中文网页搜索日志中的特殊命名实体挖掘

张 磊[1], 王 斌[1], 靖红芳[1], 吴丽辉[2]

(1. 中国科学院 计算技术研究所,100190 北京, zhanglei01@ict.ac.cn; 2. 中国科学院办公厅 信息化工作处,100864 北京)

摘 要: 利用少量具有类别信息的种子词,结合特征选择技术来提取每个类别的特征信息;再利用这些特征信息,结合文本分类等数据挖掘技术来提取特殊命名实体.过程中只有构造种子词的环节需要人工辅助,其他环节均实现自动处理. 实验证明,该系统和方法能够从查询日志中挖掘出高质量的命名实体列表,6个类别上识别结果的平均P@500达到了77%.系统的自动化程度和识别的效果均达到实用的要求.

关键词: 特殊命名实体;数据挖掘;信息检索;网页搜索日志

中图分类号: TP391.3 **文献标志码:** A **文章编号:** 0367-6234(2011)05-0119-04

Mining special named entities from Chinese Web search query logs

ZHANG Lei[1], WANG Bin[1], JING Hong-fang[1], WU Li-hui[2]

(1. Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, China, zhanglei01@ict.ac.cn;
2. Informationization Division, General Office of Chinese Academy of Sciences, 100864 Beijing, China)

Abstract: Recognizing special Named Entities from Chinese Web search query logs is very useful for practice. In this paper, a few seed words and feature selection techniques are used to extract features from the query logs first, and then the features are combined into a text classifier for NE recognition. Experiments on a handred milllion queries to mine six types of Special Named Entities show that it is able to find high quality Sepecial Named Entities from Chinese Web search query logs. The average P@500 of the six classes achieves 77%.

Key words: special named entities; data mining; information retrieval; Web search query logs

特殊命名实体在中文查询中占很大比例. Dou Shen 等[1-2]研究表明,在每天的网页搜索查询里面,有2% ~4%的查询由单独的人名组成. Javier Artiles 等[3]研究表明,大约30%的查询里包含人名.在手工标注的76 717条查询里面,人名出现6 245次,占总查询数的14.83%.从查询日志中自动挖掘这些实体能够获得更具时效性、更加全面的结果.

传统的命名实体(Named Entity,NE)识别多数是在普通文本上进行的,其识别主要以人名、地名、机构名为主.识别算法大致可以分为基于规则的方法和基于统计的方法.一般而言,基于规则的方法准确率较高,但是需要耗费大量的人力物力,可移植性不好,而基于统计的方法的健壮性和灵活性更好,且不需要太多的人为干预,当然它需要大规模的语料库训练.基于统计的方法往往使用统计模型和机器学习算法,如隐马尔可夫模型、基于记忆的学习算法(Memory-based learning)、支持向量机(Support Vector Machine,SVM)等[4-6].查询日志中蕴含了大量命名实体特别是特殊命名实体信息,然而上述方法并没有在查询日志上进行过尝试.另外查询日志和传统文本也不同,查询文本具有简短、缺少上下文、多次重复出现等特点,上述方法也很难直接应用于查询日志中的命名实体识别.

文献[7]用提取模板的方法尝试了在英文查询日志上进行命名实体识别,其基本思想是先针对每个类别设置一批已知的命名实体作为种子(Seed)来提取模板,然后根据模板来识别命名实

收稿日期: 2011-03-04.

基金项目: 国家自然科学基金资助项目(60603094);国家重点基础研究发展计划(973)资助项目(2007CB311103);国家高技术研究发展计划(863)资助项目(2006AA010105);北京市自然科学基金资助项目(4082030).

作者简介: 张 磊(1983—),男,硕士;
王 斌(1972—),男,副研究员,博士生导师.

体并计算其得分. 该方法在 10 个类别上的平均 P@250指标为 80%. 然而,由于中文的复杂性,上述方法并不能直接用于中文特殊命名实体的识别. 本文主要基于查询日志来识别中文特殊命名实体.

1 算法和系统描述

1.1 系统描述

如图 1 所示,系统首先对网页搜索日志进行预处理,包括 2 个步骤:“切分”和“过滤”,然后采用最大匹配中文切分算法来分离 tags 和 NE. 其中 tags 为 NE 的上下文,如查询“千千阙歌试听”中,“千千阙歌”为 NE,“试听”为 tags. 在分离中对不成词的字符串不做切分,因为它们可能是候选的命名实体. 构建切分词典的方法是:在搜索日志上执行 tags 选择步骤选出质量较高的 tags,加上手工构造的“种子词表(seeds)”形成词典.

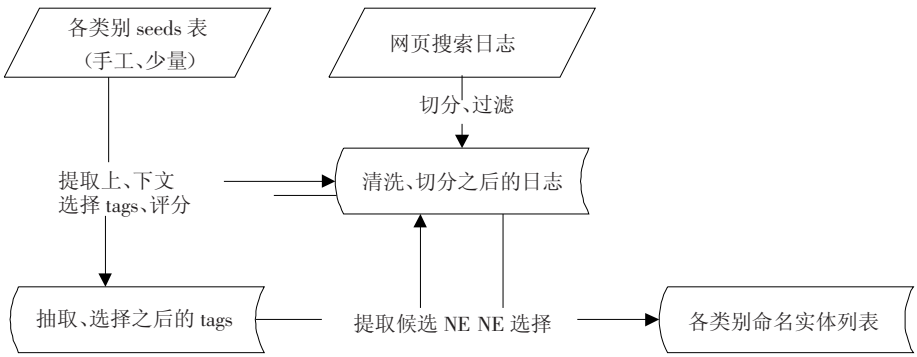

图 1 网页搜索日志中挖掘特殊命名实体的流程图

最后利用采用最大匹配中文切分算法来分离 tags 和 NE 选出的高质量 tags,先从预处理之后的日志中提取和它们共现的字符串作为候选 NE;然后执行评分算法对候选 NE 进行评分,得到按得分降序排列的每个预设类别的 NE 列表.

1.2 Tags 的抽取和选择

1.2.1 Tags 的权重计算算法

借用关联规则挖掘的支持度和置信度等[8]概念来设计 tags 的权重计算算法. 首先计算统计量:

1) SF_c (Seed Frequency in Class c):即“种子词频率”.

2) ICF(Inversed Class Frequency):对于一个 tag,它所出现过的类别个数的倒数.

3) TF_c (Tag Frequency in Class c): tag 和类别 c 的 seeds 共现的次数.

分析表明, SF_c、ICF 和 TF_c 都和 tag 的权重正相关,于是可以用乘法来综合它们的值为

$$w_{tc} = f_1(SF_{tc}) \times f_2(ICF_t) \times f_3(TF_{tct}). \quad (1)$$

经过实验对比发现, f_1 和 f_2 取线性函数, f_3 取对数函数的时候,式(1)能获得不错的识别效果,具体地, $f_1(SF_{tc}) = SF_{tc}$, $f_2(ICF_t) = ICF_t$, $f_3(TF_{tct}) = \log(TF_{tc} + 1)$.

1.2.2 基于分布的单类别特征选择算法

有的 tags 可能仅仅因为出现的次数多而对多个类别都权重很高,这种 tag 对于分类而言区分度不高. 因此本文引入了基于分布的单类别特征选择框架(Single Class Distribution Based Feature Selection)[9]来设计更有针对性的特征选择算法.

基于分布的单类别特征选择框架利用了候选特征在类内、类间的分布,着重选择那些在一个类别内部分布均匀、其他类别内出现较少的那些特征. 其表达式为

$$\text{score}(t,c_i) = \text{VAC}(t,c_i) - \text{VIC}(t,c_i). \quad (2)$$

其中:

$$\text{VAC}(t,c_i) = \frac{1}{m-1}\sum_{j=1,j\neq i}^{m} \text{sign}(F(t,c_i) - F(t,c_j))w_j(F(t,c_i) - F(t,c_j))^2, \quad (3)$$

$$\text{VIC}(t,c_i) = \frac{1}{|c_i|-1}\sum_{d\in c_i}(F(t,d) - F(t,c_i))^2. \quad (4)$$

式中: t 为一个 tag; c_i 为第 $i(1 \leqslant i \leqslant N)$ 个类别; N 为类别个数; d 为一个 seed 和它周围出现的 seeds 组成的“文档”. 采用一定的平滑策略之后, 2 种分布函数的计算方式为

$$F(t,c) = \frac{\#t_c + 1}{|c| + \text{size\_}c}. \quad (5)$$

式中: $\#t_c$ 为标签 t 在类别 c 中出现的次数; $\text{size\_}c$ 为类别 c 中去重后的 tags 个数.

$$F(t,d) = \frac{\#t_d + 1}{|d| + \text{size\_}d}. \quad (6)$$

式中: $\#t_d$ 为标签 t 在该 seed 提取候选 tags 之后形成的 d 中出现的次数; $\text{size\_}d$ 为 d 中去重后的 tags

个数.

1.3　候选命名实体的选择算法

每一个候选 NE 和它的上下文 tags 可以看作是一个待分类的文本. 其中：NE_i 为第 i 个候选实体；score_{ni} 为在第 n 类上的得分；tag_{nj} 为第 n 类的第 j 个标签；W_{nj} 为它的归一化权重. 归一化权重是把一个类别的 tags 看成一个 M 维的向量，对每一个元素进行归一化. 其中：M 为每个类别选取的 tags 数目. $\#\text{NE}_i$ 和 $\#\text{tag}_{nj}$ 分别为 NE_i 和 tag_{nj} 出现的频率；$\#\text{NE}_{i\_}\text{tag}_{nj}$ 为 NE_i 和 tag_{nj} 共现的频率. 评分算法的公式为

$$\text{score}_{ni} = \sum_{j=1}^{J} W_{nj} \times f(\#\text{NE}_{i\_}\text{tag}_j). \qquad (7)$$

其中 $f(\#\text{NE}_{i\_}\text{tag}_j) = \log(\#\text{NE}_{i\_}\text{tag}_j + 1)$.

2　实验结果及分析

本文的查询日志主要来自新浪爱问搜索引擎和搜狐搜狗搜索引擎，具体信息如表 1 所示，其中独立的查询是指不重复的查询.

表 1　网页搜索日志来源和数量统计　　条

查询方式	查询总数	独立的查询总数
新浪日志	66 043 406	19 203 787
搜狗 2007 年 03 月	44 429 353	
搜狗 2006 年 08 月	21 426 661	
总计	131 899 420	

本文选取 Disease、TV、Music、Movie、Person 和 University 等多个类别来进行实验. 实验中 tag 的数目取为 50，采用前 N 个结果中正确的比例 P@N 来衡量效果. 实验中，对每类的识别结果进行了人工标注.

表 2　SCDBFS 结合关联规则的算法(50seeds,50tags)结果(P@N)

N	Disease	Movie	Music	TV	Person	University	Average
50	1.00	0.86	0.94	0.94	0.96	1.00	0.95
100	0.96	0.80	0.84	0.96	0.91	0.97	0.91
150	0.92	0.80	0.84	0.95	0.86	0.93	0.88
200	0.89	0.76	0.85	0.94	0.83	0.88	0.86
250	0.85	0.72	0.85	0.92	0.83	0.85	0.84
300	0.84	0.67	0.85	0.91	0.80	0.83	0.82
350	0.81	0.64	0.85	0.90	0.79	0.80	0.80
400	0.79	0.62	0.85	0.90	0.79	0.76	0.78
450	0.79	0.63	0.84	0.90	0.78	0.74	0.78
500	0.77	0.61	0.84	0.89	0.77	0.71	0.77

最好的结果是采用 SCDBFS 结合基于关联规则的 NE 识别算法时取得的. seeds 数目取 50 和 100 并未对结果带来太大影响，只不过 seeds 数目取 100 在 N 比较小的时候相对于 seeds 数目取 50 有一定的优势，如表 2、3 所示. TV 类的结果总是最好的，最差的结果出现在 movie 类. 观察发现 movie 类的实体和电子游戏名很容易混淆，导致结果不好.

表 3　SCDBFS 结合关联规则的算法(100seeds,50tags)结果(P@N)

N	Disease	Movie	Music	TV	Person	University	Average
50	0.98	0.94	0.92	0.98	1.00	1.00	0.97
100	0.98	0.85	0.92	0.95	0.96	0.98	0.94
150	0.97	0.79	0.88	0.93	0.96	0.92	0.91
200	0.94	0.74	0.87	0.92	0.95	0.89	0.88
250	0.90	0.70	0.86	0.91	0.92	0.85	0.86
300	0.87	0.67	0.85	0.89	0.90	0.80	0.83
350	0.85	0.63	0.83	0.89	0.89	0.77	0.81
400	0.83	0.60	0.83	0.89	0.88	0.75	0.80
450	0.81	0.58	0.83	0.89	0.88	0.73	0.79
500	0.80	0.58	0.83	0.89	0.85	0.70	0.77

图 2 反映了使用 SCDBFS 进行 tags 选择和直接按照 tags 权重选取 tags 的识别结果比较. 可以看出,在 N 比较小的时候,SCDBFS 能够带来精确度的提高.

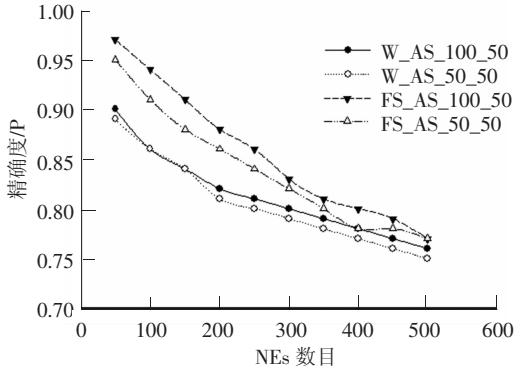

图 2 tags 选择算法对比

在最后的实体识别中,本文分别采用了关联规则方法(AS)和朴素贝叶斯方法(NB)进行了实验. 图 3 给出了这 2 种算法的比较结果. 采用关联规则方法相对于采用朴素贝叶斯方法有比较明显的优势,尤其在 N 变大的时候,两者之间的差距在拉大. 其原因可能是,基于 AS 的算法的本质是 tags 对候选 NE 的支持能力的一种累加方式,相对于一般分类算法更加适合此项任务.

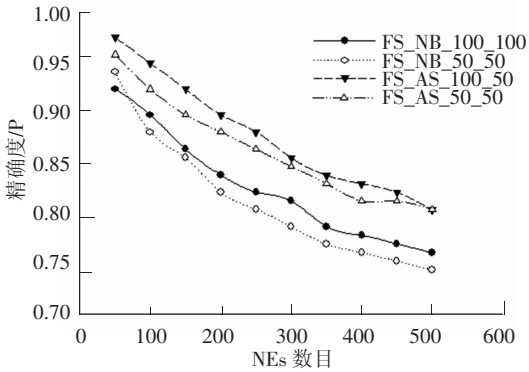

图 3 NE 识别算法效果对比

3 结 论

1) 中文查询日志中蕴含丰富的信息,可以为特殊命名实体挖掘服务;

2) 本文的方法能够达到较高的识别正确率,在6个类别上识别结果的平均P@500达到了77%,具有一定的实用价值;

3) 本文方法并不受限于类别体系,实际上它可以推广到其他类别命名实体的识别.

参考文献:

[1] DOWNEY D, BROADHEAD M, ETZIONI O. Locating complex named entities in Web text [C]//Proceedings of the 20th international joint conference on Artifical intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc. 2007: 2733 – 2739.

[2] SHEN D, WALKER T, Zheng Z, *et al.* Personal name classification in Web queries [C]//Proceedings of the international conference on Web search and web data mining. New York, NY: ACM, 2008: 149 – 158.

[3] ARTILES J, GONZALO J, VERDEJO F. A testbed for people searching strategies in the www [C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, 2005: 569 – 570.

[4] BROWN P, De SOUZA P, MERCER R, *et al.* Class-based n-gram models of natural language [J]. Journal Computational Linguistics, 1992, 18(4): 467 – 479.

[5] CHEN H, DING Y, TSAI S, *et al.* Description of the NTU system used for MET2 [C]//Proceedings of the 7th Message Understanding Conference. [S. l.]: [s. n.], 1998.

[6] JOACHIMS T. Text Categorization with support vector machines: Learning with many relevant features [J]. Springer, 1998, 1398(23): 137 – 142.

[7] PASCA M. Weakly-supervised discovery of named entities using Web search queries [C]//Proceedings of the 16th International Conference on Information and Knowledge Management. New York, NY: ACM, 2007: 683 – 690.

[8] HAN J, KAMBER M. Data Mining-Concepts and Techniques [M]. Third Edition. Beijing: Higher Education Press & Morgan Kaufmann Publishers. 2001: 227 – 228.

[9] 靖红芳. 文本分类中特征选择形式化研究 [D]. 北京:中国科学院研究生院硕士学位论文. 2009.

(编辑 张 红)