查询专指度特征分析与自动识别*

唐祥彬[1]　陆　伟[2]　张晓娟[1]　黄诗豪[1]
[1](武汉大学信息管理学院　武汉 430072)
[2](武汉大学信息资源研究中心　武汉 430072)

摘要: 【目的】基于 Sogou 查询日志构建人工标注集，实现查询专指度的特征分析与自动识别，并对识别效果进行分析与评测。【方法】选取用户查询串基本特征与内容特征进行统计分析，并分别训练决策树、SVM 和朴素贝叶斯分类器对专指度进行自动识别。【结果】使用以上特征的识别效果良好，十折交叉检验的宏平均 F-measure 均高于0.8。【局限】分类特征的选择未考虑用户点击信息；朴素贝叶斯的独立性假设在本实验中是否可以忽略仍需进一步验证。【结论】利用查询串基本特征和内容特征，可以有效识别弱、略和强专指度查询。

关键词: 查询专指度　决策树　SVM　朴素贝叶斯
分类号: G353.1

1　引　言

随着互联网的蓬勃发展，搜索引擎 Google、Yahoo!、百度等已成为用户访问网络信息资源的主流工具[1]。传统的搜索引擎通过提供整个互联网或多个主题网站上与用户提问相关的各种信息，然后由用户判断哪些返回结果相关，哪些无关。但相关并不一定使用户信息需求得到满足，若返回的结果范围过大或过小，都会导致用户花费更多时间和精力在大量繁杂信息中寻找有用信息。因此，自动识别用户查询专指度(Query Specificity)，返回与其信息需求限制范围相符的个性化查询结果，成为改善综合搜索引擎性能、提高用户检索体验的重要途径。

专指度作为用户查询意图的 10 大维度之一[2]，指的是“用户通过查询语句对其自身信息需求或某查询主题的限制范围程度”，反映用户对所检索信息的专一性、详细性和确定性要求。另一个与专指度相关的概念是查询模糊度(Query Ambiguity)[3]，但查询模糊度与专指度并不能简单地看作一个问题的正反两面，实质上，两者具有相似点但又各有侧重。查询模糊度侧重于分析查询串中是否含有一词多义的词项或者查询串本身是否有多种解释[4]，而专指度研究侧重于探讨用户对信息需求范围的界定是否明确清楚，即专指度主要是分析用户查询中使用了哪些限制，如数量限制、名字限制、时间限制、位置限制等。

然而“如何对查询语句的专指度强弱进行判别”及“如何实现专指度自动识别”等问题，尚未引起学界的足够重视。基于此，本文对查询专指度进行分类；深入分析用户查询语句查询串特征；利用综合搜索引擎 Sogou 查询日志，在人工标注数据集的基础上实现查询专指度的自动识别。

2　相关研究

查询意图分类与识别作为解决“信息过载”、“主题偏移”等问题的有效途径，已成为现代 Web 信息检索、智能搜索领域的一个研究热点。意图识别可以从多个角度进行，如查询目标、查询所涉及的主题、组合查询的多个维度等。如Broder[5]通过对用户查询及 AltaVista 日志进行分析将用户查询意图(查询目标)分

通讯作者: 陆伟，ORCID: 0000-0002-0929-7416，E-mail: reedwhu@gmail.com。
*本文系国家科技支撑计划课题“文化遗产知识本体构建存储可视化技术研究”(项目编号:2012BAH33F03)和国家自然科学基金面上项目“基于语言模型的通用实体检索建模及框架实现研究”(项目编号: 71173164)的研究成果之一。

为三类，即信息类、导航类和事务类。Web 查询意图的相关研究大多沿用 Broder 分类体系或改进该分类体系，如 Rose 等[6]认为“事务类”不足以概括网上的所有资源，提出以“资源类”将其取代，指出“资源类”不再局限于一般的 Web 活动，而是包括网页上可获取的任何资源，并在此基础上提出了更细致的层次化意图分类体系。González-Caro 等[2]提出应从题材(Genre)、主题(Topic)、任务(Task)、目标(Objective)、专指度(Specificity)、范围(Scope)等10大维度对用户查询意图进行分类和识别，本文所讨论的就是根据专指度这一维度进行意图分类。

用户向搜索引擎等信息检索系统提交的查询语句表达了其潜在的信息需求，因此，学界主要从信息需求角度对专指度进行了相关研究。Donato等[7]在研究用户信息搜寻行为时表明专指度是影响查询语句可表达性的主要因素之一。Chang等[8]认为导航类和事物类查询往往比信息类查询表达得更加详细、明确，更有可能是强专指度查询。Calderón-Benavides等[9]的研究表明当查询含有多义词时，不到 1%的查询为强专指度查询，而 38%的查询为弱专指度查询。可见，用户查询目标或查询语句类型与专指度之间存在一定的关联，如非歧义查询大都不是弱专指度查询[10]，问题类查询[11]往往是强专指度查询。然而，以上研究主要是对专指度某一类别做了关联分析，并没有专门探讨专指度识别这一问题。近年来，也有学者直接针对专指度这一维度进行研究，如Phan等[12]发现查询语句长度与用户查询语句专指度有关，且长查询串通常是强专指度查询。Hafernik等[13]仅提出了“专指度关联属性列表”并以此将专指度分为强、弱两类。该文虽然认识到专指度并不是一个简单的二值分类问题，但未做深入考虑，对于专指度识别效果也未做详细对比和分析。

以上研究表明，专指度的识别存在以下问题:

(1) 专指度尚无统一定义，以致缺乏统一的标准构建分类体系;

(2) 用以实现有效分类的属性特征有待进一步挖掘;

(3) 多数研究仅停留在专指度某一特定类别的描述上，缺乏各类别特征的综合分析;

(4) 尚未探讨查询目标与专指度之间的关联。

针对以上问题，本文经过文献调研明确了专指度分类体系，在前人研究成果的基础上完善分类属性特征。且在不借助其他外部资源的情况下，利用 Sogou 查询日志，构建查询专指度标注集，并分别利用决策树、SVM、朴素贝叶斯分类器实现专指度的自动识别。

3 查询专指度分类与特征选择

3.1 专指度分类体系

目前，学界关于查询专指度分类并无统一说法，较为常见的是将查询语句专指度划分为两类或三类。如 Igwersen 等[14]从信息需求角度出发将“知识陈述明确性(Knowledge State Specificity)”分为“宽泛的(Generic)”或“具体的(Specific)”两种。类似的，Ramíez 等[15]使用“广泛(Broad)”和“狭隘(Narrow)”两类描述信息需求类型。以上二值分类较为简单，部分查询语句可能并没有得到正确分类。为使分类体系能覆盖绝大部分的查询语句，本文采用 Calderón-Benavides 等[9]提出的分类方法，按专指度强弱程度依次分为“强专指度(Specific)”、“略专指度(Medium)”、“弱专指度(Broad)”三类。

(1) 强专指度查询：用户具有明确的信息需求，且该信息需求可以界定在某一特定范围内，可以在查询语句中表达得非常清楚。比如用户想到达某网站，想知道某话题具体信息，想对比某事物，想了解关于某一具体问题的答案、建议和方法等。如查询“baidu.com”、“哈利波特 6 最新剧照”、“番茄鸡蛋的做法”、“2008 年北京高考时间”等。

(2) 略专指度查询：自身信息需求不够明确，无法在查询语句中表达得非常清楚，或用户仅需要了解相关信息，不必在查询语句中表达清楚。比如用户想知道关于某话题的某一方面，查询语句范围限定并不严格。如“哈利波特剧照”、“骏捷+油耗”、“番茄小炒”等。

(3) 弱专指度查询：用户信息需求较为宽泛，只是想获取关于该查询的信息即可，或用户对所检索信息不熟悉，需要进一步搜索。因此，在查询语句中基本没有限定范围。如“哈利波特”、“高考招生”、“骏捷”、“家常菜”等。

3.2 专指度特征选择

专指度分类体系构建完成后，如何选取合适的特征对专指度进行自动识别，是本节需要解决的问题。

查询串是用户经过思考后提交给搜索引擎的，对其进行分析有助于识别用户的查询意图。于是，本文选取查询串基本特征和内容特征作为专指度的分类特征。

查询串基本特征主要包括查询串长度、词项个数、词项长度等信息。通常而言，查询串或词项长度越长、词项个数越多，往往是强专指度查询；反之，则为略或弱专指度查询。然而，经统计分析笔者发现，即使用户提交的查询串只有一或两个词，也有可能是强专指度查询。比如“百度”、“竹石，诗”、“中国人口”等专指性强的查询，查询串非常简短，而“长篇小说”、“高考分数线”、“赞保罗皮尔斯”等专指性不太强的查询，查询串相对较长。可见，虽然长查询串通常是强专指度查询，但反之并不成立。于是，除考虑以上基本特征外，在结合 Hafernik 等[13]研究的基础上，本文列出了 10 个与专指度相关的内容特征，并将其作为区分强、略、弱三种专指度的重要依据，如表 1 所示:

表 1　查询串内容特征及举例

编号	查询串内容特征	查询举例
1	查询为比较多个事物	“中印军事对比”、“墨水比较”
2	查询为一个含确切答案的问题	“广州越秀邮编”、“山东有什么美食”
3	查询比较多种不同的想法和话题	“天籁车优缺点”、“范跑跑先跑该不该”
4	查询包含方向、建议、指导等方面的信息需求	“如何减少腹部脂肪”、“项目报建手续”、 “孕妇食谱”
5	查询包含一个 IP 地址或 URL 或网站域名或网站名称	“http://www.huohu123.com/channel/channel_43.html”、“4399”、“起点中文网”、“sohu.com”
6	查询包含名字(人名、书名、软件名、游戏影音名等)及其他词项	“奥巴马简介”、“ie 浏览器下载”、“网游洪荒”、“视频红高粱”
7	查询包含地理坐标或地方名及其他词项	“唐古拉山风暴”、“米兰时尚”、“北纬 67 东经 140”
8	查询包含确切日期或时间及其他词项	“08 高考状元”、“Q235+价格+2008”
9	查询包含数量词及其他词项	“科比 81 分视频”、“中星 9 号卫星”
10	查询包含英文缩写词及其他词项	“ERP 定义”、“3gp 手机电影下载”

4　数据集构建与特征分析

4.1　数据集及其标注

查询日志是用户行为的载体，由一系列信息需求组成，是用户查询意图分析的重要数据来源。本实验所采用的原始数据集为 2008 年 6 月的Sogou查询日志(含 25 天)[16]，其数据格式为: 用户访问时间+用户ID+查询词+该URL在返回结果中的排名+用户点击的顺序号+用户点击的URL。为保证样本无偏性，首先用 KNIME 数据挖掘工具[17]，使用“用户 ID=1400500473350”为随机种子，以日为单位分层抽取 288 条查询语句, 25 天共计 7 200 条查询语句。然后，去除重复和无法识别的查询语句(如纯日语、纯符号等)后，得到 6 456 条查询语句作为实验数据集。

相对于时间[18-19]、地理[20-21]等维度，查询专指度的标注工作更为复杂。考虑到有些类别不易从字面上理解其具体含义，在标注界面中不仅给出三种专指度类别的定义(见 3.1 节)，且对每个意图类别也进行简单说明:

(1) 信息类(Informational)

①有指导性的(Directed): 用户想知道关于某个话题的特定信息。

②无指导性的(Undirected): 用户想了解关于某个话题的任何信息。

③建议(Advice): 用户想得到一些建议、想法、指南或其他方面的指导。

④位置(Locate): 用户想知道哪里可以获取某现实产品或服务的地理位置。

⑤列表(List): 用户想得到一组可信的网站列表，以便进一步查询。

(2) 资源类(Resource)

①获取(Obtain): 用户想获得一个不是必须通过电脑才能使用的资源，如歌词、菜谱等。

②下载(Download): 用户想获得需要安装在本地电脑或者在其他电子设备上才能使用的资源(如某软件 APP)。

③娱乐(Entertainment): 用户只是想查看或获取网页上的娱乐资源。

④交互(Interact): 用户想通过在结果页上所得到的动态程序或服务与其它资源进行交互，如地图查询、股票收益等。

(3) 导航类(Navigation)

用户想访问已知的某特定网站或主页。McCreadie等[22]研究表明，将与查询相关的内容(如点击文档)融合到标注界面中能提高标注结果的准确度。基于此，本文在标注界面添加了用户提交查询后在 Sogou 查询日志中排名前 4 的点击结果页。若通过以上信息，标注者对该查询的判定仍具有模糊性，或用户点击列表中存在死链接，可利用该界面中的“百度搜索”或“搜狗搜索”链接在搜索结果中对专指度类别进行判定。对于查询串每条内容特征(见表 1)，笔者要求标注者一一选择，若符合要求，则选择“是”选项；否则选择“否”选项。

因此，本实验标注内容包括三大部分：专指度类别标注、查询意图类别标注和查询串属性特征标注。标注工作由 10 名武汉大学信息管理学院图情专业硕士研究生分 10 个小组完成，每人负责一个小组的标注工作：前 9 个小组每组标注 646 条数据，最后一个小组标注 642 条数据，形成标注数据集。

Cohen[23]提出用 KAPPA 系数作为评价判断的一致性程度指标。为此，实验从标注数据集中以小组为单位随机抽取了 10%的查询(共计 646 条数据)由另一名图情专业学生进行标注，并计算其标注结果与抽样数据集标注结果的 KAPPA 值，以评价标注者之间对专指度分类的一致性，检验个人主观因素对分类标注的干扰程度。实验表明强专指度和弱专指度的 KAPPA 值都高于 0.85(分别为 0.912 和 0.879)，而略专指度约为 0.754，平均值为 0.8483。可见，本实验标注结果一致性较高，且 6 456 条查询语句中强、略、弱专指度查询分别占查询总数的 56.1%、25.5%和 18.4%。

4.2 专指度特征分析

(1) 查询串属性特征分析

①查询串基本特征分析。通常，在文本分类中，停用表主要包括英文字符、数学字符、标点符号以及使用频率特高的单汉字[24]。而在专指度这一特定维度中，英文字符、数学字符以及虚词等都会对分类结果产生较大影响，因此，本实验进行统计时仅将标点符号作为停用词进行过滤，其他情况不做考虑。强、略、弱专指度下查询串的基本特征如表 2 所示。由于三类专指度在查询串长度、词项长度和词项个数上的最小值均为 1，因此主要列出其平均值、最大值和标准差。

表 2 查询串基本特征分析

专指度	查询串长度(字)			词项长度(字)			词项个数		
	平均值	最大值	标准差	平均值	最大值	标准差	平均值	最大值	标准差
强	6.99	40	2.578	1.79	5	0.459	3.97	9	1.609
略	4.35	8	1.257	1.80	5	0.525	2.47	9	0.813
弱	2.45	5	0.974	1.87	4	0.694	1.39	5	0.616

分析表 2 可知，强专指度的查询串长度分布并不集中：最多 40 个字，最少 1 个字(如查询“天”)，平均约 7 个字。而略专指度的查询串长度分布相对集中：最多 8 个字(如查询“中国股份公司+中国”)，最少 1 个字，平均为 4.35 个字。弱专指度的查询串长度分布更加集中：最多 5 个字(如查询“什么都能干”)，最少 1 个字，平均 2.45 个字。而且，强、略、弱专指度查询平均分别包含 3.97、2.47 和 1.39 个词项。随着专指度强度增加，词项长度平均值和标准差都略微降低，而词项个数平均值和标准差都明显升高。也就是说，专指度与查询串长度、词项个数在一定范围内成正相关，与词项长度成负相关。因此，可以选用查询串长度(字)、平均词项长度(字)和平均词项个数作为区分强、略、弱三类专指度的特征。

②查询串内容特征分析。在查询串内容特征分析方面，首先统计分析每一类别专指度下符合某一内容特征(即被标注为“是”的查询数目)占该类专指度查询总数的比例。如在强、略、弱专指度查询中，包含“英文缩写词及其他词项”的查询语句分别占 2.71%，3.571%和 0.505%。且从图 1 可知，强、略专指度在“包含地理坐标/地方名及其他词项”、“包含数量词及其他词项”两项特征的比例接近。对于其他特征，强专指度查询所占比例显著高于略专指度，而弱专指度在以上特征中所占比例很小，换言之，不含以上 10 项内容特征的查询更有可能属于弱专指度类。

再统计分析每一类别专指度下符合某一内容特征占符合该特征的查询语句总数的比例。例如，当查询语句包含“数量词及其他词项”时，强、略、弱专指度查询约占总数的 73.735%、23.22%和 3.05%。图 2 表明随着专指度从强到弱，各项内容特征在每类查询中的百分比明显降低。其中，诸如包含内容特征第 1 项-第 5 项(见表 1)的查询很可能属于强专指度类；“包含英文缩写词及其词项”、“包含坐标/地方名及其词项”的查询较有可能属于略专指度类；而弱专指度基本不符合任一特征。可见，除查询串基本特征外，选择以上 10 项内容特征有效识别专指度是十分必要的。

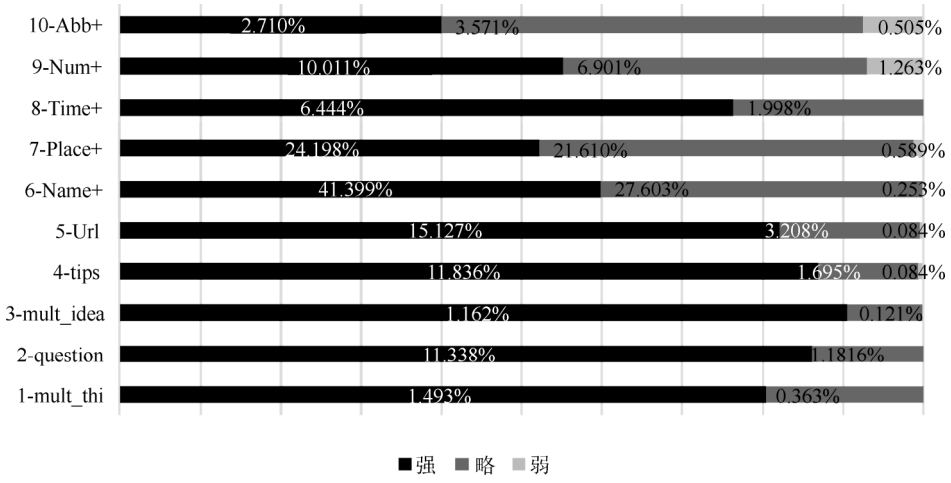


图 1　每一专指度下符合某一内容特征的查询数目占该类专指度查询总数比例

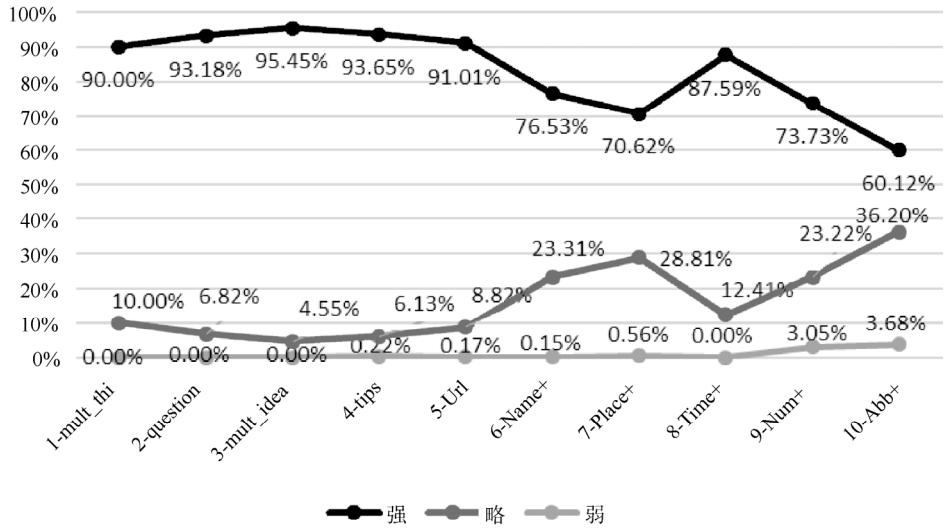


图 2　每一专指度下符合某一内容特征的查询数目占符合该特征的查询总数的比例

(2) 查询专指度与查询意图分析

Baeza-Yates 等[25]指出查询意图(即查询目标)是查询语句的最重要维度。为了更好地理解专指度与查询意图类别间的关系，进一步讨论专指度在各意图类别及其子类别下的分布情况及所占比值，分析所采用的意图类目体系为 Rose 二层查询意图类目体系，如图 3 和图 4 所示。

从图 3、图 4 分析可知，不管查询语句属于哪一种专指度类别，大多数查询的目标为信息类，资源类次之，最后为导航类。而且，随着专指性强度降低，查询目标为信息类的比例显著增加，为导航类的比例显著减少，为资源类的比例虽逐渐减少但波动不大。具体而言，有指导类、建议类、导航类在强专指度查询中所占比例明显高于略、弱专指度查询。而无指导类在弱专指度查询中所占比例显著高于其他两类专指度。

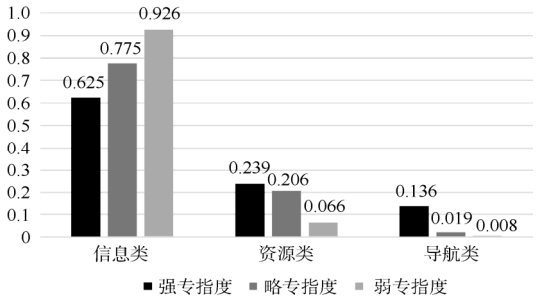
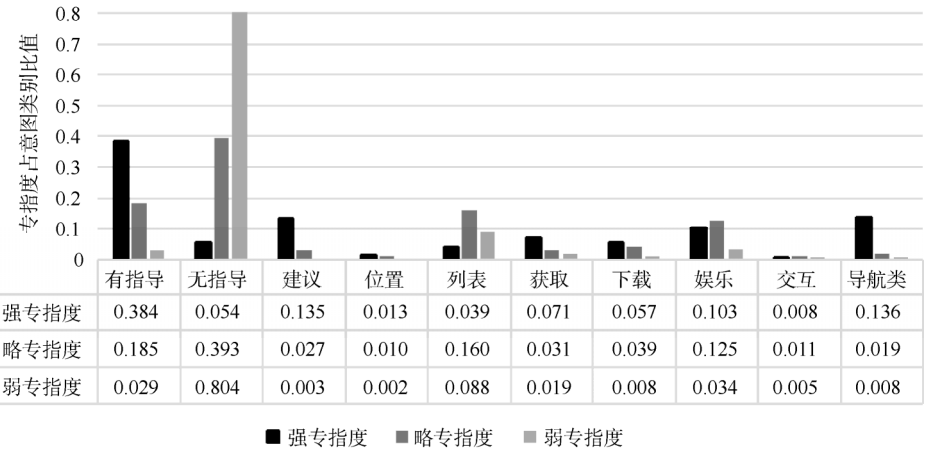


图 3　专指度查询在三大意图类别中的比值



	有指导	无指导	建议	位置	列表	获取	下载	娱乐	交互	导航类
强专指度	0.384	0.054	0.135	0.013	0.039	0.071	0.057	0.103	0.008	0.136
略专指度	0.185	0.393	0.027	0.010	0.160	0.031	0.039	0.125	0.011	0.019
弱专指度	0.029	0.804	0.003	0.002	0.088	0.019	0.008	0.034	0.005	0.008

图 4　专指度查询在子意图类别中所占比值

5　查询专指度自动识别

第 4 节主要选取查询串基本特征和内容特征共 13 个属性特征作为分类特征。在识别阶段，基于以上特征利用已标注的数据集选用常见的 Web 文本分类器实现专指度的自动识别。最后，采用十折交叉检验对实验结果进行分析、比较与评测。

5.1　专指度自动识别模型

目前，在众多的 Web 文本分类模型中，应用最为广泛的是决策树[26]、支持向量机(Support Vector Machine, SVM)[27]和朴素贝叶斯模型[28]。因此，本文分别训练上述三种分类器对专指度进行自动识别。

(1) 利用决策树算法识别专指度

采用 C4.5[29]决策树分类法对 6 456 条查询语句进行强、略、弱三类专指度自动识别。分类特征共 13 个，包括 3 个查询串基本特征及 10 个查询串内容特征(如 Mult_thiCompare, Mult_idea, IsUrl)。根据以上特征所构造的决策树分类效果良好：宏平均准确率为 0.853，强、略、弱专指度准确率分别为 0.888、0.724 和 0.927。

(2) 利用 SVM 算法识别专指度

SVM 模型建立在小样本统计理论基础上，在解决有限样本、非线形及高维模式识别问题中具有特有的优势。而且在进行文本分类时，SVM 的总体分类效果往往好于决策树[30]。因此，本文在原标注数据集上训练 SVM 分类器完成识别。实验结果表明当核函数为径向基函数，cost、gamma 值分别为 1 024、1 时，强、略、弱专指度准确率分别为 0.9、0.72 和 0.935，宏平均分类准确率为 0.86。与利用决策树算法识别专指度相比，SVM 分类效率略好于决策树，但在略专指度识别效果方面，两者表现都不够好。

为此，本文计算出专指度分类结果的混淆矩阵，表示某一类别被判别为另一类别(包括本类别)的数目，如表 3、表 4 所示。例如，对于强专指度查询，有 3 384 条被决策树算法正确分类，有 236 条被错误分类成略专指度查询，有 2 条被错误分类成弱专指度查询。分析表 3、表 4 可知，强、弱专指度之间绝大多数都能正确分类，而强、略专指度之间和略、弱专指度之间错误分类的查询数目较多。

表 3　决策树分类结果的混合矩阵

应分类 \ 被分类	强	略	弱
强	3 384	236	2
略	424	1 148	75
弱	3	201	983

表 4　SVM 分类结果的混合矩阵

应分类 \ 被分类	强	略	弱
强	3 350	270	2
略	371	1 210	66
弱	3	200	984

(3) 利用朴素贝叶斯算法识别专指度

决策树或 SVM 分类器均已取得了较好的识别效果。但为了降低略专指度被错误分类的影响，本文考虑将查询语句与各类专指度之间的映射关系以概率的形式表现出来。朴素贝叶斯算法假设实例的所有属性之间相互独立，并独立地学习每个特征在每一给定类

表 5　略专指度查询被分类为强或弱专指度查询的后验概率值

编号	查询	后验概率			分类结果
		强专指度	略专指度	弱专指度	
189	“超级 QQ”	0.496	0.461	0.044	强专指度
3495	“这一次来真的”	0.618	0.361	0.021	强专指度
6059	“优化电脑系统”	0.524	0.424	0.053	强专指度
1573	“复古 dance”	0.169	0.183	0.648	弱专指度
4793	“无法无天”	0.007	0.351	0.641	弱专指度
6153	“再世魔导”	0.098	0.39	0.512	弱专指度

别下的条件概率，分类器应用贝叶斯公式计算某特定实例在给定属性值下各类别的后验概率，并返回使后验概率最大的类别[31]。然而，该算法的独立性假设在实际研究中很少满足，但文献[32]表明朴素贝叶斯算法的表现与独立性假设是否满足没有必然联系。因此，最后选用朴素贝叶斯分类器实现专指度自动识别。

实验中，朴素贝叶斯分类器计算查询语句被判断为某一特定类别的概率，选择具有最大后验概率的专指度类别作为该查询语句所属的类别。表 5 列出了一些略专指度查询被识别成强或弱专指度类别的后验概率值。可知，以概率的形式表示查询专指度比将查询语句直接判定为某类别，能够在一定程度上降低略专指度被错误分类的影响。

5.2　实验结果分析与评测

在评估分类效果时，为减少实验误差，均采用十折分层交叉验证[33]的方法，选取准确率 P、召回率 R、F-measure[33]三个参数评估实验结果，如表 6 所示。其中 F-measure 是一个综合考虑准确率和召回率的测试参数，本文认为 P 和 R 有相等的权值，因此采用常用的 F_1。

表 6　十折交叉检验的实验结果

专指度	C4.5 决策树分类法			SVM			朴素贝叶斯		
	P	R	F-measure	P	R	F-measure	P	R	F-measure
强	0.888	0.934	0.911	0.9	0.925	0.912	0.895	0.85	0.872
略	0.724	0.697	0.663	0.72	0.735	0.727	0.637	0.68	0.658
弱	0.927	0.828	0.875	0.935	0.829	0.879	0.784	0.829	0.806
宏平均	0.853	0.854	0.853	0.86	0.859	0.859	0.809	0.803	0.806

从表 6 可知，三种分类器的 F-measure 均高于 0.8，SVM 分类效果最好，决策树次之，最后为朴素贝叶斯。具体而言，SVM 对查询专指度的分类宏平均准确率略好于决策树(约 0.7%)，且前两者都好于朴素贝叶斯(分别为 4.4%和 5.5%)。然而，朴素贝叶斯分类器计算所得的后验概率值能够更直观地表示并区别查询语句属于某一专指度类别的可能性。总之，以上三种识别方法应视具体情况进行选取。SVM 分类器鲁棒性好但内存开销大、运行时间长，而决策树和朴素贝叶斯算法简单，运行时间快但难以处理大数据[26]。在不考虑内存开销、训练时间的情况下，为保证较高的分类准确率，选用 SVM 分类器；否则选择决策树。当需要以概率的形式表示分类结果以便应用到检查模型中时，优先选择朴素贝叶斯分类器。

6　结　语

本文基于综合搜索引擎 Sogou 查询日志，人工构建查询专指度标注集，统计分析各类别下查询串基本特征及内容特征，并以此作为专指度的分类特征。选用决策树、SVM 和朴素贝叶斯分类器实现查询专指度的自动识别。通过实验发现:

(1) 专指度确实与查询串长度在一定范围内成正相关。长查询串一般是强或略专指度查询，但短查询串也有可能是强专指度查询。

(2) 当用户查询语句含有某一内容特征尤其是前5项中任一项(见表1)时，很可能是强专指度查询，当含有内容特征6、7、10时也有可能是略专指度查询，当不含上述内容特征时，有较大可能是弱专指度查询。

(3) 专指度与查询目标之间关系密切，弱专指度查询大都属于信息类，略专指度查询基本不属于导航类，而强专指度查询属于各类别均有可能。

尽管本文尝试对专指度进行全面透彻的分析，但仍存在以下不足，也是后续研究的主要方向:

(1) 除查询串属性特征外，如何借助外部资源挖

掘其他特征(如点击结果页)，进一步提高专指度识别准确率。

(2) 尽管有研究表明朴素贝叶斯分类器的表现与独立性假设是否满足没有必然联系，但在本实验中是否如此仍需验证。

(3) 除查询目标外，如何将专指度与其他维度(如时间、地理等维度)结合起来，综合分析多维度之间的相互影响。

参考文献：

[1] comScore, Inc. Global Search Market Draws More than 100 Billion Searches per Month [R/OL]. (2009-08-31). [2014-01-11]. http://www.comscore.com/Insights/Press_Releases/2009/8/Global_Search_Market_Draws_More_than_100_Billion_Searches_per_Month.

[2] González-Caro C, Calderón-Benavides L, Baeza-Yates R, et al. Web Queries: The Tip of the Iceberg of the User's Intent [C]. In: Proceedings of the 4th ACM WSDM Conference, Hong Kong, China. 2011.

[3] Nguyen B V, Kan M. Functional Faceted Web Query Analysis [C]. In: Proceedings of the 16th International Conference on World Wide Web. ACM, 2007.

[4] Song R, Luo Z, Wen J, et al. Identifying Ambiguous Queries in Web Search [C]. In: Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 1169-1170.

[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 [C]. In: Proceedings of the 13th International Conference on World Wide Web. New York: ACM, 2004: 13-19.

[7] Donato D, Donmez P, Noronha S. Toward a Deeper Understanding of User Intent and Query Expressiveness[C]. In: Proceedings of ACM SIGIR for Query Representation and Understanding Workshop. ACM, 2011.

[8] Chang Y, He K, Yu S, et al. Identifying User Goals from Web Search Results [C]. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'06). IEEE, 2006: 1038-1041.

[9] Calderón-Benavides L, González-Caro C, Baeza-Yates R. Towards a Deeper Understanding of the User's Query Intent [C]. In: Proceedings of the SIGIR 2010 Workshop on Query Representation and Understanding. 2010:21-24.

[10] Song R, Luo Z, Nie J, et al. Identification of Ambiguous Queries in Web Search [J]. Information Processing & Management, 2009, 45(2): 216-229.

[11] White M D, Iivonen M. Questions as a Factor in Web Search Strategy [J]. Information Processing & Management, 2001, 37(5): 721-740.

[12] Phan N, Bailey P, Wilkinson R. Understanding the Relationship of Information Need Specificity to Search Query Length [C]. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07). New York: ACM, 2007: 709-710.

[13] Hafernik C T, Jansen B J. Understanding the Specificity of Web Search Queries [C]. In: Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems (CHI EA'13). New York:ACM, 2013:1827-1832.

[14] Ingwersen P, Jarvelin K. The Turn [M]. Springer, 2005.

[15] Ramírez G, de Vries A P. Relevant Contextual Features in XML Retrieval [C]. In: Proceedings of the 1st International Conference on Information Interaction in Context. New York: ACM, 2006: 56-65.

[16] 用户查询日志(SogouQ) [EB/OL]. [2013-12-27]. http://www.sogou.com/labs/dl/q.html. (User Query Logs (SogouQ) [EB/OL]. [2013-12-27]. http://www.sogou.com/labs/dl/q.html.)

[17] KNIME [EB/OL]. [2012-09-24]. http://www.knime.org/.

[18] Metzler D, Jones R, Peng F, et al. Improving Search Relevance for Implicitly Temporal Queries [C]. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09). New York: ACM, 2009: 700-701.

[19] 张晓娟，陆伟，周红霞. 用户查询中潜在时间意图分析及其检索建模[J]. 现代图书情报技术，2011(11): 38-43. (Zhang Xiaojuan, Lu Wei, Zhou Hongxia. Analyzing and Retrieval Modeling on Implicit Temporal Intents in User's Queries [J]. New Technology of Library and Information Service, 2011(11): 38-43.)

[20] Ding J, Gravano L, Shivakumar N. Computing Geographical Scopes of Web Resources [C]. In: Proceedings of the 26th International Conference on Very Large Databases (VLDB'00). San Francisco: Morgan Kaufmann Publishers Inc., 2000: 545-556.

[21] Jones C B, Abdelmoty A I, Fu G. Maintaining Ontologies for Geographical Information Retrieval on the Web [M]. Springer Berlin Heidelberg, 2003: 934-951.

[22] McCreadie R M C, Macdonald C, Ounis I. Crowdsourcing a News Query Classification Dataset [C]. In: Proceedings of the

3rd Computer Science and Engineering. 2010.

[23] Cohen J. A Coefficient of Agreement for Nominal Scales [J]. Educational and Psychological Measurement, 1960, 20: 37-46.

[24] 周钦强，孙炳达，王义．文本自动分类系统文本预处理方法的研究[J]. 计算机应用研究，2005, 22(2): 85-86. (Zhou Qinqiang, Sun Bingda, Wang Yi. Study on New Pretreatment Method for Chinese Text Classification System [J]. Application Research of Computers, 2005, 22(2): 85-86.)

[25] Baeza-Yates R, Calderón-Benavides L, González-Caro C. The Intention Behind Web Queries [C]. In: Proceedings of the 13th International Conference on String Processing and Information Retrieval (SPIRE'06). Berlin, Heidelberg: Springer-Verlag, 2006: 98-109.

[26] Mitchell T M. 机器学习[M]. 曾华军，张银奎等译．北京：机械工业出版社，2008: 62-70. (Mitchell T M. Machine Learning [M]. Translated by Zeng Huajun, Zhang Yinkui, et al. Beijing: China Machine Press, 2008: 62-70.)

[27] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995.

[28] Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-one Loss [J]. Machine Learning, 1997, 29(2-3): 103-130.

[29] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Francisco: Morgan Kaufmann Publishers Inc., 1993.

[30] 邓乃扬，田英杰．支持向量机：理论、算法与拓展[M]. 北京：科学出版社, 2009: 77-85. (Deng Naiyang, Tian Yingjie. Support Vector Machine: Theory, Algorithms and Extensions [M]. Beijing: Science Press, 2009: 77-85.)

[31] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques [M]. Morgan Kaufmann Publishers, 2006.

[32] 范金金，刘鹏．朴素贝叶斯分类器的独立性假设研究[J].计算机工程与应用，2008, 44(34): 139-141. (Fan Jinjin, Liu Peng. Research on Naive Bayesian Classifier's Independence Assumption [J]. Computer Engineering and Applications, 2008, 44(34): 139-141.)

[33] Manning C D, Schutze H, Raghavan P. 信息检索导论 [M]. 王斌译．北京：人民邮电出版社，2010: 105-107, 196-200. (Manning C D, Schutze H, Raghavan P. Introduction to Information Retrieval [M]. Translated by Wang Bin. Beijing: Posts & Telecom Press, 2010: 105-107, 196-200.)

作者贡献声明：

唐祥彬：文献调研，分析数据，起草论文，论文多次版本以及最终版本修订；
陆伟：提出研究思路，论文多次版本以及最终版本修订；
张晓娟：文献调研，分析数据，论文初稿修订；
黄诗豪：标注系统构建，实验数据处理。

收稿日期: 2014-04-23
收修改稿日期: 2014-05-20

Feature Analysis and Automatic Identification of Query Specificity

Tang Xiangbin[1]　Lu Wei[2]　Zhang Xiaojuan[1]　Huang Shihao[1]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper constructs a human-annotated collection on the basis of Sogou query logs, aims at feature analysis and automatic identification of query specificity, as well as evaluates and compares the identifing results. [**Methods**] The queries' basic features and content features are selected and analyzed. And then the decision tree, SVM and Naive Bayes classifiers are built and trained to achieve the automatic query specificity classification. [**Results**] Using the features mentioned above, an effective query specificty identification is obtained. Finally, the macro average F-measures of the identification effects are all above 0.8. [**Limitations**] Users' clickthrough information is not selected during the feature selection, and the ignorance of the conditional independence assumption of the Naive Bayes classifier in this particular experiment should be further verified. [**Conclusions**] The queries' basic features and content features, by themselves, can well distinguish broad, medium, and specific queries.

Keywords: Query specificity　Decision tree　SVM　Naive Bayes