查询意图自动分类的方法改进探讨*

贺国秀[1,2]，张晓娟[3]
（1.武汉大学信息检索与知识挖掘研究所，武汉 430072；2.武汉大学信息管理学院，武汉 430072；
3.西南大学计算机与信息科学学院，重庆 400715）

摘要：本文在降低数据标注成本的基础上，提高查询意图自动分类的准确率。首先，将ODP主题类目体系映射到Rose等意图类目体系，利用启发式和匹配的方法形成标注规则，对查询日志数据进行自动标注；其次，在提取查询的统计特征、用户行为特征和基于自然语言处理的语义特征基础上，提取查询的句法依赖关系作为分类特征；最后，使用集成学习模型GBDT作为分类器，对查询意图进行分类研究。实验表明，本文提出的标注规则可以获得大量被标注的训练数据集，新增的句法依赖关系特征可以提高查询意图的分类效果，GBDT作为集成学习模型相比线性分类模型可以明显提高查询意图分类的准确率。

关键词：GBDT；机器学习；查询日志；查询意图；自然语言处理
中图分类号：G353.4
DOI：10.3772/j.issn.1673-2286.2018.01.009

随着互联网的蓬勃发展，网络信息呈现爆炸式增长，以Google、百度和Bing等为代表的搜索引擎成为辅助人们快速获取信息的主要工具。当前的搜索引擎主要采用基于关键词匹配的技术以及网页链接算法来为用户返回所需信息[1]。由于用户提交的查询一般较短，且自然语言存在模糊性，故无法清晰地表达用户意图，使返回结果无法满足用户需求。因此，通过识别用户查询关键词所包含的意图（用户的信息需求、查询目标和查询动机等）[2-4]，搜索引擎可针对不同意图的查询采取不同的处理策略以获得更好的检索结果，或通过改变检索结果布局来方便用户快速定位查询目标；同时，搜索引擎和电子商务站点也可通过用户提交查询意图来提供个性化的检索和推送服务。

查询意图识别的主流方法是将其转化为查询意图自动分类，即在给定查询意图类别体系的情况下，通过提取特征、训练分类模型来实现对不同类别查询意图的自动分类[3]。综合已有研究发现，当前查询意图自动分类研究存在如下局限：（1）大多采用人工来标记数据集，而受时间和精力限制，人工标注数据集规模有限，故分类模型可获得的训练数据规模较小，最终影响训练所得分类器的自动分类准确度；（2）在选取查询分类特征时，较少考虑查询本身包含的丰富语义特征，如查询词、词性特征，以及查询中词间依存关系等；（3）查询意图自动分类主要依赖线性机器学习模型，因此需要对不同的度量标准和调优参数进行大量研究和实验。基于此，本文尝试分别在数据集标注、特征选取和分类器使用中利用新方法来解决查询意图分类研究存在的问题，提出一种标注规则，对已有查询日志数据进行自动标注。在此基础上，利用LTP工具提取查询的句法依赖关系等特征，并将集成学习模型梯度提升树（Gradient Boost Decision Tree，GBDT）应用到分类模型的训练。

1 相关研究

综合已有研究，查询意图分类主要包括数据标注、特征提取和分类方法研究三个方面[3]。

（1）查询意图数据标注。Broder[5]通过对用户查询及AltaVista日志进行分析研究，将用户查询意图分为信息类（I）、导航类（N）与事务类（T）三类。信息类

* 本研究得到国家社会科学基金青年项目“融合用户个性化与实时性意图的查询推荐模型研究”（编号：15 CTQ019）资助。

指用户在互联网上获取的信息，无其他交互操作；导航类指用户的目的为查找某个特定的网址；事务类指用户想通过查询获取互联网资源或服务。Rose等[6]使用日志分析人为地扩展Broder思想，认为Broder提出的事务类查询意图无法包含互联网所有资源，提出资源类（R）查询意图。资源类包括互联网上除信息类外的任何可获取的资源。虽然不同类目体系的划分各有其依据和支持，但Rose等的类目体系最受推崇。Ahituv[7]采用开放式分类目录（Open Directory Project，ODP）作为主题标签，建立查询类别，其中ODP是由人工对互联网中出现的各类站点的总结分类。张晓娟[4]选定Rose等的类目体系，主要靠人工标注的方法获取实验数据，但这种方法成本很高，可获得的数量却很小。宋巍[8]采用网页分类目录资源，利用本体匹配法对查询日志数据进行自动标注，其结果过度依赖于分类目录资源，故泛化能力较差。

（2）查询意图特征提取。张森等[9]认为，特征提取的研究工作主要解决如何从用户简单的查询中获取充分、明确的特征，以此来识别查询意图。①基于查询表达式的特征提取。通过对查询词本身包含的词义[6]、词性[10]、词长[9]和在语料库中的统计信息来识别查询的潜在意图，使用各类查询的一组启发式特征来区分查询[11]，使用时间和地理等特征来表示查询。②基于用户行为的特征提取。用户行为是用户对检索结果反馈的行为，是用户目标的显示表达，主要包括用户交互行为、用户点击行为和语境变化等，如Liu等[12]提出利用点击相关的相互点击意图（MCI）和点击意图排序（CIR）等作为特征。③隐含语义特征。Mendoza等[13]基于用户日志利用PLSA提取查询的隐含主题表达作为特征。然而，由于查询一般以自然语言显示，故其语义特征还可由自然语言处理工具进行深入挖掘。

（3）查询意图分类方法。Liu[14]和Kanhabua[15]等使用典型的决策树算法执行分类任务，Ji等[16]通过用户浏览行为的动态数据来预测用户查询意图，Hu等[17]通过把查询映射到维基百科的现象空间以识别查询意图，Feng[18]利用ODP和用户查询日志构建用户兴趣模型。高景斌[19]和张杨浩[20]利用线性分类模型支持向量机（Support Vector Machine，SVM）[21]对查询意图进行分类。SVM是对逻辑回归[22]的一种优化，通过寻求结构风险最小化来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下获得良好统计规律的目的。GBDT作为集成学习算法，通过融合弱分类器来提升分类器性能[23-25]。相比线性分类器参数调优困难的局限性，GBDT可以快速提高模型性能，并在一定程度上避免模型过拟合问题。

2 查询意图自动分类方法

在构建标注集方面，本文提出一种基于ODP主题类目体系的自动标注规则，以此获得大量标注数据；在提取分类特征方面，探索词之间的句法依赖特征对意图分类效率的影响。由于集成学习通过构建并结合多个机器学习模型来完成学习任务，可以有效弥补线性分类模型的调优缺陷，所以本文使用集成学习模型中的GBDT作为分类器，对查询意图进行识别。本文的整体框架如图1所示。

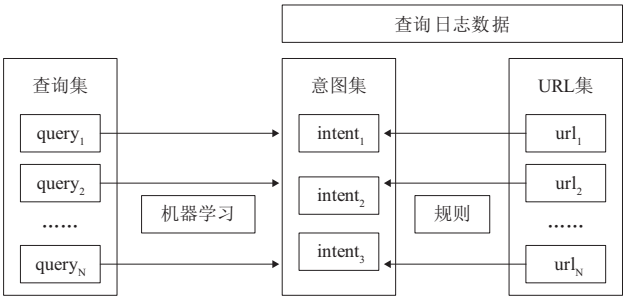

图 1 整体框架

2.1 构建标注规则

本文主要将ODP主题类目体系映射到Rose等的意图类目体系，利用启发式和匹配的方法形成标注规则，对查询日志数据进行自动标注。ODP将网络中出现的url总结为14个主题，每个主题都包含相应的url，用<url，topic>表示url和topic的对应关系，整个ODP数据集可以表示为“ODP={<url_1，$topic_1$>，<url_2，$topic_2$>，…，<url_M，$topic_M$>}”。其中，ODP的主题结构如表1所示，Rose的查询意图类目体系、相关解释与实例如表2所示。

本文将日志数据中的URL映射到Rose等的意图类目体系的过程如图2所示。首先，通过对Rose意图类目体系的分析发现，导航类的标注规则被确定为“仅包含类似‘www.baidu.com’这样以‘www.’开头，以‘.com/.cn/.org’等结尾的url所对应的查询属于导航类”。其次，结合两种方法生成资源类的标注规则。

表 1 ODP主题类目体系

主 题	数量/个	主 题	数量/个
休闲	529	新闻	552
体育	244	游戏	576
健康	750	社会	1 557
儿童	240	科学	923
参考	2 220	艺术	960
商业	5 600	计算机	1 639
家庭	114	购物	430

（1）启发式方法。通过人为对URL的归纳，发现如果url中含有“download”“game”“movie”“music”或“book”等关键词的一般为资源类url，即其对应的用户查询属于资源类。（2）匹配的方法。通过分析，ODP主题中属于休闲、商业、游戏、计算机和购物的url属于资源类。所以提取ODP中相应的url，去除“http://www.”的开头，构建匹配列表。如果用户点击的url可在匹配列表中找到，则该url对应的查询为资源类；结合以上分析以及Rose等的意图类目体系，若用户点击的url所对应的用户意图不属于导航类和资源类，则属于信息类。

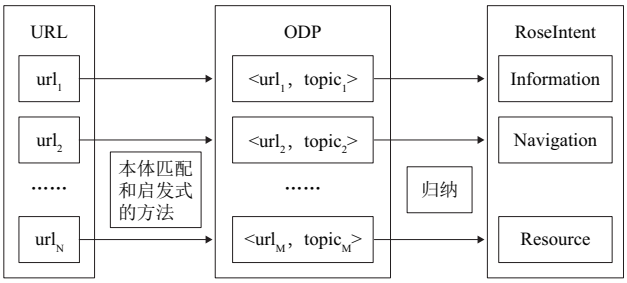

图 2 日志数据中URL与Rose意图类目体系间映射关系

表 2 Rose类目体系、相关解释与实例

层 级	解 释	实 例
导航类（N）	用户为了获得一个明确的网址	公司、学校等的主页
信息类（I）	用户为了获得数据或信息	某条法律条款的解释
资源类（R）	用户为了获得有用的资源	购买物品、游戏等
事务类导航（N.T）	用户用来处理事务的导航网址	match.com
信息类导航（N.I）	用户用来获取信息的网址	Yahoo.com
有指导性的（I.D）	用户为了获取某个特定问题的答案	哈尔滨市的邮编
无指导性的（I.U）	用户为了获取一个主题的所有信息	2016年新出电视剧的信息
发现（I.F）	用户为了获得一个产品或者服务的具体位置	哈尔滨中央大街的位置
列表（I.L）	用户为了获得一组可信的站点列表	电子商务网站有哪些
建议（I.A）	用户为了获得某个主题的建议、观点和指南等	如何高效率地学习
获取（R.O）	用户为了获得一个明确的资源或项目	某首歌的歌词
下载（R.D）	用户为了把某个资源下载到本地	电影、音乐、小说和论文等的下载
娱乐（R.E）	用户可以在网页上进行的娱乐活动	游戏、聊天等
交互（R.I）	用户与网络上的程序或者资源进行交互	在淘宝网上购买商品
确定的（I.D.C）	用户为了获得一个问题的无歧义回答	《宪法》的第3条
开放的（I.D.O）	用户为了获得两个或更多个信息	人类的免疫系统
在线的（R.O.O）	用户需要在线获取	火车票的余票信息
离线的（R.O.F）	用户可以离线获得资源	-

2.2 查询特征提取

特征提取是用户查询意图分类的关键，需要从中获取充分的特征。本文在选取查询词的统计特征（查询所包含的字长和词长）与用户行为特征（用户点击url的排名和用户最终点击的次数）的基础上，重点考虑查询中的语义特征。

本文利用哈尔滨工业大学开发的LTP[26]提取三方面的语义特征集合。（1）查询分词特征。本文对查询进行中文分词，即将汉字序列切分成词序列，以介于汉字和句子间的粒度对查询进行表示。如对于查询“年轻人住房问题”可以分词为“年轻人”“住房”和“问题”三个

词。（2）查询词性特征。词性作为对词的一种泛化，在语音识别、句法分析和信息抽取等任务中有重要作用。本文对每个查询的分词结果进行词性标注，并将词性作为查询特征，如查询“年轻人住房问题”的分词均为名词。（3）查询句法依存关系特征。通过分析语言单位成分间的依存关系，揭示其句法结构特征。如查询“年轻人住房问题”存在的句法依存关系有“ATT”“ATT”和“HED”，其中“ATT”表示“年轻人”修饰“住房”的定中关系，第二个“ATT”表示“住房”修饰“问题”的定中关系，“HED”表示“问题”为整个句子的核心。

在提取语义特征集合后，本文利用词袋模型分别对所有的分词集合、词性集合和词间的句法依存关系集合进行表示。考虑特征间的重要性随其在该查询中出现的次数而增加，但同时会随其在整个查询集合中出现的查询频次而减少，故本文引入TF-IDF对每个特征值进行加权，TF表示该特征词在查询中出现的频次，具体计算方法见公式（1）。

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

其中，$n_{i,j}$是该特征在查询中的出现次数，而$\sum_k n_{i,j}$是在查询中所有特征词出现的次数。IDF表示该特征词在查询集合中出现的查询频次，具体计算方法见公式（2）。

$$idf_i=\log\frac{|D|}{1+|\{d\in D:t\in d\}|} \quad (2)$$

其中，$|D|$表示查询集合中的查询总数，$1+|\{d\in D:t\in d\}|$表示查询集合中出现该特征词的查询数。TF-IDF表示TF与IDF的乘积，具体计算见公式（3）。

$$tfidf_{i,j}=tf_{i,j}\times idf_i \quad (3)$$

2.3 分类器构造

在3.1节标注规则得到的数据集和3.2节提出的特征集合的基础上，本文将使用集成学习模型中的GBDT[23-25]作为分类器对查询意图进行分类研究。其中，GBDT由多棵决策树组成，通过投票的方式将所有树的结论进行累加作为最终答案。原始的提升算法（Boost）为数据集中的每个样本赋一个权值。每次迭代都对算法决策的准确性进行验证，增加其中错误样本的权重，减少正确样本的权重。进行H次迭代后，将会得到H个简单的分类器，然后将其通过投票的方式集合起来，得到最终的模型。

Gradient Boost与Boost的区别是：每次迭代的目的是减少出现的错误，而不是对正确预测样本和错误预测样本进行加权。通过在误差减少的梯度方向上建立一个新模型，以迅速得到最优模型。GBDT作为集成模型，相比于线性模型，其优化速度更快，泛化能力更强。

3 实验

3.1 实验数据

本文采用搜狗实验室提供的查询日志数据[14]进行实验，数据样本格式如表3所示。从左到右分别表示访问时间（time）、用户匿名ID（user id）、查询词（query）、该url在返回结果中的排名与用户点击的顺序号（result click）和用户点击的url。本文使用Graphlab Create包中的SFrame工具对数据进行读取和初步统计分析，再将数据集随机分为训练集（training data）、验证集（validation data）和测试集（test data），分别占整体数据集的60%、20%和20%。其中，训练集用来训练分类模型，验证集用来选择最优的分类模型参数，测试集用来评估分类模型的性能。最后基于3个数据集，进行分类器的训练和评估。

本文通过网络爬虫程序自动获得ODP主题类目数据，所获得数据集的格式如表4所示。其中，url表示网址，name表示url代表的网址名称，label表示ODP主题类目的二级主题。

表 3 搜狗查询日志样本格式

time	user id	query	result click	url
00:00:00	2982199073774412	360安全卫士	8 3	download.it.com.cn/softweb/software/firewall/antivirus/20067/17938.html
00:00:00	07594220010824798	哄抢救灾物资	1 1	news.21cn.com/social/daqian/2008/05/29/4777194_1.shtml
00:00:00	5228056822071097	75810部队	14 5	www.greatoo.com/greatoo_cn/list.asp?link_id=276&title=%BE%DE%C2%D6%D0%C2%CE%C5
00:00:00	6140463203615646	绳艺	62 36	www.jd-cd.com/jd_opus/xx/200607/706.html

表 4 ODP数据的相关信息

url	name	label
http://news.jmu.edu.cn/	集美大学新闻网	大专院校
http://jjxj.swufe.edu.cn/	经济学家	出版物
http://www.jsacd.gov.cn/	江苏省农业资源开发局	江苏
http://www.yndaily.com/	云南日报网	地区
http://www.panda.org.cn/	成都大熊猫繁育研究基地	熊猫

3.2 标注数据集

在获得用户查询日志数据和ODP数据的基础上，本文随机选取查询日志中的1万条数据进行实验。基于3.1节提出的标注规则，本文首先通过启发式的方法，构建启发式列表[download，book，read，music，movie，software]，然后将ODP主题中的休闲、商业、游戏、计算机和购物等主题所包含的二级主题对应的url映射到匹配列表中。

再将上述启发式列表和匹配列表结合起来，最后得到资源列表。其列表的部分信息为“download，book，read，music，movie，software，52384.com，map.baidu.com，htffund.com”等。被标注数据集的label比例如表5所示，结果显示，使用标注规则自动标注的数据集和人工标注的数据集相比[4]，label比例基本一致，但使用标注规则的好处在于可以迅速获得大量被标注的数据集。

表 5 基于ODP的最终数据集标注结果

数据集分类	数量/个	比例/%
信息类	6 684	66.84
资源类	2 099	20.99
导航类	1 217	12.17

3.3 基准实验

为了与线性分类模型的效果进行有效对比，本文选择逻辑回归（Logistic Regression，LR）和支持向量机作为基准分类器。其中，在使用逻辑回归时，本文利用随机梯度下降作为优化器训练模型，并加入L1和L2规则来防止过拟合；在使用支持向量机时，使用RBF核函数。

3.4 实验分析

对于逻辑回归方法，本文首先通过验证集validation set选择最优的使用L1规则的惩罚L1 penalty和使用L2规则的惩罚L2 penalty；其次，使用得到的最优超参数，分别比较不同的特征选择对结果的影响，得到Model1（使用3.2所述的所有特征）和Model2（不使用词之间的句法依存关系特征）；最后，通过比较测试集，得到模型的平均准确率、精准率、召回率和F1值。需指明的是，本文首先训练Model1，且在训练该模型时，为了选择最优的L2 penalty，先保持L1 penalty=0不变，令L2 penalty在0—5以适当的间隔选择15个有代表性的值进行实验；再令L2 penalty=1.65不变，使L1 penalty从0—100选择适当间隔的15个值进行实验，实验结果如图3所示。图3的实验结果表明，当L1 penalty=0和L2 penalty=1.65时，获得最优的分类器，此时的平均准确率为0.68；图4的实验结果表明，L1 penalty=10，L2 penalty=1.65时，获得最优的平均准确率为0.69。

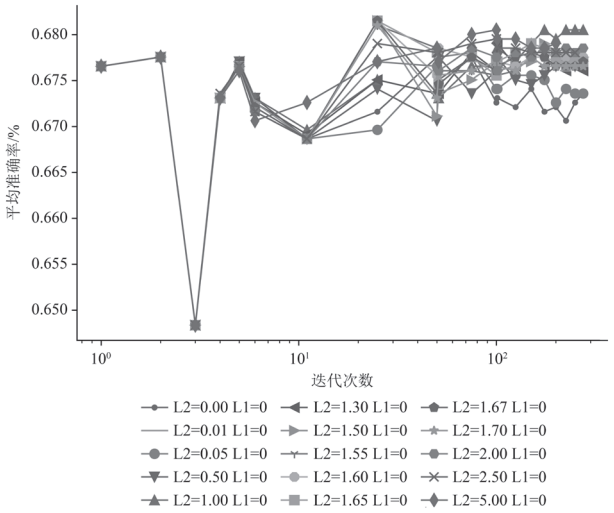

图 3 L2_penalty对查询意图分类平均准确率的影响

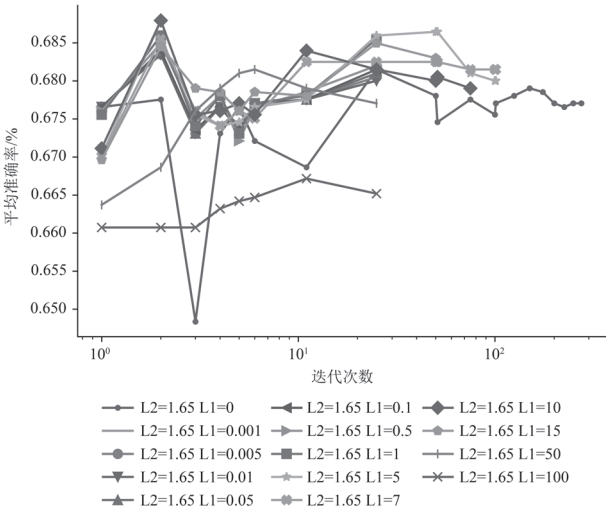

图 4 L1_penalty对查询意图分类平均准确率的影响

在训练Model2时，使用与Model1相同的L1 penalty和L2 penalty，但不考虑词之间的句法依赖关系特征，再使用测试集分别得到两个模型的各个指标（平均准确率、精准率、召回率和F1值），其对比的实验结果如图5所示。可以看出，当使用逻辑回归作为分类模型时，利用LTP充分提取查询的语义特征，可以提高查询意图的分类准确率。

支持向量机与逻辑回归的训练相同，首先找到最优的防止过拟合的惩罚参数（penalty），然后分别训练Model1和Model2，并比较这两个模型的优劣。本文先训练Model1（选择不同的惩罚参数），然后比较平均准确率，结果如图6所示。实验表明，不同的惩罚参数对性能的影响基本一致，故本文采取默认的penalty=1，并进一步对Model1和Model2的结果进行对比，结果发现充分提取语义特征的分类器明显优于另一个分类器，如图7所示。

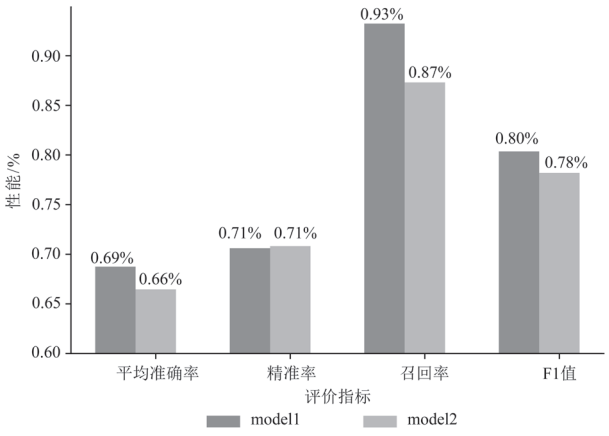

图 5 LR中Model1和Model2的查询意图分类性能比较

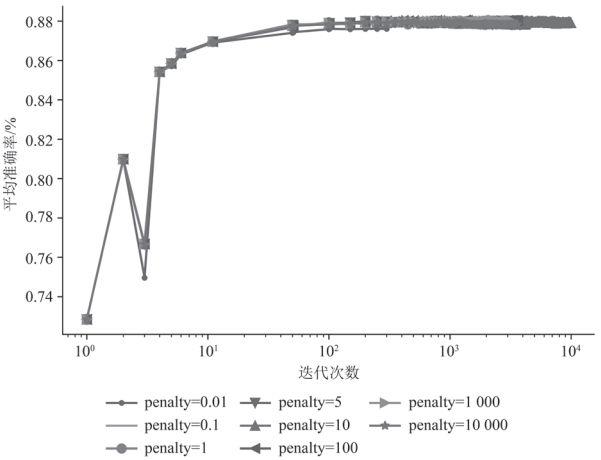

图 6 不同penalty对查询意图分类平均准确率的影响

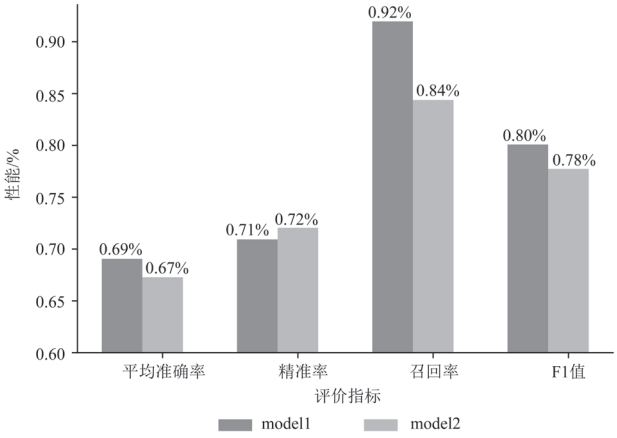

图 7 SVM中Model1和Model2的查询意图自动分类性能比较

本文将线性分类模型LR和SVM的结果作为Baseline与集成分类模型GBDT进行对比。对于GBDT，本文首先训练Model1以确定最优的超参数，令每次迭代的学习率step size=0.3不变，改变树的最大深度（max depth），其实验结果如图8所示；再令max depth=20不变，改变迭代的步长（step size），其实验结果如图9所示。图8与图9的实验结果表明，当max depth=20，step size=0.3时，获得最优的平均准确率。然后，本文再使用同样的超参数，通过提取不同特征来训练Model2，以此比较不同的特征提取对查询意图分类性能的影响，其实验结果如图10所示。实验表明，充分提取语义特征的分类器明显优于不使用词间句法依存关系特征的分类器。

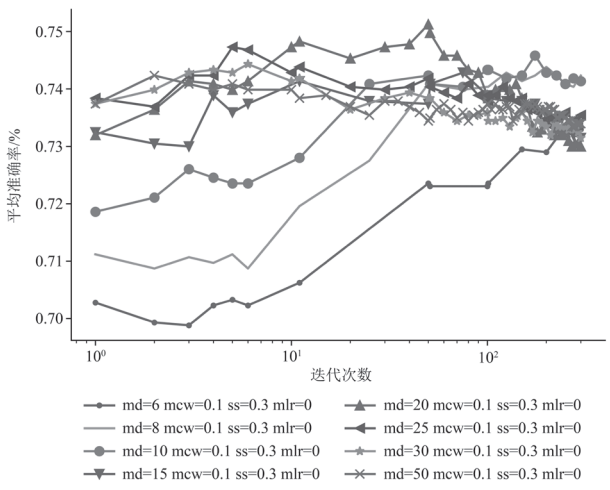

图 8 不同max depth对查询意图分类平均准确率的影响

本文再分别从平均准确率、精准率、召回率和F1值比较线性分类模型（LR和SVM）与集成分类模型（GBDT）的性能差异，结果如图11所示。实验结果表明，

使用集成学习模型对查询意图分类的性能明显优于线性分类模型；由三个分类器使用不同特征集合的实验表明，使用LTP提取查询的句法依赖关系作为查询语义特征，能够提高机器学习算法对查询意图的分类准确率。

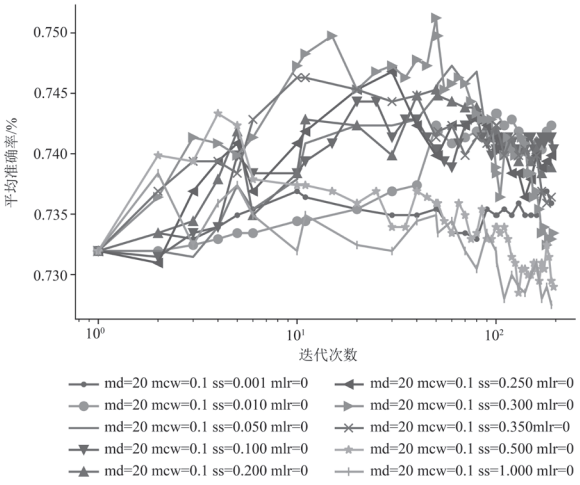

图 9 不同step size对查询意图分类平均准确率的影响

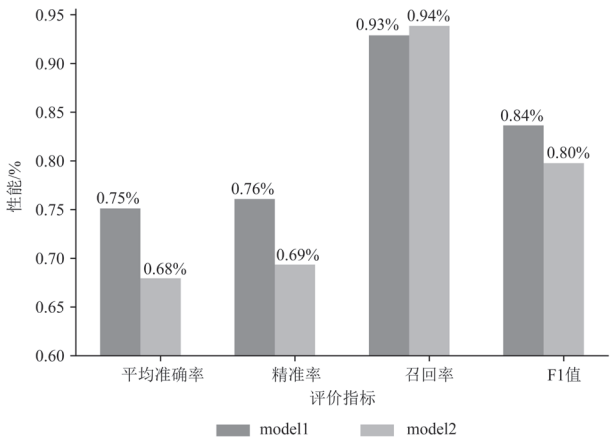

图 10 GBDT中 Model1和Model2的查询意图分类性能比较

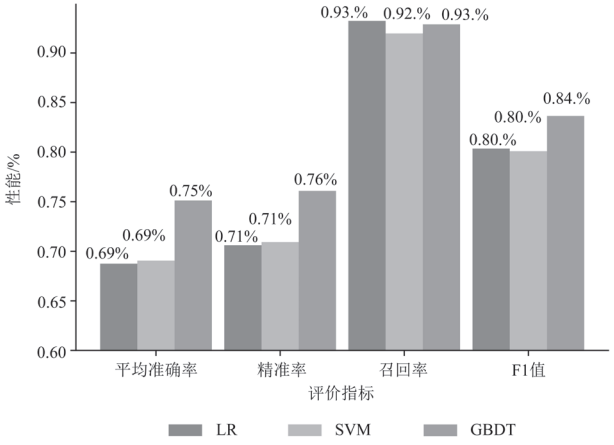

图 11 LR、SVM和GBDT的查询意图分类性能比较

4 总结与展望

本文尝试在查询意图自动分类中的标注集构建、特征选择以及在自动分类方法中采用新方法。研究工作主要包括以下方面：（1）基于Rose等提出的意图类目体系，结合ODP主题类目数据，使用启发式的方法和匹配的方法形成标注规则，对数据集进行标注；（2）从查询的语义特征、统计特征和用户行为特征三个方面进行特征提取，主要利用LTP工具提取查询的句法依赖关系特征作为语义特征；（3）使用集成学习模型GBDT作为分类器对查询意图进行分类。最终实验结果表明：（1）通过对标注的数据集标签进行统计，发现标签比例和人工标注的标签比例基本一致；（2）使用本文提出的特征集合所训练的分类器对查询意图的分类效率明显优于不使用词之间的句法依赖关系特征的分类器；（3）使用集成学习模型的GBDT对查询意图的分类效率明显优于线性分类器。本文方法虽取得较好的实验结果，但也存在不足之处，主要包括：（1）由于ODP数据收录网页数量的局限性，本文所提的标注规则仍有待优化；（2）将会在其他查询日志数据上进一步验证本文方法；（3）考虑利用词向量和递归神经网络提取查询的深度语义特征，以此提高查询的分类效率；（4）将查询意图识别结果应用到检索模型中，为查询返回更准确的查询结果。

参考文献

[1] KLEINBERG J M.Hubs,authorities,and communities[J].Acm Computing Surveys,1999,31(4es):5.

[2] NGUYEN H.Capturing user intent for information retrieval[C]// Nineteenth National Conference on Artificial Intelligence.July 25-29, 2004,San Jose: DBLP,2004,48(3):371-375.

[3] 陆伟,周红霞,张晓娟.查询意图研究综述[J].中国图书馆学报,2013, 39(1):100-111.

[4] 张晓娟.查询意图自动分类与分析[D].武汉:武汉大学,2014.

[5] BRODER A.A taxonomy of web search[J].Acm Sigir Forum,2002, 36(2):3-10.

[6] ROSE D E,LEVINSON D.Understanding user goals in web search[J]. World Wide Web,2004(11):13-19.

[7] AHITUV N.Popular searches in Google and Yahoo!: a “digital divide” in information uses[J].Information Society,2010,26(1):17-37.

[8] 宋巍.基于主题的查询意图识别研究[D].哈尔滨:哈尔滨工业大学,2013.

[9] 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.

[10] DUAN R,WANG X,HU R,et al.Dependency relation based detection of lexicalized user goals[J].Ubiquitous Intelligence and Computing Lecture Notes in Computer Science,2010,6406:167-178.

[11] JANSEN B J,BOOTH D L,SPINK A.Spink A.Determining the user intent of web search engine queries[C]//Proceedings of the 16th International Conference on World Wide Web.ACM,2007:1149-1150.

[12] LIU P,AZIMI J,ZHANG R.Contextual query intent extraction for paid search selection[C]//Proceedings of the 24th International Conference on World Wide Web.ACM,2015:71-72.

[13] MENDOZA M,ZAMORA J.Identifying the Intent of a User Query Using Support Vector Machines[C]//SPIRE.2009:131-142.

[14] LIU Y,ZHANG M,RU L,et al.Automatic query type identification based on click through information[C]//Asia Information Retrieval Symposium.Springer Berlin Heidelberg,2006:593-600.

[15] KANHABUA N,NGOC N T,NEJDL W.Learning to detect event-related queries for web search[C]//Proceedings of the 24th International Conference on World Wide Web.ACM,2015:1339-1344.

[16] JI M,YAN J,GU S,et al.Learning search tasks in queries and web pages via graph regularization[C]//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.ACM,2011:55-64.

[17] HU J,WANG G,LOCHOVSKY F,et al.Understanding user's query intent with wikipedia[C]//Proceedings of the 18th International Conference on World wide web.ACM,2009:471-480.

[18] FENG L.Novel query intent identification method based on user interest model[J].Journal of Information & Computational Science,2015,12(10):3881-3888.

[19] 高景斌.基于查询子意图识别的检索结果多样化方法研究[D].哈尔滨:哈尔滨工业大学,2012.

[20] 张杨浩.基于搜索引擎日志的查询意图分类研究[D].重庆:西南大学,2016.

[21] CORTES C,Vapnik V.Support vector machine[J].Machine learning,1995,20(3):273-297.

[22] DOMÍNGUEZ-ALMENDROS S,BENÍTEZ-PAREJO N,GONZALEZ-RAMIREZ A R.Logistic regression models[J].Allergologia et immunopathologia,2011,39(5):295-305.

[23] FRIEDMAN J H.Greedy function approximation:a gradient boosting machine[J].Annals of Statistics,2001:1189-1232.

[24] FRIEDMAN J H.Stochastic gradient boosting[J].Computational Statistics & Data Analysis,2002,38(4):367-378.

[25] JOHNSON R,ZHANG T.Learning nonlinear functions using regularized greedy forest[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2014,36(5):942-954.

[26] CHE W,LI Z,LIU T.Ltp:A Chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics:Demonstrations.Association for Computational Linguistics,2010:13-16.

作者简介

贺国秀，男，1995年生，硕士研究生，研究方向：信息检索，E-mail：guoxiu.he@whu.edu.cn。
张晓娟，女，1985年生，博士，副教授，研究方向：信息检索，E-mail：zxj0614@swu.edu.cn。

Discussion on the Improvement of Methods for Automatic Classification of Query Intent

HE GuoXiu[1,2], ZHANG XiaoJuan[3]
(1.Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China; 2.School of Information Management, Wuhan University, Wuhan 430072, China; 3.School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: On the basis of reducing the cost of data marked, the research is to improve the accuracy of query intent automatic classification. Firstly, the ODP subject category system is mapped to Rose's intent class system, the heuristic and matching methods are used to form the annotation rules, and the query log data is automatically marked; Then, based on statistical characteristics of the extracted query, the characteristics of user's behavior and natural language such as word segmentation and part of speech, the syntactic dependency of the query is extracted as the classification feature; Finally, the integrated learning model GBDT is used as a classifier to classify the query intent. The experimental results show that the proposed rules can be used to obtain a large number of trained training data sets. The new syntactic and relational feature can improve the classification result of query intent. GBDT can improve the accuracy rate of query intent classification as an integrated learning model compared with linear classification model.

Keywords: GBDT; Machine Learning; Query Log; Query Intent; Nature Language Processing

（收稿日期：2017-10-22）