## 1.2 Our Solution

在本文中，我们的目的是系统地调查Web用户分析问题。首先，我们将Web用户分解分解为三个子任务：proflle提取，名称消歧和用户兴趣发现。所有三个子任务都可以使用图形模型进行形式化。特别地，为了提取摘要，由于Web上的信息自然地以层次结构布局，我们提出在树状结构的条件随机空间中形式化问题。为了消除歧义，问题是将文件分配给具有相同名称的不同人员。我们在马尔可夫随机图中形成问题，其中每个节点表示纸张，边缘表示paper之间的关系（例如共同作者）。为了消除歧义，问题是将文件分配给具有相同名称的不同人员。我们在马尔可夫随机图中形成问题，其中每个节点表示纸张，边缘表示纸张之间的关系（例如共同作者）。对于用户兴趣发现，我们提出了一个生成图形模型，其中纸张写作过程在一系列概率性步骤中被形式化。据我们所知，我们的工作是以组合方式正式化所有用户分类的子任务，并立即解决所有问题。

我们已经在ArnetMiner.org系统中实施了提出的方法。该系统已经在互联网上运行了三年多，吸引了来自190个国家的用户访问。总共有50多万研究人员的资料被提取出来。我们进行实验以提取研究人员的素质。实验结果表明，我们的方法明显优于使用分离模型进行萃取的方法。实验结果也表明，我们的消歧方法可以胜过现有的方法。我们将提出的方法应用于专家报告。实验结果表明，我们的提取方法，名称消歧和用户兴趣分析确实可以提高专家的销售量（MAP方面为26％）

我们在本文中的贡献包括：（1）用户沟通问题的形式化，（2）提出了一种提取用户污名的单一标记方法，（3）提出了一个概念性方法来排除歧义，（4）主题模型的建议，用于对用户兴趣进行局部分析，以及（5）对所提出方法的有效性的实证验证。 本文提出的方法是一般的，可以应用于许多应用，例如社交网络提取和信息集成。

本文的其余部分安排如下。在第2节中，我们正式化了Web用户分析的问题。在第3节中，我们概述了我们的方法。在第4节中，我们解释了我们的提取方法，在第5节中我们将描述在整合提取的概念时如何处理名称模糊问题。在第6节中，我们介绍了用户兴趣发现的方法。第7节给出了实验结果。第8节描述了一个演示系统。最后，在第10节结束之前，介绍相关工作。

# 2.问题形式化定义

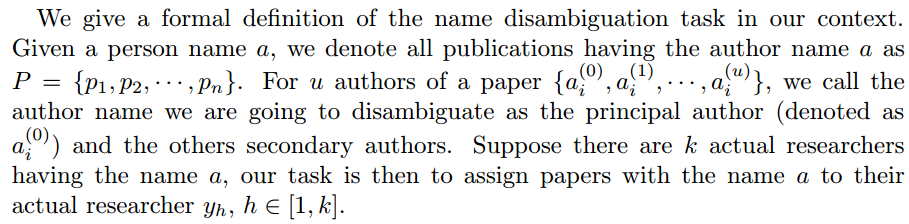
在不同的应用程序中，侧写模式的区别可能是不同的。在本文中，我们以研究者文摘为例进行了说明。研究人员的侧写和提出的用户评价方法可以很容易地扩展到其他应用程序。

我们deflne研究者proflle的模式（如图2所示），通过扩展FOAF本体[Brickley和Miller 2004]。在模式中，4个概念，29个属性和4个关系被排除。社会网络表示与当前研究者有关的子社会图。兴趣表示语义主题方面，稍后将详细介绍。该出版物表示由研究人员共同撰写的文献。

我们在这里描述我们要处理的三个关键问题：简档提取，名称消歧和用户兴趣。

（1）profile提取

（2）名称消歧。我们不直接从主页提取出版物。相反，我们将现有在线数据源的发布数据进行整合。我们选择了DBLP参考书目（dblp.uni-trier.de/），这是最好的格式化和有组织的参考书目数据集之一。DBLP涵盖了主要计算机科学出版物的大约120万篇论文。在DBLP中，作者以他们的名字来识别。为了整合研究人员和出版物数据，我们使用研究员姓名和作者姓名作为识别者。该方法不可避免地具有名称模糊性问题。



# 3.我们的方法概述

我们提出一种组合方法来解决用户分析问题。图3显示了我们的方法概述。 主要有两个部分：提取和整合，用户兴趣分析。第一个组件的目标是提取和整合来自Web的信息; 而第二部分的目标是分析用户的兴趣。

在proflle提取和集成组件中，给出研究人员名称，我们使用Google API来检索包含研究人员名称的文档列表。然后我们使用分类模型来识别列表中的文档是主页还是研究人员的介绍页面。接下来，我们使用提取模型从识别的页面中提取信息。特别地，我们将问题视为将标签分配给输入文本，每个标签表示一个proflle属性。

我们从几个在线数字图书馆（例如DBLP）中检索发布信息。我们整合发布信息和提取的信息。**我们提出一个概率模型来处理提取的用户资源的名称模糊问题。该模型可以将任何类型的域背景知识或监督信息（例如，用户反馈）作为特征，以改善消歧的性能。**在用户兴趣分析组件中，我们使用概率主题模型来发现与每个研究者相关联的潜在主题分布。然后我们使用发现的主题分布作为研究者的兴趣。

在本文中，我们的主要技术贡献在于我们提出的处理两个组件中的三个子任务的方法：提取，集成和用户兴趣发现。理论上，所有这三种方法都是基于概率图形模型。更具体地说，为了提取和整合，我们的方法是基于马尔科夫随机场的理论[Hammersley和Clifiord 1971]。马尔可夫随机场（MRF）是符合马尔科夫属性的标签（隐含变量）的概率分布。它可以正式地被排除如下。

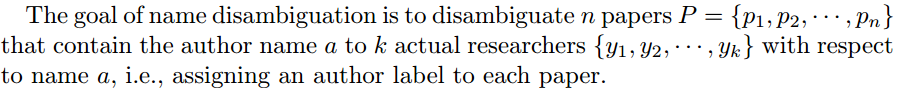
**提出的提取模型是树结构条件随机场（TCRFs），提出的名称消歧模型是基于隐马尔可夫随机场（HMRFs）[Basu et al。2004年]。我们使用这两个模型的原因是：（1）这样一个模型可以描述信息之间的依赖关系，从而提高提取的精度和名称消歧。（2）对于抽取提取，我们可以为监督学习标注一些培训数据; 而为了消除歧义，提供足够的培训数据是很不寻常的。因此，我们提出使用歧义模型（TCRF）进行摘要提取和生成模型（HMRF）作为消歧任务。（3）两种型号都可以轻松扩展，因此对于不同的应用，我们可以根据应用特征来扩展模型。**

对于用户兴趣分析，提出的模型是一个**多层次的贝叶斯网络**，它通过跟随一个随机过程来对每个论文进行建模：论文的作者将根据他/她的研究兴趣（即主题分布）fP（zja）gz来决定要写的主题。然后，根据主题fP（wjz）gw的字分布，从主题z中采样一个字wdi。这一系列的概率性步骤可以很好地掌握作者撰写论文的过程。此外，可以以无监督的方式估计参数（主题分布和字分布）。使用贝叶斯网络进行用户兴趣分析的另一个原因是我们可以轻松地将不同类型的对象（例如研究人员，出版场所和论文）合并到一个模型中，从而我们可以发现异构对象之间的潜在依赖关系。

# 4.profile提取

# 5.名称消歧

我们从现有的在线数据源中抓取出版物数据。为了整合研究人员和出版物数据，我们使用研究员姓名和作者姓名作为识别者。该方法不可避免地具有名称含糊的问题。



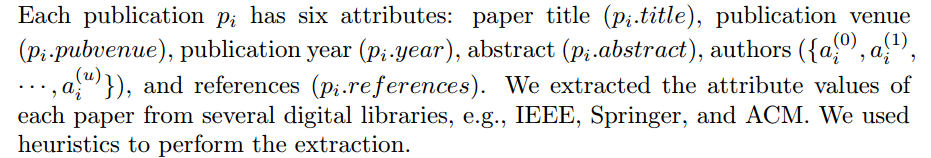
名称消歧的目的是消除n篇论文P，其中包含作者姓名a至k的实际研究人员关于姓名a，即将作者标签分配给每篇论文。

我们提出一个解决问题的概率模型。我们在这种方法中的直觉是基于两个观点：（1）具有相似内容的论文倾向于具有相同的标签（属于同一作者）; 和（2）具有较强关系的论文往往具有相同的标签，例如两篇论文由同一作者撰写。

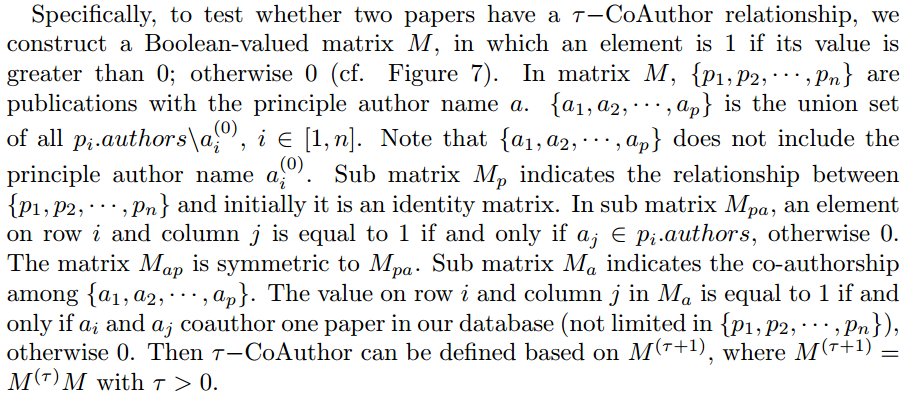
我们的方法是基于隐马尔可夫随机场（HMRF）模型，MRF的一个特例。我们选择HMRF的原因是由于其自然的优势。首先，与所有MRF家族成员一样，HMRF可用于建立观察结果之间的依赖关系（例如CoAuthor）（每篇论文被视为观察）。第二，HMRF支持无监督学习，监督学习以及半监督学习。在本文中，我们将重点关注使用HMRF进行名义消歧的无监督学习，但很容易将一些先前/监督的信息纳入模型，从而将拟议的方法扩展到半监督学习。第三，在HMRF模型中进行模型选择是很自然的。HMRF模型中的目标函数是给定观测值的隐含变量的后验概率，可用作模型选择的标准。

在本节的其余部分，我们将介绍隐马尔科夫随机场模型，然后对名义消歧问题的目标函数进行排序。

## 5.1 Data Preparation



我们定义了五种文件之间的关系（表二）。关系r1代表两篇论文在同一地点发表。关系r2意味着两篇论文具有相同的第二作者，而关系r3意味着一篇论文引用了另一篇论文。关系r4表示通过用户反馈提供的基于约束的关系。例如，用户可以指定将两篇论文消除歧义给同一作者。我们用一个例子来解释关系r5。假设pi有作者\ David Mitchell“和\ Andrew Mark”，pj有作者\ David Mitchell“和\ Fernando Mulford”。我们要消除“米切尔大卫”的消息。如果\ Andrew Mark“和\ Fernando Mulford”也合着另一篇文章，那么我们说pi和pj有一个2-CoAuthor关系。



## 5.2 Formulation using Hidden Markov Random Fields