文章编号：***

一种挖掘网页标题中命名实体的远距监督方法*

胡腾[1,2]，王厚峰[1]，赵世奇[2]，张超[2]

（1.北京大学，北京 100871； 2.百度公司，北京 100193）

摘要：本文提出了一种利用百度百科自然标注数据来远距监督网页标题中命名实体挖掘的新方法。首先，通过关联规则挖掘，从百科词条标签数据集中挖掘出层次化的类别结构；然后，以特定类别下百科实体的参考资料网址和相应网页标题作为训练数据，利用一种基于跳跃二元语法模型的贪心策略，学习得到网址和网页标题的混合模板，用于从网页标题中挖掘对应类别的命名实体。实验结果表明，相较于其他使用同类数据源的挖掘方法，我们的方法在挖掘效率、召回量以及部分类别的准确率上都有显著提升。

关键词：命名实体挖掘；关联规则；远距监督；跳跃二元语法模型；混合模板

中图分类号：TP391　　**文献标识码：**A

A Distant Supervision Approach for Named Entity Mining from Webpage Titles

Teng Hu[1,2], Houfeng Wang[1], Shiqi Zhao[2]，Chao Zhang[2]

（1.Peking University，Beijing 100871；2.Baidu Inc.，Beijing 100193 ）

Abstract: We proposed a novel approach for named entity mining from webpage titles, using naturally labeled data of Baidu Baike as distant supervision. Fist, we carry out association rules mining in the data set of Baike entities' multi labels, to build a hierarchical category structure. Then, we take reference URL and related titles of specified category as training set, and use a greedy method based on skip bigrams to learn URL and title's hybrid patterns for named entity mining. The results show, our method greatly improved the mining efficiency, the recall volume, and also the precision of a few categories, compared to methods reported for the same data source.

Key words: Named Entity Mining；Association Rule；Distant supervision；hybrid pattern；Skip Bigram

1 引言

命名实体是特定类别实体的名称表示，如，人名、地名、公司名、电影名等等。命名实体挖掘是信息抽取的一项基本子任务，是自然语言处理、信息检索的重要内容。在知识图谱的构建过程中，实体属性和实体关系抽取，都以命名实体挖掘为前提；搜索引擎为了使用户有更好地体验，不再是仅仅依靠关键词的匹配，而是更多的借助实体知识库来识别需求类别；当前流行的用户兴趣精准建模也常常将各种类别的命名实体词典作为重要的特征。因此，通过命名实体挖掘技术来构建或不断完善一个命名实体知识库具有重要意义。

命名实体挖掘难点在于实体名称边界的识别，现有挖掘方法中候选的产生方式主要有以下几种：（1）基于上下文模板的抽取方法[1,2,3,4]，该方法假设同类实体名称通常具有相似上下文。模板生成又可以分为手动生成[1]和自动生成[2,3,4]，前者抽取准确率高，但需要大量的人力，不适合大批量类别的挖掘任务；而后者又可以分为无指导和弱指导两种自动生成方式，文献[4]利用句子对齐在无指导条件下生成开放领域的挖掘模板，弱指导则常以少量特定类别的种子实体指导生成挖掘模板[3]。（2）基于并置关系的抽取方法，利用 html 表格或文本中的并置关系（如：中文的顿号）抽取具有某种相似属性的短语作为候选[5]。（3）基于句法规则的抽取方法，利用词性标注、依存分析等工具获取句子的句法信息，抽取符合

*收稿日期：****定稿日期：****

基金项目：本研究为百度公司自然语言处理部支持项目，受到国家 863 项目（No. 2015AA015402），国家自然科学基金项目（No.61370117 & No.61333018）和国家社科基金重大项目（No.12&ZD227）资助。

特定句法规则的句子片段作为候选实体[6]。（4）基于平行语料的抽取方法，同时利用多种语言固有特征（如：英文中专名首字母大写）进行候选实体抽取[7]。（5）基于序列标注的抽取，类似于命名实体识别(NER)过程，效果受训练语料规模的制约，存在召回率低的问题[8]。以上方法中，基于模板挖掘的方法准确率高，使用广泛，模板生成方式又以基于种子的弱指导方式最为常见，最符合挖掘任务的应用场景——已有指定类别和少量样本。然而，当待挖掘的类别规模很大，并与其他类别之间存在大量的同名歧义实体时（如：“小说”与“音乐”），如何选择种子来避免模板的语义漂移（如图 1 所示）又成了棘手问题[9]。文献[4]利用多类协同挖掘，通过模板与类别之间的相似性在一定程度上限制了模板语义漂移，但如何确定多类的类别数，什么类别之间适合协同，以及众多类别的种子选择都成为难题。

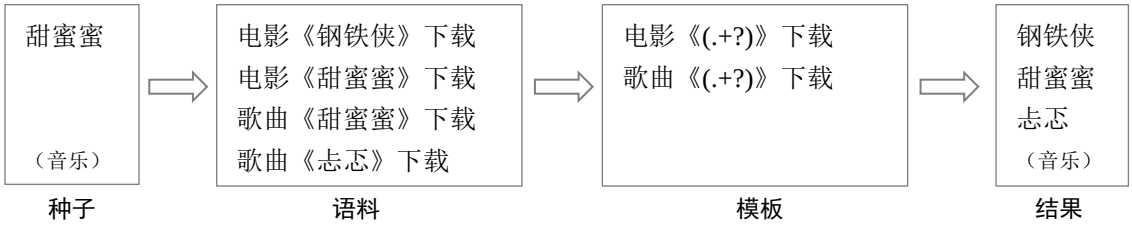


图 1. 命名实体抽取过程中的语义漂移示意图

本文提出了一种远距监督方法来挖掘网页标题中的命名实体，充分利用百度百科中的自然标注数据，自动构建层次化的多类别体系以及高质量的训练语料，用于指导混合模板的生成。本方法有效解决了多类协同挖掘中类别确定问题，避免了种子选择的难题，也缓解了模板语义漂移问题。同时，通过层次化的类别体系来适应不同粒度的语料集合使得挖掘更充分，类别匹配更准确。此外，本文还提出了一种基于跳跃二元语法模型的贪心策略快速生成非连续的上下文模板，相比传统连续模板而言，使用了更多的上下文信息，精度更高。

2 研究动机

文献[3,4]探索了以网页 URL 和 TITLE 为数据源的命名实体挖掘方法，结果显示挖掘效果优于以搜索日志为数据源的挖掘方法。垂直领域网站的 URL 集合往往具有相同的结构形式，而其 TITLE 集中则常含有大量类别明确的命名实体资源，且通常具有相同或高度相似的上下文，非常适合基于模板的挖掘方法。文献[3]采用了基于种子匹配的弱指导方式自动生成模板，用于挖掘新候选，并通过 Bootstrapping 的方法迭代扩增。为了避免迭代过程中模板的语义漂移，该文使用多个类别协同挖掘、交叉验证的方式提升了结果准确率，但同时也降低了召回率（尤其是多歧义类别，如，音乐）。文献[4]采用基于多序列对齐的无指导方式自动生成模板，是一种开放领域的挖掘方法，召回量大，但要获取特定类别专名还需要后续利用特定类别的种子进行相似度计算，准确率不如文献[3]中的方法。

观察发现，百度百科页面中多数实体页面下的“参考资料”栏中都罗列了数量不等的参考网页，其主要内容多是对该实体的描述，网页 TITLE 一般会包含对应的实体名称，这样网页的 URL 与 TITLE 可以用该实体名称和其所属类别进行自然标注。如此，收集所有自然标注的 URL 与 TITLE 可以作为训练数据集，用于指导模板生成，从而有效避免基于种子匹配方式中的语义漂移问题。文献[10]中将这种利用现有知识库和特定假设生成训练数据的监督学习方式称为“远距监督”（Distant Supervision）。

传统的信息抽取模板采用连续的上下文结构，通过指定上下文窗口或公共片段最长化[1,2,3]的方式产生模板。当窗口过小时生成的模板容易发生错误匹配，而窗口过大时生成的模板数量较少。文献[4]中使用句子对齐的方式生成一种非连续的模板，保留了更多的上下文信息，提高了模板匹配精度，但由于采用的是无指导的方式，需要借助其他数据的指导来确定有意义的抽取槽位（SLOT），而且，对于具有相同前缀或后缀的实体类别（如学校、企业）难以产生边界适合的模板。本文提出一种在有指导的情况下非连续模板的生成方法，既

保证模板边界的正确，同时保留更多上下文信息。

本文将尝试从三个方面来改进基于网页标题的命名实体挖掘方法：

1) 使用远距监督的方法来代替基于种子的弱监督或无监督，避免种子选择和语义漂移；
2) 构建完善的层次化类别体系，适应不同语料集合的类别粒度，使挖掘更充分；
3) 利用基于跳跃二元语法模型的贪心算法快速生成高精度的非连续挖掘模板。

本文提出的挖掘系统流程如图 2 所示。

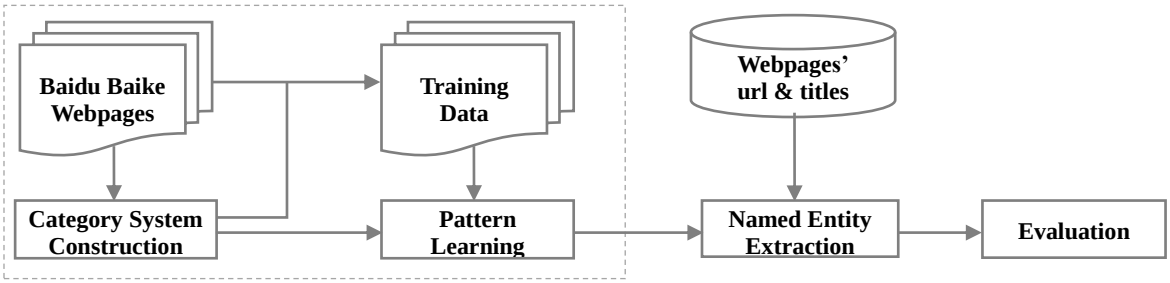


图 2. 基于远距监督的挖掘系统流程图

3 基于远距监督的命名实体挖掘方法

3. 1 基于关联规则挖掘的层次化类别体系构建

非开放领域的命名实体挖掘需要事先指定待挖掘的类别，这些类别是根据具体的挖掘需求而指定的离散集合。事实上，不同来源数据类别粒度的组织形式因编者思维方式的不同而呈现差异。优质语料集合的类别粒度不一定能匹配上指定的挖掘类别，从而导致挖掘过程中资源被丢弃或者分类错误。本文根据百科中词条类别标签间的共现关系，利用关联规则挖掘构建出一种层次化的类别结构体系，用于实现挖掘过程中语料集合类别粒度的自适应。

在百度百科页面中，多数词条页面底端都标有表示词条所属类别的多个“词条标签”，如表 1 所示。容易发现，每个实体的多个标签之间常常具有明显的层次关系。

表 1. 百科实体词条标签示例

编号	实体名称	词条标签
1	李娜	运动员、体育人物 、人物
2	李娜	音乐人物、演员、歌手、人物
3	复仇者联盟	科幻电影、美国电影、电影、漫画、影视作品
4	斯坦福大学	大学、高等教育、教育机构、教育
5	华西村	自然地理、村庄、地理、地点、中国其他行政区划

3. 1.1 关联规则挖掘

将每个百科实体的词条标签集合看作一条记录，记录总条数记为 N，所有标签的集合为 L (l_1，l_2，l_3⋯ l_m)，其中 m 为不同标签总数。标签的出现的频次为 F（f_1，f_2，f_3⋯f_m），标签 l_i，l_j 的共现频次为 B（l_i，l_j），r_{ij} 表示标签 l_i，l_j 之间的一个两两关联规则 $l_i \rightarrow l_j$，α 为频次阈值，β 为置信度阈值，按文献[11]中的计算方式，标签支持度、关联规则置信度计算方法分别为：

$$support(l_i) = \frac{f_i}{N} \quad (1) \qquad confidence(r_{ij}) = \frac{B(l_i,\ l_j)}{f_i} \quad (2)$$

标签之间存在的两两关联规则挖掘步骤如下：

1) 根据记录统计所有标签的出现频次存入 F，统计所有两两共现频次存入 B；
2) 过滤掉频次<α 即支持度<α / N 的标签，得到新的标签集合 L'；
3) 计算标签集合 L'中所有两两关联规则的置信度，将置信度>=β 的规则存入 R；
4) 当 R 中同时存在 r_{ij} 和 r_{ji} 两个规则时，如果 f_i>f_j 则去掉 r_{ij}，否则去掉 r_{ji}。例如，根据置信度计算，“文学作品”→“小说”，与“小说”→“文学作品”都被纳入 R 中，而 $f_{文学作品}$>$f_{小说}$，所以应保留“小说”→“文学作品”这条规则。

3. 1. 2 基于关联规则构建类别森林

上面挖掘得到的标签间两两关联规则可以看作是子类标签结点指向祖先类标签结点的边，可按如下规则逐层构建类别森林（也可以认为是一个简单的 Ontology），其中 η 为相似结点合并的置信度阈值，γ 为孤立结点删减阈值：

1) 逐层构建——剩余标签按结点出度由低到高的顺序进行处理，出度为 0 的结点作为树的根结点，出度为 i 的结点作为第 i 层候选结点，当且仅当候选结点所指向的结点集合完全匹配森林中第 i-1 层的某结点到根结点的路径时，将候选结点作为该结点的子类结点加入森林，否则，抛弃该结点标签；
2) 相似合并——每构建完一层结点，检查该层同父兄弟结点两两关联规则置信度，如果>= η，则合并两个标签，以 *f* 值大的标签为合并结点名称，只要合并结点内其中一个结点与其他兄弟结点满足合并条件可以继续合并。
3) 孤立结点删减——对于森林中的孤立根结点若其 *f* 值<γ 则从森林中移除。

3. 2 训练样本的生成

3. 2. 1 类别标签的统一化

为了生成训练样本，必须统一样本的标签集合。以 3.1 部分中生成的类别森林为标准，根据百科实体 *ne* 的词条标签，将百科实体映射到类别森林的特定结点，得到统一标记的样本集合 S_{NE}。记 *T*（T_1，T_2，T_3…）为类别森林，T_i 是其中一棵树的标签集合，P_i（P_{i1}，P_{i2}，P_{i3}…）是树 T_i 中所有的路径集合，其中 P_{ij} 是树 T_i 从根结点到第 j 个叶子结点路径上的标签集合，*X* 是实体 *ne* 的词条标签集合，按如下步骤确定其在类别森林中的标签；

1) 找到与 *X* 交集最大的树 T_i，如果不存在，说明该实体标签出错风险大，丢弃 *ne*；
2) 找到 P_i 中与 $X \cap T_i$ 交集最大的路径集合 P_{ij}，将 P_{ij} 中最靠近叶子端的标签 *c* 作为 *ne* 的类别标签，将（*c*, *ne*）存入 S_{NE}，如果不存在，丢弃 *ne*。

3. 2. 2 训练样本的抽取

本文“远距监督”的假设为：若百科实体 *ne*（字面形式为 *m*）对应的类别为 *c*（由 3.2.1 确定），则其百科参考网页 TITLE 中所包含的子串 *m* 为 *c* 类实体名，且上下文边界正确。

对于 S_{NE} 中的每个实体样本 *ne*（字面形式为 *m*），获取其百科参考网页的 URL 和对应 TITLE，筛选出包含子串 *m* 的 TITLE，将 TITLE 中 *m* 子串所在位置泛化成特定的 SLOT(抽取槽)标记，把泛化后的 TITLE、对应 URL 和标记类别 *c* 作为一条训练数据存入训练集 S_{Train}。

3. 3 混合模板学习

3. 3. 1 基于 Skip Bigram 的 URL 模板生成与类别自适应

URL 模板是基于 URL 集合泛化得到的，使用 URL 模板的目的是从海量网页中过滤出指定类别的垂直网站页面用于信息抽取。垂直网站的 URL 集合通常具有相似的结构，易于泛化生成统一的 URL 模板，例如，豆瓣网书籍类目下的网址都可以泛化为：

^http://book\.douban[1]\.com/subject/\d+$

本文通过基于 Skip Bigram 模型的贪心策略发现 URL 集合中的易变部分，进行泛化，得到 URL 模板。其中，URL 非域名部分用字符‘/’和‘.’进行切分，并添加开始符‘^’和结束符‘$’（padding）。Skip Bigram 记录任意两个有序的连续或非连续片段在集合中共现的频次。模板生成按从左到右的方向进行，具体步骤如算法 1 所示。

为了使生成的 URL 模板更加适合于挖掘任务，可以充分利用训练数据指导 URL 集合的划分：对于每个域名下的所有 URL 集合，首先，按其在训练语料中对应类别 *c* 划分成多个子集；然后，再根据其对应 TITLE 的 SLOT 边界是否一致，进一步划分成相似性更高的子集，将最终子集 U 作为算法 1 的输入集合。

算法 1. 基于 Skip Bigram 的 URL 模板生成

[1]本文中提到的所有正则表达式均采用 python 语言中的正则表达式规范.

Input：URL 集合 U，泛化参数 ρ
Output：URL 的模板字典 UPT

```
初始化空字典 Skip_Bigrams，UPT
if length( U ) < 2 then
return UPT
end if
for each url in U do
    tokens ← segment( url )
padding(tokens )
add(Skip_Bigrams, tokens)
end for
for each url in U do
    tokens ← segment( url )
last_token ← '^'
upt ← url
for each token in tokens do
if Skip_Bigrams( last_token, token ) < max( ρ*length( U ), 2 ) then
upt ← generalize( upt , token )
else last_token← token
        end if
end for
UPT ←UPT∪ (url，upt)
end for
return UPT
```

通过上述算法，每个可泛化 URL 都能获得一个与之对应的 URL 模板，每个 URL 模板对应一个或多个类别。训练语料中存在大量综合类的新闻、论坛、百科等非垂直网站的 URL，生成了大量对挖掘无用的 URL 模板。为了筛选出垂直网站 URL 模板，定义如下两个指标：

1) 重要度——URL 模板在其域名下所占比重；

$$\text{importance}(UPT_i) = \frac{\#(UPT_i)}{\#(D[UPT_i])} \tag{3}$$

其中，#(・)表示特定条件下匹配的网页数，UPT_i 表示某个 URL 模板，$D[UPT_i]$表示模板 UPT_i 对应的域名。

2) 专一度——URL 模板所匹配到类别的纯度，即模板匹配到最大类别的网页数量占匹配到的所有网页数量的比重；

$$\text{specificity}(UTP_i) = \frac{\#(UPT_I[C_{max}])}{\#(UPT_i)} \tag{4}$$

其中，

$$C_{max} = \underset{c \in C[UPT_i]}{argmax} \#(\, UPT_i[\text{c}]\,) \tag{5}$$

C[UPT_i]为模板 UPT_i 对应的类别集合，UPT_i [c]表示模板在特定类别下的匹配范围。容易发现垂直网站的 URL 模板的专一度接近于 1，重要度也不会过低。通过设置专一度阈值 λ、重要度阈值 μ 可以过滤掉大部分非垂直网站的 URL 模板。但是，专一度的计算与用来衡量 URL 模板的类别粒度有关。例如，豆瓣电影、电视剧的信息页面具有相同 URL 模板（^http://movie\.douban\.com/subject/\d+$），如果以电影、电视剧这一类别层次来衡量，其专一度小于 0.7，如果增大类别粒度，将电影、电视剧回溯到其父类（影视类）则其专一度变为 0.9 左右。为了避免过滤掉垂直网站 URL 模板，同时尽可能保留其分类信息，需要找到与之适应的最小类别粒度。算法 2 通过类别逐层回溯（backtrack）的方法来适应垂直网站

URL 模板的最佳类别粒度。

算法 2. URL 模板类别粒度自适应

Input：模板 UPT_i，专一度阈值 λ、重要度阈值 μ
Output：模板 UPT_i 对应的类别 c ，若为非垂直网站模板，则返回 None

```
if importance( UPT_i ) < μ then
return None
end if
while length(C[UPT_i]) >0 do
if specificity( UPT_i ) >= λ then
return max( C[UPT_i] )
else backtrack( C[UPT_i] )
end if
end while
return None
```

3. 3. 2 基于 Bidirectional Skip Bigram 的 TITLE 模板生成

与 URL 类似，垂直网站的 TITLE 集合也常常具有相似的结构，易于泛化成统一的 TITLE 模板。不同的是，垂直网站 URL 模板是用来确定网页是否属于某类垂直网站，而 TITLE 模板则用于从该网页 TITLE 中抽取命名实体。为了准确识别命名实体的边界，TITLE 模板中 SLOT 两端应该存在非泛化的边界字符，同时，为了提高模板匹配精度，需要保留更多远距离固定搭配信息。SLOT 两侧上下文的泛化过程与算法 1 类似，由于离 SLOT 越近的上下文与 SLOT 的相关性越高，所以，泛化方向以 SLOT 位置为起点，分别向 TITLE 的两端进行。

利用垂直网站 URL 模板对训练集 S_{Train} 中的所有 TITLE 进行划分，得到多个子集，对每个子集根据其 SLOT 边界字符是否一致进一步划分得到更细粒度的子集 T_{SLOT}。使用算法 3 处理每个子集，其中，函数 generalize_right(•)与算法 1 相同，按从左到右的方向生成 SLOT 右侧部分的模板，generalize_left(•)函数只是与前者生成方向相反，按从右到左的方向生成 SLOT 左侧部分的模板，两者的泛化参数都设置为 θ。TITLE 的切分采用单个汉字、单个英文单词、数字的简单切分方式。merge(•)函数合并两侧模板得到 TITLE 的完整模板。

算法 3. 基于 Bidirectional Skip Bigram 的 TITLE 模板生成

Input：TITLE 子集 T_{SLOT}，对应 $SLOT_i$，泛化参数 θ
Output：TITLE 的模板字典 TPT

```
初始化空集合 LEFT, RIGHT
初始化空字典 TPT, L_TPT, R_TPT
if length(T_SLOT) < 2 then
return TPT
end if
for each title in T_SLOT do
LEFT ← LEFT ∪ left( title, SLOT_i)
RIGHT ← RIGHT ∪ right( title, SLOT_i)
end for
L_TPT ← generalize_left( LEFT, T_SLOT,θ)
R_TPT ← generalize_right( RIGHT , T_SLOT,θ)
TPT ← merge( L_TPT, R_TPT, SLOT_i)
return TPT
```

3. 3. 3 混合模板的验证

通过前面两个部分的模板学习，可以得到大量(URL 模板，TITLE 模板，所属类别)三元组数据，每个三元组称为一个混合模板(Hybrid_PT)。其中，每个 URL 模板只对应一个类别，但可以对应多个 TITLE 模板。混合模板用于从网页 TITLE 中抽取命名实体。

由于生成的候选模板数量众多，泛化程度和质量各异，在正式用于挖掘之前需要进行验证和过滤。利用每条候选混合模板从训练数据集 S_{Train} 中抽取命名实体，抽取过程同 3.4.1，当抽取结果类别和边界都与训练数据一致时才认为抽取正确。据此，计算如下两个指标：

1) 准确率——每条混合模板的准确抽取数量/抽取总数量，用来衡量模板的现实风险；
2) 转化率——每条混合模板的准确抽取数量/模板匹配到的 URL 条数，用来衡量模板的潜在风险，转化率低的模板往往匹配到的是领域相关的新闻或论坛站点，虽然在训练集中准确率高，但在大规模的测试集中存在很大的误匹配风险。

过滤掉准确率低于 σ，转化率低于 ω 的混合模板，得到用于挖掘任务的混合模板集合 H。

3.4 命名实体的挖掘与排序

3.4.1 命名实体的挖掘

设 W 为网页资源库，可利用学习得到的混合模板集 H 从 W 挖掘出指定类别的命名实体，挖掘过程与文献[3, 4]中一致：

1) 对 W 中的每个网页，以混合模板中的 URL 模板匹配网址，筛选出垂直领域网页；
2) 对筛选出网页的 TITLE，以相应混合模板中的 TITLE 模板匹配抽取候选实体。

3.4.2 基于 HITS 算法的结果排序

HITS(Hyperlink Induced Topic Search)[12]是信息检索领域中基于超链接分析网页与话题相关度的常用算法，算法假设网页可以被分为权威网页（类似热门门户网）和枢纽网页（类似导航网），通过链接权重转移迭代计算每个相关网页的权威值（Authority Scores）和枢纽值（Hub Scores）直到收敛，最后实现相关网页在两个维度上的排序。

对于大规模的命名实体挖掘，很难做到 100%的准确率，因此，有必要对结果进行排序，来帮助用户进行筛选。常用方法是根据候选实体被抽取到的频次进行排序，尽管结果中高频部分准确率极高，但由于信息的长尾分布特征，很大一部分长尾实体的频次都为 1，虽然来自于不同质量的网站，却无法通过排序的方式进行区分。更合理的排序方法是，来自不同质量网站的候选实体应赋予不同的权重，而不同的网站应根据其挖掘到的高质量实体的比重来进行赋权，这种相互赋权的方式正是 HITS 算法的思想。对于某一类别下挖掘到的候选结果，由于不同模板可能对应同一域名，其挖掘结果具有同质性（正确和错误的情况均相似）而影响排序结果。这里将所有混合模板 Hybrid_PT 映射到对应的域名，得到域名集合 D 和候选实体集合 NE 之间的对应关系，构成典型的二部图。其中，域名、候选实体可以分别看作是 HITS 中的枢纽结点、权威结点，将候选实体的权威值均初始化为 1，通过下面公式进行迭代直到收敛，即得稳定的域名枢纽值和候选实体权威值，从而分别对两者进行排序。

$$\text{Hub\_Score}(D_i) = \sum_{ne}^{NE[D_i]} \frac{1}{\#(NE[D_i])} Authority\_Score(ne) \tag{6}$$

$$\text{Authority\_Score}(NE_i) = \sum_{d}^{D[NE_i]} \frac{1}{\#(D[NE_i])} Hub\_Score(d) \tag{7}$$

其中，NE_i，D_i 分别指某个候选实体和域名，$D[NE_i]$表示指向候选实体 NE_i 的域名集合，$NE[D_i]$表示域名 D_i 指向的候选实体集合。

4 实验设置与结果评估

4.1 类别体系构建的参数设置与构建结果

在类别标签的关联规则挖掘过程中，用于标签过滤的频次阈值 α 设为 1000，关联规则抽取的置信度阈值 β 设为 0.8，从百科网页库中获取原始标签 3204 个，经过 α 过滤后剩下 431 个，根据其共现关系挖掘得到 571 条关联规则用于构建类别森林。

类别森林构建中相似结点合并的置信度阈值 η 设为 0.4，孤立结点过滤的频次阈值 γ 设为 4000，最终得到的类别森林共包含 34 棵树、147 个结点。部分树的结构如表 2 所示。

表 2. 构造产生的类别森林中部分树结构

一级类	二级类	三级类
人物	体育人物	足球运动员、篮球运动员
	艺人	演员、歌手、模特、乐队
	医生、学者、历史人物、虚拟人物、经济人物、政治人物、军事人	——
影视	电影、电视剧	——
医药	中药、西药	——
文学作品	小说、诗词、文章	——

4. 2 训练语料生成结果

从 1,002,021 个百科实体页面中所列的参考资料中抽取训练样本 1,337,145 条，参考数据在前二十大类中的分布如表 3 所示，其中，前十个类别共占据了 81.1%的训练数据，可见由于不同类别实体热度不一，自动生成的训练语料在不同类别上的分布极不均匀。为避免训练语料过于稀疏，将类别粒度小于二级类别的训练样本标签统一回溯到二级类别标签。

表 3. 训练样本的类别分布

书籍	人物	游戏	影视	地名	电器	音乐	食物	机构	文学作品
241489	234611	122628	96667	93371	90464	61829	58157	57514	27742
企业	**植物**	**软件**	**生物**	**药品**	**汽车**	**活动**	**事件**	**设备**	**疾病**
21517	11231	7371	6372	5676	5069	4648	2901	2647	2249

4. 3 模板学习中的参数设置与结果

URL 模板生成过程中泛化参数 ρ 设置为 0.3，专一度阈值 λ 设为 0.9，重要度阈值 μ 设为 0.1，生成 URL 模板总数 11831 个。生成 TITLE 模板时泛化参数 θ 设为 0.9，得到混合模板为 13572 个。验证过程中，准确率阈值 σ 设为 0.9，转化率阈值 ω 设为 0.2 ，过滤得到最终的混合模板 8070 个。在前二十个一级类别上分布如表 4 所示。

表 4. 混合模板的类别分布

书籍	人物	游戏	影视	地名	电器	音乐	食物	机构	文学作品
168	2312	657	328	1758	75	87	247	1274	238
企业	**植物**	**软件**	**生物**	**药品**	**汽车**	**活动**	**事件**	**设备**	**疾病**
332	82	26	70	51	27	26	6	30	21

4. 4 挖掘结果与效率评估

本文挖掘数据源采用与[3,4]中相同的网页库，约含 300 亿网页。[3]中采用的是基于种子弱监督的模板学习和多类协同的筛选过滤方法，这里简称其为 SMCL，[4]中则采用基于多序列对齐方法 MSA 来构造模板进行开放领域的命名实体挖掘，前者通过抽取后的大量后处理侧重提高准确率，后者则注重召回。本文分别按照两文中的评估方式进行对比评估。

表 5. 与 SMCL 方法的效果对比

Category	SMCL			Ours			
	P@all	C_{hot}	Volume	P@same	P@all	C_{hot}	Volume
艺人	0.99	0.90	7630	0.99	0.91	0.92	484448
电影	0.95	0.86	24183	0.97	0.92	0.99	251492
电视剧	0.92	0.93	21655	0.93	0.85	0.92	75156
音乐	0.96	0.33	11011	0.97	0.90	**0.82**	463178
游戏	0.96	0.75	14049	1.0	0.97	0.76	446958
均值	0.96	0.75	15706	**0.97**	**0.91**	**0.88**	**344246**

对挖掘到的候选实体，SMCL 方法通过多类协同进行了严格的筛选，本文则直接截取 HITS 排序结果前 80%作为评估集合。准确率 P@all 参照[3]中随机抽样人工评估的方式，两人同时独立标注，标注不一致的样本通过共同讨论决定标注结果，取 3 次随机抽样评估的平均值作为最终结果。P@same 是在本实验排序结果 top 端抽取与 SMCL 召回等量的数据进行准确率评估，评估方式与 P@all 相同。C_{hot} 是指对特定类别热门实体的覆盖率，热门实体集合为最近一月内每日搜索次数均在 10 次以上的实体名称集合。对比评估结果如表 5 所示，

其中 Volume 表示召回量。数据显示，在等量召回的情况下，我们的方法在 5 个类别的准确率都有一定提升；从全量召回上看，在平均准确率损失较小的情况下，平均热门覆盖率的提升主要体现在高歧义类别（音乐），平均召回量显著提升（成对样本 t 检验，双侧检验概率 P=0.018<0.05），为 SMCL 方法的 20 多倍。其主要原因是我们对挖掘模板的验证是在训练集上进行，挖掘后不再进行交叉验证和冗余验证，从而使得大量的歧义、低频实体得以保留。相反，SMCL 需要在挖掘后进行多类协同验证和冗余验证，侧重召回热门无歧义的实体。

按照 MSA 方法原文[4]中召回率的估算方式，随机抽取百科词条中带标注的词条 110 万，将其类别标签与挖掘结果的类别标签均回溯到一级类别标签，估算一级类别的召回率。

表 6. 与 MSA 方法的召回率对比

	人物	文学作品	机构	影视	食物	生物	植物	音乐	游戏	疾病
MSA	54.9%	40.9%	53.1%	69.8%	51.2%	51.0%	51.8%	58.8%	62.7%	66.1%
Ours	62.6%	78.1%	58.3%	79.1%	61.6%	14.6%	21.6%	81.3%	67.4%	13.2%

结果显示，相较基于 MSA 的开放领域挖掘，我们的方法对其中 7 个热门类别的召回都有明显的提升。而其他三个类别的召回率远远低于 MSA 方法，主要因为百科中对冷门实体类别能提供的有效训练语料较少，生成的挖掘模板数量不足，难以覆盖大量网页。

由于本文采用层次化的类别体系，挖掘结果分布在类别体系的不同层级结点标签下，下面不考虑类别粒度，对挖掘数量大于 5000 的主要类别进行分段抽样评估，结果如表 7 所示。

表 7. 主要召回类别的分段抽样评估

评估指标	企业	书籍	小说	艺人	食物	音乐	游戏	电影
P@100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@1000~1100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@10000~10100	1.00	1.00	1.00	0.99	0.99	0.99	1.00	0.98
P@100000~100100	1.00	1.00	1.00	0.99	0.95	0.96	1.00	0.95
P@1000000~1000100	0.98	1.00	0.95	——	——	——	——	——
P@all	0.97	0.99	0.98	0.91	0.92	0.90	0.97	0.92
Recall	0.59	0.62	0.78	0.62	0.69	0.81	0.67	0.77
Volume	20890803	1909408	1358140	519770	484448	463178	446958	251492
评估指标	**学校**	**医生**	**电视剧**	**体育人物**	**医药**	**植物**	**动物**	**医院**
P@100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@1000~1100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@10000~10100	0.90	1.00	0.99	0.96	1.00	0.91	——	——
P@100000~100100	**0.00**	1.00	——	——	——	——	——	——
P@all	0.83	1.00	0.85	0.90	0.97	0.82	0.81	0.99
Recall	0.30	0.61	0.76	0.38	0.35	0.22	0.17	0.23
Volume	109299	167489	75156	62624	45136	17401	6444	5675

可以看出所有类别 top1000 的准确率高达 100%，说明挖掘方法的可靠性。其中，企业类召回高达 2000 多万，准确率高，但召回率不到 0.6，分析发现挖掘得到的企业名称都以规范的全称出现，而百科词条中存在大量企业简称。医生类准确率高达 1.0，是因为该类仅学习得到 2 个适用于优质垂直网站的混合模板，且站点不含噪音信息。学校类最后一个抽样集合准确率为 0.0，内容全为“**年**学校招生信息”的形式，是误将招生信息发布网站作为学校介绍的垂直网站，虽然，在模板学习过程中未能过滤掉，但其均出现在数据集的尾端，也证明了基于 HITS 算法排序的有效性。此外，我们的方法也有很高的效率，表 8 是效率的对比。

表 8. 与 SMCL 和 MSA 方法的效率对比

步骤	MSA	SMCL	Ours
训练语料生成	——	——	~ 300 seconds
URL 模板生成	~ 2 days	~ 2 days	68 seconds
TITLE 模板生成	~ 2hour	~ 1 hour	126 seconds
候选 NE 抽取	~ 5 hours	~ 5 hours	~ 1 hour
Bootstrapping	——	~ 5 * n hours	——

5 结论

本文描述了一种利用百度百科资源中的参考资料链接作为自然标注数据，自动生成训练集，来指导网页标题中命名实体挖掘的方法。结果显示，与基于种子弱监督的挖掘方法相比，在准确率基本不变条件下，本方法提升了多歧义实体类别（如，音乐）的召回量与覆盖率。与无监督的开放类别挖掘方法相比，在百科热门类别上的召回率上均有较大提升。本方法主要优点包括：1）通过远距监督的方式避免了模板生成过程中的语义漂移，使得大规模挖掘多歧义类别实体成为可能；2）将混合模板的学习从全量数据上的无监督或弱监督迁移到训练数据上的有监督，大幅提升了模板的学习效率；3）完善的层次化类别体系灵活适应不同粒度的语料集合，进一步提升召回；4）产出数据中包含大量的高质量混合模板可用于实时更新挖掘，此外，大量优质的实体-URL 对为实体属性挖掘提供了重要线索。

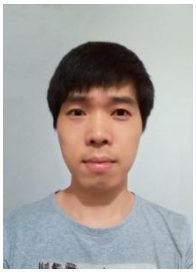
致谢

在此，感谢宋巍老师为本研究提供的参考数据以及其他同学和同事提供的帮助。

参考文献：

[1] Wang R C, Cohen W W. Language-Independent Set Expansion of Named Entities Using the Web[J] IEEE Computer Society, 2007:342-350.
[2] Pasca M. Acquisition of categorized named entities for web search.[J]. Proceedings of the AcmCikm International Conference on Information & Knowledge Management, 2004:137-145.
[3] C. Zhang, S. Zhao, and H. Wang. Bootstrapping large-scale named entities using url-text hybrid patterns[J]. IJCNLP 2013:293-301
[4] Wei Song, Shiqi Zhao, Chao Zhang, et al. Exploiting Collective Hidden Structures in Webpage Titles for Open Domain Entity Extraction. Proceedings of ACM WWW Conference, 2015.
[5] Dalvi B B, Cohen W W, Callan J. WebSets: Extracting Sets of Entities from the Web Using Unsupervised Information Extraction[J]. In WSDM., 2013:243-252.
[6] Li X L, Zhang L, Liu B, et al. Distributional similarity vs. PU learning for entity set expansion[J]. In Proceedings of the ACL 2010: 359-364
[7] A Kotov，CX Zhai，R Sproat. Mining named entities with temporally correlated bursts from multilingual web news streams[J].In WSDM'11: 237-246
[8] Bing, Lidong, W. Lam, and T. L. Wong. Wikipedia entity expansion and attribute extraction from the web using semi-supervised learning[J]. In WSDM'13:567-576.
[9] Zhenyu Qi,Kang Liu,Jun Zhao.Choosing Better Seeds for Entity Set Expansion by Leveraging Wikipedia Semantic Knowledge [J].Proceedings of CIKM 2007, Portugal:683-690
[10] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data.[C]. In Proceedings of the ACL，2009:1003-1011 .
[11] Pasquier N, Bastide Y, Lakhal T L. Discovering Frequent Closed Itemsets for Association Rules[J]. Lecture Notes in Computer Science, 1999, 1540:398--416.
[12] Kleinber J. Authoritative sources in a hyperlinked environment[J]. In 9th ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604--632.

作者简介:



胡腾（1987—），男，在读硕士，主要研究领域为自然语言处理与知识挖掘。
Email:huteng@baidu.com

王厚峰（1965—），男，教授，主要研究领域为自然语言处理与机器学习。
Email: wanghf@pku.edu.cn

赵世奇（1981—），男，博士后，主要研究领域为自然语言处理与知识挖掘。
Email: zhaoshiqi@baidu.com

张超（1984—）男，硕士，高级软件工程师，主要研究领域为自然语言处理与知识挖掘。