文章编号:1671-9352(2012)05-0032-06 DOI:10.6040/j.issn.1671-9352.2012.05.006

基于二部图半监督方法的查询日志实体挖掘

曹雷[1,2],郭嘉丰[1],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 100190)

摘要:基于用户查询日志的命名实体挖掘的目标是从用户查询日志中挖掘一组具有指定类别的命名实体。为解决已有用户查询日志实体挖掘研究工作中的种子实体不充分的问题,提出了一种基于二部图的半监督排序方法,利用实体之间的关系(实体共享查询模板)来改善实体排序效果。该方法首先基于候选实体和查询模板构建一个二部图,然后基于二部图将种子实体的类别相关性传播到其他候选实体,最后按照类别相关性得分对候选实体进行排序,并进一步给出方法中迭代过程的等价优化框架。实验结果表明本文提出的方法优于基准方法,具有较好的挖掘效果。

关键词:用户查询日志;命名实体挖掘;半监督方法;二部图
中图分类号:TP391 **文献标志码**:A

Bipartite graph based semi-supervised method for entity mining from the query log

CAO Lei[1,2], GUO Jia-feng[1], CHENG Xue-qi[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 100190, China)

Abstract: Named entity mining from query log aims to mine a list of named entities with the specific type from the query log. A bipartite graph based semi-supervised ranking method, which leverages the relationship between the entities (i.e. entities share common templates) to help improve the ranking, was proposed to resolve the scarcity of seed entity in existing work about named entity mining from the query log. First, a bipartite graph based on the candidate entities and templates was constructed. Then, the relevance score was propagated from the seed entities to other candidate entities. Finally, the candidate entities were ranked according to the relevance score. An optimization framework for the iterative process was further developed in this ranking method. Experimental results show the effectiveness of the proposed method.
Key words: query log; named entity mining; semi-supervised method; bipartite graph

0 引言

命名实体是现实世界中抽象或者具体的实体,常见的实体包括人、地点、机构和专有名称等。命名实体是文本的基本信息元素,是人们正确理解文本的基础。从文本中识别和挖掘命名实体,是知识库建立、关系挖掘和机器翻译等研究领域的基础,在自然语言理解领域占有重要的地位。针对命名实体识别和挖掘的研究工作已经开展了很多年,然而大部分研究工作都是面向传统的文本领域。作为一类重要的用户使用数据——用户查询日志,同样包含了丰富的实体信息,这些实体构成了用户查询的核心语义单元,是人们正确表达查询意图的基本元素。

收稿日期:2011-11-30; **网络出版时间**:2012-04-27 17:04
网络出版地址:http://www.cnki.net/kcms/detail/37.1389.N.20120427.1704.015.html
基金项目:国家自然科学基金资助项目(60903139,60873243,60933005);国家"八六三"高技术研究发展计划基金资助项目(2010AA012502,2010AA012503)
作者简介:曹雷(1977-),男,博士研究生,研究方向为实体检索与挖掘. Email: leicao@software.ict.ac.cn

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。不同于传统的文本,用户查询通常比较简短(2 ~3 个词),书写也不规范(例如,查询词通常是小写的),并且实体的上下文特征不丰富,使得传统命名实体挖掘和识别方法不能直接应用到用户查询这样的环境下,也给本文的命名实体挖掘研究工作带来了挑战。

早期的命名实体识别和挖掘研究工作主要是面向文本领域,作为自然语言处理领域内的一项重要技术,已经取得了很多成果。最初命名实体识别主要是基于人工编写规则的方式[1]。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[2-4]、半监督式学习[5-7]和无监督式学习[8-10]。基于用户查询日志的命名实体识别和挖掘研究工作开展比较晚,相关工作比较少。Paşca[11]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[12]采用弱指导的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。

Guo 等人[13]针对命名实体的类别歧义性的问题,提出了一个弱监督式 LDA 模型框架来识别用户查询中的命名实体。Xu 等人[14]结合用户查询中的上下文和"click-through data"信息,从用户查询中来识别命名实体。Du 等人[15]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[13-15]主要关注如何从单条用户查询中识别出命名实体,而本文工作是从整个用户查询日志中挖掘命名实体。Jain[16]等人基于用户查询日志研究了开放式命名实体挖掘问题。

另外一部分相关工作就是基于图的半监督学习。很多基于图的半监督学习方法[17-19]都可以看成是基于图来估计一个函数 f, 函数 f 需要满足以下两个条件:(1)对于图上的已标记的节点的 f 取值, 同给定的标记尽可能地保持一致;(2)f 在图上是平滑的。基于图的半监督学习已经得到广泛的应用, 包括用户查询分类[20],Web 图片标注[21], 用户查询推荐[22]等。其中, Li 等人[20]利用用户查询日志中的"click-through data"来提升用户查询意图分类器的效果。本文中使用的模型同文献[20]中模型相似,区别是文献[20]中的模型是本文提出的模型的特例,即本文提出的模型是文献[20]中模型的泛化形式。

本文的主要贡献:(1)提出利用候选实体之间的关系来对候选实体进行排序,进而提升命名实体挖掘的效果;(2)采用一个基于二部图的半监督学习方法来对实体进行排序;(3)给出学习方法中的迭代过程所对应的优化框架。

1　问题描述

我们用 $Q = \{q_1, q_2, \cdots, q_n\}$ 表示用户查询数据集,用 $S = \{s_1, s_2, \cdots, s_k\}$ 表示隶属于类别 T 的种子实体集合。用 $E = \{e_1, e_2, \cdots, e_m\}$ 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:给定类别 T 和具有类别 T 的种子实体集合 S,目标是从用户查询数据集 Q 中挖掘得到具有类别 T 的实体集合 E。

已有研究工作[13]通过对用户查询日志分析表明,大约有 70% 的查询包含有单个命名实体,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,命名实体"the matrix",对于用户查询"the matrix movie"和"the matrix cast",相应的查询模板分别是"# movie"和"# cast"。

2　命名实体挖掘流程

基于用户查询日志的命名实体挖掘主要分为以下几个阶段:

(1) 获取种子实体对应的查询模板。利用目标类别下的一组种子实体,采用字符串匹配的方式从用户查询日志中得到与种子实体相对应的查询模板。这里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,从用户查询日志中匹配得到与查询模板相对应的候选实体(包括种子实体)。

(3) 获取候选实体对应的查询模板。利用获取得到的候选实体,从用户查询日志中得到候选实体所对应的查询模板。该阶段完成后,每个候选实体都会对应一组查询模板集合;同样,每个查询模板也会对应一组候选实体集合。

(4) 基于候选实体集合和查询模板集合,构建一个二部图。在该二部图上进行类别相关性的传播,即将种子实体的类别相关性传播到其他候选实体;最后每个候选实体都会被赋予一个类别相关性

的得分,根据这个得分对候选实体进行排序。具体排序方法在下节中进行详细说明。

3 基于二部图的排序方法

3.1 二部图的构建

我们基于候选实体集合和查询模板集合,构建一个二部图。将候选实体集合和查询模板集合看成是二部图的两个节点集合,如果某个候选实体和某个查询模板相互对应(即利用该查询模板通过遍历用户查询日志可以得到该候选实体),则在二部图上与它们相应的节点之间就会有一条边相连。如图1所示,用户查询日志中存在用户查询"the matrix cast",该查询可以分解为候选实体"the matrix"和查询模板"#cast",则在二部图中,表示候选实体"the matrix"和查询模板"#cast"的两个节点之间就会有一条边相连。

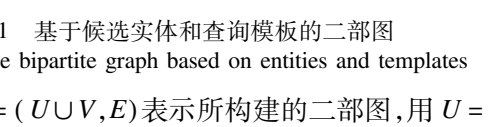

图1 基于候选实体和查询模板的二部图

Fig. 1 The bipartite graph based on entities and templates

用 $G=(U\cup V,E)$ 表示所构建的二部图,用 $U=\{u_1,u_2,\cdots,u_m\}$ 表示候选实体对应的节点集合,用 $V=\{v_1,v_2,\cdots,v_n\}$ 表示查询模板对应的节点集合,对于某条 (u_i,v_j),其权重定义如下:

$$w_{ij}=\begin{cases}1, & \text{if } u_i \text{ isconnected to } v_j,\\ 0, & \text{otherwise}。\end{cases} \tag{1}$$

二部图中的每个节点都有先验的分数值。对于候选实体,如果该实体是该类别的种子实体,则先验分数值设置为1;如果不是该类别的种子实体,则先验值设置为0。针对查询模板,每个节点的先验分数值设为0。

3.2 基于二部图的类别相关性传播

我们基于二部图在候选实体之间进行类别相关性的传播。传播过程用公式(2)表示:

$$\begin{cases}\boldsymbol{X}_{t+1}=\alpha\boldsymbol{M}_1\boldsymbol{Y}_t+(1-\alpha)\boldsymbol{X}_0,\\ \boldsymbol{Y}_{t+1}=\boldsymbol{M}_2\boldsymbol{X}_{t+1}。\end{cases} \tag{2}$$

其中 $\boldsymbol{M}_1=\boldsymbol{D}^{-\lambda}\boldsymbol{A}$,$\boldsymbol{M}_2=\boldsymbol{A}^T\boldsymbol{D}^{-(1-\lambda)}$,$\boldsymbol{A}$ 是二部图的邻接矩阵,$\boldsymbol{A}^{\mathrm{T}}$ 是 $\boldsymbol{A}$ 的转置矩阵;$\boldsymbol{D}$ 是对角矩阵,其(i,i)元素等于矩阵 $\boldsymbol{W}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 第 i 行元素之和;$\boldsymbol{X}_0$ 是候选实体的初始分值向量,$\boldsymbol{X}_t$ 和 $\boldsymbol{Y}_t$ 是经过 t 步迭代后候选命名实体和查询模板所对应的分数值的向量;α 是取值范围在0和1之间的权重变量;λ 也是取值范围在0和1之间的变量。公式(2)中的第一行体现了候选实体分数值的更新,第二行体现了查询模板的分数值的更新。

3.3 迭代过程的收敛性

因为最终需要的是候选实体的排序结果,所以更加关注候选实体对应的分数值向量 $\boldsymbol{X}_t$。本节证明迭代过程中产生的序列$\{\boldsymbol{X}_t\}$收敛。通过公式(2)可以得到如下迭代公式:

$$\boldsymbol{X}_{t+1}=\alpha\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{X}_t+(1-\alpha)\boldsymbol{X}_0。 \tag{3}$$

随着迭代的不断进行,可以得到

$$\boldsymbol{X}_{t+1}=(\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^{t+1}\boldsymbol{X}_0+(1-\alpha)\boldsymbol{X}_0\sum_{k=0}^{t}(\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^k。 \tag{4}$$

因为 $0<\alpha<1$,并且 $\boldsymbol{M}_1\boldsymbol{M}_2$ 特征值落在区间$[-1,1]$(矩阵 $\boldsymbol{M}_1\boldsymbol{M}_2$ 与矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{W}$ 相似,$\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{W}=\boldsymbol{D}^{-(1-\lambda)}\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{D}^{1-\lambda}$),所以得到

$$\begin{cases}\lim\limits_{t\to\infty}(\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^{t+1}=0,\\ \lim\limits_{t\to\infty}\sum\limits_{k=0}^{t}(\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^k=(\boldsymbol{I}-\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^{-1}。\end{cases} \tag{5}$$

最终得到

$$\boldsymbol{X}^*=\lim_{t\to\infty}\boldsymbol{X}_t=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{M}_1\boldsymbol{M}_2)^{-1}\boldsymbol{X}_0=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{D}^{-\lambda}\boldsymbol{W}\boldsymbol{D}^{-(1-\lambda)})^{-1}\boldsymbol{X}_0。 \tag{6}$$

这样可以不用迭代过程,而是直接计算得到收敛后的结果 $\boldsymbol{X}^*$。

3.4 优化框架

因为本文主要关注候选实体对应的分数值向量 $\boldsymbol{X}_t$,所以这里给出针对公式(3)所对应的优化框架。将该公式中的 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 分别替换成 $\boldsymbol{D}^{-\lambda}\boldsymbol{A}$ 和 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}^{-(1-\lambda)}$,得到如下公式:

$$\boldsymbol{X}_{t+1}=\alpha\boldsymbol{D}^{-\lambda}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}^{-(1-\lambda)}+(1-\alpha)\boldsymbol{X}_0。 \tag{7}$$

从公式(7)中可以看出,当 $\lambda=0$ 时,表示为个性化的随机游走传播过程;当 $\lambda=1$ 时,表示为带有先验的类似热传导的传播过程;当 $\lambda=0.5$ 时,表示为文献[19]中的传播过程。由此可见,参数 λ 不同的取值对应着不同的传播方式。

给出公式(7)所对应的优化框架。类似于文献[19]中的优化框架,公式(7)中的 $\boldsymbol{X}$ 对应的代价函数具有如下形式:

$$Q(\boldsymbol{X})=\frac{1}{2}\sum_{i,j=1}^{m}w_{ij}\left\|\frac{\boldsymbol{x}_i}{d_i^{1-\lambda}}-\frac{\boldsymbol{x}_j}{d_j^{1-\lambda}}\right\|^2+$$

$$\frac{\mu}{2}\sum_{i=1}^{m}\frac{1}{d_i^{1-2\lambda}}\|\boldsymbol{x}_i-\boldsymbol{x}_i^0\|^2。\tag{8}$$

在公式(8)中,w_{ij}是矩阵 $\boldsymbol{W}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 的第$(i;j)$分项,d_i是矩阵 D 的第$(i;i)$分项,$\boldsymbol{x}_i^0$向量 $\boldsymbol{X}_0$ 的第 i 个分项,μ 是正则化参数因子。因而公式(7)中的迭代过程可以转化成如下的优化问题:

$$\boldsymbol{X}^*=\arg\min Q(\boldsymbol{X})。\tag{9}$$

将代价函数 $Q(\boldsymbol{X})$对 $\boldsymbol{X}$ 求导,得到

$$\frac{\partial Q(\boldsymbol{X})}{\partial \boldsymbol{X}}\Big|_{\boldsymbol{X}=\boldsymbol{X}^*}=(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{D}^{-(1-\lambda)}\boldsymbol{X}^*+\mu\boldsymbol{D}^{\lambda}(\boldsymbol{X}^*-\boldsymbol{X}^0)=0,\tag{10}$$

如果令

$$\alpha=\frac{1}{1+\mu},\tag{11}$$

则得到

$$\boldsymbol{X}^*=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{D}^{-\lambda}\boldsymbol{W}\boldsymbol{D}^{-(1-\lambda)})^{-1}\boldsymbol{X}_0。\tag{12}$$

基于以上的推导过程,我们证明了公式(8)是公式(7)中迭代过程的等价优化框架。从公式(8)中可以看到,当 $\lambda=0.5$ 时,就是文献[19]中给出的优化框架形式。所以这里给出的是更加一般化的优化框架。

4　实验

4.1　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方法的有效性。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,其他信息暂时不考虑。

在实验中主要考虑了 3 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)等。我们尽量挑选受到关注度较高的命名实体作为种子实体,因为受到关注度较高的实体在用户查询日志中对应的查询模板也相对较多,这样在一定程度上保证了构建的二部图的连通性。种子实体列表如表 1 所示。

针对每个类别,返回前 500 个候选实体的列表,列表中的候选实体按照所赋予的类别相关性得分由高到低进行排列。请 3 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,本文采用 $P@N$(前 N 个结果的准确率)来评估算法性能。在公式(2)中设置 $\alpha=0.5$,$\lambda=0.2$。我们将在节 4.4 中分析参数 α 和 λ 的敏感度。

表 1　类别和种子命名实体
Table 1　The class and seed entities

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

4.2　参评方法

在实验中,将本文方法与 Paşca 的方法[11]进行对比。Paşca 的方法作为基准方法,简称 Paşca,而基于二部图的方法简称为 BGQER(bipartite graph based query entity ranking)。Paşca 方法主要是利用候选实体与类别的查询模板分布相似性来对候选实体进行排序的。然而 Paşca 方法仅仅考虑候选实体与类别之间的相似性,忽略了候选实体之间的关系。方法 BGQER 则是利用候选实体之间的关系来改善候选实体排序的效果。

4.3　实验结果及分析

基于上面所描述的实验数据和度量,我们对方法 Paşca 和 BGQER 进行了比较,结果见表 2。从实验结果中可以看出,在每个度量指标上,方法 BGQER 基本上都优于 Paşca 方法。利用候选实体之间的关系,确实可以改善实体排序的效果。

需要指出的是,尽管方法 BGQER 在改善实体排序效果方面,相对于 Paşca 的方法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”,“ps2”或者“xbox”等也经常同上述查

询模板共现。这样,就会将“ps1”、“ps2”和“xbox” 误判为“Game”类别。

表2 方法 BGQER 和 Paşca 性能($P@N$)

Table 2 The performance for methods BGQER and Paşca ($P@N$)

$P@N$	Movie		Game		Food		Average $P@N$	
	Paşca	BGQER	Paşca	BGQER	Paşca	BGQER	Paşca	BGQER
$P@25$	0.88	0.92	0.68	0.76	0.68	0.84	0.75	0.84
$P@50$	0.72	0.96	0.70	0.76	0.68	0.72	0.70	0.81
$P@100$	0.70	0.94	0.64	0.79	0.71	0.71	0.68	0.81
$P@150$	0.71	0.90	0.63	0.73	0.63	0.73	0.66	0.79
$P@250$	0.73	0.82	0.65	0.72	0.60	0.78	0.66	0.77
$P@500$	0.66	0.73	0.42	0.54	0.54	0.58	0.54	0.62

4.4 参数敏感度分析

在迭代过程中(请见公式(2)),有两个参数 α 和 λ。本小节以类别“Movie”为例,来分析 α 和 λ 的变化对候选实体排序结果的影响。

参数 α 是一个平衡因子,用来平衡候选实体的先验类别信息和从其他候选实体传播过来的类别信息之间的贡献。这里令 $\lambda=0.2$,α 取值从0到1,观察类别“Movie”的候选实体排序效果的变化,见图2。从图中可以看到,当 α 在区间$[0.1,0.9]$变化时,$P@N$ 的变化是比较平滑的。

图2 方法 BGQER 的 $P@N$ 随 α 变化趋势图

Fig. 2 The performance of BGQER with respect to α

参数 λ 控制着传播方式。当 $\lambda=0$ 时,传播过程采取的是个性化的随机游走方式;当 $\lambda=1$ 时,传播过程采取的带有先验信息的类似“热传播”过程方式;当 $\lambda=0.5$ 时,采取的就是文献[19]中的传播方式。这里令 $\alpha=0.5$, λ 取值从0到1,观察类别“Movie”的候选实体排序效果的变化,见图3。从图中可以看到,当 λ 在区间$[0.1,0.5]$变化时,$P@N$ 的变化基本上是平滑的。

图3 方法 BGQER 的 $P@N$ 随 λ 变化趋势图

Fig. 3 The performance of BGQER with respect to λ

5　结论

本文通过采用基于二部图的半监督排序方法,利用实体之间的关系来改善对实体的排序效果,以解决已有用户查询日志实体挖掘研究工作中的种子实体不充分的问题。我们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。

参考文献:

[1] LISA F RAU. Extracting company names from text [C]//Proceedings of the 7th Conference on Artificial Intelligence Applications. Washington: IEEE Computer Society, 1991:29-32.

[2] HAI LEONG CHIEU, HWEE TOU NG. Named entity recognition: a maximum entropy approach using global information[C]//Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 1-7.

[3] KOICHI TAKEUCHI, NIGEL COLLIER. Use of support vector machines in extended named entity recognition [C]//Proceedings of the 6th Conference on Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2002: 1-7.

[4] HOIFUNG POON, PEDRO DOMINGOS. Joint inference in information extraction[C]//Proceedings of the 22nd National Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2007:913-918.

[5] COLLINS MICHAEL, SINGER YORAM. Unsupervised models for named entity classification[C]//Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. [S. l.]:[s. n.], 1999:100-110.

[6] WHITELAW CASEY, KEHLENBECK ALEX, PETROVIC NEMANJA, et al. Web-scale named entity recognition[C]//Proceeding of the 17th ACM Conference on Information and KNOWLEDGE Management. New York: ACM Press, 2008:123-132.

[7] ETZIONI OREN, CAFARELLA MICHAEL, DOWNEY DOUG, et al. Unsupervised named-entity extraction from the web: an experimental study [J]. Artificial Intelligence, 2005, 165(1):91-134.

[8] ENRIQUE ALFONSECA, SURESH MANANDHAR. An unsupervised method for general named entity recognition and automated concept discovery[C]//Proceedings of the 1st International Conference on General WordNet. [S. l.]:[s. n.], 2002:1-9.

[9] DOUG DOWNEY, MATTHEW BROADHEAD, OREN ETZIONI. Locating complex named entities in web text [C]//Proceedings of the 20th International Joint Conference on Artifical Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, 2007:2733-2739.

[10] RICHARD EVANS. A framework for named entity recognition in the open domain[C]//Proceedings of the Recent Advances in Natural Language Processing. [S. l.]:John Benjamins Publishing Company, 2003:137-144.

[11] MARIUS PAŞCA. Weakly-supervised discovery of named entities using web search queries[C]//Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2007:683-690.

[12] 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.

[13] GUO Jiafeng, XU Gu, CHENG Xueqi, et al. Named entity recognition in query[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2009:267-274.

[14] XU Gu, YANG Shuanghong, LI Hang. 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. New York: ACM Press, 2009:1365-1374.

[15] DU Junwu, ZHANG Zhimin, YAN Jun, 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. New York: ACM Press, 2010:765-766.

[16] ALPA JAIN, MARCO PENNACCHIOTTI. Open entity extraction from web search query logs[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010:510-518.

[17] AVRIM BLUM, SHUCHI CHAWLA. Learning from labeled and unlabeled data using graph mincuts[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001:19-26.

[18] ZHU Xiaojin, GHAHRAMANI ZOUBIN, LAFFERTY JOHN. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 20th International Conference on Machine Learning. [S. l.]: AAAI Press, 2003: 912-919.

[19] ZHOU Dengyong, BOUSQUET OLIVIER, LAL THOMAS NAVIN, et al. Learning with local and global consistency[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2003:321-328.

(下转第 42 页)

可考虑利用现有的资源,通过领域适配,获得一批训练样本,尽量减少人工标注的代价;可综合考虑多个特征,以减少噪音对性能的影响;还可考虑利用微博本身之外的背景知识(例如相关微博)来增强上下文,以获得更丰富、更鲁棒的特征。

6 总结与展望

本文提出构建基于语义分析的微博搜索以克服现有微博搜索的不足。微博语义搜索从大量的微博中提取出信息点,以搜索和分类浏览的方式允许用户快捷地访问这些信息点。不同于现有的微博搜索,微博语义搜索对单个和一组微博做了深度的语义分析,提供了超越关键字搜索的高级搜索(例如搜索事件和搜索观点)和导航功能,并在一定程度上支持商业智能。本文分析了语义微博搜索的主要挑战和对策,并介绍了其参考实现和相关的关键技术。特别是,本文以语义角色标注为例,讨论了面向微博的语义分析技术如何克服微博本身特点带来的挑战。我们已经构造一个针对英文的微博语义搜索原型系统,下一步将扩展这一原型系统以支持中文、日文等其他语言。

参考文献:

[1] 周明,刘晓华,蒋龙,等. 语义分析和搜索教程[R]. 北京:微软亚洲研究院,2010.

[2] LIU Xiaohua, ZHOU Ming, LI Kuan. SRL for news tweets[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010.

[3] 曹国伟. 新浪微博注册用户数已经突破1.4亿[EB/OL]. (2011-05-13)[2011-06-09]. http://news.ccidnet.com/art/3709/20110513/2389053_1.html.

[4] Jameson Berkow. FP tech desk: twitter now fastest-growing search engine[EB/OL]. (2010-07-08)[2011-05-24]. http://business.financialpost.com/2010/07/07/fp-tech-desk-twitter-now-fastest-growing-search-engine/

[5] Kelly Ryan. Twitter study reveals interesting results about usage[R]. San Antonio, Texas: Pear Analytics, 2009.

(编辑:许力琴)

(上接第37页)

[20] LI Xiao, WANG Yeyi, ALEX ACERO. Learning query intent from regularized click graphs[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2008:339-346.

[21] RUI Xiaoguang, LI Mingjing, LI Zhiwei, et al. Bipartite graph reinforcement model for web image annotation[C]//Proceedings of the 15th International Conference on Multimedia. New York: ACM Press, 2007:585-594.

[22] DENG Hongbo, LYU MICHAEL R, IKING RWIN. A generalized co-hits algorithm and its application to bipartite graphs[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009:239-248.

(编辑:许力琴)