# 1.简介

Web用户分析（也称为用户简档提取和用户简档挖掘）长期以来一直被视为Web挖掘和自然语言处理中的一个重要且具有挑战性的问题。相关研究[1]，[2]，[3]，[4]，[5]，[6]可追溯到30年前。 Web用户分析的一般任务是从非结构化Web提取“基于语义”的用户简档属性（例如联系信息，教育史，经验和传记）。Web用户分析可以应用于许多应用，并且在大多数社会相关系统中变得必要。通过大型和高质量的个人资料数据库，我们可以轻松识别我们应该向特定用户推荐哪种信息。在电子商务中，还可以利用配置文件信息来定位新产品的目标客户。还有几个产品，如Email Breaker1，Email Hunter2和Sidekick3，提供服务，帮助用户从Web上查找目标人员的电子邮件地址

尽管在这一领域进行了许多研究，使自动化侧写提取过程，但问题仍然很大程度上尚未解决[7]，[8]，[5]。要为特定用户生成配置文件，通常的方法是首先从Web查找此用户的相关网页，然后使用机器学习模型（例如，CRF）从页面中提取配置文件属性。两个阶段的最先进精度达到了90.0％左右.例如，Tang等人的主页查找任务的F1得分为92％。 [9]，87％用于从主页提取配置文件属性[5]。从独立观点来看，表现听起来不错。然而，考虑到误差传播，组合两个阶段的方法的总体精度降低到80％。更严重的是，随着Web上数据量的快速增长，问题变得更加具有挑战性，因为大型Web数据包含更多的嘈杂和冗余数据.

在本文中，我们尝试从大数据的角度重新审视这个问题。 具体来说，为了避免错误传播，我们提出了一个统一的方法框架，以一步处理所有的提取子任务。而且，除了找出减少噪声（或冗余）的新方法之外，我们开发了一个简单但非常有效的方法来提取用户配置文件属性，使用大型Web数据中的冗余信息。为了整合冗余信息以提高提取精度，我们提出了一种马尔科夫逻辑因子图（MagicFG）模型，将人类知识形式化为模型中的一级逻辑。

我们将提出的模型应用于学术搜索和挖掘系统AMiner.org4，以从Web上提取研究人员的资料。图1显示了AMiner.org中研究员简介的一个例子。简介包含韩家伟博士的基本信息，如联系，职位，图片，电子邮件和主页。随着提出的模式，我们已经成功地提取了超过10万的研究人员资料。我们还对地面实数数据集上的拟议模型进行了定量评估。 所提出的方法通过几种替代方法获得显着改善（F1分数为+ 4-6％; p = 0:01，t检验）。

# 2.方法框架

在本节中，我们首先给出提出的框架的基本思想来解决分析问题，然后介绍从Web中提取概要文件属性的三种方法。

## 2.1基本理念

给定一个人v，被称为查询人，我们的目标是提取个人的个人资料属性，并构建研究员资料。例如，在学术搜索系统中，研究人员简介包括位置，图片，地址，电话，电子邮件，主页，研究兴趣等。详细的定义可以在[5]中找到。我们的目标是设计一种从网络中高精度地自动提取配置文件属性的通用方法。该方法还应具有足够的灵活性以扩展以处理新的配置文件属性。

传统方法通常首先从Web查找查询人员的相关网页，然后使用SVM或CRF等模型从页面中提取所需的配置文件属性来处理此问题。在这两个步骤中，通过传统方法实现的最先进的表现约为90％[5]，[9]。然而，由于两个步骤之间的误差传播，通过组合两个步骤的总体精度不可避免地下降到80％。同时，所需的配置文件属性可能分布在不同的Web页面中，这将导致两个新的挑战：1）从分布式页面提取，2）提取数据冗余。

为了解决错误传播和数据冗余问题，我们提出了一个统一的框架，从大型Web数据处理所有提取子任务。这种方法很简单但非常有效。具体来说，对于每个配置文件属性，我们首先构造一个“智能”查询，并使用搜索引擎来检索与查询相关的片段，最后将提取模型应用于返回的片段来提取配置文件属性。背后的想法是利用数据冗余来帮助提取。假设我们要提取“菲利普·苏”的隶属关系。 构造的查询可以是“Philip S. Yu隶属关系”。以类似的方式，为了提取“菲利普·余玉”的电子邮件地址，我们可以构建“菲利普·俞·电子邮件”。图3显示了Google使用两个构造的查询返回的两个示例代码段。我们可以看到，从3（a）我们可以很容易地识别两个不同的隶属关系：“芝加哥伊利诺伊大学”和“IBM TJ沃森研究中心”（归一化后），从3（b）我们还可以识别两个电子邮件地址： “psyu@cs.uic.edu”和“hanj [at] cs.uiuc.edu”。我们将所识别的关联/电子邮件从片段称为候选人/电子邮件。现在的问题是如何对所识别的信息进行排名。我们的基本思想是利用冗余信息，例如，“芝加哥伊利诺伊大学”在代码片段中出现四次，“IBM T. J. Watson研究中心”两次。更准确地说，我们提出了一个MAkov loGIC因子图（MagicFG），通过利用冗余信息对获得的候选者进行排名。该模型是灵活的，可以轻松地纳入任何领域的人类知识，以进一步提高提取精度。

值得注意的是，我们仅限于两个示例配置文件属性。实际上，提出的方法一般是灵活的。要提取新的配置文件属性，我们需要做的是构建“智能”查询并训练MagicFG模型。对于某些配置文件属性，可以轻松构建查询。例如，我们发现，对于电子邮件，我们可以通过使用姓名+电子邮件来实现高精度。对于某些其他配置文件属性，例如“性别和职位”，情况可能会更加复杂。我们将介绍如何在第II-B节中构建一般配置文件属性的智能查询。还请注意，通常有两种类型的配置文件属性：分类属性和非分类属性。例如，“性别”是一个分类属性。职位也是具有多重价值的分类属性，如教授，学生，研究员和工程师。而电子邮件和年龄是两个非分类属性。在提出的框架中，两种属性将略有不同。

## 2.2智能查询构造

我们以不同的方式构建了分类和非分类属性的查询。对于分类属性，我们通过自动识别每个候选类别中的代表性关键字构建查询，并将它们组合在一起作为查询。要找到每个类别的代表性关键字，我们首先从AMiner和LinkedIn等专业网站收集每个类别的几个人名（例如1000）。然后，我们将相应的人名作为查询提交给像Google这样的搜索引擎，以获取top-k（例如10个）片段。在片段中的所有单词中，我们将最具代表性的关键字识别为具有最高TF-IDF分数的关键词[11]。类别c中的词w的TF-IDF得分如下计算。

## 2.3基线提取模型

我们首先引入两个基线模型来提取侧写属性。

1）基于规则的模型：在基于规则的模型中，为了提取查询人v的电子邮件，我们简单地通过组合人名和单词“电子邮件”来构造查询。一旦从搜索引擎（例如Google）获取了返回的代码段，我们可以使用基于规则的启发式方法来提取查询人的电子邮件。规则定义如下：我们首先从搜索到的片段中提取候选电子邮件地址，如果我们发现查询人姓名的名字或姓氏包含在候选电子邮件地址的前缀中，则将提取提取的电子邮件作为结果。值得注意的一件事是识别电子邮件地址，因为在许多网页，特别是某些人的主页中，电子邮件地址可能会以特殊模式编码，例如“firstname [dot] lastname [at] cmu [dot] edu”。 定义启发式规则以识别潜在的电子邮件地址。

我们的初步实验表明，这么简单方法可以达到88％的准确度 - 相当于获得的最先进的表现传统的两步法（详见第三节比较）。

2）基于分类的模型：我们利用Logistic回归（LR）作为分类模型。让我们先考虑一个二类分类问题。令f（x1; y1）; （xN，yN）g是训练数据集，其中xi表示候选信息的特征向量，yi 2 f-1; + 1g表示分类标签（候选者是否正确）。基于分类的提取模型由学习和提取两个阶段组成。在学习中，尝试找到最佳权重配置以最大化观察到的实例的对数似然函数）。在提取中，我们使用学习模型来分类我们要提取哪些候选信息。

关于分类模型中的特征，我们使用与我们提出的模型中定义的属性因子相同的属性特征（详见第II-C节）。分类可以调整不同特征的权重，将特征组合在一起，从而获得比基于规则的方法更好的表现（比F1分数为90％）。然而，如图3所示，返回的片段通常包含可能对提取有帮助的冗余信息。基于规则和基于分类的模型都将每个候选实例视为独立的，因此不能利用这种冗余信息。

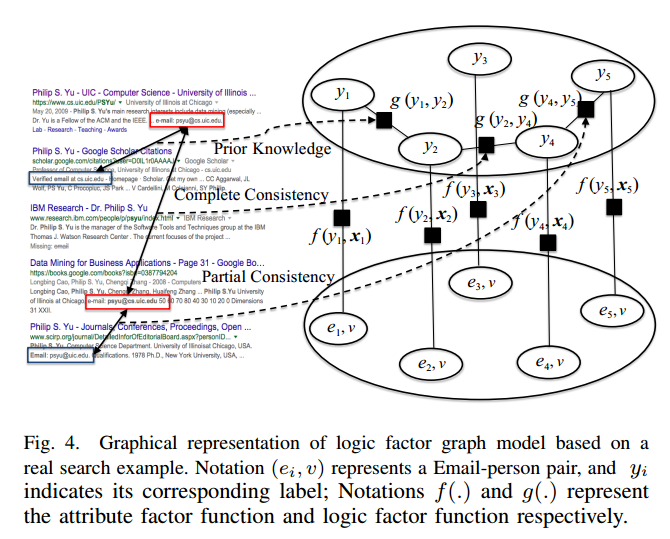
## 2.4马可夫逻辑因子图（MagicFG）模型

基于规则的方法与传统方法相当，基于分类的方法胜过许多现有方法。上述两种方法都将候选电子邮件地址视为独立的分类对象，而基于规则和基于分类的方法都忽略了在返回的片段中标识的候选实例之间的相关性。但是，可以利用这些冗余信息。 实际上，驻留在代码段中的冗余信息可以有助于提高提取精度。例如，在图3（b）中，对于电子邮件提取，两个候选电子邮件地址“psyu@cs.uic.edu”和“psyu@uic.edu”之前的“@”之前的相同前缀“psyu”表示 两个电子邮件地址属于同一个人。

现在，问题变成如何将这种基于冗余的相关性编码并结合到统一的提取模型中。我们提出了一种新颖的马尔可夫逻辑因子图（MagicFG）模型，将相关性建模为一级逻辑，并利用逻辑来提高提取性能。我们分别介绍如何为非分类和分类属性建立数据冗余。

### 建模非分类属性

当提取非分类属性时，对于每个查询人员，我们构建一个因子图模型，每个节点表示候选实例，每个边对应于两个候选实例之间的依赖关系。我们为所有查询者同时优化因子图模型。我们以电子邮件为例说明分类属性的建模过程。对于查询人v，我们将每个候选电子邮件表示为ei。如图4所示，我们可以提取四个候选人e1;E2;E3;e4。对于每个候选人电子邮件，我们创建一个实例（ei; v）并与潜在变量yi相关联。为了模拟候选实例之间的相关性，我们可以构建如图4所示的因子图模型。该模型称为马尔科夫逻辑因子图（MagicFG）模型。在MagicFG中，相关性被表示为一阶逻辑，因为它将先验知识和所有其他相关性定义为一阶逻辑。我们将解释我们之后如何定义第一个基于逻辑的相关关系。在高级别，在MagicFG中，我们定义了两种类型的函数。



属性因子函数：它捕获电子邮件人对的特征，并将其定义为指数其中φk（:)是v之间定义的第k个特征函数和ei相对于i的价值; αk是重量相应的属性特征; xi是第i个特征向量。 Za是归一化因子。

逻辑因子函数：它捕获潜变量之间的相关性。 它也被定义为指数函数：其中m（:)是根据一阶逻辑知识库在yi和yj之间定义的第m个相关因子函数; βm是相应相关系数的权重。

对于属性因子函数，我们可以定义多个特征函数fφk（yi; xi）gk来表征每个候选实例。为了提取电子邮件，我们定义了一些功能，例如v的名字，姓氏或全名是否包含在ei的前缀中人v与提取候选ei的上下文ci之间定义了另一种特征。例如，v的名字，姓氏或全名是否包含在上下文ci中，以及v的隶属关系是否包含在上下文ci中。在这里，**我们使用affiliation信息来消除同名的人。**

关于逻辑因子函数，我们主要考虑潜在变量之间的三种一阶逻辑关系：完全一致性，部分一致性和先验知识。一阶逻辑是将数学形式化为公理的标准，并在数学基础上进行了研究。在我们的问题中，我们使用一阶逻辑来对候选实例之间的用户特定关联进行编码，并提取关于提取的领域人类知识。有关一级逻辑的一般介绍，请参考[12]。

### 建模分类属性

在处理分类属性时，对于所有查询者，我们构建一个因子图，每个节点表示一个查询者，边缘表示两个查询者之间的依赖关系。我们以性别为例来说明分类属性的建模过程。与非分类属性不同，每个人只能有一个性别，男性或女性。因此，在此任务中，我们直接为每个查询人员分配一个标签。我们通过梳理每个性别的人名和代表性关键字来构建查询，如前所述，“男”为女，“女”为“她”。查询终于看起来像“他的名字”。然后我们根据返回的片段来制定MagicFG。

MagicFG模型的配置与非分类属性的配置也有一点不同。我们将模型与每个观察变量一起作为一个人vi。对每个人的相应潜变量yi表示vi的属性值，例如，vi是男还是女。对于属性因子函数，我们首先从他/她的搜索上下文中提取每个人的特征。例如，搜索结果中的片段是否包含人名和单词“他/她”，片段是否包含关联和“他/她”一词，无论“他/她”是否出现在代码段中 的前3个返回的搜索结果，以及所有搜索结果中的数字“他/她”。对于逻辑因子函数，我们定义完整一致性逻辑类型的相关特征如下.

这个逻辑表明，如果两个人的姓氏相同，那么两个人的性别更有可能是一样的。总而言之，为非分类和分类属性构建的因子图略有不同。对于非分类属性，构建多个图，其中每个图形为每个人构建，每个候选属性被形成为一个节点，并且两个候选属性之间的依赖关系被形成为边缘。 而对于分类属性，仅构建一个图，每个人被形成为一个节点，两个人之间的依赖关系被形成为边缘

# 3.实验

在本节中，我们展示了我们的方法对分类和非分类属性的有效性。对于定量评估，我们以性别为例，将分类属性和电子邮件作为非分类属性的一个例子。请注意，我们的框架非常灵活，已经应用于在线学术搜索和挖掘系统AMiner.org，以提取研究人员的资料。本工作中使用的所有数据集和代码都是公开的。

## A．实验设置

### 数据集

为了构建定量评估的地面实况数据集，我们从AMiner.org [9]随机选择了2000名研究人员。具体来说，为了提取每个研究人员的电子邮件，我们通过查询人名和“电子邮件”这个词搜索引擎。这样可以得到4,528个电子邮件候选人。应用人体注释来识别正确的电子邮件地址。以类似的方式推断性别，我们通过查询人名和单词“他”或“她”来搜索网页。还进行了人类注释，以确定这2,000名研究人员的性别。对于注释中的分歧，我们进行了“多数投票[Majority Vote Algorithm](http://blog.csdn.net/feliciafay/article/details/18876123)”。最后，对于2000名研究人员，我们确定了34％的研究人员是女性研究人员; 约40％的电子邮件候选人是正确的电子邮件，这意味着我们的框架可以为超过90％的用户找到正确的电子邮件。

### 评估指标

为了定量评估我们的模型，我们将数据集划分为训练集和测试集。我们执行五次交叉验证，并在精度，召回和F1分数方面报告提取性能

### 比较方法

我们将MagicFG与以下用于在真实数据集上提取Email和Gender的方法进行比较。

• 规则。 使用几个简单的定义规则来提取配置文件属性。例如，为了提取性别，我们计算女孩和男孩的常用名称数量。为了提取电子邮件，我们发现电子邮件的前缀是否包含人名。

•支持向量机（SVM）。 使用与特征相同的属性因子，并采用SVM-Light [15]来训练和预测电子邮件和性别。

•随机森林（RF）。 使用与特征相同的属性因子，并采用sklearn软件包进行训练和预测。

•逻辑回归（LR）。 使用与特征相同的属性因子，并采用sklearn软件包进行训练和预测

## B．提取性能

电子邮件提取 在与我们提出的框架相同的情况下，MagicFG模型优于最佳提取方法，即射频（F1 = 2.12％）。这是因为除了独立的属性功能之外，MagicFG通过将不同的电子邮件候选者之间的依赖关系作为一级逻辑的特征来捕获。

性别推断 在我们提出的相同框架下，MagicFG模型优于最佳方法LR（F1分为2.21％）。因为MagicFG还包含完整一致逻辑的一个逻辑关系，并捕获不同人选之间的依赖关系。

因素的影响 我们进一步深入分析了不同的逻辑相关因素如何影响用户分析的性能。图5显示了通过考虑不同级别的逻辑因素，提出的MagicFG的不同评估指标。从图5（a）可以看出，对于电子邮件提取，精度性能明显下降，没有逻辑关系。另外，添加先验知识逻辑的因素可以进一步提高性能。图5（b）也表明完全一致性逻辑的因素显着提高了性别推理的表现。

## C．与现有方法比较

我们现在将我们的方法与电子邮件提取和性别推断任务的几种最先进的方法进行比较：

TCRF对于电子邮件提取，我们使用[5]中提出的方法作为基准（以下称为：TCRF），这是从Web提取主页和电子邮件的最新方法之一。该方法有两个步骤，首先找到用户的主页，然后使用名为TCRF的提取模型以高精度从主页提取电子邮件。

Facebook生成的名称列表预测器（FGNL）

对于性别推断，我们使用[10]提出的方法作为基线（以下称为：FGNL）。用于推断性别的最先进的方法取决于男性和女性的常用名称列表。在[10]中，作者提出了一种使用Facebook数据构建扩展和高质量名单的方法。他们将用户的名字与列表进行匹配以进行推理。如果名字与其中一个男性名称相匹配，则该用户被视为男性，反之亦然。而如果姓名不是男性名字，也不是女性名字，或者姓名列表中的名字，都会随机猜测用户的性别。

表三显示了不同方法对性别推理的分类表现。我们可以看到，我们的方法在F1-分数上优于基线（FGNL）+ 6.49％。我们的方法比召回的FGNL方法好（+ 13.13％）。 这是因为FGNL方法在很大程度上取决于名称列表。但是，您无法列出所有这些名称，无论列表多大。

相反，我们的方法可以自动找到描述具有特定性别的用户的文档的代表性关键字，并从较小的Web数据推断性别限制。所以我们很少有FGNL在他们的列表中找不到名字时必须面对的问题。如表所示，FGNL方法的精度略高于（+ 1.22％），这显然是使用名单的优点。然而，我们的方法在将召回提升到不同程度的同时，实现了一个紧密的精确。考虑到FGNL方法的局限性，我们的方法更广泛。