基于Web的实例扩展与属性值扩充方法

李 贵 陈韶刚 韩子扬 李征宇 孙 平 孙焕良
(沈阳建筑大学信息与控制工程学院 沈阳 110168)

摘 要 实例扩展与属性值扩充是Web抽取与集成领域中的一个重要研究课题,将Web数据列表和实例建模成二分图,根据扩展实例的质量分数,对扩展集合进行迭代更新直到扩展集合的质量分数最大,且扩展集合不再更新来实现实例的扩展。同时,为了完善扩展实例的属性信息,对结构化数值属性或离散属性进行抽取,提出了基于整数线性规划的属性值扩充方法。实验表明,与以前的方法相比,本方法能更好地处理含有噪声数据的Web网页,并提高了抽取的准确率和召回率。

关键词 实例扩展,属性值扩充,整数线性规划
中图法分类号 TP391 **文献标识码** A

Entities Expansion and Attribute Values Discovery Method Based on Web

LI Gui CHEN Shao-gang HAN Zi-yang LI Zheng-yu SUN Ping SUN Huan-liang
(Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract Entities expansion and attribute values discovery has been an important research topic in the field of Web data extraction and integration. In this paper the Web table and domain entity were modeled as bipartite graph. Based on quality score, the expansion entity set will be update iteratively until the expansion entity set's quality score reaches a local maximum and the expansion entity set will not update. To collect structured numerical or discrete attributes of the entities, we presented a method based on ILP to complete the attribute values discovery of the entities. Experiment results show that the proposed approach outperforms previous techniques in terms of both precision and recall.

Keywords Entity expansion, Attribute values filling, Integer linear program

1 引言

随着Web数据的不断增加,访问Web数据已成为获取信息的重要手段。网页上的信息往往包含了现实生活中的某些实例,例如一部电影、一本书籍、一件商品等,它们由若干属性来描述(例如上映时间、作者名、价格等)。用户在访问Web数据时,对页面中包含的实例信息更感兴趣。因此,能够从页面中自动地获取有价值的实例数据信息,将大大减少用户进行筛选、比较的负担。在Web数据抽取领域,多年来一些研究学者致力于研究如何获得同类实例信息,即如何解决实例集合扩展问题。

实例集合扩展(简称"实例扩展")指的是通过发掘其他属于相同概念集合的实例,把种子实例集合扩展成为一个更完整的实例集合。其中,相同概念集合是指某一应用领域中语义相同、属性值不同的实例的集合,种子是指集合中的一些实例,即相同概念集合的某一子集。

传统的用于实例扩展的随机游走算法的基本思想是[2]:在图结构中,候选实例越接近给定种子越有可能和种子属于相同的概念集合。在处理含有大量噪声数据的Web数据源时,该算法效果并不理想。本文使用普通的Web数据作为数据源,不采用根据特定种子抽取的在线数据,提出了一种更简单有效的方法来度量实例扩展集合的质量。直观上一个较好的实例扩展集合要满足两个基本的条件:

(1)产生的实例集合必须与种子集合语义相近;

(2)产生的实例集合在概念上是一致的。

本文研究列表数据与领域实例之间的关系,将其关系建模为二分图,并在二分图模型之上给出实例间相似性的算法、候选实例质量的评估标准以及实例扩展算法。

在实例扩展时,为了完善实例的属性信息,需要解决以下问题:给定一个实例的类别和它的模式,以及一个领域中含有该实例的网页集合,如何根据给定的模式抽取属性值,尤其是当属性是非结构化数值或离散的形式时。

以往的算法都将该问题看作单个文档信息抽取问题,并且没有针对数值属性进行特殊处理。本文的属性值扩充方法综合了实例相关的所有网页上下文,结合了领域中真实值的约束信息,采用一个整数线性规划(ILP)将收集的信息整合,用以对所有的属性进行统一优化分配。

2 相关工作

目前,解决实例扩展问题的方法大体可分为基于模板、基

本文受国家自然科学基金(61070024),辽宁省自然科学基金(2014020068)资助。

李 贵(1964-),男,博士,教授,主要研究方向为Web数据挖掘与信息集成、分布对象技术、软件工程,E-mail:Ligui21c@sina.com;陈韶刚(1990-),男,硕士生,主要研究方向为Web数据挖掘与集成;韩子扬(1979-),男,讲师,主要研究方向为Web数据挖掘与信息集成;李征宇(1980-),男,讲师,主要研究方向为Web数据挖掘与信息集成、分布对象技术;孙 平(1980-),女,硕士,副教授,主要研究方向为Web数据挖掘与信息集成、推荐系统;孙焕良(1969-),男,博士,教授,主要研究方向为数据仓库与数据挖掘。

于分布以及基于融合等3大类。

基于模板的方法，代表性工作包括文献[2-4]等。该类方法的核心思想是通过某种方式得到模板，并利用模板抽取候选实例，最后对其进行评分排序得到结果。这里的模板可以是预先定义的文本包装器，也可以是种子在语料中出现的高频上下文。

基于分布的方法，代表性工作包括文献[5-7]等。该类方法的核心思想是统计语料库中每个词项的上下文分布并构造词项分布矩阵，利用该矩阵计算每个词项与种子的相似度，以此作为评分和排序的标准。例如，文献[5]提出的Know-ItAll系统能自动地抽取文本中的候选实例。

基于融合的方法，文献[8]提出的方法使用多种类型的数据(比如普通网页文本、网页表格、查询日志等)，对不同类型的数据采用不同处理方法(基于模板或基于分布)，并对各自的结果进行融合。这种方法可以降低单一方法产生的错误对总体结果的影响。

实例扩展中较突出的成果是SEAL[2]系统。在第一次提取阶段，为每一个网页进行了个性化包装，尽可能用上下文涵盖所有给定种子，这些种子轮流应用于构造的网页上来提取候选项。在第二个排序阶段，将网页、包装器、候选项模拟成图中的点，根据他们与图中种子的相似度，用随机游走技术给候选项排序。

目前，属性值扩充方法可以分为基于规则的方法和基于学习的方法。

采用包装器技术抽取Web页面信息时，包装器可以看作一个处理器，定义为一组指令集，通过执行这些指令抽取Web页面中感兴趣的数据，将隐藏在Web页面中的信息转换成结构化数据。MDR算法抽取Web页面中的半结构化数据，在忽略数据项语义的情况下，利用页面标签结构信息完成数据的抽取。

文献[12]进行了基于半监督的方法抽取产品描述中的属性/值对的算法研究。由于任意给定一个属性，该算法需要花费额外的步骤去解决值的奇异性问题，因此该算法在实际应用中的效果并不理想。文献[13]进行了属性抽取(不抽取值)的相关工作研究。文献[14]提出了领域比较学习的算法。

以往这些数据获取模型都有某些局限性。手工方法不足以解决大规模的实例抽取问题，同时也不能及时发掘新出现的实例。基于包装器的技术是一种相对有效的解决方法，然而并不能自适应抽取，且只能应用到那些结构化的网站，因此抽取性能不稳定，需要持续维护。

3 实例扩展

本文提出的实例扩展方法是基于二分图的实例扩展模型，定义实例间相似性计算方法，然后通过定义质量评估方法，选出最优候选实例。

3.1 实例扩展模型

本文以列表页中的数据作为研究对象。列表页中的数据通常是从后台数据库中获取并根据固定的模板展现在网上的，列表页中通常都含有若干实例对象，如图1所示。

序号	行政区	公示标题
1.	武胜县	武胜县公共资源交易中心国有土地使用权招拍挂出让成交公示
2.	宝山区	宝山区规划和土地管理局国有土地划拨用地批前公示
3.	宝山区	宝山区规划和土地管理局国有土地划拨用地批前公示

图1 网页列表数据

结合列表数据的特点，将列表数据和相关的领域实例建模成二分图，如图2所示。

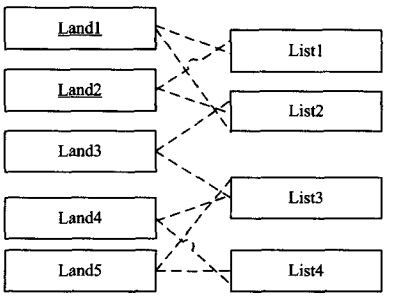


图2 网页列表的二分图数据模型

在抽取过程中，每个列表都被建模成一个图中右边的节点，每个出现在这些网页列表中的领域实例都被建模成左边的节点。图2中，左边这些带下划线的点“Land1”和“Land2”是种子实例，余下的“Land3”，“Land4”和“Land5”是可能的候选实例。如果某个实例属于某列表，就有一条边连接这一个实例和这个列表。例如列表“List2”与“Land1”、“Land2”和“Land3”相连，表明这3个实例都是“List2”的成员，“List2”很可能是列举相关土地信息的一个列表页。在实验中为了简单起见，将二分图模型中每条边的权重都置为1。

3.2 实例间相似性的计算

根据给定的种子实例找出相似实例的问题，可以看成二分图数据模型中以右边的节点作为特征去发现与种子节点相似的实例节点的问题。为计算实例节点间的相似性，本文使用Jaccard相似性[9]计算方法和余弦相似性[9]计算方法。

定义1 令$n1$，$n2$为二分图模型中左边的两个领域实例节点，同时S_{n1}和S_{n2}是二分图模型中连接节点$n1$和$n2$的网页列表节点的两个集合，那么$n1$和$n2$的Jaccard相似性表示为$Sim_{Jac}(n1,n2)$：

$$Sim_{Jac}(n1,n2)=\frac{|S_{n1}\cap S_{n2}|}{|S_{n1}\cup S_{n2}|} \tag{1}$$

定义2 令$n1$，$n2$为二分图模型中左边的两个领域实例节点，用权重向量V_{n1}、V_{n2}表示二分图模型中连接网页列表节点与实例节点$n1$和$n2$的边的权重，那么$n1$和$n2$的余弦相似性可以表示成$Sim_{Cos}(n1,n2)$：

$$Sim_{Cos}(n1,n2)=\frac{V_{n1}\cdot V_{n2}}{\|V_{n1}\|\ \|V_{n2}\|} \tag{2}$$

下面用两个例子来简单地解释Jaccard相似性和余弦相似性。首先解释二分图模型中两个实例节点的Jaccard相似性计算。

例1 在图2中，与实例节点“Land1”连接的列表节点L(“Land1”)={“List1”，“List2”}；与实例节点“Land3”连接的列表节点L(“Land 3”)={“List2”，“List3”}。通过定义2可知，种子实例节点“Land1”和实例节点“Land3”的Jaccard相似性为0.33。同理，种子实例节点“Land2”和实例节点“Land3”的Jaccard相似性也是0.33。

实例节点“Land4”与两个种子实例节点“Land1”、“Land2”的Jaccard相似性是0。因此，在采用Jaccard进行相似性评估时，实例“Land3”与种子集合的相似性比实例“Land4”更高。

接下来举例说明如何计算实例节点间的余弦相似性。

例2 同样在图2中，实例节点“Land1”连接到

L("Land1")={"List1","List2"},那么它的边的权重向量V("Land1")=(1,1,0,0)。同理,"Land3"的边权重向量V("Land2")=(0,1,1,0)。根据定义2可知,种子实例节点"Land1"和实例节点"Land3"的余弦相似性为0.5。同理,种子实例节点"Land2"和实例节点"Land3"的余弦相似性也是0.5。

因为没有共同的网页列表节点,"Land4"、"Land5"与两个种子实例节点"Land1"和"Land2"的余弦相似性是0,再次证明实例"Land3"与种子实例的相似性比"Land4"更高。

3.3 质量评估

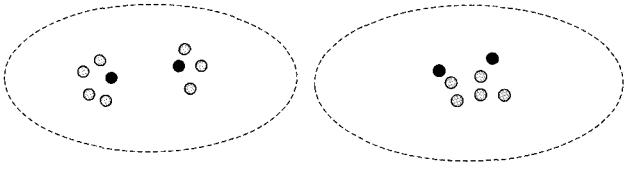
基于随机游走的算法基于单个实例与给定的种子的相似性来评估实例的质量。将扩展的实例集看成一个整体,并提出一个直观的计算方法来评估扩展实例的质量。

实例扩展的任务就是发现与种子集合在相同概念集合中的其他实例,所以扩展实例与种子实例相似性越大,扩展实例的质量越好。下面给出计算扩展实例与种子实例之间的相似性定义。

定义3 令E为全部实例集合,$X(X\subseteq E)$为扩展集合,$S(S\subseteq E)$为种子集合,$Sim:E\times E\to[0,1]$为评估两个实例的相似性的矩阵,那么X与S的相似性被定义为$Srel(X,S)$:

$$Srel(X,S)=\frac{1}{|X|\cdot|S|}*\sum_{x\in X}\sum_{s\in S}Sim(x,s) \tag{3}$$

为方便理解,下面使用图形来说明扩展集合的质量。在图3(a)和图3(b)中,两个在中间深色的点代表给定的种子集合S,这两个点周围的是扩展实例集合X,图中任意两点的距离表示这两个实例间的相似性。



(a)一个相似性较高的集合　　(b)一个一致性较高的集合

图3

很显然,两个图中椭圆包围的实例扩展集合和两个给定的种子是很相似的。使用定义3的相似性来度量,两个图中的集合与给定种子都有很高的相似性。

然而,相似性定义不能充分展示扩展集合的质量。集合扩展的目的是发现与给定种子相似的、一致的"概念集合"。有些情况下,虽然扩展实例和种子相似,但它们不是一个一致的"概念集合"。

例如,在图3(a)中,尽管用小圆表示的扩展实例与种子实例很接近,但是它们在空间上分散,不能形成一致的概念集合,而图3(b)中的扩展实例不仅和种子实例很接近,它们彼此之间的距离也很小,可以形成一致的"概念集合"。因此,图3(b)中的扩展实例显然比图3(a)中的更适合作候选实例。

为了表示图3(b)中扩展实例之间的距离越小,一致性越好,整个集合的质量也越好的情况,下面给出一致性概念的定义。

定义4 令E为实例集合,X为扩展集合$(X\subseteq E)$,$Sim:E\times E\to[0,1]$为评估两个实例的相似性的矩阵,那么X的一致性表示为:

$$S_{coh}(X)=\frac{1}{|X|\cdot|X|}*\sum_{i=1}^{|X|}\sum_{j>i}^{|X|}Sim(r_i,r_j) \tag{4}$$

其中,$r_i,r_j\in X$。

由于扩展集合的质量需要用相似性和一致性一起表示,这里将扩展集合的质量用相似性和一致性的加权和表示。

定义5 令E为实例总集,X为扩展集合$(X\subseteq E)$,S为种子集合$(S\subseteq E)$,$Sim:E\times E\to[0,1]$为计算任意两个实例的相似性的矩阵,$\alpha(0<\alpha<1)$是权重因子常量,那么扩展集合X的质量分数表示为:

$$Q(X,S)=\alpha*S_{rel}(X,S)+(1-\alpha)*S_{coh}(X) \tag{5}$$

其中,α作为参数调节相似性和一致性的权值,本文取$\alpha=0.5$。

3.4 实例扩展算法

根据质量计算方法的定义,实例扩展的问题可以被重新描述成如下的一个求解过程:已知全部候选实例E和种子实例$S(S\subseteq E)$,令$Sim:E\times E\to[0,1]$为评估两个实例相似性的矩阵,扩展集合X($X\subseteq E$,X的大小为K),求目标函数$Q(X,S)$的最大值。

扩展集合(以下简称ESS)由候选集合中质量较高的实例组成。若质量分数很高,该ESS就是"好"的。一旦推导出一个好的ESS(用X表示),就可以在X和种子集合S的基础上,用函数$g(t,X,S)$对每个实例t进行排序,其中$r_i\in X,s_i\in S$。

$$g(t,X,S)=\frac{a}{|S|}\sum_{i=1}^{|S|}Sim(t,s_i)+\frac{(1-a)}{|X|}\sum_{i=1}^{|X|}Sim(t,r_i) \tag{6}$$

寻找大小为K且质量分数最大的最优X是NP-hard问题[7]。为求ESS的最优值,本节使用了两个贪心算法:静态阈值算法和动态阈值算法,通过迭代地计算候选的ESS(即X),来获得最大的$Q(X,S)$函数值。静态阈值算法首先固定ESS、X的大小,然后迭代更新X中的实例,使得$Q(X,S)$最大;而动态阈值算法每次迭代都会修改X的大小和内容。

3.4.1 静态阈值算法

静态阈值算法首先给ESS设定一个初始值,然后通过替换上一代ESS中的一个实例来迭代地提高质量分数。直到ESS的计算收敛,并达到质量分数的局部最优值。算法伪代码如图4所示。

```
static_Thresholding (seeds,graph)
for each entity in graph. entitys do
Rel_Score[i] ← Srel(entity_i,seeds)
end for
sort entity_i by Rel_Score[i] desc
K← Pick_Threshold(Rel_Score[i])
X_0← the top K ranked terms by Rel_Score[i]
Iter← 1
while true do
    for each entity_i in graph. entitys do
        Sim_Score[i] ← Srel(entity_i,R_iter−1)
        g(term_i) a * Rel_Score[i]+(1−a) * Sim_Score[i]
    end for
    sort term_i by g(term_i) desc
    R'_iter←the top K terms by g(term_i)
    if R'_iter R'_iter−1 then
        let r ∈ R'_iter be the top ranked term not in R_iter−1
        let q ∈ R_iter be the last ranked term in R_iter−1
        R_iter(R_iter−1 ∪{r})−{q}
    else
        R_iter R_iter−1
        break
```

```
    end if
        iter ++
end while
return R_iter
```

图 4　静态阈值算法

静态阈值算法有两个输入参数：种子集 seeds 和二分图（所有候选实例作为左边节点）。在首轮循环中，首先计算每个实例与种子集的相似性 $S_{rel}(entity, seeds)$，然后根据实例的相似性分数对其排序，并挑出序列顶部的 K 个实例初始化种子扩展集合 R_0。

while 循环中，首先根据先前的 R_{iter-1} 计算出新的候选 ESS 即 R_{iter}，然后逐步提高实例扩展集合的质量分数直到达到极大值。在每次迭代中，首先计算每个候选实例 $entity_i$ 和 R_{iter-1} 的相似性分数 $S_{rel}(entity_i, R_{iter-1})$、相应的候选实例质量排序函数 $g(entity_i)$，以及实例和 R_{iter-1} 的相似性分数的加权和，然后通过 $g(entity_i)$ 对实例排序。令序列顶部的 K 个实例序列为 R'_{iter}，如果 $R'_{iter} \neq R_{iter-1}$，用 R'_{iter} 序列顶部的且 R_{iter-1} 中不存在的实例 r 来代替 R_{iter-1} 中序列底部的实例，否则，本算法收敛并返回 R_{iter-1} 作为最终的 ESS。

3.4.2　动态阈值算法

静态阈值算法第一次迭代算出的静态阈值 K，可能没有真正反映 ESS 的实际大小。为了解决这个问题，又使用了一个动态阈值算法，在每次迭代时使用当前分数概率分布的新阈值来调整 ESS 的大小。动态阈值算法和静态阈值算法相似，区别是在每次的迭代过程中通过排序函数 $g(entity_i)$ 给所有实例排序，根据 $g(entity_i)$ 得出的新分数概率分布重新调用自动阈值函数来确定 ESS 的大小。基于新的分数概率分布重新计算阈值。算法伪代码如图 5 所示。

```
dynamic_Thresholding (seeds,graph)
for each entity in graph. entitys do
Rel_Score[i] Srel(entity_i,seeds)
end for
K_0 Pick_Threshold(Rel_Score[i])
X_0 the top K ranked terms by Rel_Score[i]
Iter 1
while iter≤Max_iter do
    for each entity_i in graph. entitys do
        Sim_Score[i] ← Srel(entity_i,R_iter−1)
        g(term_i) a * Rel_Score[i]+(1−a) −Sim_Score[i]
    end for
    sort term_i by g(term_i) desc
    R'_iter←the top K terms by g(term_i)
    if R'_iter R'_iter−1 then
        let r∈R'_iter be the top ranked term not in R_iter−1
        let q ∈ R_iter be the last ranked term in R_iter−1
        R_iter(R_iter−1 ∪ {r}) −{q}
    else
        R_iter R_iter−1
        break
    end if
    iter ++
end while
return R_iter
```

图 5　动态阈值算法

动态阈值算法适应了分数概率分布的变化，并且能够更准确地反映出 ESS 的大小。实验中显示该算法的性能略好于静态阈值算法。不过，由于动态地改变了阈值，不能像静态阈值算法那样保证收敛性，因此设置了最大迭代次数作为一个循环终止条件，来兼顾执行效率，在本实验中，设置迭代次数为 5。

4　属性值扩充

属性值扩充可以理解为对一个实例的属性及属性值进行填充，扩充结果的每一行对应一个实例，实例的属性被表示成扩充结果表的列。

考虑到现有属性值扩充方法的局限性，一个好的扩充方法应满足以下要求：

第一，属性值扩充应该是自动的；

第二，属性值扩充方法不应该对网页格式和结构有很高的要求。

扩充结果的属性值都看作结构化数据。本文的方法支持类别属性和数值属性，主要研究数值属性，为简单起见，本文假设没有多值属性。

本文的属性值扩充方法可以根据给定的领域实例去学习如何获得结构化数据。对于一个类别，给出一个模式（包含一个属性名的集合）、一个种子实例集合以及它们在模式中相应的属性值、相关网页的集合（例如，商家提供的关于产品的网页）。属性值扩充方法将会自动地去抽取该类别中其它实例的属性值。

属性值扩充方法在抽取过程中需要考虑的因素包括：1）大多数网页中实例的候选属性值是一致的，例如，如果在一个网页中有抽取错误，可以通过抽取其他网页的相同属性来调节恢复；2）利用值的分布来减少错误抽取；3）能够根据领域知识生成各种约束。

整合全局约束和局部上下文信息是一个组合最优化问题，本文采用 ILP 来解决。

4.1　属性值扩充方法概述

属性值扩充分两个阶段进行：

第一阶段——训练阶段。本阶段的任务是训练生成局部分类模型与类别约束，作为第二阶段的输入。

第二阶段——部署阶段。本阶段的任务是根据上一阶段得到的结果，与实例约束一起作为输入，通过 ILP 进行属性分配，输出最优分配结果。

以下分别介绍这两个阶段。

4.1.1　训练阶段

在训练开始前，实例集合中有一些种子实例有全部的属性值。每一个待扩充的实例都和一个上下文集合相关。

如图 6 所示，从 Web 网页中收集上下文，收集的网页首先通过片段抽取器，定位期望的属性所在的文本片段，正确分类数据包含结构化、正确的属性值。在收集的片段中找到这些属性值来自动地生成训练数据。

本文使用了一个关于一个片段和一个属性相容性的局部模型（4.2.3 节将做详细介绍）。该局部模型使用正确分类数据训练，用来关联属性与文本片段。另外，正确分类数据还可以通过密度估计生成类别约束，在部署阶段作为 ILP 的输入。

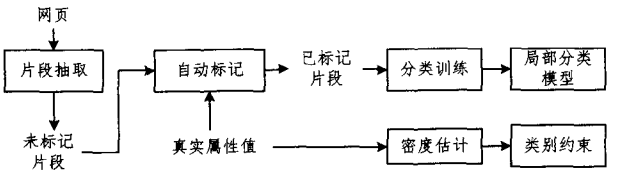


图 6　训练阶段示意图

4.1.2　部署阶段

实例集合中待完善的实例都有对应的非结构化的上下文。部署阶段(如图 7 所示)的任务便是通过挖掘上下文来填充这些属性值。

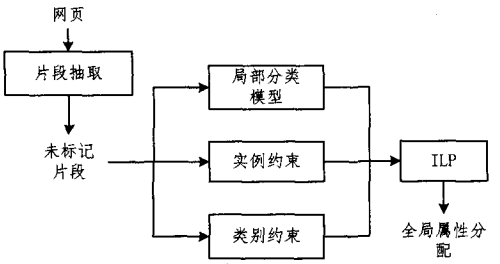


图 7　部署阶段示意图

在部署阶段,局部模型、实例约束和训练阶段生成的类别约束都被输入到一个 ILP 中,从而进行全局属性分配。

4.2　基于 ILP 的属性值扩充方法

本节详细地解释基于 ILP 的属性值扩充方法解决全局优化问题。建立 ILP 模型有 3 个基本步骤:

1. 找出问题的决策变量;

2. 找出问题中所有约束;

3. 找出目标函数。

下面介绍模型中涉及到的相关概念和术语。

上下文 c 是指一个给实例提供属性值的相关数据的文本片段。一个上下文可能跨越一个网页或者一个网页的一个区域。片段 c. s 是含有一个属性值 c. s. v 的上下文 c 中的文本片段。c. s. v 是实例某个属性相关的候选者。通常,一个上下文包含多个片段。

本文使用以下规则创建片段。一旦检测到属性值,查看该属性值左右两边的 6 个词。若存在数字或一些特殊的 HTML 标签(如〈title〉、〈li〉等),就收录所有词。如果遇到属性名集合中包含的属性名,就把它加入片段中,忽略其他的词。

实例表的模式由属性的集合组成。直观地说,一个属性就是实例表中的一个列 ID,且属性有单位。

假设用 a 来表示一个属性,用 e 表示一个实例,用 $a_1,\cdots,a_n$ 来表示其属性。属性值扩充方法的目标就是生成一个元组$\langle v_1,\cdots,v_n\rangle$,其中 v_i 和属性 a_i 相关。片段 c. s 分配给属性 a 表示片段中 c. s. v 是属性 a 的候选值。属性值由候选值和其他因素一起决定。若属性没有数值对应,该属性值将被置为无值属性"NA"。

在 ILP 中把分配决策编码成变量 0/1。从一些片段和值的分配约束中得到 ILP 的约束条件。目标函数是一个片段与一个属性局部相容性的度量。

4.2.1　决策变量

在属性扩充中使用的决策变量有两种。第一种表示片段到属性的分配关系。片段 c. s 到属性 a 的分配记录用变量 $x(c.s,a)\in\{0,1\}$表示。如果片段 c. s 被分配给属性 a,则 x 等于 1,否则为 0。所有的属性都有一个范围,包括 NA。第二种表示值到属性的分配关系,用 $z(v',a)\in\{0,1\}$表示。如果属性 a 分配给上下文 c. s,其中 $c.s.v=v'$,那么 $z(v',a)=1$,否则为 0。

4.2.2　约束

在属性扩充中主要考虑两种约束:实例约束和类别约束,实例约束保证值到属性的全局统一分配,类别约束防止给属性分配异常值。

实例约束又分为两种类型:一致约束和属性之间的相互约束。前者对相同属性多个值进行统一,通常所有的属性和类别使用相同的约束。后者应用领域知识对不同的属性值进行约束。

(1)实例一致性约束

式(7)是一致性约束的表示:

$$\sum_a x(c.s,a)=1,\forall c,\forall a \tag{7}$$

该约束确保了每个片段 c. s 最多分配给一个属性。网页中的每个值都有相对应的一个属性。如果一个值在网页中多次出现,那么每一次出现应该与不同的属性相关。对于多个属性可以这样表示:

$$\sum_v z(v,a)\leqslant 1,\forall a \tag{8}$$

该约束说明了一个属性最多应该分配到一个值。有时也有例外,例如不同的商品房可以有相同的楼栋。在这种情况下,把式(8)右边的 1 可以替换成一个合适的期望值上界。

片段分配与起值分配的关系是片段分配是值分配的必要条件但不是充分条件,表示为:

$$z(v',a)\geqslant x(c.s,a),\forall a,\forall v',\forall c.s:c.s.v=v' \tag{9}$$

该约束保证了如果片段 c. s 是分配给属性 a 的,就意味着把片段中的值 c. s. v 分配给这个属性,也就是把约束从片段转移到属性。

如果值到属性存在分配,那么 a 至少分配了一个包含该值的片段。

$$z(v',a)\leqslant \sum_{c.s;c.s.v=v'} x(c.s,a),\forall a,v' \tag{10}$$

这个约束保证如果属性 a 分配值 v',那么属性 a 应该至少分配了一个 c. s,$c.s.v=v'$。

(2)实例属性的相互约束

一致性约束是通用的,而属性的相互约束则与领域相关。例如,一个"商品房"的合理约束是建筑面积大于使用面积。这些领域知识可以表示成:

$$\sum_v v\cdot z(v,\text{"建筑面积"})\geqslant\sum_v v\cdot z(v,\text{"使用面积"})$$

根据常识,这些约束都潜在地排除了错误值,保证了属性值的正确性。

对于上面的约束,如果"建筑面积"未分配属性值,而"使用面积"分配了属性值,则会发生冲突。为了解决这一问题,改进上面的约束为:

$$v_{\max}[1-\sum_v z(a_b,v)]+\sum_v vz(a_b,v)\geqslant\sum_v vz(a_s,v) \tag{11}$$

(3)类别约束

假设要对楼盘户型的属性值进行扩充。户型的面积大小一般是在 20～300m^2。如果一个网页中提到面积大小为 10000m^2,那么肯定不是表示户型的面积。类别约束应考虑这些基本的领域实例属性值分布。

对于一个属性 a，令 $V_a=\{v_{a1},v_{a1},\cdots,v_{am}\}$ 表示训练数据中属性 a 的值所组成的集合，测试片段 c. s 的值 c. s. v，如果值 c. s. v 接近 V_a，那么该值更有可能被分配给 a。

在考虑 c. s. v 是否是属性 a 的一个合理的属性值时，如果只计算 c. s. v 与 V_a 的标准差的大小，在数据出现多峰分布时效果不好。因此使用一个更有效的方法，即建立 V_a 的核密度模型，把 c. s. v 的值作为在 V_a 点处的密度：

$$S(c.s.v,V_a)=\frac{1}{Z}\sum_{k=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(c.s.v-v_{ak})^2}{2\sigma_{ak}^2}) \quad (12)$$

其中，σ_{ak}^2 是第 K 次的高斯变量，Z 是用来保证 $S(c.s.v,V_a)\in[0,1]\sqrt{2}$ 的常量。对于多峰分布的数据，可以简单地置 $Z=1$，称为非正规支持度。

在 c. s. v 有 V_a 充分支持时将 c. s 分配给 a。设定一个特定属性的支持度阈值为 τ_a，那么可以采用下面简单的约束增加集合。

$$(S(c.s.v,V_a)-\tau_a)x(c.s,a)\geqslant 0 \quad (13)$$

阈值 τ_a 的取值为：$\tau_a=\min_i\{S(v_{ai}-\alpha\sigma_a,V_a)\}$。

其中 $\alpha\geqslant 0$ 是一个常量，在实验中通常取 $\alpha=4$。通常，与 V_a 中值的标准差不超过 α 值都可以。

4.2.3 目标函数

选择满足以上所有约束条件的属性值分配方式取决于片段 c. s 和属性 a 相容性的局部模型，可表示为一个条件概率模型 $\Pr(a|c.s)$。该模型采用训练学习的方式，称之为局部模型是因为片段与属性的关联强度只依赖于片段本身的内容，而与其他的片段无关。

ILP 的目标是使满足约束的局部相容性达到最大。目标函数如下：

$$\sum_{c.s,a}x(c.s,a)\log\Pr(a|c.s) \quad (14)$$

4.3 局部模型

4.3.1 特征描述

为了关联片段与属性，模型需要考虑片段中文本与不同元素的匹配方法，如片段中的文字分布、属性元数据。

每一个匹配信号都称为一个特征，并成为一些 d 维向量空间中一个特征向量 $f(c.s,a)\in R^d$ 中的元素。上下文 c. s 和属性 a 之间的整体相容性是通过局部模型 $w\in R^d$ 特征的线性组合获得的，记为内积 $w^tf(c.s,a)$。

为了联合各种 $(c.s,a)$ 的局部相容性，本文使用一个多类逻辑表达式来校准这些概率：

$$\Pr(a|c.s)=\frac{\exp(w^tf(c.s,a))}{\sum_{a'}\exp(w^tf(c.s,a'))} \quad (15)$$

其中，a 是类标签。通常特征元素是非负的，如果 f 值是正的，那么这个特定的元素非 0。

属性值扩充方法主要使用以下几种特征：

(1)基于属性名的特征

对于某些属性，在不同网页中通常都有不同的叫法。令 $Names(a)$ 为属性 a 的名称集合，如果输入片段中包含 $Names(a)$ 中的一个名称，函数将返回 1，否则返回 0。

(2)基于词分布的特征

基于属性名称的特征的方法无法保证属性的名称集合包含所有的名称，有时候属性值出现的地方却不会提到属性名称，因此，需要其他特征方法的补充。本文也采用计算片段 c. s 之间文本相似度的方法，即训练上下文中的片段被分配给了其他实例的属性 a。

本文采用 TF-IDF 和 Jensen Shannon divergence 两种方法进行特征计算。

TF-IDF 方法将训练数据中包含属性 a 的片段看作包含 a 的文档，然后计算它的 TF-IDF 分数。对于一个片段 c. s，属性 a 的 TF-IDF 值为 c. s 中词 w_i 的分数的总和，表示为：

$$tf\text{-}idf(c.s,a)=\sum_{w\in c.s}tf\text{-}idf(w,a) \quad (16)$$

其中，$tf\text{-}idf(w,a)$ 是属性 a 相对于训练片段的值。

计算基于 Jensen-Shannon divergence 的特征：对于每个属性 a，收集所有属性 a 的相关训练片段为词袋，用词的分布 P_a 表示。对于一个词 w，$Pa(w)$ 的计算方法如下：

令 x 是文本片段中属性 a 出现的次数，y 是这些片段中标记的总数，那么 $P_a(w)=x/y$，同样地，可以计算目标片段的分布，并表示为 $P_{c.s}$，那么基于词分布的特征可以用 $JS(P_a\parallel P_{c.s})$ 来计算。

(3)基于属性单位的特征

数值属性的单位也是一个很重要的局部特征，大多数目标数值都有单位如价格、面积，因此增加了检测典型单位的特征。与属性值 c. s. v 相关的基本特征有：标记是否为(小数)整数？标记是否含小数部分？标记是否为字母？应注意一些属性值如 49.32hm2 可能不是一个完全的数值型。

然后利用典型的单位添加特征。例如，商品房面积可以用平方米表示。对于土地面积，则用公顷表示。某些属性是没有单位的，如宗地编号。考虑到这些情况，使用如下特征：片段中是否含有单位？对于每一个单位 u，是否包含于某个片段？

4.3.2 自动训练

属性值扩充方法用种子实例，以及相应的上下文作为训练集。为了训练局部模型，必须将属性标签分配给上下文的片段。属性值扩充方法自动地去训练局部模型。属性值扩充方法计算每个文档中值的出现次数，并将它们与相应实例关联起来。将片段 c. s 分配给实例 e，属性值扩充方法要考虑 3 种情况：

1. 片段中数值 c. s. v 的单位是 u，如果种子实例 e 包含属性值对 $\langle a,v'\rangle$，$v'=$c. s. v，v' 的单位为 u，就用属性 a 标记片段 c. s。

2. 若 c. s. v 与单位无关，属性值扩充方法就查找 e 的上下文中 c. s. v 的出现次数，并在片段 c. s 中查找合适的属性名作为备用。具体地说，假设种子实例 e 的上下文中有属性值对 $\langle a,v'\rangle$，$v'=$c. s. v。备用方案就是检查 a 的其他名字是否出现在值 c. s. v 所在片段 c. s 中。如果是，就用属性 a 标记片段 c. s。

3. 若以上的两种情况都未出现，就用无值属性 NA 标记片段。

5 实验

5.1 实验数据

在实验中，主要采用了 Jsoup 工具包来进行网页源码获取和解析工作。为了验证实例扩展的有效性，本文对中国土地市场网(www. landchina. com)中几个城市的土地信息进行了抓取实验。为进行对比，实验还使用了通过网页搜索得到的有噪声的数据信息。

5.2 评估标准

本文采用以下指标来评估算法效率：召回率(Recall Rate)、准确率(Precision Rate)及召回率和准确率的调和平均值。

召回率等于系统抽取到的正确数据记录占所有正确数据记录的比例：

$$R=\frac{\text{抽取到的正确记录数}}{\text{所有正确的记录数}}$$

准确率等于系统抽取到的正确数据记录占所有抽取到的记录的比例：

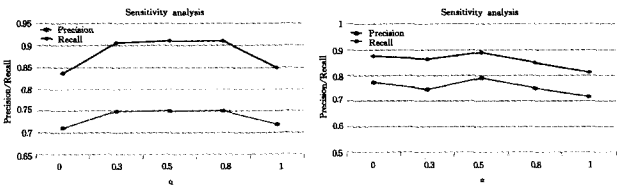
$$P=\frac{\text{抽取到的正确记录数}}{\text{所有抽取到的记录数}}$$

为了综合评价系统的性能，通常还需计算召回率和准确率的调和平均值，即 F-score(F)，它的计算公式如下：

$$F=\frac{2PR}{P+R}$$

5.3 参数 α 的确定

本实验确定 3.3 节中参数 α 的值，使用动态阈值算法为例来说明如何调节相似性和一致性的权值。当 $\alpha=0$ 时表示只考虑一致性，当 $\alpha=1$ 时表示只考虑相似性。实验中以大连与哈尔滨的数据为例(其他城市数据也有类似的结果)，观察发现当 $\alpha=0.5$ 时算法性能最好(见图 8)，故接下来的实验中采用 $\alpha=0.5$。



(a) 不同 α 的值对结果的影响(Dalian) (b) 不同 α 的值对结果的影响(Harbin)

图 8

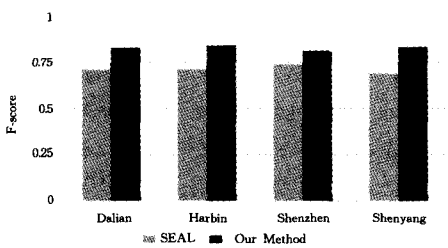
5.4 实验结果与分析

表 1 中给出了本文方法与 SEAL 方法的对比实验结果，采用了中国土地市场网的数据。其中“A”代表抽取到的正确的记录数，“A+B”指抽取到的记录数，“A+C”指正确的记录数。从表 1 中可以看出，与 SEAL 算法相比，本文的方法在抽取的效率上均有所提升，P 值平均提升了 11%，R 值平均提升了 3.7%，F 值提升了 8.4%。通过实验验证了本文提出的算法的有效性。

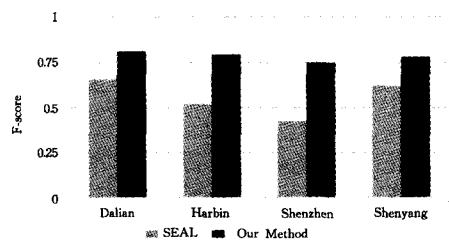
表 1 抽取结果对比

City	SEAL					Our Method					Comparison
Symbol	Dalian	Harbin	Shenzhen	Shenyang	Average	Dalian	Harbin	Shenzhen	Shenyang	Average	
A	1870	1922	401	1835		1918	1966	418	1978		
A+B	2915	3016	652	2580		2552	2119	586	2515		
A+C	2103	2208	449	2140		2103	2208	449	2140		
P	0.6415	0.6373	0.6150	0.7112	0.6513	0.7515	0.7937	0.7133	0.7865	0.7613	11%
R	0.8892	0.8705	0.8933	0.8575	0.8776	0.9120	0.8904	0.9310	0.9243	0.9144	3.7%
F	0.7453	0.7358	0.7285	0.7775	0.7468	0.8240	0.8393	0.8077	0.8498	0.8302	8.4%

为展示本文方法处理噪声数据的性能，在中国土地市场网数据和网页搜索得到的数据上对比了本文方法与 SEAL 方法，如图 9 所示。观察发现，本文方法要优于 SEAL，在有噪声的数据上两种方法性能都有所下降，但本文方法的下降幅度明显小于 SEAL，因此本文方法能更好地处理有噪声的数据。



(a) 使用 landchina 数据的实验结果对比



(b) 使用有噪声数据的实验结果对比

图 9

结束语 本文给出了基于二分图的实例扩展模型，定义实例间相似性计算方法，然后通过定义的质量评估方法来获取最优的候选实例。属性值扩充部分是对实例扩展的补充，使用基于 ILP 的方法解决属性分配问题，从而进行实例的属性值扩充。在此基础上给出了基于 Web 的实例扩展与属性值扩充方法。

实验结果表明：本文提出的方法有效解决了有噪声的数据抽取和针对数值属性值的抽取问题，提高了准确率与召回率。在未来工作中，可以考虑提升抽取系统的健壮性，来应对网站信息的频繁变更。

参 考 文 献

[1] 刘兵. Web 数据挖掘[M]. 俞勇，薛贵荣，韩定一，译. 北京：清华大学出版社，2013

[2] Wang R, Cohen W. Iterative set expansion of named entity using the Web[C]// Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. 2008: 1091-1096

[3] Lin Xi-de, Zhao Bo, Weninger T, et al. Entity RelationDis-covery from Web Tables and Links[C]// Proc. WWW. 2010: 1145-1146

[4] Wang R, Cohen W. Character-level analysis of semi-structured documents for set expansion[C]// EMNLP. 2009

[5] Etzioni O, Cafarella M, Downey D, et al. Web-scale information extraction in KnowItAll[C]// WWW. 2004: 100-110

[6] Pantel P, Crestan E, Borkovsky A, et al. Web-Scale Distributional Similarity and Entity Set Expansion[C]// Proceedings of EMNLP2009. Singapore: ACL, 2009: 938-947

[7] He Ye-ye, Xin Dong. Set Expansion by Iterative Similarity Aggregation[C]// Proc of WWW 2011. dia: ACM, 2011: 427-436

[8] Pennaechiotti M, Pantel P. Entity Extraction via Ensemble Semantics[C]//Proc of EMNLP2009. Singapore: ACL, 2009: 238-247

[9] Tan Pang-ning, Kumar V. Introduction to Data Mining [M]. 2005

[10] 李贵,张淼,李征宇,等. 基于领域模型的 Web 数据抽取与集成[J]. 微电子学与计算机,2012,29(9):152-156

[11] 马安香,张斌,高克宁,等. 基于结果模式的 Deep Web 数据抽取[J]. 计算机研究,2009,46(2):280-288

[12] Probst K, Ghani R, Krema M, et al. Semi-supervised learning of at-tribute-value pairs from product descriptions[C]// Proceedings of the 20th International Joint Conference on Artifical Intelligence. 2007: 2838-2843

[13] Pasca M. Organizing and searching the world wide web of facts-step two: harnessing the isdom of the crowds[C]// Proceedings of the 16th International Conference on World Wide Web. 2007: 101-110

[14] Wick M, Culotta A, McCallum A. Learning Field Compatibilities to Extract Database Records from Unstructured Text[C]// EMNLP. 2006: 603-611

(上接第 402 页)

[5] Dandekar T, Snel B, Huynen M, et al. Conservation of Gene Order: A Fingerprint of Proteins that Physically Interact[J]. Science, 1998, 23(9): 324-328

[6] Marcotte E M, Pellegrini M, Ng H L, et al. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences [J]. Science, 1999, 285(5428): 751-753

[7] Enright A J, Iliopoulos I, Kyrpides N C, et al. Protein Interactions Maps for Complete Genomes Based on Gene Fusion Events [J]. Nature, 1999, 402(6747): 86-90

[8] Pellegrini M, Marcotte E M, Thompson M J, et al. Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles[J]. Proc. Natl. Acad. Sci. USA, 1999, 96 (8): 4285-4288

[9] Pazos F, Valencia A. In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs[J]. Proteins: Structure, Function and Genetics, 2002, 47(2): 219-227

[10] Goh C S, Bogan A A, Joachimiak M, et al. Co-evolution of Proteins with their interaction Partners[J]. J Mol Biol, 2000, 299 (2): 283-293

[11] Martin S, Roe D, Faulon J L. Predicting Protein-Protein Interactions Using Signature Products[J]. Bioinformatics, 2005, 21(2): 218-226

[12] Shen J, Zhang J, Luo X, et al. Predicting Protein-Protein Interactions Based Only on Sequences Information[J]. PNAS, 2007, 104(11): 4337-4341

[13] Guo Y, Yu L, Wen Z, et al. Using Support Vector Machine Combined with Auto Covariance to Predict Protein-Protein Interactions from Protein Sequences[J]. Nucleic. Acids. Res., 2008, 36 (9): 3025-3030

[14] Gomez S M, Lo S H, Rzhetsky A. Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks[J]. Genetics, 2001, 159(3): 1291-1298

[15] Gomez S M, Noble W S, Rzhetsky A. Learning to Predict Protein-Protein Interactions from Protein Sequences[J]. Bioinformaties, 2003, 19(15): 1875-1881

[16] Deng M, Mehta S, Sun F, et al. Inferring Domain-Domain Interactions from Protein-Protein Interactions[J]. Genome Research, 2002, 12(10): 1540-1548

[17] Ryan N. Lichtenwalter. New Precepts and Method in Link Prediction[C]//Proceedings of ACM KDD'10. 2010: 243-252

[18] Lv Lin-yuan, ZhouTao. Link Prediction in Complex Networks: A survey[J]. Physica A, 2011, 390: 1150-1170

[19] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39 (5): 651-661

[20] Bader J S, Chaudhuri A, Rithberg J M, et al. Gaining Confidence in High-Throughput Protein Interaction Networks [J]. Nature Biotechnology, 2003, 22(1): 78-75

[21] Asthana S, King O D, Gibbons F D, et al. Predicting Protein Complex Membership using Probabilistic Network Reliability [J]. Genome Research, 2004, 14(6): 1170-1175

[22] Suthram S, Shlomi T, Ruppin E, et al. A Direct Comparison of Protein Interaction Confidence Assignment Schemes [J]. BMC Bioinformatics, 2006, 7(1): 360

[23] Jensen L J, Kuhn M, Stark M, et al. STRING 8-a Global View on Proteins and Their Functional Interactions in 630 Organisms [J]. Nucleic Acids Research, 2009, 37: 412-416

[24] Erdos P, Renyi A. On the Evolution of Random Graphs [J]. Publ. Math. Inst. Hung. Acad. Sci., 1960, 5: 17-60

[25] Zou Z, Li J, Gao H, et al. Mining Frequent Subgraph Patterns from Uncertain Graph Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1203-1218

[26] Zou Z, Gao H, Li J. Discovering Frequent Subgraph over Uncertain Graph Database under Probabilistic Semantics[C]// ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). New York, USA, ACM, 2010: 633-642

[27] Li J, Zou Z, Gao H. Finding Top-k Maximum Cliques in an Uncertain Graph[C]// Proceedings of 26th International Conf. on Data Engineering. 2010: 649-652

[28] Parapetrou O, Ioannou E, Skoutas D. Efficient Discovery of Frequent Subgraph Patterns in Uncertain Graph[C]// Proceedings of the 14th International Conf. on Extending Database Technology. New York, USA, CAN, 2011, 355-366

[29] Jin R, Liu L, Aggarwal C C. Discovering Highly Reliable Subgraphs in Uncertain Graphs[C]// ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD). New York, USA, ACM, 2011, 992-1000

[30] Kollios G, Potamias M, Terzi E. Clustering Large Probabilistic Graph[J]. IEEE Transactions on Knowledge and Data Engineering, 2012

[31] Yan Xi-feng, Zhou X J, Han Jia-wei. Mining Closed Relational Graphs with Connectivity Constraints[C]// Proc of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York: ACM Press, 2005, 324-333

[32] Krogan N J, Cagney G, et al. Global Landscape of Protein Complexes in theYeastSaccharomyces Cerevisiae[J]. Nature, 2006, 440(7084): 637-643