文章编号:1006-2467(2011)02-0164-04

基于转移学习的命名实体挖掘技术

翟海军[1,2], 郭 勇[3], 郭嘉丰[2], 程学旗[2]
(1. 中国科学技术大学 计算机科学与技术学院, 合肥 230027;
2. 中国科学院计算技术研究所, 北京 100190; 3. 北京系统工程研究所, 北京 100101)

摘 要: 研究了针对大规模查询日志中丰富的命名实体的挖掘技术,通过利用 Wikipedia 数据,结合转移学习方法构建目标类别的分类器.该技术很好地利用了监督学习的优越性能以提高查询日志中命名实体挖掘的准确性,同时也解决了监督学习方法中大规模标注的问题.实验结果表明,基于转移学习的命名实体挖掘技术具有优越的命名实体挖掘性能.

关键词: 转移学习; 命名实体挖掘; 正例学习
中图分类号: TP 391.43 **文献标志码**: A

A Named Entity Mining Method Based on Transfer Learning

ZHAI Hai-jun[1,2], *GUO Yong*[3], *GUO Jia-feng*[2], *CHENG Xue-qi*[2]
(1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China; 2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 3. Beijing Institute of System Engineering, Beijing 100101, China)

Abstract: This paper addresses the problem of mining named entities from query logs. A novel scheme was introduced based on transfer learning, which trains classifier for target category by leveraging Wikipedia data source. In this way it can greatly make use of supervised learning and also deal with the large scale labeling problem. The experiment results show the effectiveness of the novel scheme based on transfer learning.

Key words: transfer learning; named entity mining; one class learning

随着互联网络的急剧发展,搜索引擎成为人们获取日常所需信息的重要工具.长期以来,各大商业搜索引擎聚积了大量的用户查询行为日志.用户查询日志作为一类富含大众智慧的海量数据资源,成为了数据挖掘领域广泛关注的研究对象.从查询日志中获取的各种知识不仅可为信息检索领域所用,还可成为机器翻译、自然语言处理等领域的基础.

以往对命名实体挖掘的研究主要集中在文本领域中[1-3],至今已有近20年的发展历史.命名实体挖掘作为自然语言处理领域的一项重要技术,已经取得了很多成果.早期命名实体挖掘的技术通常依赖于人工指定规则.近年来,机器学习的方法也开始被应用于命名实体挖掘,包括监督学习[1],半监督学习[2]和无监督学习[3].

Wikipedia 作为一类新型的数据资源近年来在工业界和研究领域得到了极大的关注,其为一个聚合大众智慧的知识库,该知识库是大量用户长期编辑的结果,它包含大量的命名实体以及相应的描述

收稿日期:2010-03-02
作者简介:翟海军(1982-),男,博士生,湖南永州市人,研究方向:信息检索. E-mail:zhhj@mail.ustc.edu.cn.
郭 勇(联系人),男,研究员,博士生导师,电话(Tel.):010-66356580;E-mail:guoy@public.bise.ac.cn.

文档，且每个文档都有人工标注的详细的类别信息.

与文本领域中的命名实体挖掘不同，用户查询通常都很简短(往往只有2～3个词)，并且不具备严格的语法，语义很模糊，因此，文本领域中的命名实体挖掘技术不能直接有效地应用到查询上. 这给基于用户查询的命名实体挖掘的研究工作提出了新的挑战. 已有的研究表明，用户查询数据具有一些独有的分布特性，分析这些特性有助于从用户查询日志中挖掘命名实体. Pasca[4]提出了一种利用查询模板从用户查询日志中挖掘命名实体的技术，其将查询分解为两部分：某个类别的实例(即命名实体)和查询模板(即查询上下文). 在此基础上，通过人工给定目标类别下的一组种子(命名实体)作为指导，估计目标类别下的查询模板的分布，从用户查询日志中挖掘目标类别下的新命名实体.

区别于已有的工作，本文提出了一种新颖的基于转移学习的命名实体挖掘技术，该技术借助已有的知识库Wikipedia数据源指导基于查询日志的命名实体挖掘，同时提出了一种转移学习技术解决Wikipedia数据和查询日志数据特征分布的差异性问题，该技术能够充分利用已有的知识库Wikipedia数据结合转移学习方法来构建查询日志上的分类器，该分类器具有监督学习的优越性能，可以高效地应用到查询日志领域.

本文收集了来自一个商业搜索引擎的15×10^6条真实用户查询数据作为实验数据，实验结果表明，基于转移学习的命名实体挖掘方法在实体挖掘性能上显著优于其他作为对比的方法.

1　命名实体挖掘框架

给定一组目标类别，命名实体挖掘的目标是从大规模用户查询日志中挖掘出目标类别相关的命名实体. 本文的命名实体挖掘框架可以分为3个阶段：

(1) 构建目标类别的训练语料. 根据给定的目标类别，在Wikipedia数据中匹配所有标注了该类别的文档，比如目标类别为Person，可以通过在Wikipedia数据中匹配所有标注类别为Living people的文档，来获取目标类别Person的训练语料(命名实体及其Wikipedia描述文档). 此外，每篇匹配到的文档的标题作为下一步处理的种子命名实体. 通常，可以通过简单地理解目标类别的语义在Wikipedia数据中找到合适的类别标签来描述该目标类别，确保了训练语料的构建可以简单进行.

(2) 针对第1阶段中匹配到的种子命名实体获取相应查询描述文档，然后采用转移学习方法训练分类器. 即对每个种子命名实体，通过遍历整个查询日志，收集所有包含该种子命名实体的用户查询. 本文将查询分解为2部分：命名实体和查询模板. 例如命名实体harry potter，对于查询harry potter poster，相应的查询模板为#poster. 通过组合该命名实体的所有查询模板，得到该命名实体的查询描述文档. 可以像普通文档一样来处理查询描述文档. 因此，对所有种子命名实体，通过遍历整个查询日志，匹配得到相应的描述文档集合，结合第1阶段中从Wikipedia数据获取得到的训练语料，就可针对每个目标类别采用转移学习方法训练相应的分类器.

(3) 获取候选命名实体，并采用第2阶段中训练好的分类器来挖掘候选命名实体. 将第2阶段中所获取的所有种子命名实体的查询模板作为目标串，遍历整个查询日志通过匹配获取候选命名实体及相应的查询描述文档. 利用前面学习得到的每个目标类别的分类器，来分类候选命名实体，从而挖掘得到各个目标类别下的新命名实体.

2　转移学习方法

由于Wikipedia数据和查询日志数据特征分布存在很大的差异，不能直接采用Wikipedia数据构建分类器来挖掘查询日志中的命名实体，因此提出了3种不同的转移方法，尝试解决不同领域上数据特征分布的差异性.

2.1　问题定义

对每个目标类别C，有Wikipedia文档集，记为

$$D_{\mathrm{P}}=\{d_{\mathrm{P},1},d_{\mathrm{P},2},\cdots,d_{\mathrm{P},n}\}$$

及查询日志文档集，记为

$$D_{\mathrm{L}}=\{d_{\mathrm{L},1},d_{\mathrm{L},2},\cdots,d_{\mathrm{L},m}\}$$

其中，m、n为文档集合的大小. 本文采用词袋模型表示文档，每篇文档都可以用词空间的一个向量表示. D_{P}对应的词空间记为

$$W_{\mathrm{P}}=\{w_{\mathrm{P},1},w_{\mathrm{P},2},\cdots,w_{\mathrm{P},s}\}$$

D_{L}对应的词空间记为

$$W_{\mathrm{L}}=\{w_{\mathrm{L},1},w_{\mathrm{L},2},\cdots,w_{\mathrm{L},t}\}$$

其中，s、t为词集合的大小.

2.2　简单转移方法

简单转移方法(STM)的基本假设为：对每个C的文档集D_{P}和D_{L}，其对应的词空间W_{P}和W_{L}具有相同的分布，并且$W_{\mathrm{P}}=W_{\mathrm{L}}$.

2.3　基于重叠特征的转移方法

基于重叠特征的转移方法(Overlap Based Transfer Method，OBTM)的基本假设为：对每个C的文档集D_{P}和D_{L}重叠的特征为

$$W_O = W_P \cap W_L$$

2.4 基于差异性的转移方法

基于差异性的转移方法(Difference Based Transfer Method, DBTM)的基本假设为:对每个 C 的文档集 D_P 和 D_L 重叠的特征为

$$W_O = W_P \cap W_L$$

具有不同的分布.基于该假设本文提出了FCE(Frequently Co-occurring Entropy)来进行特征选择.FCE的定义为:

$$\mathrm{FCE}(w) = \lg(P_{D_P}(w)P_{D_L}(w) + \alpha) \quad (1)$$

式中:$P_{D_P}(w)$和$P_{D_L}(w)$分别为词 w 在文档集合 D_P 和 D_L 中出现的概率,

$$P_{D_P}(w) = \frac{N_{D_P,w} + \beta}{|D_P| + 2\beta} \quad (2)$$

$$P_{D_L}(w) = \frac{N_{D_L,w} + \beta}{|D_L| + 2\beta} \quad (3)$$

$N_{D_P,w}$和$N_{D_L,w}$分别为在文档集合 D_P 和 D_L 中包含词 w 的文档数目,$|D_P|$、$|D_L|$分别为文档集合 D_P 和 D_L 的大小;α、β为防止计算溢出而设置的平滑参数,实验中分别设置为 2×10^{-4}和 10^{-4}.由FCE的定义可知,该定义偏向于选择在2个文档集合中都很一般化的特征.

3 实验结果及其分析

本文收集了来自一个商用搜索引擎的真实用户查询数据,通过实验来验证本文方法的有效性.将3种转移学习方法与直接在 D_L 语料上训练的分类器(Directly Training Classifier, DTC)进行比较.需要特别指出的是,本文的实验采用 One-class SVM[5] 作为基本分类器.

3.1 实验数据

本文所采用的实验数据集包含了从 MSN Live Search 搜索引擎随机采样得到的 15×10^6 条真实用户查询数据.该数据集包含6 623 961个不同的用户查询,用户查询的平均长度为2.423个英文单词.本文中,用户查询被认为是互相独立的,不考虑它们是否来自同一用户.此外,本文下载了整个 Wikipedia 数据,共计2 948 000多篇英文文档.

本实验共考虑了5个不同语义类别,综合考虑了不同粒度的类别,其中包括命名实体挖掘和挖掘中经常会涉及的粒度较粗的类别.对于每个目标类别,从 Wikipedia 数据集合中随机地选择5 000个文档作为该目标类别的训练样本.

3.2 评价方法

为了确保评价的可靠性,对每个目标类别,采用文献[6]中挖掘结果的前250个作为候选命名实体集合进行测试(丢弃描述文档长度少于50个词的候选命名实体).通过人工标注这些候选命名实体,对属于该类别的结果命名实体判定为 true,否则判定为 false.另外,为了确保结果的准确性,本实验同时召集了3个研究生做评价,对评价结果不一致的命名实体,采用投票的方式确定命名实体最终的评价结果(多于2个人同时认同的结果将被接受).基于上述人工评价后的结果数据,采用准确率 A 来评价分类器的性能:

$$A = N_c/N_e \quad (4)$$

式中:N_c 为正确预测的样本数;N_e 为总测试样本数.

3.3 实验结果及其分析

采用上节描述的实验数据和度量,对4种方法进行比较.实验结果如表1所示.由表1可见:

(1) 除类别 Location 外,DTC 的分类性能都很差($A\leqslant0.211$).这是由于用户查询中命名实体往往语义很模糊,同一个命名实体串在不同查询中可能指代不同类别的实体.比如,命名实体 harry potter,在用户查询 harry potter poster 中为电影 harry potter;而在用户查询 harry potter author 中为书 harry potter;直接从查询日志中获取命名实体的查询描述文档作为训练数据会含大量噪音,因此,利用查询日志数据训练分类器很难获得很好的分类性能.

(2) 在所有类别下,STM 的准确率都低于 DTC.这是由于查询日志数据的词特征分布与 Wikipedia 数据的词特征分布存在很大的差异,基于 Wikipedia 数据训练得到的分类器不能直接应用到查询日志数据.

(3) DBTM 的性能远远优于其他3个方法.DBTM 在各类别上的平均准确率相对于 OBTM、STM 和 DCT 分别提高了47%,672%和228%.这表明查询数据具有自己独有的分布特性,正确地分析这些特性有助于准确地挖掘出命名实体,同时也说明了 Wikipedia 数据作为外部知识的指导作用.

表1 各方法的准确率

Tab.1 Accuracy of different methods

方法	Location	Person	Movie	Music	Game	平均值
DTC	0.828	0.186	0.052	0.009	0.211	0.257
STM	0.388	0.013	0.008	0.011	0.167	0.117
OBTM	0.984	0.762	0.989	0.005	0.130	0.574
DBTM	0.995	0.783	0.995	0.628	0.817	0.844

DBTM 能够取得更好的挖掘效果的原因，主要是由于 DBTM 通过考虑查询日志数据的词特征与 Wikipedia 数据的词特征分布上的差异，更加准确地提取出 2 个领域之间共同的词特征，从而实现了从 Wikipedia 数据到查询日志数据的有效转移学习. 表 2 中以 Person 为例，分别给出了其对应的 D_P 和 D_L 下前 10 个词(分别根据词 w 在 D_P 和 D_L 中出现的概率 $P_{D_P}(w)$ 和 $P_{D_L}(w)$ 的从大到小取前 10 个，分别记为 PT(i) 和 LT(j)，$i,j=1,2,\cdots,10$)，以及 D_P 和 D_L 共现词集合中取前 10 个词(按词在 D_P 和 D_L 中共现的信息商从大到小取前 10 个记为 CT(k)，$k=1,2,\cdots,10$). 表 2 在一定程度上反映了 CT(k)($k=1,2,\cdots,10$)相对于 PT(i)和 LT(j)类别相关性更加明确并且更加一般化，而这些一般化的特征正是转移学习成功的基础.

表 2　Person 类别下 PT(i)，LT(j)和 CT(k)对比

Tab. 2　Comparison of PT(i), LT(j) and CT(k) for Person

i,j,k	PT10	LT10	CT10
1	born	pictures	bio
2	links	www	office
3	time	lyrics	webpage
4	career	photos	scholarships
5	life	pics	works
6	american	biography	picture
7	made	music	song
8	university	videos	pics
9	school	video	son
10	world	bio	family

图 1 所示为 DBTM 在各类别下的准确率随选取特征比例变化的曲线. 由图可见，初始时 DBTM 的准确率随特征比例的增加逐渐增大，并到达最大值；再继续增大选取特征比例，准确率逐渐下降，这是由于选用更多的特征会引入噪音. 该图说明了特征选择在转移学习中的重要性，同时也说明了 DBTM 方法的有效性.

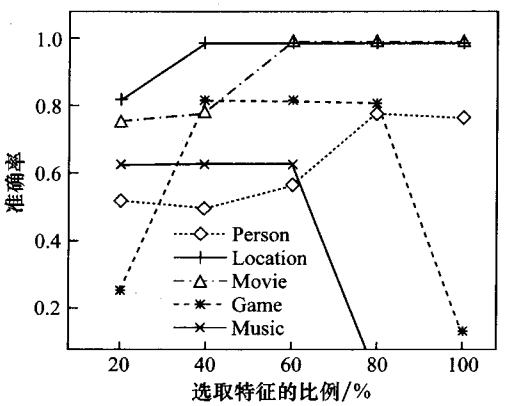

图 1　DBTM 在各类别下的准确率随选取特征比例变化的曲线

Fig. 1　Accuracy curves of DBTM vs. proportion of features selected

4　结　语

本文提出了一种新颖的基于转移学习的命名实体挖掘框架，该框架通过利用 Wikipedia 数据，结合转移学习的方法来构建目标类别的分类器，很好地利用了监督学习的优越性能，同时也解决了监督学习方法中大规模标注的问题. 实验结果表明，基于转移学习的命名实体挖掘方法具有优越的命名实体挖掘性能. 下一步期望在中文查询日志数据集上验证本文所提出方法的可行性，同时探索挖掘的大量命名实体在检索方面的应用.

参考文献：

[1] Andrew B, John S, Eugene A, *et al*. NYU: Description of the MENE named entity system as used in MUC-7[EB/OL]. (2001-01-12)[2010-01-26]. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[2] Alessandro C, Velardi P. Unsupervised named entity recognition using syntactic and semantic contextual evidence[J]. **Computational Linguistics**, 2001, 27(1): 123-131.

[3] Richard E. A framework for named entity recognition in the Open Domain[C]//**Proc Recent Advances in Natural Language Processing**. Philadelphia: John Benjamins Publishing Company, 2003: 267-276.

[4] Pasca M. Weakly-supervised discovery of named entities using web search queries[C]//**CIKM 07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management**. New York: ACM, 2007: 683-690.

[5] Chang Chih-chung, Lin Chih-jen. LIBSVM : A library for support vector machines[CP/OL]. (2001-09-13)[2010-02-06]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] 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[C]//**全国第十届计算语言学学术会议**. 烟台：中文信息学会，2009：563-569.