文章编号：1003-0077(2012)05-0026-07

基于半监督话题模型的用户查询日志命名实体挖掘

曹 雷[1,2]，郭嘉丰[1]，白 露[1,2]，程学旗[1]

（1. 中国科学院 计算技术研究所 网络数据科学与工程研究中心，北京 100190；
2. 中国科学院 研究生院，北京 100049）

摘 要：基于用户查询日志的命名实体挖掘，目标是从用户查询日志中挖掘具有指定类别的命名实体。已有研究工作提出一种基于种子实体的挖掘方法，利用实体类别与候选实体之间的模板分布相似性来对候选实体进行排序。然而该挖掘方法忽略了命名实体具有歧义性、查询模板具有多义性和未标注实体信息，因而不能够有效的对候选实体进行排序。该文采用半监督话题模型，利用查询模板之间的关系来学习实体类别的模板分布，进而改善候选实体的排序效果。实验结果表明了该文提出方法的有效性。

关键词：用户查询日志；命名实体挖掘；半监督话题模型

中图分类号：TP391 **文献标识码**：A

Named Entity Mining from Query Log through Semi-supervised Topic Modeling

CAO Lei[1,2], GUO Jiafeng[1], BAI Lu[1,2], CHENG Xueqi[1]

(1. Research Center of Web Data Science & Engineering, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Named entity mining from query log aims to mine a list of named entities with the specific type from the query log. Previous work proposed a seed-based method which ranked the candidate entities based on the similarity between the template distribution of the specified class and that of the entities. However, it doesn't take into account the ambiguity of named entity, the polysemy of the template and the unlabeled data. In this paper, we propose a semi-supervised topic model, which leverages the relationship between the templates (i.e. the co-occurrence between templates) to learn the template distribution of the specified class so as to improve the entity ranking. Experimental results show the effectiveness of the proposed method.

Key words: query log; named entity mining; Semi-supervised Topic Model

1 引言

伴随着万维网的迅速发展和搜索引擎的广泛使用，产生了一类重要的用户数据，即用户查询日志。用户查询日志不仅记录了用户使用搜索引擎时的行为，同时还反映了用户的喜好、兴趣和查询习惯等非常重要的隐形知识，是一类富含“大众智慧”的海量数据资源。近年来，用户查询日志已经成为相关研究领域广泛关注的对象。其中在用户查询日志中，含有丰富的实体信息。这些实体是用户查询的核心语义单元，也是人们正确表达查询意图的重要元素。从用户查询日志中挖掘命名实体，对于垂直搜索、知识库建立和查询意图理解等方面有着积极的作用。

基于用户查询日志的命名实体挖掘研究工作开展较晚。已有研究工作[1]采用一种弱指导方式的挖

收稿日期：2011-09-05 **定稿日期**：2012-04-05

基金项目：国家自然科学基金资助项目(60903139, 60873243, 60933005)；国家 863 计划重点资助项目(2010AA012502, 2010AA012503)

作者简介：曹雷(1977—)，男，博士研究生，主要研究方向为实体检索与挖掘；郭嘉丰(1980—)，男，博士，助理研究员，主要研究方向为信息检索与数据挖掘；白露(1983—)，男，博士研究生，主要研究方向为文本检索与挖掘。

掘框架从用户查询日志中挖掘指定类别的命名实体，该挖掘框架利用类别和候选实体之间的模板分布相似性来对候选实体进行排序。该工作主要是基于单类别假设，即一个命名实体只属于一个类别。然而，命名实体可能隶属于多个类别，即命名实体具有歧义性；查询模板也可能来自于多个类别，即查询模板具有多义性。例如，对于命名实体“star wars”，即可能隶属于类别“Movie”，也可能隶属于类别“Game”；同样对于查询模板“# download”，即可能来自于类别“Movie”，也可能来自于类别“Software”。同时该工作也忽视了未标注实体的信息（即种子实体以外的候选实体），因而不能够准确地估计得到类别的真实模板分布，进而不能有效地对候选实体进行排序。

本文综合考虑命名实体歧义性、查询模板多义性和未标注实体信息，采用半监督话题模型，利用查询模板之间的关系来学习得到类别的模板分布，并基于学习得到的类别模板分布对候选实体进行排序。最后在一个真实的用户查询日志数据集上来验证本文提出的方法。实验结果表明，本文的方法优于基准方法，具有较好的挖掘效果。

本文的主要贡献：(1)综合考虑命名实体歧义性、查询模板多义性和未标注实体信息，利用查询模板之间的关系来学习类别的模板分布，进而改善候选实体排序的效果；(2)采用半监督话题模型来学习类别的模板分布。

本文内容组织如下：第二节介绍相关工作；第三节给出问题描述；第四节介绍基于用户查询日志的命名实体挖掘流程；第五节详细阐述采用半监督话题模型来学习类别的模板分布；第六节介绍如何对候选实体进行排序；第七节给出实验结果和分析；第八节对工作进行总结。

2　相关工作

命名实体识别和挖掘研究工作已经开展了 20 多年，但主要是面向传统文本领域。早期的命名实体识别主要是基于人工编写规则的方式。随着机器学习的发展，越来越多的基于统计学习的方法被引入到命名实体识别领域，包括监督式学习、半监督式学习和无监督式学习。

基于用户查询日志的命名实体挖掘研究工作开展较晚，相关工作较少。Paşca[1] 首先提出基于用户查询日志来挖掘指定类别的命名实体，该工作采用了一种弱指导方式的挖掘框架。翟海军等人[2] 利用弱指导的关联话题模型，来解决用户查询日志命名实体挖掘所面临的类别模糊性的问题。该工作同本文的工作相近，但是存在以下区别：(1)文献[2]主要是基于种子实体来学习类别的模板分布；而本文的工作是基于所有候选实体来学习类别的模板分布；(2)加入指导信息的方式不同，文献[2]以种子实体的类别标注信息作为指导，加入到关联话题学习过程中；而本文是以种子实体对应的查询模板信息作为先验知识，加入到 PLSA 模型的学习过程中。

Guo 等人[3] 针对命名实体的类别歧义性的问题，提出一个弱监督式 LDA 模型框架来识别用户查询中的命名实体。Xu 等人[4] 结合用户查询中的上下文（即查询模板）和“click-through data”信息，从用户查询中识别命名实体。Du 等人[5] 从查询会话中提炼更加丰富的上下文特征，并基于提取的特征分别采用条件随机场模型和话题模型来从用户查询中识别命名实体。文献[3-5]更加关注如何从单条用户查询中识别命名实体，而本文工作是基于整个用户查询日志来挖掘命名实体。

另外一部分相关工作是话题模型（Topic Model）。话题模型通过模拟文档的生成过程来对文档进行建模，经典的话题模型包括 PLSA[6] 和 LDA[7]。话题模型被广泛地应用在各个领域，包括文本分析挖掘、文档检索、社会网络分析以及情感分析等。近年来，已有一些研究工作利用监督式话题模型[8] 和半监督式话题模型[9] 进行建模。其中，文献[9]采用半监督话题模型来解决用户观点集成问题，而本文工作则是利用半监督话题模型解决用户查询日志命名实体挖掘问题。

3　问题描述

我们用 $Q=\{q_1,q_2,\ldots,q_n\}$ 表示用户查询数据集合，用 $S^T=\{s_1^T,s_2^T,\ldots,s_k^T\}$ 表示隶属于类别 T 的种子实体集合，用 $E^T=\{e_1^T,e_2^T,\ldots,e_m^T\}$ 表示具有类别 T 的命名实体集合。则基于用户查询日志的命名实体挖掘问题可以形式化为：给定一组类别 T_1，$T_2,\ldots,T_n$，以及每个类别下的一组种子实体集合 S^{T_i}，目标是从 Q 中挖掘得到隶属于类别 T_i 的实体集合 E^{T_i}。

已有研究工作[3] 通过对用户查询日志分析表明，大约有 70% 的用户查询包含单个命名实体，因而本文主要关注包含单个命名实体的用户查询。基

于上述的分析结果，针对于包含命名实体的用户查询 q 就可以表示成为一个二元组，即 $q=(e,t)$。其中，e 表示命名实体；而 t 表示用户查询 q 中实体 e 的上下文，即查询模板。例如，对于用户查询“starcraft walkthrough”，就可以表示成(“starcraft”，“# walkthrough”)。

4 命名实体挖掘流程

本文基于用户查询日志的命名实体挖掘可以分为以下几个阶段。

(1) 获取种子实体对应的查询模板

利用所有指定类别下的种子实体，采用字符串匹配方式从用户查询日志中获取种子实体对应的查询模板集合。在进行字符串匹配的时候，忽略了大小写。

(2) 获取候选实体

利用得到的查询模板，从用户查询日志中获取查询模板所对应的候选实体集合。

(3) 为候选实体建立“描述文档”

利用得到的候选实体，从用户查询日志中获取候选实体所对应的查询模板集合。每个候选实体都会对应一组查询模板。这组查询模板可以看成是候选实体的“描述文档”，与普通文档的区别是，每个候选实体的“描述文档”中的“词语”是查询模板。

(4) 学习目标类别的模板分布

基于所有指定类别下的候选实体及其“描述文档”，采用半监督话题模型(SS-PLSA)学习得到目标类别的模板分布。具体的学习过程将在小节 5 中进行详细阐述。

(5) 对候选实体进行排序

最后，利用候选实体和类别之间的模板分布相似性对候选实体进行排序。

5 半监督话题模型(SS-PLSA)

每个候选实体都可以通过一组查询模板来描述，这组查询模板形成了候选实体的“描述文档”。例如，对于实体“star wars”，在查询日志中可能对应一组相应的查询模板(图 1)，这组查询模板可以看成是实体“star wars”的“描述文档”。在这个“描述文档”中，包含了类别“Movie”和“Game”的信息，其中，查询模板“# movie review”、“# cast”和“# trailer”对应着类别“Movie”，而查询模板“# walkthrough”、“# game”和“# cheats code”对应着类别“Game”，这就相当于一个文档具有多个话题。另外，查询模板“# download”和“funny #”都对应着类别“Movie”和“Game”，这就相当于一个词语可能来自于多个话题。这样，就可以将类别模板分布的求解问题转化成传统文档的话题模型学习问题。具体而言，候选实体对应于传统文档，查询模板对应于传统文档中的词语，而候选实体所隶属的类别对应于话题。具体的映射关系如表 1 所示。

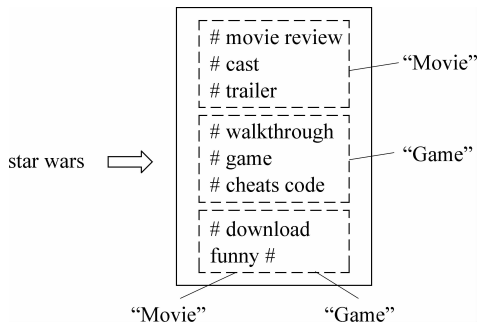

图 1 实体“star wars”对应的模板集合

表 1 对象映射表

用户查询领域	传统文本领域	符号
查询模板	单词	w_j
命名实体	文档	d_i
实体类别	话题	z_k

5.1 传统 PLSA 模型

我们首先介绍一下 Zhai 等人[10]提出的话题模型。假设给出一组文档集合 $D=\{d_1,d_2,\ldots,d_N\}$，文档中词语都是来自于词典 $V=\{w_1,w_2,\ldots,w_M\}$。这些文档共同分享 $K+1$ 个不同的话题，话题集合用 $Z=\{z_1,z_2,\ldots,z_K,z_B\}$ 来表示，其中 $z_1,z_2,\ldots,z_K$ 表示所关注的 K 个话题，这些话题同指定的类别一一对应；z_B 表示背景话题，引入背景话题的目的是为了加强其他 K 个话题的区分能力；一般而言，背景话题 z_B 对于那些没有区分度或者没有信息含量的词语，会给予较高的概率值。我们假定每篇文档的产生过程如下：

1. 以概率 $p(d_i)$选择一个文档 d_i；
2. 对于文档 d_i 中每一个词语 w_j，
   (a) 以 λ_B 的概率选择背景话题 z_B；然后再以 $p(w_j|z_B)$的概率产生词语 w_j；
   (b) 以 $1-\lambda_B$的概率不选择背景话题 z_B；然后再以 $p(z_k|d_i)$的概率选择话题 z_k；最后以 $p(w_j|z_k)$的概率产生词语 w_j。

给定参数集合 $\Lambda=\{p(w_j|z_k),p(w_j|z_k)|w_j\in V,d_i\in D,z_k\in Z\}$，单个文档生成概率如下：

$$p(d|\Lambda)=\prod_{j=1}^{M}p(d,w_j)^{n(d,w_j)}=p(d)^{n(d)}\prod_{j=1}^{M}\left\{\lambda_B p(w_j|z_B)+(1-\lambda_B)\sum_{k=1}^{K}p(w_j|z_k)p(z_k|d)\right\}^{n(d,w_j)} \quad (1)$$

在式(1)中，$\lambda_B\in[0,1]$是人工给定的经验值；$p(w_j|z_B)$表示背景话题 z_B 产生词语 w_j 的概率，该分项是基于整个文档集 D 上估计得到。λ_B 和 $p(w_j|z_B)$在迭代求解过程中保持不变。整个文档集合的对数似然函数如下所示：

$$\log p(D|\Lambda)=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,w_j)\cdot\log\left\{\lambda_B p(w_j|z_B)+(1-\lambda_B)\sum_{k=1}^{K}p(w_j|z_k)p(z_k|d)\right\}+\sum_{i=1}^{N}n(d_i)\log p(d_i) \quad (2)$$

采用标准的 EM 算法来求解式(2)中的参数 Λ。以下分别是 EM 算法的 E-Step 和 M-Step。

E-Step：

$$p(z_k|d_i,w_j)=\frac{p(w_j|z_k)p(z_k|d_i)}{\sum_{l=1}^{K}p(w_j|z_l)p(z_l|d_i)} \quad (3)$$

$$p(z_B|d_i,w_j)=\frac{\lambda_B p(w_j|z_B)}{\lambda_B p(w_j|z_B)+(1-\lambda_B)\sum_{k=1}^{K}p(w_j|z_k)p(z_k|d_i)} \quad (4)$$

M-Step：

$$p(z_k|d_i)=\frac{\sum_{j=1}^{M}n(d_i,w_j)(1-p(z_B|d_i,w_j))p(z_k|d_i,w_j)}{\sum_{l=1}^{K}\sum_{j=1}^{M}n(d_i,w_j)(1-p(z_B|d_i,w_j))p(z_l|d_i,w_j)} \quad (5)$$

$$p(w_j|z_k)=\frac{\sum_{i=1}^{N}n(d_i,w_j)(1-p(z_B|d_i,w_j))p(z_k|d_i,w_j)}{\sum_{l=1}^{M}\sum_{i=1}^{N}n(d_i,w_l)(1-p(z_B|d_i,w_l))p(z_k|d_i,w_l)} \quad (6)$$

5.2　PLSA 模型的半监督学习方式

在本文问题中，直接使用一个无指导的话题模型是不可行的。这是因为本文问题要求得到的隐话题能够和预先定义好的类别一一对齐，因而需要在话题模型的学习过程加入指导信息，即先验知识。在本文中，先验指导信息来自于给定类别的种子实体所对应的查询模板。

对于某个指定的实体类别 T，利用该类别下的一组种子实体来遍历用户查询日志，得到一组对应的查询模板。在这组查询模板中，如果查询模板 t 对应类别 T 的种子实体数目越多，对应其他类别的种子实体数目越少，则表明该查询模板 t 与类别 T 的关联越紧密。我们将这些与类别 T 关联紧密的查询模板称为类别 T 的特征模板，这些特征模板在实体的类别判定上优于其他普通模板。本文通过简单的启发式策略来获取类别的特征模板，并将这些特征模板作为先验知识，加入到 PLSA 的学习过程中。启发式策略如下：(1)查询模板对应的目标类别种子实体数目较多；(2)查询模板对应其他类别中的种子实体数目较少。

这里通过扩展传统 PLSA 模型，加入特征模板来达到学习的话题与指定类别对齐的目的。具体而言，就是为每一个多项式分布的话题定义一个共轭先验，即狄利克雷先验(Dirichlet Prior)，$Dir(\{1+\mu_k p(w_j|\theta_k)\}_{w_j\in V})$。其中，$\theta_k$ 定义为与话题 z_k 相对应的先验话题；μ_k 表示了对先验知识的可信程度，这里可以看作是一个虚文档；而 $\mu_k p(w_j|\theta_k)$表示为当估计 $p(w_j|z_k)$时候，额外增加的词语 w_j 的个数；特别的，如果 $\mu_k=0$，则表示没有任何先验知识。相应的词语产生过程可以用图 2 来表示。

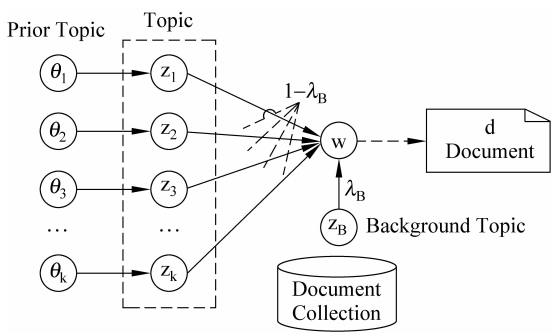

图 2　词语产生过程

将参数集合的先验表示为 $p(\Lambda)$，则 $p(\Lambda)$具有如下形式，

$$p(\Lambda)\propto\prod_{k=1}^{K}\prod_{j=1}^{M}p(w_j|z_k)^{\mu_k p(w_j|\theta_k)} \quad (7)$$

根据上述定义的先验，采用 MAP(Maximum a Posterior)来估计所有参数，如式(8)所示。

$$\hat{\Lambda} = \arg\max_{\Lambda} \log p(D \mid \Lambda) p(\Lambda) \quad (8)$$

同样采用 EM 方法来求解式(8)中的参数，这里 EM 的更新公式和 5.1 节中的更新公式基本一致，唯一变化是式(6)中的更新过程加入了先验指导信息，新的更新式如(9)所示。可以看到，传统的 PLSA 和半监督 PLSA 区别就在于对分项 $p(w_j|z_k)$的计算。

$$p(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)(1 - p(z_B \mid d_i, w_j)) p(z_k \mid d_i, w_j) + \mu_k p(w_j \mid \theta_k)}{\sum_{l=1}^{M} \sum_{i=1}^{N} n(d_i, w_l)(1 - p(z_B \mid d_i, w_l)) p(z_k \mid d_i, w_l) + \mu_k} \quad (9)$$

6 候选实体排序

在第五节中，本文通过采用半监督 PLSA 模型学习得到话题 z_k 中词语的分布 $p(w|z_k)$，即实体类别所对应的模板分布。在本小节，利用类别的模板分布作为参照，来计算类别模板分布与候选实体模板分布的相似性，并基于该相似性来对候选实体进行排序。(候选实体模板分布计算如式(10)所示)。为了方便同 Paşca 工作[1]的比较，本文同样采用 Jensen-Shannon Divergence 作为相似度量。候选实体与类别越相似，表示该候选实体隶属于指定类别的可信度就越大。

$$p(w_j \mid d) = \frac{count(w_j, d)}{\sum_{w' \in d} count(w', d)} \quad (10)$$

7 实验

7.1 实验设置

我们基于一个真实的用户查询日志数据集来验证本文提出的方法。该数据集是从一个商用搜索引擎查询日志中随机采样得到的，大约含有 1 500 万条用户查询。这里主要是利用查询词语本身，不考虑其他信息。

在实验中主要考虑了三个语义类别，分别是“Movie”、“Game”和“Food”。针对类别“Movie”和“Game”，分别从不同的网站人工采集了相应的种子命名实体，网站包括维基百科(www.wikipedia.com)，豆瓣电影(movie.douban.com)和 GameSpot(www.gamespot.com)。种子实体列表如表 2 所示。

表 2 类别和种子命名实体

目标类别	种子命名实体
Movie	sin city, just my luck, miami vice, see no evil, stick it, the breakup, the lake house, the new world, an american haunting, animal farm, ghost rider, say anything, spiderman 3, star wars, over the hedge, silent hill, the da vinci code
Game	doom 3, empire earth, half life 2, the sims, super mario brothers, age of empires 2, destroy all humans, command and conquer generals, call of duty, the legend of zelda, need for speed most wanted, devil may cry 3, kingdom hearts 2, drivers ed, star wars, over the hedge, silent hill, the da vinci code
Food	apple, beef, candy, chicken, coffee, eggplant, hamburger, meat, pasta, pork, potato, rice, shrimp, tofu, tomato, tuna, vegetables

针对每个类别，返回前 500 个候选实体。我们请三个研究人员对这 500 个候选实体进行标注，如果某个候选实体属于指定类别，则标注为正确(分值为 1)；否则，标注为不正确(分值为 0)。对于标注不一致的候选实体，则采取投票的方式来确定候选实体的标注值。基于上述人工标注后的结果，采用 P@N(前 N 个结果的准确率)来评估算法性能。

在 SS-PLSA 模型的学习过程中，有两类参数需要人工预先指定。一个是式(2)中的参数 λ_B，它表示选择背景话题来产生词语的概率，实验中设置 $\lambda_B = 0.98$。另外一类是式(9)中的 μ_k，在实验中我们赋予 μ_k 较大的初始值(>10 000)，而后采用 Tao 等人[11]提出的策略来不断减少 μ_k 的取值。

7.2 参评方法

为了评价基于半监督话题模型方法的性能，我们与 Paşca 提出的方法[1]进行了比较。Paşca 提出的方法作为基准方法，简称 Determ；基于半监督话题模型的方法简称为 SS-PLSA(Semi-Supervised PLSA)。

Determ 方法主要是基于命名实体单类别假设，不能够准确地估计类别的真实模板分布，进而无法有效的对候选实体进行排序。而 SS-PLSA 方法综合考虑了命名实体歧义性，查询模板多义性和未标注实体的信息，利用模板之间的关系能够较好地学习得到类别的模板分布；而后利用学习得到的类别模板分布来对候选实体进行排序，进而改善了命名实体挖掘的效果。

7.3　实验结果及分析

基于上面所描述的实验数据和度量，本小节对方法 Determ 和 SS-PLSA 进行了比较。实验结果在表 3 中列出。从表 3 中可以看到，在每个度量指标上，SS-PLSA 方法基本都优于 Determ 方法。

我们进一步分析了方法 SS-PLSA 能够取得较好挖掘效果的原因，这主要是由于 SS-PLSA 综合考虑了命名实体具有歧义性、查询模板具有多义性和未标注实体信息，充分利用查询模板之间的关系，更加准确地估计出各目标类别的查询模板分布。表 4 中给出了 SS-PLSA 和 Determ 估计的各类别下前 10 个查询模板。从表 4 可以直观的看到，在方法 Determ 中，各类别下的前 10 个查询模板中往往混杂有一些噪音模板或者其他类别的查询模板，例如，类别“Game”的第 1 个模板“# movie”。

表 3　方法 SS-PLSA 和 Determ 的性能（P@N）

	Movie		Game		Food		Average P@N	
	Determ	SS-PLSA	Determ	SS-PLSA	Determ	SS-PLSA	Determ	SS-PLSA
P@25	0.88	0.92	0.68	0.80	0.68	0.92	0.75	0.88
P@50	0.72	0.92	0.70	0.74	0.68	0.92	0.70	0.86
P@100	0.70	0.82	0.64	0.77	0.71	0.91	0.68	0.83
P@150	0.71	0.75	0.63	0.70	0.63	0.90	0.66	0.78
P@250	0.73	0.74	0.65	0.71	0.60	0.73	0.66	0.73
P@500	0.66	0.66	0.42	0.43	0.54	0.63	0.54	0.57

表 4　各个类别下前 10 个查询模板

Movie		Game		Food	
Determ	SS-PLSA	Determ	SS-PLSA	Determ	SS-PLSA
# movie	# movie	# movie	# cheats	# recipes	# recipes
# the movie	# the movie	# cheats	# game	# vacations	# recipe
# trailer	# trailer	# game	# walkthrough	condoleezza #	# salad
what is #	# soundtrack	# games	# cheat codes	kentucky fried #	# receipes
lego #	# trailers	what is #	cheats for #	# ipod	# recipies
# toys	# cast	# walkthrough	# mods	# salad	# cake
# soundtrack	# characters	lego #	# 2 cheats	# diet	# sauce
# games	# hentai	# 2 cheats	# the game	# pox	grilled #
# galaxies	# campers	# toys	# demo	# computers	baked #
# game	the movie #	# galaxies	# roms	# stand	# dip

需要指出的是，尽管方法 SS-PLSA 在改善实体排序效果方面，相对于方法 Determ 确实有一定程度的提升。但是从绝对效果上来说，还有很大的提升空间。特别是针对类别“Game”，候选实体的排序效果还有待进一步改进。这是因为用户查询日志中，仍然存在着大量的噪音数据；而查询模板有时不具有区分性。例如，针对类别“Game”，隶属于“Game”的命名实体经常与“# cheats”，“# game”或者“# cheat codes”等查询模板共现，但是某些游戏平台，例如“ps1”和“xbox”等也经常同上述查询

模板共现。这样，就会将“ps1”和“xbox”误判为“Game”类别。

8 结论

用户查询日志蕴含了丰富的实体信息，这些实体是用户查询的核心语义单元，也是人们正确表达查询意图的基本元素。从用户查询日志中挖掘命名实体，对于垂直搜索、知识库建立和查询意图理解等方面都有着积极的作用。然而，用户查询通常比较简短，书写不规范并且具有歧义性，给基于用户查询日志的命名实体挖掘研究工作带来了巨大的挑战。本文综合考虑命名实体歧义性，查询模板多义性和未标注实体信息，采用半监督话题模型，充分利用模板之间的关系学习得到类别的模板分布，并基于学习得到的类别模板分布来对候选实体进行排序，进而提升了命名实体挖掘效果。实验结果表明了本文方法的有效性。

参考文献

[1] Marius Paşca. Weakly-supervised discovery of named entities using Web search queries[C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management, 2007: 683-690.

[2] 翟海军，郭嘉丰，王小磊，等. 基于用户查询日志的命名实体挖掘 [J]. 中文信息学报，2010，24 (1) :71-76.

[3] Jiafeng Guo, Gu Xu, Xueqi Cheng, et al. Named entity recognition in query[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 267-274.

[4] Gu Xu, Shuang-Hong Yang, Hang Li. Named entity mining from click-through data using weakly supervised latent dirichlet allocation[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009: 1365-1374.

[5] Junwu Du, Zhimin Zhang, Jun Yan, et al. Using search session context for named entity recognition in query[C]// Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, 2010: 765-766.

[6] Thomas Hofmann. Probabilistic latent semantic indexing[C]// Proceeding of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999: 50-57.

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] David M. Blei, Jon D. McAuliffe. Supervised topic models[C]// Proceedings of the 21st Annual Conference on Neural Information Processing Systems, 2007.

[9] Yue Lu, Chengxiang Zhai. Opinion integration through semi-supervised topic modeling[C]// Proceeding of the 17th International Conference on World Wide Web, 2008: 121-130.

[10] ChengXiang Zhai, Atulya Velivelli, Bei Yu. A cross-collection mixture model for comparative text mining [C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004: 743-748.

[11] Tao Tao, ChengXiang Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 162-169.