## 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，其中每个节点表示论文，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，其中每个节点表示论文，边缘表示论文之间的关系（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，其中论文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

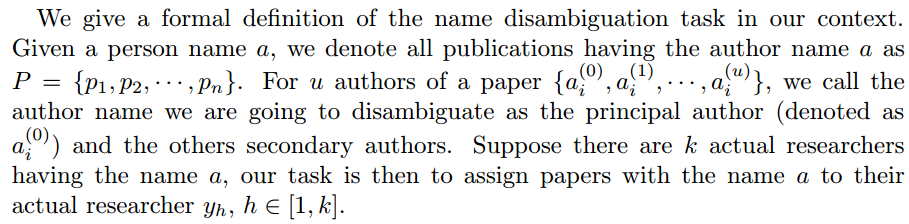
在不同的应用程序中，侧写模式的区别可能是不同的。在本文中，我们以研究者文摘为例进行了说明。研究人员的侧写和提出的用户评价方法可以很容易地扩展到其他应用程序。

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。该论文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论文。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论文的大约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论文数据，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。**



我们在我们的上下文中对名称消歧任务进行正式的定义。给定一个人名字a，我们将所有作者姓名为a的论文表示为P.对于一篇论文的u个作者，我们称之为我们将消除歧义的作者姓名作为主要作者（表示为（0））和其他次要作者。假设有k个实际的研究人员有一个名字，我们的任务就是**给他们的实际研究人员指定一个名字a的文件**

# 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。

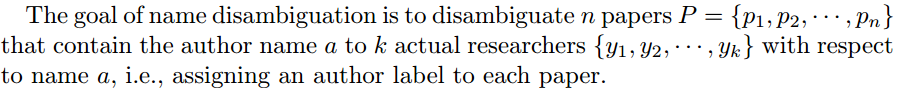
**提出的提取模型是树结构条件随机场（TCRFs），提出的名称消歧模型是基于隐马尔可夫随机场（HMRFs）[Basu et al。2004年]。我们使用这两个模型的原因是：（1）这样一个模型可以描述信息之间的依赖关系，从而提高提取的精度和名称消歧。（2）对于抽取提取，我们可以为监督学习标注一些培训数据; 而为了消除歧义，提供足够的培训数据是很不寻常的。因此，我们提出使用歧义模型（TCRF）进行摘要提取和生成模型（HMRF）作为消歧任务。（3）两种型号都可以轻松扩展，因此对于不同的应用，我们可以根据应用特征来扩展模型。**

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论文数据。为了整合研究人员和论文数据，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。



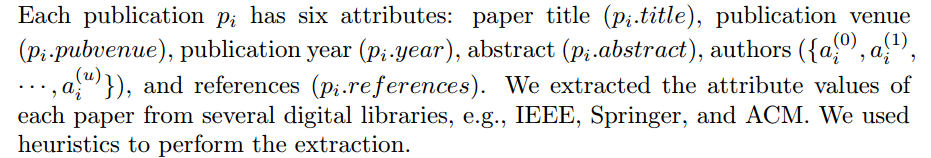
名称消歧的目的是消除n篇论文P，其中包含作者姓名a至k的实际研究人员关于姓名a，即将研究者标签分配给每篇论文。

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。

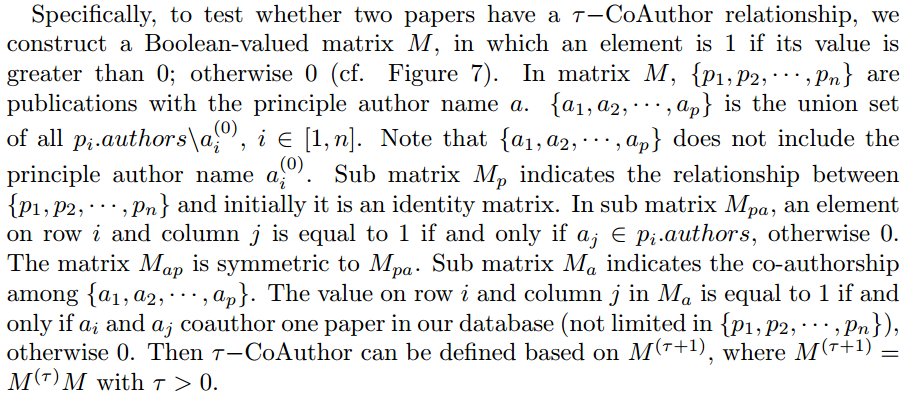
在本节的其余部分，我们将介绍隐马尔科夫随机场模型，然后对名义消歧问题的目标函数进行排序。

## 5.1 Data Preparation



每篇文章6种属性，**文章名、出版地、出版年份、摘要、作者（多位作者），参考文献。**我们从几个数字图书馆，例如IEEE，Springer和ACM中提取了每篇论文的属性值。我们用启发式来进行抽取。

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。



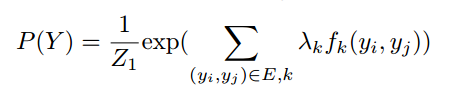
## 5.2 Formulation using Hidden Markov Random Fields

隐马尔可夫随机场（HMRF）是MRF家族的成员，其概念源自隐马尔可夫模型（HMM）[Ghhramani and Jordan 1997]。HMRF主要由三个组成部分组成：可观察的随机变量集X = fxign i = 1，隐藏的随机变量Y = fyign i = 1，隐藏的变量中每对变量之间的邻域。

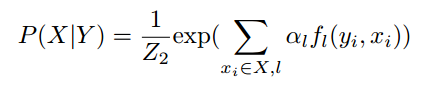
我们将消歧问题正式化为将关系论文分成不同的集群。让隐藏变量Y成为论文中的集群标签。每个隐藏变量yi从集合f1获取一个值 1到K，是分类的标签。观测变量X对应于论文，其中每个随机变量xi由相应隐藏变量yi确定的条件概率分布P（xi|yi）生成。

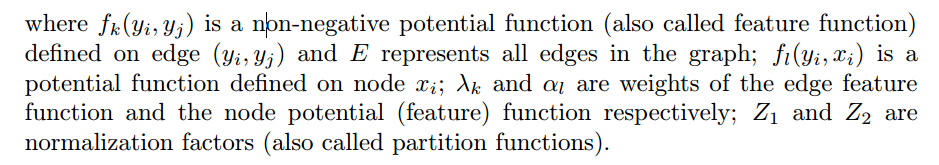
图8示出了HMRF的示例图形表示。观测变量xi对应于论文，隐藏变量yi对应于分配结果。隐藏变量之间的相关边界对应于论文之间的关系（参见表II关系的定义）。

通过随机场的基本定理[Hammersley和Clifiord 1971]，标签组合Y的概率分布形式如下：

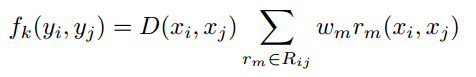


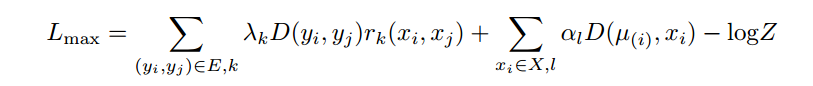
我们假设发布数据是在球面高斯分布下生成的，所以我们有





边特征函数





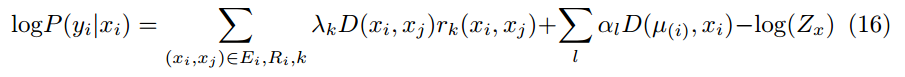
## 5.3 参数估计

参数估计问题是确定参数的值并确定所有论文的分配。更准确地说，我们针对条件模型P（Y jX;£）优化对数似然目标函数。

参数估计算法主要包括三个步骤：初始化，论文分配和参数更新。基本思想是我们**首先选择参数的初始化，并为每个集群选择一个质心**。然后，我们每个文件分配到其最近的聚类，根据分配计算每个集群的质心。之后，我们通过最大化目标函数来更新每个特征函数的权重。迭代持续到收敛。

我们现在详细介绍算法中的三个步骤。在初始化步骤中，我们首先根据它们之间的关系将论文集中到不相交的群体中，即如果两个论文有关系，那么它们被分配给同一个研究者。因此，我们得到了n组。如果，n等于我们实际的研究人员k，那么这些组被用作我们的初始赋值。如果，n<k，我们选择k-n随机分配。如果n> k，我们聚集最近的组，直到只剩下k个组。

在分配步骤中，将每篇论文xi分配给“（h）”以使局部logP（yijxi）最大化，



其中Ei表示与xi有关的所有关系。等式（16）中的第二项是区域相似度函数D（xi;“（h））和关系相似函数D（xi; xj）的多项式组合，其可以在多项式时间中计算。Zx是归一化因子，可以近似地被视为一个常数。

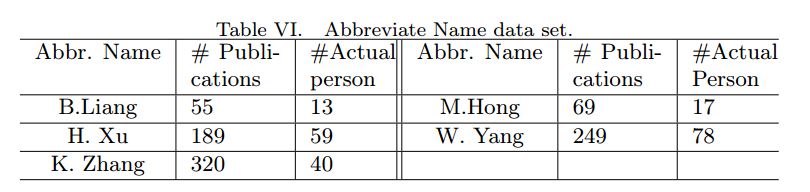
同时保持固定的其他论文的任务是执行文件的分配。使用贪心算法来顺序更新每张纸的分配。该算法以随机顺序执行所有论文的分配。在分配所有文件后，重新分配过程。这个过程运行直到没有论文在两次连续的迭代之间改变其分配。

# 7. EXPERIMENTAL RESULTS

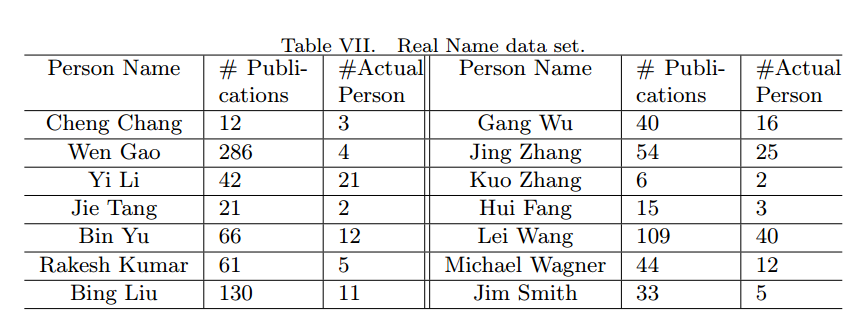
## 7.2 Name Disambiguation Performance

### 7.2.1Data sets and evaluation measure

为了评估我们的方法，我们从ArnetMiner创建了两个数据集，即缩写名称数据集和实名数据集。通过在我们的数据库中查询五个缩写名称来收集第一个数据集。所有这些缩写名称都是通过将原始名称简化为其名称的初始姓氏和姓氏来生成的。例如，“程昌”简化为\ C。Chang“。该数据集的统计数据如表VI所示。



另一个数据集包括14个真实姓名。在这些名字中，一些名字只对应几个人。例如“程昌”对应三名实习人员和“文高”四人; 而一些名字似乎很受欢迎。例如，有25名名为“京Zhang”的人，“雷王”为40人。 该数据集的统计数据见表七。



对论文进行消歧五个人工注释。创建了一个规范来指导注释过程。每篇论文都标有一个数字，表示实际的人。标签工作是根据作者主页，资料和电子邮件地址的论文列表进行的。为了在注释中进一步分歧，我们进行了“多数投票”。

从统计数据我们发现消歧结果是非常不平衡的。例如，“文高”共撰写了286篇论文，其中282篇由中国科学院计算研究所文高教授撰写，其余三篇文献均由“文高”撰写。

我们基于分层聚类算法来排除了基线方法。该方法类似于[Tan et al。 2006]，除了[Tan et al。 2006]也利用搜索引擎来帮助消歧。我们还将我们的方法与[Yin et al。2007年]。在所有的实验中，我们假设手动提供人数k。

### 7.2.2 Results

我们评估了我们的方法的性能和两个数据集的基线方法。表八显示结果。可以看出，我们的方法胜过名义消歧的基线方法（根据平均F1分数，真实姓名数据集中+4：54％，实名数据集+10：75％）。基线方法有两个缺点：1）它不能利用论文之间的关系，2）它依赖于固定的距离度量。我们的框架受益于分配结果之间的依赖关系建模的能力。

我们将我们的方法与[Yin等人2007]中提出的DISTINCT方法进行了比较。我们使用了[Yin et al。2007]和我们的实验比较中使用的人名。图10显示了比较结果。可以看出，对于一些名称（例如，“惠芳”和“Rakesh Kumar”），DISTINCT和拟议的方法均实现了高性能。这是因为这些名称的文件明确分开。（关于更详细的分布分析，请参见第7.2.4节）。在某些名称上，我们的方法胜过DISTINCT（例如，\ Michael Wagner“）;在其他名称上，我们的方法落后于DISTINCT（例如\ Bin Yu”）。这是因为DISTINCT通过从自动构建的训练集的监督学习确定每个链接的重量（例如共同作者关系）。权重对应于等式（14）差异在于我们的方法以无监督的方式学习权重。所学习的权重对具有不同分布的数据集具有不同的偏好。我们手动调整重量，发现性能相应变化。目前在我们的方法中的参数混合可能导致平均接近最佳的性能。

# 9.2 Name Disambiguation Relate work

已经提出了许多方法来指定不同领域的歧义。例如，[Bekkerman和McCallum 2005]试图将网页与具有相同名称的不同个体区分开来。他们提出了两个无监督的框架来解决这个问题：一个是基于网页的链接结构，另一个是使用聚集/集合聚类方法。该方法基于无监督聚类，无法描述数据点之间的关系。

还有许多作品着重于论文数据上的名称消歧。例如，Han等人提出了使用K-way谱聚类方法的无监督学习方法[Han et al 2005]。他们为每个名称数据集计算一个Gram矩阵，并将K方法谱聚类算法应用于Gram矩阵得到结果。On和Lee [On和Lee 2007]提出了一个可扩展的名称消歧问题的算法。它们适应多级图分区技术，以解决大规模名称消歧问题。他们的算法在效率方面可以有很大的提高。 Bhattacharya和Getoor [Bhattacharya和Getoor 2007]提出了一种使用属性和关系信息来消除歧义的关系聚类算法。另见[Tan et al.2006]。这种类型的方法通常在其聚类算法中使用参数变换距离度量，而在消歧时可以学习我们的距离度量的参数。

[Han et al。 2004]基于朴素贝叶斯和支持向量机。对于给定的作者姓名，该方法从列车数据中学习一个具体的模型，并使用该模型来预测作者是否创建了一个新的论文。但是，该方法是用户依赖的。为大型数字图书馆中的所有个人培训数千种模型是不切实际的。与监督方法相比，我们的方法更具可扩展性。另一种类型的相关工作是半监督聚类，例如 [Basu et al 2004] [Cohn et al 2003] [Zhang et al 2007a]。[Basu et al 2004]提出了一种基于隐马尔可夫随机场的半监督聚类概率模型。他们的模型结合了基于约束和基于距离的方法。

# 10. CONCLUSION AND FUTURE WORK

在本文中，我们调查了Web用户分析的问题。我们将分解问题正式化为几个子任务。我们提出了组合方式来应对这些问题。 具体来说，我们提出了一种树结构条件随机场（TCRF），从网页中提取出一个信息，并提出了一个概率模型来解决名称模糊问题，用于整合来自不同来源的信息。此外，我们提出了一个主题模型来发现用户兴趣。实验结果表明，我们提出的方法优于基线方法。专家评论的实验也表明，提取的用户界面可以用于提高专家发布的准确性。 我们已经基于提出的方法开发了一个示范系统。用户反馈和系统日志显示系统的用户认为系统是有用的。

这项工作有几个潜在的改进。首先，一般的网页可能包含很多噪音，如何从嘈杂的数据中提取准确的信息是一个具有挑战性的问题。其次，通过结合其他关系或人类背景知识，可以进一步提高名称消歧的表现。第三，所提出的用户兴趣发现方法是一种无监督的方法，不考虑任何域名知识。实际上，对于特定领域（例如，计算机科学），人们可能已经建立了一些分类法（例如，ACM类别）来描述域中的子字段，其可以用于指导用户兴趣的发现。

这项工作还有许多其他未来的发展方向。调查如何提取基于部分标签数据的专业术语将是有趣的。机器学习的数据标签通常是冗长乏味的。如何减少标签工作是一个具有挑战性的问题。调查动态问题也很有趣。一个研究人员的研究人员可能会在多年之后发生变化，例如转移到一家新公司。此外，深入分析用户界面也很重要。