计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(6): 1298－1305, 2012

基于 URL 主题的查询分类方法

张 宇 宋 巍 刘 挺 李 生
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)
(yzhang@ir.hit.edu.cn)

Query Classification Based on URL Topic

Zhang Yu, Song Wei, Liu Ting, and Li Sheng
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract Many online resources contain crowd intelligence. Categorized website directory is one kind of resources constructed and maintained manually. It aims to organize websites according to a topical taxonomy. Based on the URLs with topical labels in website directory, a URL topical classifier could be designed. Together with pseudo relevance feedback technique and search engine query logs, an automatic, fast and efficient query topical classification method is proposed. In detail, the method combines two strategies. Strategy-1 is to predict a query's topic by computing the topic distribution among the returned URLs of a search system. Strategy-2 is to train a statistical classifier using the automatically labeled queries in query logs based on the topic of clicked URLs. The experimental results show that our method can achieve better precision compared with a state of the art algorithm and is more efficient for online processing. It has good scalability and can construct large scale training data from query logs automatically.

Key words query classification; URL classification; query log; pseudo relevance feedback; statistical learning

摘 要 互联网上很多资源蕴含人类群体智慧.分类网站目录人工地对网站按照主题进行组织.基于网站目录中具有主题标注的 URL 设计 URL 主题分类器,结合伪相关反馈技术以及搜索引擎查询日志,提出了自动、快速、有效的查询主题分类方法.具体地,方法为 2 种策略的结合.策略 1 通过计算搜索结果中 URL 的主题分布预测查询主题,策略 2 基于查询日志点击关系,利用具有主题标注的 URL,对查询进行标注获取数据并训练统计分类器预测查询主题.实验表明,方法可获得比当前最好算法更好的准确率,更好的在线处理效率并且可基于查询日志自动获取训练数据,具有良好的可扩展性.

关键词 查询分类;URL 分类;查询日志;伪相关反馈;统计学习

中图法分类号 TP391; TP18

查询分类意在将互联网用户提交的查询按照主题或意图分配到预先定义的类别体系中.查询分类具有广泛的应用前景,如计算广告[1]、搜索[2]、用户建模和查询意图分析[3],因此成为学术界与工业界关注的研究课题.查询分类任务的定义与传统的文档或网页分类类似,但处理的对象是用户提交的关键词查询,这使该任务面临新的挑战.首先,查询长度较短,缺少上下文特征,很大比例查询具有歧义;其次,分类通常被视为机器学习问题,基于机器学习的方法需要一定规模的训练数据.人工地对查询进

收稿日期:2010-11-25;修回日期:2011-04-25
基金项目:国家自然科学基金重点项目(60736044);国家自然科学基金面上项目(61073129);语言语音教育部-微软重点实验室开放基金项目(HIT.KLOF.2009020);"核高基"国家科技重大专项基金项目(2011ZX01042-001-001);国家"八六三"高技术研究发展计划基金项目(2011AA01A207)

行类别标注是十分耗时且昂贵的. 以上原因促使查询分类技术一方面侧重于利用各种手段和资源对查询进行扩展、丰富查询的特征表示,从而在文本分类的框架下解决该问题;另一方面致力于如何自动对查询进行类别标注,扩充训练数据. 查询分类的应用场景通常需要迅速地对用户请求进行响应,因此在线处理效率等因素是评价算法的重要参考. 查询分类的性能与效率之间往往存在冲突. 例如,基于搜索引擎搜索结果中的文本(标题、摘要或网页正文)对查询进行扩展的方法被证明是有效的策略. 然而,查询扩展过程需要在线地从搜索引擎获取搜索结果,并对其中的文本进行自然语言处理通常很耗时. 无法快速地回应用户请求,将影响类似方法在实际系统中的应用.

互联网上的网站分类目录人工地将网站根据主题进行组织,相当于对收录网站的 URL 进行了人工的主题标注. 基于这一资源,我们分别利用搜索结果中具有标注的 URL 在主题上的分布预测对应查询的主题,以及通过查询日志中查询与用户点击 URL 构成的对应关系对查询进行主题标注,并训练统计分类器进行查询分类. 两种分类策略都不需要在线进行复杂的文本处理,具有较好分类性能且兼顾在线处理的效率.

1　查询分类相关工作

Broder 按照意图将查询分为信息类、导航类和事务类[4]. 其应用主要是导航类查询的识别,该类查询通常具有明确的搜索目的,如找到某个网站的首页. 此类问题相关技术可参考文献[5]. 本文讨论基于主题的查询分类,即将查询按照主题进行划分. 主题可包括文化、体育、经济等.

查询分类的主要难点在于关键词查询包含的词汇少且有歧义. 仅针对查询本身,传统的文本分类方法难以构造足够的特征,面临特征稀疏问题. 解决该问题的思路是通过查询扩展,构建更丰富的查询特征. 例如利用搜索引擎返回的文档,从中抽取与查询共现的词汇扩展查询,构建更丰富的特征表示[6-7]. 此类方法通常要在线对搜索结果进行复杂处理,由此导致的延时将影响实际应用. Broder 等人将查询分类问题转化为文本分类问题,离线对搜索引擎索引的网页按照目标类别体系进行分类,在线通过投票方法获得各个类别的概率,可高效地对查询进行分类[1]. 然而,搜索引擎索引的网页数量巨大且不断更新,不同应用需要的类别体系也可能不同,对所有网页按照文本分类方法进行分类代价大且影响扩展性. 本文提出的 URL 主题结合伪相关反馈技术的方法受文献[1]启发,利用网页分类目录训练 URL 分类器,对搜索结果中的 URL 进行分类,进而预测查询的主题. 方法不需要预先对所有网页正文进行分类,而仅根据网页的 URL 自动进行分类.

查询分类的另一难点在于数据稀疏问题. 机器学习方法依赖于训练数据. 有学者利用文本分类的语料训练分类器[1,6]. 但用户构造查询所使用的词汇与撰写文章的语言习惯有较大区别. 还有学者利用手工标注类别的查询训练分类器[8],但使用的训练数据规模较小,有些测试的查询可能从未出现在训练数据中. 为缓解数据稀疏问题,Beitzel 等人利用自然语言处理中的选择倾向性方法[8],从未标注的查询日志扩充训练数据;Li 等人利用搜索引擎日志建立查询与被点击文档构成的点击图,将种子查询标签通过图传播给未标注的查询实例[9];Hu 等人利用少量种子获得 Wikipedia 中的相关概念,利用概念的类别实现查询分类[10]. 这些基于半指导学习的方法仍然需要人工干预. Shen 等人利用翻译模型将产品名称自动转化为"伪查询"作为训练语料用于产品类别分类[11],但其并未在一般的查询数据上进行验证. Fuxman 等人从浏览器插件中自动抽取出具有商业意图的查询构造训练语料预测查询意图[12]. 本文利用查询日志中查询与点击 URL 的对应关系,利用具有主题标注的 URL 对查询日志中的查询进行主题标注,并从中选择合适的查询训练统计分类器进行查询分类. 由于查询日志是海量的,可以自动构建大规模的训练数据,文中也讨论了训练数据规模对方法的影响.

2　基于 URL 主题的查询分类方法

2.1　基本思想

分类网站目录按照主题,组织与维护网站,相当于对收录网站的 URL 进行了主题标注,如 http://sports.sina.com.cn/通常被划分为体育类,那么以该 URL 为前缀的网页也多属于体育类. 来源于专业垂直网站的页面也具有明显的主题,如来自于 http://www.autohome.com/的网页属于汽车类. 本文利用分类网站目录,构建基于前缀的 URL 主题分类器,结合伪相关反馈技术与查询日志提出快速、高效的自动查询分类方法. 图 1 给出了方法的主要处理过程.

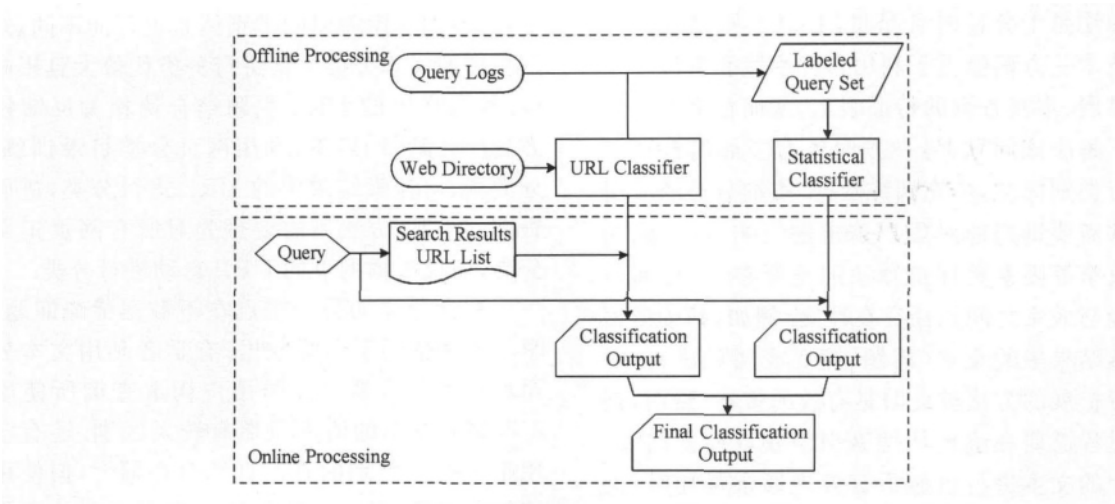

Fig. 1 Procedure of the URL topic based query classification method.

图 1 基于 URL 主题的查询分类方法框架

2.2 URL 主题分类器

本文通过 URL 建立查询与主题之间的联系. 由图 1 可见,URL 主题分类器是方法的重要组成部分,在离线训练与在线预测阶段均扮演重要角色. 本节介绍如何基于网站目录对 URL 进行分类.

2.2.1 分类网站采集

网站目录[①]是按主题来归类网站的目录结构. 网站目录并不下载或者索引网站下的网页,仅收录网站地址、名称、相关介绍等信息. 著名的网站目录(如 Yahoo 和 ODP[②] 网站目录)人工地收录网站以确保网站被正确划分到相应主题,保证收录网站的质量.

本文处理中文查询分类,采用 Yahoo[③],Google[④],Baidu[⑤]的中文分类网页目录,以 Yahoo 网页目录前 2 层作为标准,将 Google 和百度网页目录下的前 2 层节点手工映射到 Yahoo 的类别体系中,对部分子节点按照意义进行融合与增删. Google 目录中 3 层以上的节点保持与其父节点一致的主题. 最终主题体系第 1 层有六大类,第 2 层有 97 个小类. 表 1 给出了部分统计信息.

Table 1 Statistic of Category Hierarchy for URLs and Its Coverage on Query Logs

表 1 URL 类别体系相关统计及其对查询日志的概括情况

Entry	Number
# URL	55900+
# Category of first level	6
# Category of second level	97
Coverage of URLs in query logs/%	44.54
Coverage of queries in query logs/%	35.02

2.2.2 基于前缀匹配的 URL 分类器

网页目录中通常仅收录网站或网站主要版块的 URL,精确匹配 URL 难以保证足够的覆盖率. 采取回退法,利用前缀匹配对 URL 进行分类. 首先将整个 URL 与网页目录收录的 URL 进行匹配,如果完全匹配则返回网页目录中对应的主题,否则对 URL 后缀剪枝(按"/"划分),继续进行匹配,直至可以与网页目录中的 URL 匹配或者返回无法分类的标记. 例如:

http://sports.sina.com.cn/t/2010-06-06/23335022558.shtml

经过 3 次后缀剪枝,与网页目录中标注为"休闲娱乐/体育"的 http://sports.sina.com.cn/匹配,则该 URL 被标注为"休闲娱乐/体育".

对搜狗 2006 年 8 月的查询日志[⑥]的统计显示,44.54%的用户点击的 URL 可通过前缀匹配方式进行分类;若日志中查询对应的点击的 URL 中至少有 1 个可被分类,称该查询可被标注,可被标注的查询占整个查询数的 35.02%. 数据表明,相当比例

① http://en.wikipedia.org/wiki/Web_directory
② http://www.dmoz.org/
③ http://site.yahoo.com.cn/
④ http://www.google.com.hk/dirhp?hl=zh-CN
⑤ http://site.baidu.com/
⑥ ftp://ftp.labs.sogou.com/Data/SogouQ/SogouQ.tar.gz

的 URL 可被自动标注主题;查询日志包含大量可被标注的查询.基于以上观察,采取伪相关反馈技术利用搜索结果中 URL 的主题分布,以及利用有标注的查询训练统计分类器进行查询主题分类.

2.3　基于 URL 主题与伪相关反馈的查询分类方法

伪相关反馈[13],假设搜索系统返回的排序靠前文档与查询相关.这里假设排序靠前文档与查询具有相近主题.形式化地,将待分类查询 q 提交到搜索引擎获取 URL 列表 $U=\{u_1,\cdots,u_N\}$,目标类别体系为 $C=\{c_1,\cdots,c_K\}$,其中 N 表示搜索引擎返回的前 N 条结果,K 为主题数.分类过程为:对主题 k 计算 $p(c_k|q)$,即给定查询 q 其主题为 k 的概率,最终分类结果为 $c^*=\arg_{c_k}\max_{c_k}p(c_k|q)$.

为计算 $p(c_k|q)$,对每个主题 k 计算分数 $Score(c_k|q)$.首先利用基于前缀的 URL 主题分类器对 U 中所有 URL 进行分类,得到主题标注序列 $L=\{l_1,l_2,\cdots,l_N\}$,若 u_i 无法被分类,l_i 设为 null.$Score(c_k|q)$的计算公式为

$$Score(c_k\mid q)=\sum_{i=1}^{N}Indicator(c_k,l_i)\times pos(i),\quad(1)$$

其中,$Indicator(c_k,l_i)=\begin{cases}1, & \text{if } l_i=c_k\\ 0, & \text{else}\end{cases}$,$pos(i)=\begin{cases}1.0, & \text{if } 1\leqslant i\leqslant 10\\ \dfrac{1.0}{\log(i+1)}, & \text{if } 1>10\end{cases}$.$Score(c_k|q)$可视为对结果集合中所有属于主题 k 的 URL 对应的权值进行求和.$pos(i)$表示排序为 i 的 URL 的权重.通常 i 越小表明该文档排序越高,与查询越相关,应给予更高的权重.这里将排序前 10 个结果视为同等重要并给予最高权重,排序靠后结果的权重随 i 增加而减小.赋予前 10 个结果相同权重的另一个原因是我们使用商用搜索引擎结果,其首页顶部有时会植入广告或多媒体信息.这些内容对应的 URL 主题可能与查询并不完全相关,过分强调排序靠前的结果会使这种噪声的影响放大.最终 $p(c_k|q)$的计算公式为

$$p(c_k\mid q)=\frac{Score(c_k\mid q)}{\sum_{k=1}^{K}Score(c_k\mid q)}.\quad(2)$$

注意,如果搜索结果中没有任何 URL 能够被分类,则随机选择 1 个主题作为输出.

2.4　基于 URL 主题与查询日志的查询分类方法

查询日志是用户与搜索引擎的交互记录,包括用户输入的查询、点击的 URL、点击序列及相应的时间.查询日志中查询和点击的 URL 构成点击图,如图 2 所示:

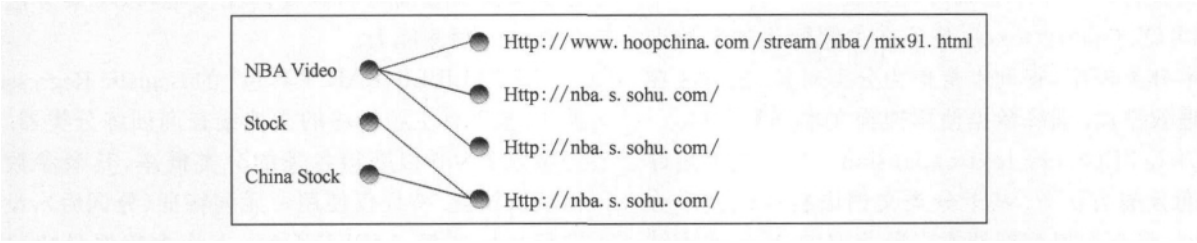

Fig. 2　An example of click graph based on query log.

图 2　查询日志中查询和 URL 构成的点击图示例

用户点击作为隐式反馈暗示查询与对应 URL 相关.很自然地,可认为查询主题和对应 URL 的主题也具有相关性.查询日志中每个条目对应 1 个〈查询,URL〉对.对每个 URL 用基于前缀的 URL 分类器进行分类.对于可分类的条目,用 URL 的主题标注对应的查询.我们利用被标注的查询训练统计分类器进行查询分类.同一个查询在日志中可能对应多个 URL(如图 2 中的查询"NBA Video"),这些 URL 可能属于不同的主题.此外,查询日志中具有噪声点击,因此需要选择具有高置信度的查询作为统计分类器的训练数据.本文假设一个查询只属于一个类,对查询 q,设其主题标签为 $C(q)=\{(c_1,n_1),\cdots,(c_K,n_K)\}$,$n_i$ 为查询 q 对应的所有 URL 中被标注为主题 c_i 的数目.最终,查询 q 可被标注为主题 c_i,需要满足:$N_L=\sum_j n_j>S$ 且 $\frac{n_i}{N_L}>T$,N_L 表示查询 q 具有的主题标签总数;T 表示主题 c_i 拥有的标签应大于整个主题标签的比例;S 和 T 作为阈值用来控制候选查询的频度和置信度要求.最终,被标注为主题 c_i 的查询将作为主题 c_i 的训练数据,训练基于统计的查询分类器.

选取查询作为训练语料时并没有将主题标签在点击图上进行多步传播,这是因为搜索引擎查询日志是海量的,当数据量足够大时,利用已有的有标注的 URL 已经可以对足够的查询进行主题标注.半指导学习方法也可以很轻易地融入,获取更多的有主题倾向的 URL 以及更多证据来获取有标注的查询,我们将此作为未来工作.

3 实验设置

3.1 评价方法

借用文本分类评价中采用的准确率、召回率和 $F1$ 值以及精确率进行评价. 对于类别 C,分类的结果可分为以下几种情况:

1) 原本为 C 类被划分为 C 类,数量记为 a;

2) 原本为非 C 类被划分为 C 类,数量记为 b;

3) 原本为 C 类被划分为非 C 类,数量记为 c;

4) 原本为非 C 类被划分为非 C 类,数量记为 d.

则针对类别 C 的评价指标为:准确率 $P=\frac{a}{a+b}$,召回率 $R=\frac{a}{a+c}$,$F1$ 值 $F1=\frac{2PR}{P+R}$. 3 个指标在各个类别的平均值称为宏平均,分别记作 $Precision$,$Recall$ 和 $F1$. 精确率定义为:$Accuracy=\frac{\text{正确分类的查询数}}{\text{总查询数}}$.

3.2 数据准备

主题分类体系采取搜狗提供的分类语料的类别体系①,共分为 10 个大类. 将 2.2.1 节基于网页目录建立的类别体系下第 2 层的 97 个小类手工映射到目标体系中,无法映射的舍弃,最终映射到目标体系中的 URL 共有 40 895 个. 从搜狗 2006 年 8 月份的查询日志中采样并标注了 2 240 个查询作为测试数据. 每个查询只标注一类作为标准答案. 使用搜狗 3 个月的查询日志用于选取具有主题标注的查询. 表 2 给出了使用的主题、映射的 URL 分布以及测试数据的统计:

Table 2 Statistic of Topic Hierarchy and Test Queries

表 2 测试用主题及测试查询统计

Topic	Economy	IT	Health	Sports	Travel	Education	Job	Art	Military	Auto
# URL	15 205	5 312	2 353	1 502	1 856	7 560	276	6 162	255	414
# Test Query	250	254	250	201	246	237	102	249	201	250

3.3 对照算法

实验将比较 4 个查询分类算法.

1) QT(query text):基于有主题标注的文档训练文本分类模型,查询本身作为分类对象. 该方法在线处理效率高. 训练数据使用搜狗文本分类语料.

2) QTE(query text expansion):当前效果最好的查询分类方法[6]. 基于分类文档语料训练文本分类模型,将查询提交到搜索引擎获取前 N 条搜索结果的摘要,将所有摘要合并成一个文档作为分类对象. 训练数据使用搜狗文本分类语料. 实验中尝试使用不同数目搜索结果扩展查询,设置为 80 时分类效果最好. 该方法在线效率低.

3) URL-PRF(URL based pseudo relevance feedback):本文 2.3 节介绍的方法. 该方法只需获取搜索结果中的 URL,不对文本进行处理,效率优于 QTE. 使用百度搜索引擎获取搜索结果.

4) URL-LOG(URL and query Log):本文 2.4 节介绍的方法,在线处理效率与 QT 相当. 希望验证,在对在线处理效率要求较高的情况下,使用有主题标注的文档以及有主题标注的查询分别作训练数据对统计查询分类器的影响. 使用搜狗公布的 3 个月查询日志用于基于 URL 主题的查询主题标注,并确保测试查询与训练查询无交集,以考察方法对新查询的预测能力.

采用 LIBLINEAR 工具包②的 Logistic Regression 模式,基于有主题标注的文档或查询训练分类器. 设置参数 s 为 0 以获得各类的分类概率,其余参数使用默认设置. 特征仅使用一元词特征(分词后),没有进行特征选择. LIBLINEAR 支持大数据量情况下的训练,在文本分类任务中可比支撑向量机模型获得相当甚至更好的分类效果和更快的训练速度[14].

4 实验结果与分析

本节首先介绍 2 种基于 URL 主题的方法的实验结果并讨论参数对系统的影响. 随后给出与其他方法的比较与分析.

4.1 URL-PRF 的实验结果

图 3 给出了 URL-PRF 使用不同 URL 数目时的表现. 算法随着 URL 增加表现呈上升趋势. 当使用 URL 数大于 50 后,结果基本趋于稳定. 这不难理解,当返回搜索结果较少时,可被分类的 URL 数更少,

① ftp://ftp.labs.sogou.com/Data/SogouQ/SogouC.reduced.atr.gz

② http://www.csie.ntu.edu.tw/~cjlin/liblinear/

因此导致算法的召回率较低,而且易受噪声的影响.我们使用商用搜索引擎获取搜索结果,结果页面通常会包括广告推送或其他嵌入的多媒体结果,它们对应的 URL 主题有可能对查询的主题预测造成影响.当使用 URL 数目较多时,URL 集合的主题分布趋于稳定.统计显示使用 100 条返回结果时,约 40% URL 可被基于前缀的 URL 分类器分类,提供较为稳定的证据.但使用较大数目的 URL 有可能引入非相关文档,损害分类的效果.式(1)中对排序靠后的 URL 赋予较低权重正是基于这种考虑.当使用 80 和 100 个 URL 时,使用平均加权(各个位置 URL 权重相同)与使用式(1)按位置加权的方法比较,$F1$ 值分别平均下降 0.68 和 0.98.说明使用位置加权有利于去除噪声影响.

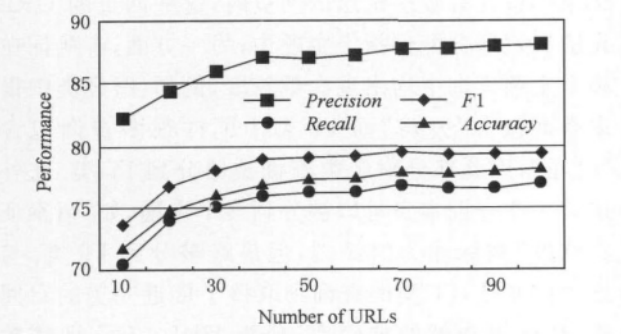

Fig. 3　The performance of URL-PRF with different number of URLs.

图 3　URL-PRF 的实验结果随使用 URL 数目的变化

4.2　URL-LOG 的实验结果

URL-LOG 使用查询日志中能够通过 URL 进行主题标注的查询训练统计分类器,因此我们关心如何选取查询作为训练数据.为了获得更多有标注的长尾查询将 2.4 节介绍的阈值 S 设为 0,考察阈值 T 对分类的影响.T 越大表明约束越严格,只有在主题分布上属于某一类的概率足够高时,才用来作为一个训练实例,因此获得训练数据也越小.表 3 给出了 T 分别取 0.3,0.5 和 0.8 时,从查询日志中获取的训练用查询的数量以及最终的分类效果:

Table 3　The Performance of URL-LOG with Different Values of Threshold T

表 3　URL-LOG 的实验结果随阈值 T 的变化

T	#Query	*Precision*	*Recall*	$F1$	*Accuracy*
0.3	$5.6\cdot10^6$	0.8979	0.8122	0.8356	0.8147
0.5	$3.6\cdot10^6$	0.8897	0.804	0.8274	0.8067
0.8	$3.2\cdot10^6$	0.8818	0.7945	0.8179	0.796

从表 3 可知,当 T 取 0.3 时共有约 560 万有标注的查询,获得最好的分类效果.这一方面说明,查询日志中通过 URL 进行主题标注的查询是比较可靠的,即使较小的阈值 T 也能获得较高标注质量的查询.另一方面说明,在不进行查询扩展前提下,训练数据规模对基于统计的学习算法的影响较大.使用更多的查询会涵盖更多的用户查询语言习惯,提高召回率,带来性能提升.进一步讨论训练数据规模对算法的影响.将 T 为 0.3 时获得数据(记为 *data*)随机平分为 5 份,考察使用不同数据规模下算法的性能.图 4 给出了实验结果:

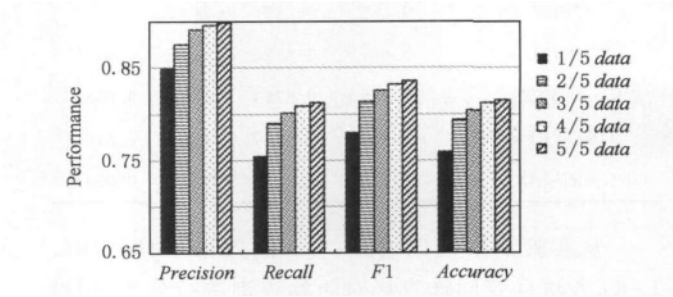

Fig. 4　The performance of URL-LOG with different ratio of training data.

图 4　URL-LOG 的实验结果随不同规模训练数据的变化

由图 4 可见,随着训练数据规模的增加,算法的表现呈上升趋势.进一步说明训练数据的规模对查询分类具有重大影响.基于 URL 主题对查询日志中的查询进行主题标注,进而训练统计分类器的方式具有很强的可扩展性.对于搜索引擎公司来说,查询日志可视为是海量的,因此可以自动获取海量的有标注的查询作为训练数据.实际上,即使将阈值 T 设置得较高获得更可靠的标注,依然能够获得大量的训练数据.这对训练和考察更好的查询分类器提供了很好的资源.

进一步将 URL-PRF 与 URL-LOG 再进行线性组合,记为 URL-PRF-LOG.URL-PRF-LOG 赋予每个主题一个分数,该分数为 URL-PRF 和 URL-LOG 对该主题分类概率的加权求和.根据前面的实验结果,URL-LOG 性能优于 URL-PRF.此外,实验过程中发现 URL-LOG 在各主题上的分类概率相对比较平滑,而 URL-PRF 则相对"陡峭",若采取平均加权,URL-PRF 将占据主导作用.基于以上 2 点考虑,赋予 URL-LOG 更高的权重.最终分数最高的主题作为 URL-PRF-LOG 的输出.URL-PRF-LOG 的实验结果将在下节一并讨论.

4.3 算法比较

表 4 给出了所有系统的实验结果. URL-PRF 以及 URL-PRF-LOG 均使用 100 条 URL; URL-LOG 与 URL-PRF-LOG 均使用 $T=0.3$ 时全部有标注的查询作为训练数据; URL-PRF-LOG 组合时赋予 URL-LOG 的权重设为 0.8.

Table 4 Experimental Results of Comparison Among Different Algorithms

表 4 查询分类算法比较的实验结果

Methods	*Precision*	*Recall*	*F*1	*Accuracy*
QT	0.6737	0.5021	0.5533	0.5468
URL-PRF	0.8803	0.7696	0.7945	0.7821
URL-LOG	0.8979	0.8122	0.8356	0.8147
QTE	0.8881	**0.8701**	**0.8749**	**0.8752**
URL-PRF-LOG	**0.9109**	0.8347	0.8548	0.8401
URL-PRF-LOG+QTE	**0.9354**	**0.9179**	**0.9266**	**0.9245**

从实验结果可以看出, QT 的效果不佳. URL-LOG 方法尽管同样直接对原始查询进行分类, 但性能显著优于 QT. 其原因可能是, 人们写某一主题文章时采用的词汇与用户查询某一主题信息时使用的词汇是不同的. 以查询作为训练数据训练分类器得到的特征空间, 更有利于区分不同主题的查询. URL-PRF 尽管十分简单, 但其获得相当好的准确率, 说明搜索结果中 URL 的主题分布与查询的主题分布具有很强的相关性. URL-LOG 比 URL-PRF 的性能更好, 说明文本信息依然是最有效的特征, 同时与 URL-PRF 相比, URL-LOG 有着更好的在线处理效率. QTE 不出意料在 4 个基本算法的比较中依然获得了最佳 $F1$ 值. 将查询扩展为文档之后, 问题转化为文本分类问题, 文本分类模型与待分类文档处于类似的特征空间, 而且通过扩展获得更多的信息与更丰富的特征表示有助于查询分类.

URL-PRF-LOG 获得了比单独分类器更好的实验结果. 这很容易解释, 两者具有相互补充的作用. 例如: 当查询的搜索结果 URL 能够被分类的比例较低时, URL-PRF 的可靠性降低, 此时 URL-LOG 可以提供更多的证据; 类似地, 当新查询没有出现在过去的查询日志中, 例如: 用户查询一款最新的汽车型号, 而搜索引擎的索引中可能已经存在相关信息, 并将相关网页返回. 此时 URL-PRF 将发挥更大的作用. URL-PRF-LOG 的准确率优于 QTE, 召回率比 QTE 差, $F1$ 值与 QTE 较为接近. 然而, URL-PRF-LOG 的处理效率会大大优于 QTE. 将 URL-PRF-LOG 与 QTE 组合(平均加权)之后, 获得的分类器在各个指标上都取得了最佳的效果. 综上, 基于 URL 主题获取的资源对查询分类算法具有正面影响; 在对在线响应要求较高情况下, URL-PRF-LOG 甚至 PRF-LOG 是 QTE 理想候选.

图 5 进一步给出了 URL-PRF-LOG 在不同主题上的宏平均指标, 精确率具有相似趋势. 我们注意到, 对于一些主题的查询, 如旅游、教育、汽车、体育等, URL-PRF-LOG 既获得了好的准确率又获得了好的召回率. 这些主题的查询比较具体, 搜索引擎的返回结果质量较高且被网页目录收录. 对于财经和军事类, URL-PRF-LOG 可以取得好的准确率却获得差的召回率. 一方面由于网页目录收录的网站对这些主题查询的搜索结果覆盖率较低. 例如, 当返回结果中包含有较多论坛的网页时, 这些网页的 URL 无法有效地分配到指定主题中; 另一方面, 有些词在多个主题查询中均出现次数较多, 例如: 财经类中很多查询包含"公司"而 IT 类中同样很多查询包含"公司", 因此部分财经类查询被错分到 IT 类. 此外还有一些查询本身可以被分到多个主题, 如"恒泰证券软件"被标注为财经类, 但最终被分到 IT 类. 与之对照的是, IT 类的查询则取得了接近完美的召回率, 却获得较低的准确率. 分析 URL-LOG 训练数据时发现, 获得标注的 IT 类查询数目要明显多于其他主题的查询数目. 由于我们没有进行特征选择, 造成了训练时具有类别偏置, 将在未来工作中改进.

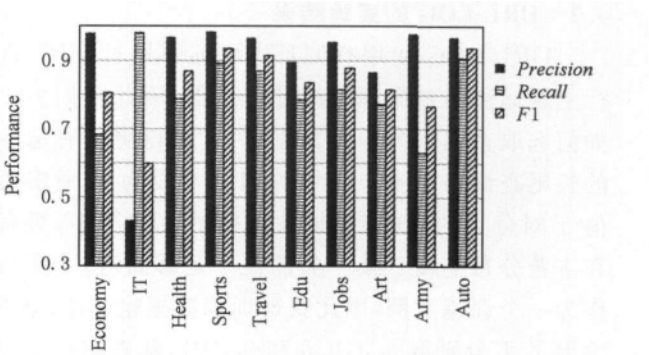

Fig. 5 The performance of URL-PRF-LOG over different categories.

图 5 URL-PRF-LOG 在不同主题上的实验结果

5 结束语

本文利用网络上人工收集的网站分类目录中有主题标注的 URL, 结合伪相关反馈技术与查询日志, 设计了自动、快速、有效的查询分类方法. 基于伪

相关反馈的方法利用搜索结果中 URL 主题的分布预测查询主题,基于查询日志的方法将 URL 的主题映射到对应的查询,从而获得大量有标注的查询作为数据训练统计分类器进行查询分类.实验表明,2 种算法都能获得较高的分类准确率.当将两者结合时,可获得比已有最佳算法更好的准确率和接近的 $F1$ 值.本文提出的方法具有较好的在线处理效率,不必在线进行复杂的文本处理,这是已有方法不具备的特点.实验发现,训练数据的规模对基于统计的查询分类算法具有很大影响,分类性能随训练数据规模增大而提升.URL 与查询日志结合的方法可以自动获取海量的训练语料,具有很好的可扩展性.

本文提出的方法依然有待改进的地方,例如:需要将网站分类目录的类别体系映射到待分类的类别体系,在训练查询分类器时并没有进行特征选择,这将在未来工作中进一步研究.

参 考 文 献

[1] Broder A, Fontoura M, Gabrilovich E, et al. Robust classification of rare queries using Web knowledge [C] //Proc of ACM SIGIR 2007. New York: ACM, 2007: 231－238

[2] Bennett P N, Krysta S, Dumais S T. Classification-enhanced ranking [C] //Proc of ACM WWW 2010. New York: ACM, 2010: 111－120

[3] Ryen W W, Peter B, Chen L. Predicting user interests from contextual information [C] //Proc of ACM SIGIR 2009. New York: ACM, 2009: 363－370

[4] Broder A. A taxonomy of web search [J]. ACM SIGIR Forum, 2002: 36(2): 3－10

[5] Zhang Sen, Wang Bin. A survey of Web search query intention classification [J]. Journal of Chinese Information Processing, 2008, 22(4): 75－82 (in Chinese)
(张森, 王斌. Web 检索查询意图分类技术综述[J]. 中文信息学报, 2008, 22(4): 75－82)

[6] Shen Dou, Pan Rong, Sun Jiantao, et al. Query enrichment for Web-query classification [J]. ACM Trans on Information Systems, 2006, 24(3): 320－352

[7] Li Ying, Zheng Zijian, Dai Honghua. KDD CUP-2005 report: Facing a great challenge [J]. ACM SIGKDD Explorations, 2005, 7(2): 91－99

[8] Beitzel S M, Jensen E C, Lewis D D, et al. Automatic classification of web queries using labeled and unlabeled training data [J]. ACM Trans on Information Systems, 2007, 25(2) (Article No. 9)

[9] Li Xiao, Wang Yeyi, Acero A. Learning query intent from regularized click graphs [C] //Proc of ACM SIGIR 2008. New York: ACM, 2008: 339－346

[10] Hu Jian, Wang Gang, Fred L, et al. Understanding user's query intent with Wikipedia [C] //Proc of ACM WWW 2009. New York: ACM, 2009: 471－480

[11] Shen Dou, Li Ying, Li Xiao, et al. Product query classification [C] //Proc of ACM CIKM 2009. New York: ACM, 2009: 741－750

[12] Fuxman A, et al. Improving classification accuracy using automatically extracted training data [C] //Proc of ACM SIGKDD 2009. New York: ACM, 2009: 1145－1154

[13] Alan J. Relevance feedback with too much data [C] //Proc of ACM SIGIR 1995. New York: ACM, 1995: 337－343

[14] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. Journal of Machine Learning Research, 2008, 9: 1871－1874

Zhang Yu, born in 1972. PhD, associate professor of Harbin Institute of Technology. Senior member of China Computer Federation. His current research interests include information retrieval, question & answering and natural language processing.

Song Wei, born in 1983. PhD candidate. His current research interests include query understanding, personalized search and social computing.

Liu Ting, born in 1972. PhD, professor of Harbin Institute of Technology. Senior member of China Computer Federation. His current research interests include natural language processing, information retrieval, and social computing.

Li Sheng, born in 1943. Professor of Harbin Institute of Technology. Member of China Computer Federation. His current research interests include natural language processing, information retrieval and machine translation.