两种方法在论文和作者权威度排序中的性能

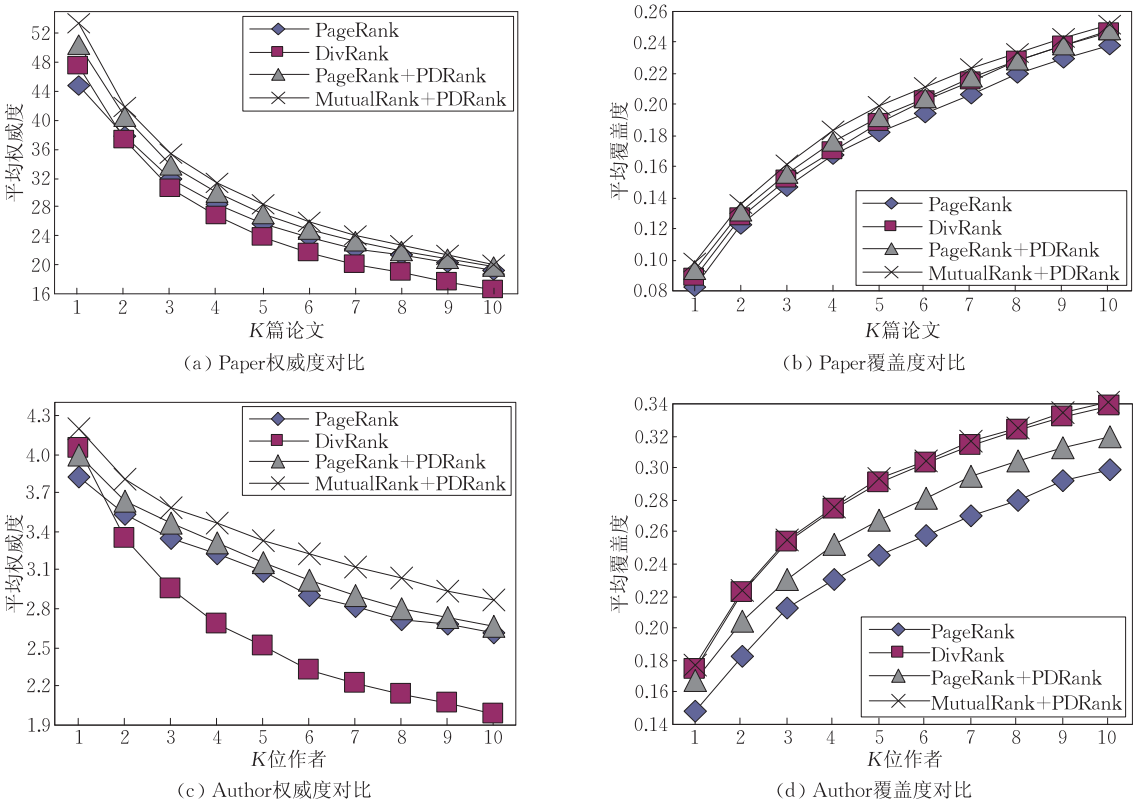
我们接着实验了不同的参数对于 MutualRank 性能的影响. 图 1 中的(c)和(d)的 $P@k$ 是 $prestige@k$ 的缩写，其中(c)反应了当 β 固定为 0.8 时，随着 α 变化，论文权威度的变化情况，从图中可知，当 α 在 0.2～0.5 区间时，得到的论文权威度达到最高；(d)反应了当 α 固定为 0.3 时，随着 β 变化，作者权威度的变化情况，从图中可知，当 β 在 0.7～0.8 区间时，得到的权威度最高. 从这组实验中，我们也可以发现一个有趣的现象，即论文关系信息对作者的权威度提升强于作者关系信息对论文的权威度提升.

在这里，我们进一步对学术调研结果进行评估. 我们基于随机选择的 200 个查询，得到相关检索结果后，分别使用 PageRank、DivRank、PageRank＋PDRank 和 MutualRank＋PDRank(我们的方法)对检索结果进行排序，取排名最高的 K 条结果作为调研结果，分别对其权威度和覆盖度进行评价. 实验中我们根据经验设置 PDRank 的参数 λ 为 0.85.

图 2 对比了 4 种方法得到的论文结果的权威度、覆盖度以及作者结果的权威度和覆盖度. 从实验结果可以看到，使用 MutualRank＋PDRank 的两阶段排序模型在论文的权威度及覆盖度上都优于其他三个方法；在作者的覆盖度上与 DivRank 的计算结果接近，但权威度大大优于 DivRank 的结果.

4.4 自动调研效果实例

我们通过一个实例来对大家对调研结果有一个更直观的认识. 表 1 展示了查询“Learning To Rank”的调研结果，这里我们展示了使用传统的 PageRank 得到的结果与使用我们两阶段排序模型的调研结果.



(a) Paper权威度对比　(b) Paper覆盖度对比

(c) Author权威度对比　(d) Author覆盖度对比

图 2　4 种方法得到的论文结果的权威度、覆盖度以及作者结果的权威度、覆盖度

表 1　话题“Learning To Rank”的调研结果对比

	PageRank	MutualRank+PDRank
论文	1. Learning to rank using gradient descent(pairwise 方法) 2. Adapting ranking SVM to document retrieval(pairwise 方法) 3. LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval(数据集) 4. Learning to rank: from pairwise approach to listwise approach (listwise 方法) 5. FRank: a ranking method with fidelity loss(pairwise 方法)	1. Learning to rank using gradient descent(pairwise 方法) 2. LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval(数据集) 3. Learning to rank: from pairwise approach to listwise approach (listwise 方法) 4. Learning to rank for information retrieval(综述) 5. Learning to Rank with Nonsmooth Cost Functions(pairwise 方法)
作者	1. Tie-Yan Liu(Microsoft) 2. Hang Li(Microsoft) 3. Christopher J. C. Burges(Microsoft) 4. Tao Qin(Microsoft) 5. Ari J. Lazier(Microsoft)	1. Tie-Yan Liu(Microsoft) 2. Christopher J. C. Burges(Microsoft) 3. Thorsten Joachims(Cornell University) 4. Gordon Sun(Yahoo!) 5. Patrick Gallinari(Laboratoire d'Informatique de Paris 6)

从表 1 的结果中我们可以看到，仅仅利用 PageRank 得到的调研结果覆盖面比较窄，有 3 篇论文都是关于 pairwase 方法，1 篇文章关于 listwise 方法，1 篇文章关于数据集. 而利用我们两阶段模型求得的论文结果囊括了多种经典的排序学习方法(pairwise 和 listwise)、数据集以及综述性文章，更符合调研需求. 同时，PageRank 求得的作者均来自于同一个研究机构 Microsoft，他们的研究工作其实存在着非常大的冗余，而我们的方法得到的作者结果来自于更广泛的科研机构，他们的研究工作覆盖了更多的范畴.

根据以上的实例可知，使用我们提出的面向权威度和多样性的两阶段排序模型得到的论文及作者结果权威度高同时覆盖度广，推荐的结果更适合科研人员进行调研工作.

5　结　论

本文提出了一种面向权威度和多样性的两阶段排序模型. 首先，我们提出了 MutualRank 模型来对权威度进行建模，该模型同时考虑论文及作者两类信息来学习它们的权威度. 接着，我们利用

PDRank 模型融合权威度和差异性两个因素对论文和作者进行排序.通过实验我们证明了 MutualRank 对于权威度的学习效果优于传统的 PageRank,同时基于我们两阶段排序模型得到的调研结果也优于已有的基准方法.

在接下来的工作中,我们准备引入基于论文内容的相似性及差异性来更准确的计算多样性,从而得到更好的自动学术调研效果.同时,也可以引入 learning-to-rank 来学习更好的自动调研模型.

参 考 文 献

[1] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web//Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia, 1998: 161-172

[2] Berkhin P. Survey: A survey on PageRank computing. Internet Mathematics, 2005, 2(1): 73-120

[3] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008

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

[5] Carbonell J G, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 1998: 335-336

[6] Zhu Xiao-Jin, Goldberg A, Gael J V, Andrzejewski D. Improving diversity in ranking using absorbing random walks//Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL). New York, USA, 2007: 97-104

[7] Du Pan. Multi-Objective Ranking with Respect to Diversity [Ph. D. dissertation]. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 2011(in Chinese)
(杜攀. 保持信息多样性的多目标排序技术研究[博士学位论文]. 中国科学院计算技术研究所, 北京, 2011)

[8] Mei Qiao-Zhu, Guo Jian, Radev D R. DivRank: The interplay of prestige and diversity in information networks//Proceedings of the Knowledge Discovery and Data Mining (KDD). New York, USA, 2010: 1009-1018

[9] Du Pan, Guo Jia-Feng, Cheng Xue-Qi. Decayed DivRank: Capturing relevance, diversity and prestige in information networks//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2011: 1239-1240

[10] Han Jia-Wei, Sun Yi-Zhou, Yan Xi-Feng, Yu P S. Mining knowledge from data: An information network analysis approach//Proceedings of the International Conference on Data Engineering (ICDE). Washington DC, USA, 2012: 1214-1217

[11] Sun Yi-Zhou, Yu Yin-Tao, Han Jia-Wei. Ranking-based clustering of heterogeneous information networks with star network schema//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 797-806

[12] Deng Hong-Bo, Zhao Bo, Han Jia-Wei. Collective topic modeling for heterogeneous networks//Proceedings of the 34th International ACM SIGIR Conference of Research and Development in Information Retrieval. Beijing, China, 2011: 1109-1110

[13] Nie Zai-Qing, Zhang Yuan-Zhi, Wen Ji-Rong, Ma Wei-Ying. Object-level ranking: Bringing order to Web objects//Proceedings of the World Wide Web Conference Series. Chiba, Japan, 2005: 567-574

[14] Wan Xiao-Jun, Yang Jian-Wu, Xiao Jian-Guo. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, 2007: 552-559

[15] Hirsch J E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences (PNAS), 2005, 102(46): 16569-16572

HAN Xiao, born in 1988, M. S. candidate. Her research interests include information retrieval and text mining.

GUO Jia-Feng, born in 1980, Ph.D., associate professor. His research interests include Web search and mining, user data mining, machine learning and social networks.

DU Pan, born in 1981, Ph. D., assistant professor. His research interests include information retrieval, text mining, machine learning.

CHENG Xue-Qi, born in 1971, Ph. D., professor, Ph. D. supervisor. His main research interests include network science, Web search and data mining, P2P, information security and distributed system.

Background

This paper focuses on the automatic literature survey technology. Literature survey of domains or topics is the foundation of scientific research. Traditional literature survey is made manually. Researchers construct keywords based on the surveyed topic or problem, and then search for the relevant information (e. g., papers) using search engine or other academic services. But, along with more and more researchers devoting themselves to their work, plenty of academic achievements come out continuously, which brings more difficulties to effective and efficient manual surveys.

This paper aims at developing an automatic literature survey framework to help researchers survey effectively. This framework recommends papers and authors which are most worthwhile surveyed based on the topic given by the user. These recommended papers and authors must be prestigious and cover different aspects of the domain or problem. The authors propose a two-phase ranking model by simultaneously exploring prestige and diversity. Firstly they introduce MutualRank to learn the prestige of the papers as well as the authors by leveraging the two heterogeneous types of information. Then they rank the authors and papers by using PDRank model which combines the prestige and diversity. Finally, they provide users with recommended survey results with high prestige and diversity. Experiments show that MutualRank is better than PageRank on modeling prestige, and the survey results based on two-phase ranking model is superior to the existing baseline methods.

This work is funded by the National Nature Science Foundation of China under Grant Nos. 2013CB329601 and 61100175, National Information Security Plan (242 Plan) under Grant Nos. 2011F45 and 2012G129, National Key Technology R&D Program under Grant Nos. 2012BAH39B04 and 2012BAH39B02, and EU FP7 Project under Grant No. FP7-PIRSES-318939.