## 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

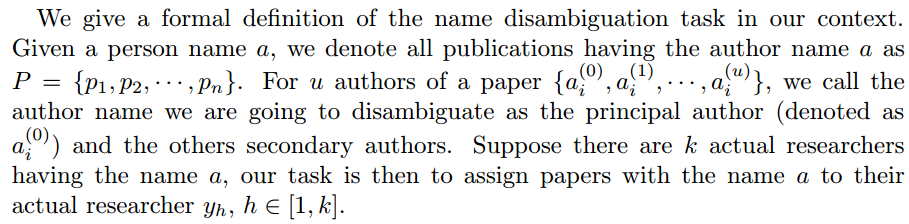
在不同的应用程序中，侧写模式的区别可能是不同的。在本文中，我们以研究者文摘为例进行了说明。研究人员的侧写和提出的用户评价方法可以很容易地扩展到其他应用程序。

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。