AMiner背后的技术细节与挑战

2015-06-12 CSDN大数据

学术文献记载着科学的发展和进步，在科技日新月异高速发展并成为“第一生产力”的今天，学术信息，包括：论文，作者和会议，以及这些实体之间的相互关系，对研究界和企业界都起着越来越重要的作用。有效进行科技论文的组织与管理不仅可以有效提高论文质量与共享方式，还能有效帮助研究人员进行学术交流，缩短科研成果产业化周期。然而，另一方面随着互联网技术的应用和普及，学术网络信息爆炸式增长，这对学术信息检索、挖掘、共享、评价等各个方面带来全新的挑战。

针对这一问题，近年涌现一些相关的学术搜索系统，如Google Scholar、Citeseer、微软的Libra和马萨诸塞州-阿默斯特大学的Rexa。然而大部分已有系统仅提供论文检索服务，例如专家推荐等高层次挖掘搜索服务方面还存在很多不足。总地来说还有许多问题亟需进一步深入研究，尤其是在研究者的脉络分析和可视化方面，目前还缺少成熟的技术方案和可用的实际系统。具体难点体现在：（1）如何从互联网自动获得研究者的语义描述信息，目前虽然已经有一些系统自动建立研究者信息，但目前语义信息抽取的精度还远不能满足实际应用的需求；（2）如何提高专家搜索的精度和推荐效果，这不仅需要对学术文献的内容进行语义分析，更需要对网络结构的分析；（3）如何对研究者网络进行深层分析和挖掘。研究者之间的合作关系多样，如何有效地实时发现研究者之间的关联网络是一个难点；（4）如何构建大规模学术知识库，构建学术知识点的发展脉络。

AMiner利用数据挖掘和社会网络分析与挖掘技术，提供研究者语义信息抽取、面向话题的专家搜索、权威机构搜索、话题发现和趋势分析、基于话题的社会影响力分析、研究者社会网络关系识别、即时社会关系图搜索、研究者能力图谱、审稿人推荐在内的众多功能，为研究者提供更全面的领域知识，和更具针对性的研究话题和合作者信息，为科研的更好发展提供服务。系统自2006年上线以来，已集成来自多个数据源的近8千万学术文献数据。这些文献数据是构建AMiner上层服务的基石。从海量文献及互联网信息中，AMiner利用信息抽取方法自动获取研究者相关信息（包括：教育背景、基本介绍）并建立研究者描述页面，提供搜索、学术评估、合作者推荐、审稿人推荐、话题趋势分析等多样化的服务。

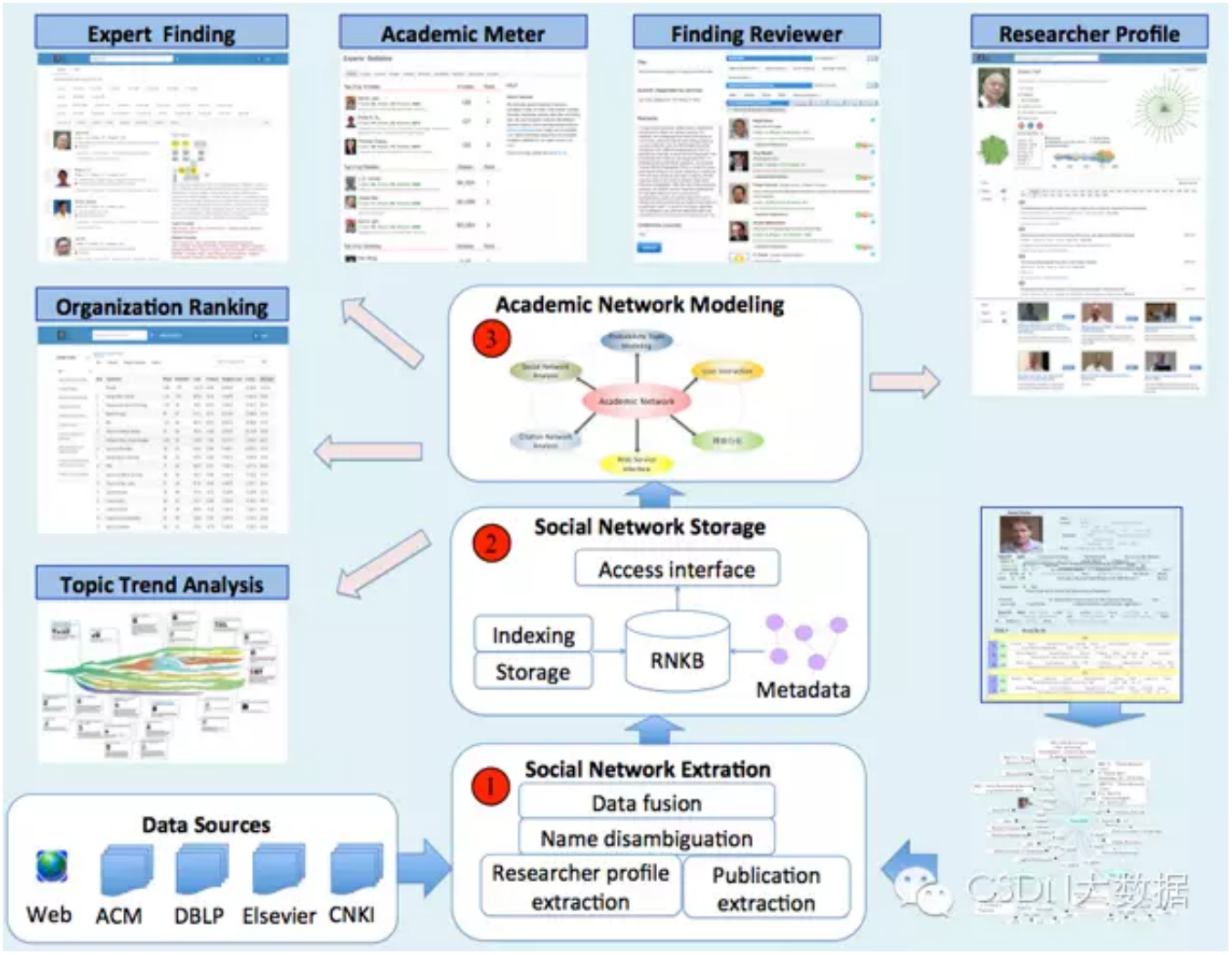


图1给出AMiner系统的核心架构和主要功能。基于自动获取的语义信息，AMiner系统主要包括以下功能：

1. 语义数据抽取：研究者描述信息抽取、研究者兴趣挖掘、研究者账号关联、同名排歧等；
2. 搜索：研究者搜索、论文搜索、综述文献搜索、关联关系搜索以及基于话题的子图搜索；
3. 学术推荐：权威审稿人推荐、优秀论文推荐、“伯乐”推荐等；
4. 深层分析/挖掘功能：领域专家发现、热点话题发现以及论文引用模式挖掘等；
5. 知识库构建与链接：扩语言的学术知识库构建以及扩语言知识库之间的链接构建等。

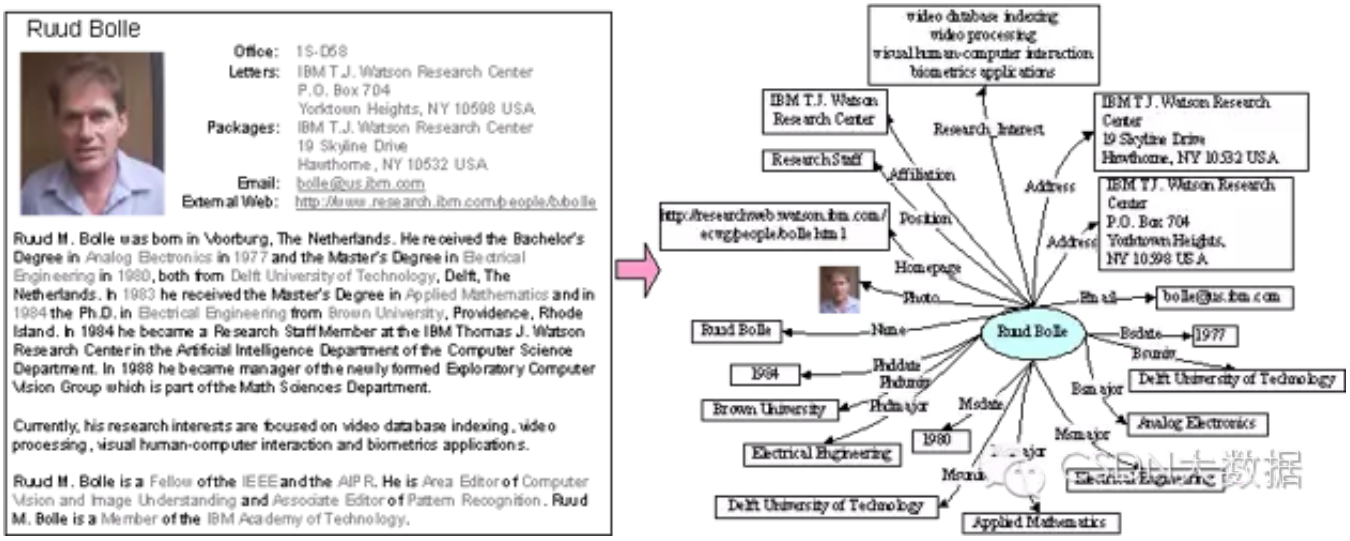
截至目前，AMiner系统已收集了7900多万论文信息、3900多万研究者信息，1.3亿论文引用关系、780万知识实体以及3万多学术会议/期刊。吸引了全球220多个国家的600多万用户访问。本文主要从自动信息抽取、账号自动关联、重名排岐、专家发现以及跨语言联系来讲述AMiner所使用的核心技术。

自动信息抽取

AMiner自动从互联网中发现作者的个人主页，并从个人主页中自动抽取单位、邮箱、个人经历以及头像等信息。抽取的个人信息是基于学术网络挖掘的基础。例如，我们可以实现面向

研究者的垂直搜索，比如查在UIUC读过(或在读)PhD的所有研究者。同时，利用个人的信息，如个人研究兴趣，个人社会关系，可以提高专家发现的准确度。下面从一个例子入手，介绍个人信息抽取的任务，然后给出解决方案。

首先定义研究者个人信息的描述结构（也称为研究者本体），研究者的属性包括：研究者的基本信息，如研究者的名字、照片、职位、工作单位，研究者的联系信息，如研究者的电话、传真、通讯地址、Email等，研究者的教育经历，如研究者毕业的学校、获得某个具体学位的时间、专业等，以及研究者发表的论文。具体来说，对于每个研究者，我们首先通过搜索引擎用其姓名做关键词搜索相关网页，然后利用一个二分类器判断返回的网页是否是该研究者的个人主页或者是该研究者的介绍性网页。最后通过信息抽取算法从该网页抽取研究者的个人信息，构造研究者本体的实例。图2给出了一个研究者个人主页的示例，其中包含了研究者的各种信息。例如：图的上部包含研究者照片、两个通信地址和他的Email地址，图的中间部分用自然语言描述了研究者的教育经历，图的下部提供了研究者的一些任职和所在组织的信息，图的右边显示了理想的结构化的抽取结果。



分析发现，个人信息的各个属性之间有很强的依赖关系。举例来说，研究者的名字可以帮助识别研究者的照片，因为照片的命名可能是研究者的姓或名。在描述个人的教育经历时，比如研究者获得了博士学位(Phd)，那么获得博士学位(Phdmajor)的专业，获得博士学位的日期(Phddate)很可能出现在同一句话中，或者一个列表中。比如从“He received the Bachelor’s Degree in Analog Electronics in 1977”，识别出学士学位的专业会提高识别获得学士学位时间的精度。

手工标注研究者的个人信息比较繁琐，耗时耗力。最近的研究工作验证了自动标注的可行性和有效性，已有技术能够从网页中提取有效信息。这些技术一般都利用一个预先制定的模板，或者针对每个属性学出一个特定的模型来解决各个属性值的提取问题。但是，用这种方法分别提取单个属性效率很低，因为：(1)对于个人信息的每一个属性，如果要使用这些方法，我们必须定义一个特定模板，或者学习一个特定模型。这些模板和模型比较难维护，训

练时间也会很长（实验证明这些针对每个属性的模型训练时间要长于我们提出的统一模型）；(2)这些特定的规则和模型不能够利用各个属性之间的依赖关系，而我们的数据特点是各个属性之间存在很强的依赖关系。通过以上分析，我们可以看出从网页中准确有效地提取各种信息是一个难题，这要求我们提出的方法必须克服以前模型的缺点才能提高语义标注的准确度。

提出的方法包括三个主要步骤：主页发现，预处理和信息标注。在主页发现中，给定研究者的名字，通过搜索引擎我们得到一系列网页。而后，我们训练一个分类器来判定这些网页是否是个人主页或者包含很多研究者信息的介绍性网页（主页发现问题已经在已有的研究中被深入研究过了，这里就不作为我们系统的重点了）。我们把确认的网页URL作为个人信息的属性Homepage的值。

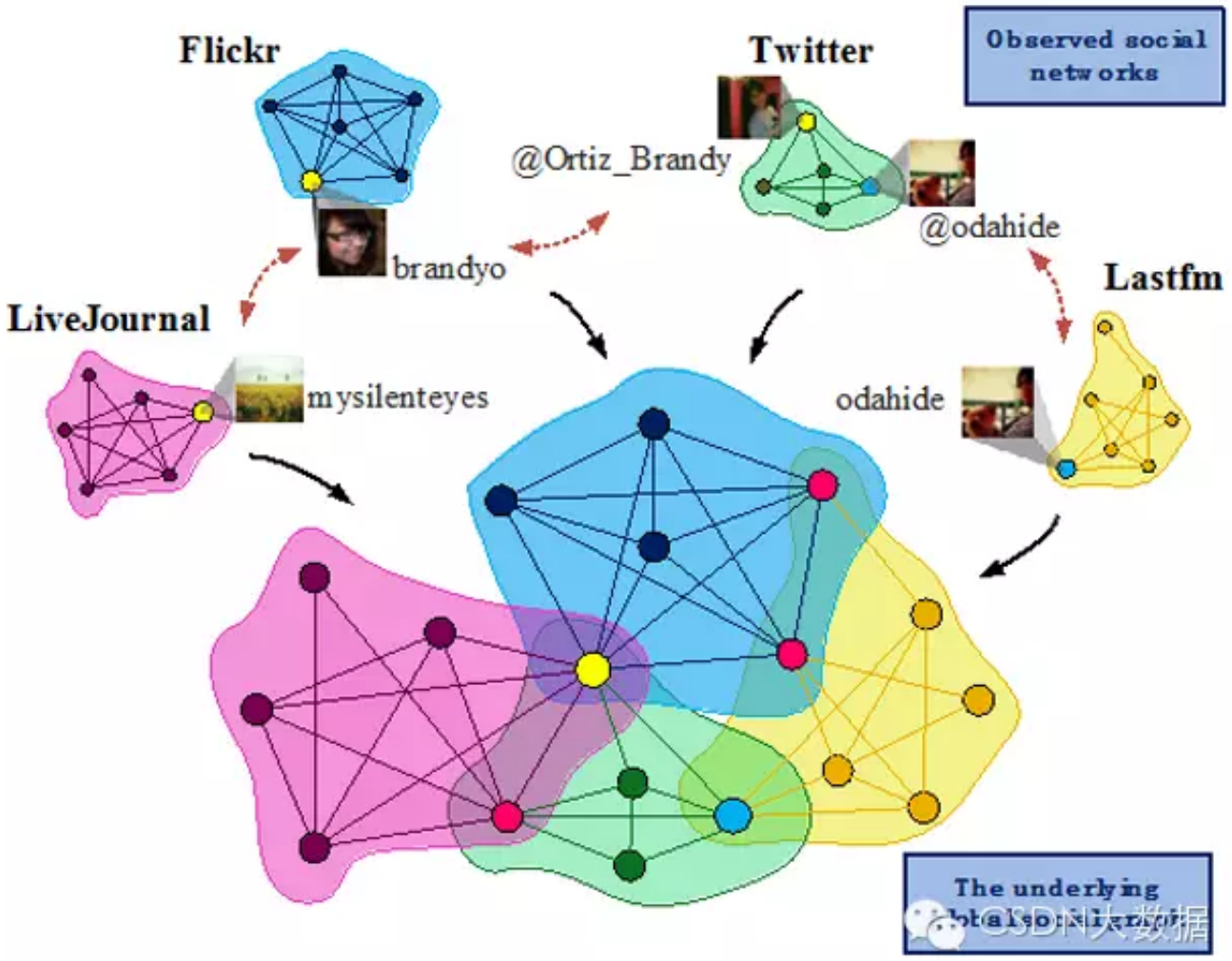
预处理可分为两大步骤：(A)把网页文本分成一个个token，这些token分属于不同的类别。(B)对于不同类型的token，我们给他们设定不同的标签（也就是个人信息的属性）。每个网页的token相当于序列模型每一个观察到的对象，一个网页可以看作一个序列。这样，个人信息的语义标注就可以表示为token的标注。

我们定义特征，从标注好的训练样本学习标注模型，利用学到的标注模型标注新的样本。当新的未标注文本被分成token并生成这些token序列的特征后，我们利用训练好的模型，寻找最好的模拟这个token序列的标签序列，也就是序列模型中状态空间的一个取值作为标注结果。条件随机场是比较流行的序列标注模型，这里我们选择用它做个人信息的语义标注。模型的特征对模型质量有重要的影响，下面我们介绍特征的定义。条件随机场模型的一个好处是对于某个观察值，它可以引入任何形式的特征。对于每一个token单元，我们定义了四种特征，包括基于内容的特征，基于模式的特征，基于term的特征和基于格式的特征。例如，单词的形态：当前单词头一个字母是否大写，单词的词缀等；图像颜色特征：图像中有多少种不同的颜色，图像中每个像素用多少个二进制位表示；格式特征：当前token是否是黑体等。（详细算法请参考[Tang, 2010]Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages。）

用户多账号关联

随着社交网络快速发展，不断涌现的大规模社交网络（如Facebook，LinkedIn，新浪微博等）吸引了数以亿计用户。不同的社交网络在其功能，用户体验，目标用户群等各个方面都有不同的特点，例如Facebook是真实社交网络的线上版本，其内部的好友关系大多正是用户在线下的真实好友关系，且其好友关系是双向的；在Twitter或新浪微博上用户则更趋和自己的偶像或是意见领袖建立关系，这种关系是单向的；LinkedIn是职业化的社交网络，以便于用户更新自己展示自己的工作经历和能力水平，Google Scholar和AMiner等学术合作网络

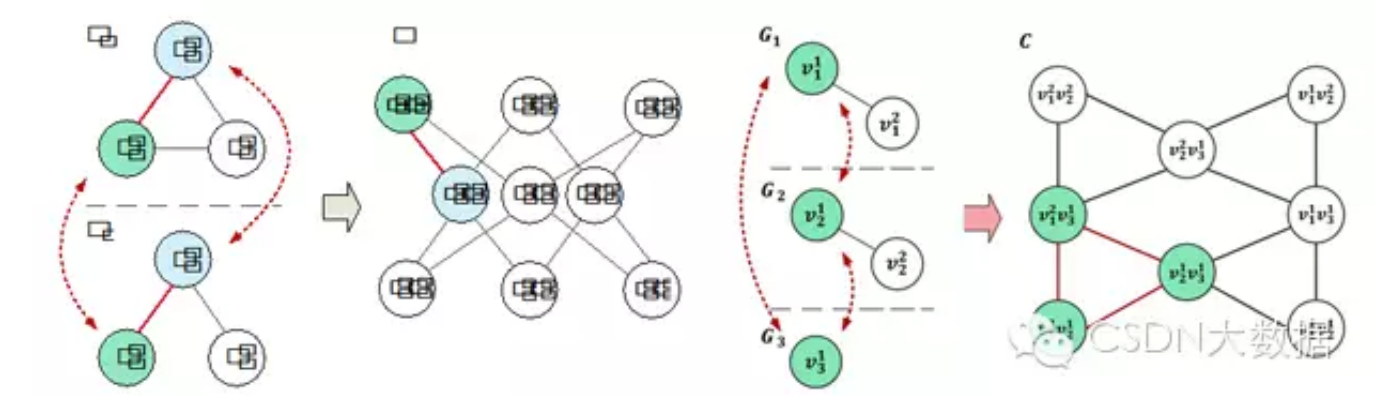
则反映了学者在发表学术论文时的合作关系。正因为每种社交网络在用户的工作和生活中都各自扮演着不同的角色，用户常常在不同的社交网络上都拥有账户。每个账户都是用户完整形象的一个局部缩影，很显然，由于分散在各个不同的社交网络，这些局部是不相互连通的。因此，AMiner通过机器学习手段，自动将多个社交网络的账户进行自动关联。



对于这一问题，我们面临许多难点。首先，获取社交网络数据很困难，鉴于这一信息的重要价值，且涉及用户隐私问题，各大互联网公司对自己拥有的社交网络数据都保持非常谨慎的态度，我们只能通过公共API获得少量不完整的数据；其次由于用户会有意或无意地在账户中略去部分个人信息，我们可以观察到的用户特征非常稀疏；同时，各个不同社交网络的用户账户信息条目是异构的，条目不能一一对应，且条目的内容表达方式也不尽相同，因而不同账户之间的相似度也无法直接度量。此外数据存在噪声信息，例如用户在一个社交网络可能存在多个账户，以及账户信息中的错误拼写甚至刻意错填的信息等等。

对于这一问题，我们需要考虑三个层面的因素。首先是用户之间的相似性，对于不同社交网络中的两个用户，我们可以从他们的用户名，账户信息，以及发表内容等方面，判断其是否是现实中的同一个人，我们将这一信息称为节点相似度；同一用户在不同社交网络中常常会有朋友圈的重叠，例如同时在微博和微信上均与某一用户是好友，我们将此称为边相似度；

此外，由于两个用户是否是同一个人这一关系为对称关系，当有多个社交网络存在时，我们还应考虑逻辑传递性，即若已知A是B且B是C，则A必然是C。AMiner采用的账户自动关联算法正是综合考虑了节点相似度，边相似度以及逻辑传递性这三个层面的因素。首先，我们将不同社交网络中的账户两两配对，将问题转化为二分类问题（即判断任意配对中的两两账户是否属于同一用户）。我们将每个配对表征为特征向量，用于刻画节点之间的相似度。为了引入边相似度和逻辑传递性，我们考虑使用马尔科夫随机场对问题进行建模。图4给出了对边相似度和逻辑传递性的建模方法示意，对于两个不同社交网络之间的两个配对，若其两两在各自网络中互为好友，则在模型中倾向于使这两个配对的判断结果相同。对于任意三个社交网络中的三个存在传递关系的配对，模型倾向于使得三个配对的判定结果不违背逻辑传递性。



重名排歧

从海量文献中自动建立研究者账户是AMiner的核心功能，其中最大挑战之一即是作者的重名排歧问题。现实世界对于实体的描述是充满二义性的，人的名称指代也是其中之一。同一个人名可能被不同人使用，例如王伟、张静、李刚等。预测同时人名可以有各种变形，如缩写，前后名倒置，中间名，以及加入前后缀等等。此外AMiner同时处理中英文双语数据，这也带来了一些独特的挑战。

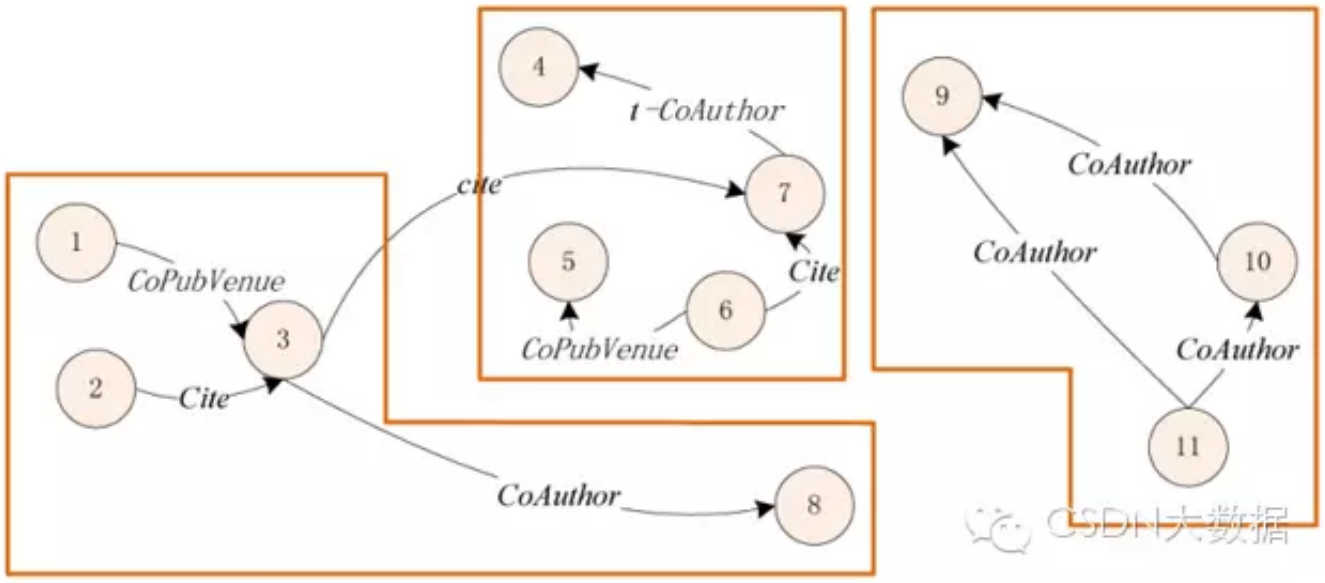
同名异义是电子数据库和语义社会网络中普遍存在的问题。比如：在查询一个研究者所发表的文章时，现有的系统会将所有与该研究者同名的作者的文章返回给用户，这样无疑会使用户产生混淆。而语义社会网络中，同名者的个人社会网络往往会出现错误的重叠或合并。针对这些问题，同名排歧的研究工作就显得非常重要。

目前，同名作者文章的排歧工作主要有以下难点：(1)每篇文章的信息量有限，往往只有文章作者的名字，文章的题目，发表会议和发表时间。(2)即使有关于文章作者的描述，比如：学校或组织机构，也会因为作者自身职位的变化而产生歧义。现有的研究工作中，有指导的学习算法要对每个排歧目标的数据进行学习和训练，方法的可扩展性差；无指导的学习方法受到可利用信息量的限制，又没有人工的指导，所以排歧效果有待提高。

针对这些问题，我们提出了基于约束的概率模型框架。首先，利用隐马尔可夫随机场理论构造目标函数，将整个问题转化为最小化目标函数问题。这里，目标函数主要包含两个部分：一部分是聚类的每个类别中数据点之间的距离，用来衡量每个聚类结果的紧密程度；另外一部分为当前聚类结果所违背的所有约束的惩罚值之和。所以，整个算法的目标就是要找到内部紧密而且尽量少违背约束的聚类结果来作为同名排歧的结果。而算法中生成约束的方法非常灵活，可以是人工的指导，也可以是通过社会网络找到文章作者之间的关系。也就是说，基于约束的概率模型框架可以灵活的将各种知识以约束的形式放到算法中，从而可以很好地利用各种指导和数据来提高精度。

在求解该问题时用到以下主要约束。所有这些约束都是定义在两篇文章之间的。第一个约束指的是两篇文章的首要作者都来自同一个组织，比如：同一个学校或者同一个研究单位。定义约束一的直观想法是来自于同一个单位而且同名的作者很可能就是同一个人，那么它们发表的文章也应该聚到一起。约束二指的是两篇文章除了首要作者名字相同之外，还有至少一个次要作者的名字也相同。定义约束二的直观想法是和同一个人合作的两个同名者很可能就是同一个人。约束三指的是两篇文章中，一篇文章引用了另外一篇文章。定义约束三的直观想法是研究者往往喜欢引用自己的文章，那么，如果两篇文章的首要作者名字相同而且存在引用关系，那么这两篇文章很可能就是同一个作者发表的。约束四指的是两篇文章的首要作者使用同一个电子邮件地址。可以看出约束四是一个很强的约束，因为电子邮件可以唯一地对作者进行标识。约束约束五指的是由用户反馈得到的约束，当用户指定两篇文章属于同一个作者时，这两篇文章之间就形成约束五。约束五可以看作是将人工指导以一种约束的形式加入到算法框架中，将算法由无指导变为半指导学习算法。

图5给出了一个重名排岐的实例。图中每个点表示一篇论文，每个有向边表示两篇论文之间的不同类型的关系，这些关系即可以转化为上述约束。两个点之间的距离反应了它们内容的相似度。实线框表示论文属于同一个作者（聚类类别）。可以非常直观地看出，仅根据内容相似度不能取得很好的聚类效果。但是不同类型的关系对于区分不同的作者非常有效。例如，根据节点3和8之间的合作关系，很容易将它们分配到同一个类别，尽管它们之间的内容相似度很低。（算法细节请参考[Tang, 2012] Jie Tang, A.C.M.Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for
Name Disambiguation in Digital Library. IEEE Transaction on Knowledge
and Data Engineering (TKDE), Volume 24, Issue 6, 2012, Pages 975-987。）



专家发现

专家搜索是AMiner提供的主要服务之一，其根据用户查询的话题找出在相关领域的权威专家。与传统文献检索相比，专家搜索的不同之处在于，搜索对象由传统的文档变成人，一个人关联的信息相比于一个文档来说，不但数量上大幅增加，而且类型上由单一的文本扩展出非文本的信息。例如，一个研究者可以关联多篇论文，论文有文本内容信息，也有非文本的发表会议以及杂志和合作者等非文本信息。因此，信息异构化带来的挑战是，依靠传统的文本检索中使用的文本匹配方法很可能造成语义缺失、检索不够准确的问题。例如，想查找“自然语言处理”方面的专家。结果发现大多数专家不会在自己的论文中撰写“自然语言处理”的字样，因
为仅依靠关键词进行匹配几乎不能返回有效的结果。而如果我们知道自然语言处理领域的权威会议是“ACL”等，根据研究者发表的会议信息可以很容易判断出他是否是该领域的权威专家。因此，需要设计一种方法有效地利用研究者的异构关联信息来发现领域专家。

我们首先建立研究者异构信息网络。与同构网络不通，异构网络中可能存在多种不通类型的网络对象，网络链接也呈现日益复杂的关系。图6给出了一个具体的研究者网络实例。在该网络中，异构实体包括：论文、研究者和会议/期刊等，网络关系包括：论文之间的引用关系，论文发表在会议/期刊上的关系，研究者撰写论文的关系等。然后基于主题模型LDA对研究者异构信息网络统一进行建模，从中估计出不同类型的实体，包括研究者、会议、关键词以及论文在不同隐含话题上的概率分布。有了这些概率分布，用户给定一个查询词，就可以推断与之概率分布相近的专家，进一步，还可以推断出相关的会议和论文等异构网络中存在的各种实体类型。具体地，建模时对于每篇论文，根据当前论文对话题的概率分布，为之生成一个隐含话题，然后根据话题对各实体的概率分布，生成该论文关联的每个单词、作者以及会议的实体。求解模型参数（各实体对话题的概率分布）可采用与LDA方法相同的Gibbs sampling算法。（相关研究请参考[Tang, 2008] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner:

Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'08). pp.990-998，[Tang, 2011] Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. Machine Learning Journal, Volume 82, Issue 2 (2011), Pages 211-237。）实践中，我们采用主题模型加权语言模型的方法进行检索。

跨语言的知识链接

AMiner正在构建和集成学术领域的知识图谱，从文献中抽取只的是概念，并与知识库进行连接，挖掘相关概念并分析知识概念的上下位关系。同时，AMiner还通过机器学习手段针对跨语言的知识库进行自动链接。

当前各类百科资源存在不同语言的知识分布极不平衡的问题。如果能够在英文维基百科和中文百度百科（或互动百科）之间有效地建立跨语言知识链接，将大大提高中英文知识的跨语言共享。图7展示了一个扩语言知识链接的实例。左边是英文维基百科上的“Anaerobic exercise”，右边是百度百科上的“无氧运动”。很多关键特征可以用来帮忙建立中英文维基之间的关联。例如，图中标出了一些有用的特征，包括标题，出链，类别和作者等。

我们充分利用维基类知识资源中的上述特征，提出基于链接因子图的异构知识库的知识链接方法和基于链接标注的增量式跨语言知识链接方法，在异构百科之间发现大规模跨语言知识链接。模型的目标是判断一个给定中英文维基页面对是否所指相同。基于链接因子图的异构知识库的知识链接方法采用链接关系的相似度进一步使用链接因子图模型对跨语言知识链接任务统一建模。具体地，如果一个中英文维基页面对被预测为相同事物，则它们各自出链的页面所组成的对也有很大概率所指为相同事物。考虑到基于链接因子图的知识链接方法主要依赖于初始种子跨语言链接集合以及词条之间的链接关系，进一步提出了基于链接标注的增量式跨语言知识链接方法，以提高跨语言知识链接的可用性。最终实验证明链接标注和增量式方法，均可有效提高跨语言知识链接的精度。（相关研究请参考[Wang, 2012] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. Cross-lingual Knowledge Linking Across Wiki Knowledge Bases. In Proceedings of the Twenty-First World Wide Web Conference (WWW'12). pp. 459-468。）

经验总结与未来展望

总之，在学术研究数据规模不断增长的今天，从海量数据中挖掘有价值的知识使用户真正获益具有极大的挑战。下面从上述四个技术点分别阐述存在的挑战以及未来可能提高的方向。

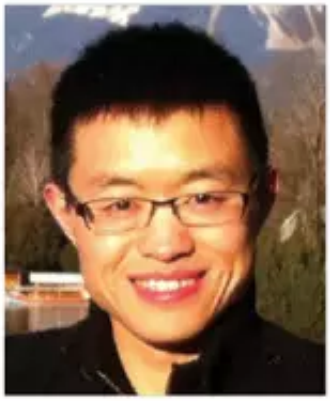
首先、异构数据提高有效信息提取的难度。例如研究者的个人主页格式五花八门，有个人撰写的，有单位统一制作的，还有Google Scholar生成的，这些不同格式要求自动抽取器能够像人脑一样非常智能地识别有效个人信息。目前处理抽取主要依靠大量训练数据来提高抽取模型的精度，未来希望能够从用户反馈的个人信息中自动识别有效特征来进一步提高抽取模型的精度。

其次、数据规模大，以及跨领域、多语言等特征造成数据合并的难度。目前收集的数据源有专业计算机领域的数据库，包括ACM和DBLP，也有面向全领域的数据源，如英文的Elsevier和中文的CNKI。这些规模庞大，来源各异的数据导致同名不同人、不同名同人、不同语言同人等问题日趋严重。有些常见人名，例如“王伟”，甚至包括跨多个领域的上千个真实个体。尽管重名排岐在过去的多年中一直有研究者不懈努力研究，但是在如此大规模数据上进行排岐还未见真实成效。Google Scholar甚至也回避此问题，简单地将所有同名的人归在一起。由此可见该问题的难度。未来可能的提高点有两个，一是在模型中加入人名常见度这一先验知识，使采用不同模型处理不同人名；二是依靠用户的个人反馈自动修正关联的错误合并结果。

再次、海量数据加大搜索有效信息的难度。目前的系统采用主题模型平滑传统语言模型来客服主题漂移的问题，但主题模型归根到底仍然逃脱不了对词共现的依赖。如果整个数据集中从来没有出现过或者极少出现过某个查询词，那么用该查询词进行检索效果依然不会很好。因此该问题仍然有待进一步提高。其可能的解决方案是让用户互打标签，标识其研究兴趣，搜索时推荐相关标签，按照标签进行搜索。

最后、知识库质量影响用户体验。目前学术知识库的构建仅称得上初见端倪。这其中仍然有很多挑战，除了之前提及的跨语言链接的问题，还存在概念上下位关系识别以及不同源之间概念链接，例如论文数据库到维基百科的链接等诸多问题。这些问题都需要深入到对信息分门别类，各个建模。

作者简介：



唐杰

清华计算机系副教授、博士生导师。于2006年6月在清华大学计算机系获得博士学位，曾在康纳尔大学、伊利诺伊香槟分校、香港中文大学、香港科技大学进行学术访问。主要研究兴趣包括：社会网络分析、数据挖掘、机器学习和知识图谱，发表论文100多篇，Google引用5000余次，申请专利12项。

张静

清华大学计算机系博士三年级学生。研究兴趣包括社会影响力分析与度量、社会网络结构相似度度量等。曾在香港科技大学、美国伊利诺伊香槟分校以及比利时鲁汶大学访问。担任ICDM2014、ASONA 2015的PC Member以及WSDM2015的Proceeding Chair。

张宇韬

清华大学计算机科学与技术系博士生。2013年6月在南京航空航天大学计算机科学与技术学院获得学士学位，曾在台湾清华大学进行学术访问。主要研究内容包括：数据挖掘、异构数据集成及信息可视化分析，在重要国际会议（包括KDD、CIKM等）发表论文4篇。作为主要技术负责人参与研发AMine[illegible]。

阅读原文

微信扫一扫
关注该公众号