DOI:10.3969/j.issn.1673-4785.201101001
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120218.1616.001.html

智能文本搜索新技术

王占一[1,2],徐蔚然[1,2],郭军[1,2]
(1.北京邮电大学 模式识别与智能系统实验室,北京 100876;2.北京邮电大学 信息与通信工程学院,北京 100876)

摘 要:面对当今互联网上海量的信息,以及搜索信息准确、高效、个性化等需求,提出了一套包括信息检索、信息抽取和信息过滤在内的智能文本搜索新技术.首先举荐了与信息检索新技术相关的企业检索、实体检索、博客检索、相关反馈子任务.然后介绍了与信息抽取技术相关的实体关联和实体填充子任务,以及与信息过滤技术相关的垃圾邮件过滤子任务.这些关键技术融合在一起,在多个著名的国际评测中得到应用,如美国主办的文本检索会议评测和文本分析会议评测,并且在互联网舆情、短信舆情和校园网对象搜索引擎等实际系统中得到了检验.

关键词:智能文本搜索;文本检索;文本分析
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-4785(2012)01-0040-10

New technologies of intelligent text search

WANG Zhanyi[1,2], XU Weiran[1,2], GUO Jun[1,2]
(1. Pattern Recognition and Intelligent System (PRIS) Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: To adapt to the massive amount of information on the internet and the need for accuracy, efficiency, and individualization, a set of technologies of intelligent text search including information retrieval, extraction, and filtering were proposed. First, new technologies of information retrieval were illustrated including the subtasks of enterprise retrieval, entity retrieval, blog retrieval, and relevance feedback. Second, the subtask of entity linking and slot filling related to information extraction was introduced. Finally, the subtask of spam e-mail filtering related to information filtering was described. These technologies were converged for application in many well-known international evaluations. These include the text retrieval conference (TREC) and text analysis conference (TAC) sponsored in the USA, and these technologies of intelligent text search were proven in practical applications such as public opinions on the Internet, short message opinions, and the campus object search engine (COSE).

Keywords: intelligent text search; text retrieval; text analysis

随着互联网技术的飞速发展,网络上的信息呈爆炸式增长.用户需要在这些海量信息数据中找到自己需要的内容,不是简单定位到某一个网站或网页,而是越精准、全面越好.同时他们希望使用尽量少的描述就可以找到自己感兴趣的内容,不带有任何垃圾信息.如何满足用户对这些信息的高精度、高效率、个性化、完备性等需求,是当前信息检索和数据挖掘面临的新问题.

传统的文本搜索基于数据库查询、关键词搜索等技术,有很强的局限性.而智能文本搜索解决的是数据海量、数据稀疏、大量并发请求、数据特征演进、主客观交叉等困难问题,从技术角度来说,智能文本搜索融合了信息的检索、抽取、过滤等方面.检索是由用户提出查询请求,系统根据这个需求对Web信息进行查询并给出结果.抽取是把文本里包含的信息进行结构化处理,变成表格一样的组织形式.过滤是系统根据预先设定的条件,对Web中与该条件相符的信息进行获取、隔离或封堵[1].

为了探索前沿技术,解决上述问题,各国学术界、产业界和政府部门都给予了高度关注,一系列评

收稿日期:2011-01-02. 网络出版时间:2012-02-18.
基金项目:国家自然科学基金资助项目(60905017);高等学校学科创新引智计划项目(B08004).
通信作者:王占一. E-mail:wangzhanyi@gmail.com.

测活动应运而生. 文本检索会议(text retrieval conference, TREC)作为文本检索领域最权威的评测会议,关注着检索技术的最新发展,比较客观地反映了十几年来的研究趋势. TREC 是由美国国家技术标准局(NIST)和美国国防部(DOD)联合主办,创立于 1992 年,主要目的是通过提供评价大型文本检索方法所必需的基础设施来支持对信息检索的研究[2]. 关注 TREC,有利于加强各个科研机构和企业之间的交流,有利于评价检索方法在实际问题中的效果,也有利于加快实验室的技术商品化的速度.

TREC 的参赛队伍从开始的 22 个发展到 2010 年的 75 个. 北京邮电大学模式识别实验室多年来致力于模式识别和网络搜索技术,从 2005 年开始参加 TREC 的多项评测并取得了较好的成绩,如垃圾邮件过滤、企业检索、博客检索、实体检索、相关反馈等. 同时,该团队还参加了国家“863”计划项目中文本分类、SigHan 分词、TAC 和中文倾向性分析等评测. 评测中涉及的任务除了用于新技术的研究,也是为了解决实际问题. 基于评测中的智能文本搜索新技术,一些实际系统也相应地被开发出来,并在实际应用中得到了检验.

本文以权威评测为主线,详细介绍智能文本搜索新技术. 第 1 部分以企业检索、实体检索、博客检索和相关反馈为例介绍信息检索新技术;第 2 部分以文本分析会议评测为例介绍信息抽取新技术;第 3 部分以垃圾邮件过滤为例介绍信息过滤新技术;第 4 部分介绍以上述技术为核心的实际应用系统,如互联网舆情系统、短信舆情系统、校园对象搜索引擎系统等;最后是总结和展望部分.

1　信息检索

1.1　企业检索

文本检索会议从 2005—2008 年制订了企业检索(enterprise track)评测任务[3],企业检索的目的是研究在企业内部数据中的用户检索行为,主要包含邮件检索(2005—2006 年)[4-5]、文档检索(2007—2008 年)[6]和专家检索(2005—2008 年)任务. 其中,专家检索是重点和难点,它的目的是寻找企业中关于某一主题的专家. 具体地,专家检索需要分成两部分来解决:一是确定所给语料集中的专家,二是计算查询与专家的相关度. 专家的标识主要是姓名和邮箱,定位专家的方法主要有命名实体识别、查询人名列表、匹配邮箱、称谓、职务等. 在实际中,这些方法经常综合运用.

1.1.1　二阶排序模型

二阶排序模型的主要思路是通过文档为桥梁,计算查询和专家的相关度. 如式(1),检索的第 1 阶段是普通的文档检索,找出一定数量的相关文档,计算出查询 Q 和文档 D_i 的相关度 $S_{\text{core}}(D_i,Q)$;第 2 阶段计算事先确定好的专家 E_j 和这些文档的相关度 $S_{\text{core}}(E_j,D_i)$;最后综合文档和查询的相关度得到查询和专家的相关度 $S_{\text{core}}(E_j,Q)$,就可以对和查询相关的专家排序了.

$$S_{\text{core}}(E_j,Q) = \sum_{i=1}^{N_r}(S_{\text{core}}(D_i,Q) \times S_{\text{core}}(E_j,D_i)). \tag{1}$$

式中:N_r 表示第 1 阶段得到的文档中,用于第 2 阶段的文档数量.

文档检索使用的算法包括语言模型、KL 距离、BM25 等. 计算专家 E_j 和这些文档的相关度 $S_{\text{core}}(E_j,D_i)$ 可以使用式(2):

$$S_{\text{core}}(E_j,D_i) = n(f_{ij}) \times \log\frac{N+1}{d(f_j)+0.5}. \tag{2}$$

式中:$n(f_{ij})$ 表示文档 D_i 中某一专家的名字和邮箱出现的次数,N 是语料集中文档的数目,$d(f_j)$ 是出现该专家名字和邮箱的文档数目.

二阶排序模型思路清晰,有理论依据且易于实现,但它以整篇文档为桥梁,单纯以专家名或邮箱代表全部的专家信息,方法较为粗糙,没有在文档中做更细致的挖掘.

1.1.2　专家经验模型

专家经验模型的主要思路是提取专家在文档中的上下文组成该专家的“经验”,再计算专家经验的概率. 提取上下文的过程相当于为该专家开了一个“窗口”,因此也叫作专家窗口模型. 笔者认为专家名或邮箱的上下文是与该专家密切联系的信息,那么在确定一个专家的同时将其前后一定数量的词也提取出来组成新的文档,这个文档就是包含该专家相关信息的文档. 因此只要检索到这个文档就认为该专家和查询是相关的. 这个过程表示为

$$P(E/Q) = \sum_d P(E_d/Q).$$

式中:E_d 表示由专家经验组成的文档. 另外,经过反复的实验发现,窗口的长度取专家前后各 150 个词效果最好. 表 1 给出了二阶排序和专家经验 2 种模型的性能比较.

表 1　2 种专家检索模型的对比

Table 1　Comparison of two kinds of expert track model

模　型	MAP	Bpref	P@10
二阶排序	0.183 3	0.418 2	0.308 0
专家经验	0.216 0	0.518 0	0.340 0

1.2　实体检索

实体检索,或称实体追踪(entity track)是 2009

年 TREC 评测新增加的一项任务[7]. 它可以看作是从2005—2008年的专家检索任务发展而来. 与专家检索相比,它具有更新更丰富的内容. 许多使用搜索引擎的用户本意并不是找出各种各样的文档,而是想知道答案是哪些具体的实体,因此,文本搜索的核心任务是相关实体查找(related entity finding, REF). REF 需要解决的问题是:给出一个输入实体,连同它的名字、主页、目标实体的类型,还有描述它们之间关系的文本,找出与目标类型相符的实体,这些实体能够表示前面要求的与输入实体的关系. 对于每个查询,要求输出实体的排序,且每个实体必须有惟一的主页. 笔者的工作主要关注3个方面:针对每个查询,找出相关的实体;依据检索模型,对实体进行排序;为每个实体赋予一个主页.

1.2.1 实体抽取

与专家检索首先要定位专家相似,实体检索的前提是必须找出与查询相关的实体,而且尽量提高查准率和查全率,这就要用到实体抽取的技术. 通常,实体抽取主要分为基于统计和基于规则2种. 基于统计的方法例如最大熵(maximum entropy)[8]或条件随机场(conditional random field)[9]将人名、地名等命名实体标识出来. 基于规则的方法例如构建命名实体词典,用词典过滤出符合要求的实体.

为了更准确、更全面地抽取实体,可以将几种方法混合使用,即规则-统计-规则. 首先通过观察语料集、构造查询在搜索引擎或维基百科中查找特殊网页,这种网页多数以表格的方式呈现,或者有其他明显的特征. 然后通过适当的规则将这些可信度较高的实体抽取出来. 这种方法可以保证准确率,但是实体的数量不够. 接下来使用文档检索得到相关度最高的前 $N(N=5)$ 篇文档,使用基于统计的命名实体识别工具抽取出与目标实体类型相同的实体. 调整 N 可以保证实体的数量,但是准确率不高,这就又要用到基于规则的方法. 利用维基百科中每个词条的语义标签建立各种实体类型的映射规则,如对于组织名(organization),以"组织"、"公司"等开头的标签,采集这些标签对应的实体,建立实体词典,前面用工具抽取出的"实体"再经过词典过滤,添加到实体列表中.

1.2.2 检索模型

有了实体列表就可以依据检索模型对实体排序了. 在实体检索任务中,根据查询、文档、实体三者的关系,形象地构建了2种模型:文档中心模型和实体中心模型.

文档中心模型将文档 d 看作查询 q 和实体 e 的桥梁,查询和实体的相关度由合并 q、d 的相关度和 e、q 的相关度得到. 文档中心模型借鉴了专家检索中的二阶思路,不同之处在于专家换成了实体. 第1阶段计算查询和文档的相关度使用的是语言模型和推理网络. 第2阶段计算实体和文档的相关度也是一个检索的过程,可以采用概率模型等,将实体转换成查询后就和第1阶段相同了.

实体中心模型是实体处在结构的中层,文档或文档的片断在底层支撑实体,实体与顶层的查询直接相连. 与文档中心模型不同,实体中心模型只需要1次检索过程.

单纯用文档支持实体过于粗糙,参考专家经验模型,取实体的上下文作为与实体相关的信息. 这里的上下文称为片断,同样也取实体前后的150个词,将某个实体的各个片断汇集在一起,形成一个新的文档. 实体与实体文档一一对应,利用查询与这些文档的相关度就可以直接对实体进行排序. 排序的具体算法有前面提到的语言模型、BM25等.

1.2.3 确定主页

与专家不同,实体需要一个主页与之对应,也是在网络上的惟一标识. 为实体分配主页的方法主要有3种:1)计算实体和各相关文档的相关度,取相关度最高的作为主页,这种方法依赖于文档的内容;2)制定规则,将实体与文档的URL作比较,找出相似度最高的作为主页;3)利用已有的外部资源,如搜索引擎排序靠前的网页、维基百科的参考链接等. 实际应用中混合使用这3种方法,相互补充,达到尽量准确分配主页的目的.

1.3 博客检索

文本检索会议 TREC 从2006年起制定了博客检索任务(Blog track),最初只对博客的观点度及其与查询的相似性进行研究. 博客检索从2008年起开始关注对博客倾向性的分析,并于2009年提出博客精选任务,该任务将博客的倾向性分为3类:"个人的(personal)"或"官方的(official)";"深入分析的(in-depth)"或"浅层描述的(shallow)";"表达观点的(opinionated)"或"描述事实的(factual)",其目的是在博客关于查询的相似性检索的基础上进一步对博客的倾向性进行检索和排序. 笔者参加了2007—2010年的博客检索任务,并于2009年在多项评测指标中都取得了第1名的优异成绩.

1.3.1 博客精选(Blog distillation)

随着各大博客网站的推出和兴起,网络上涌现出海量的博客用户,这些博客内容丰富多彩,种类多样,同时也充斥着各种感情色彩,可谓鱼龙混杂. 在信息如此泛滥的情况下来判断相对比较具体的一些话题的倾向性是有困难的,因此有必要事先挑选出一些与话题相关性大的博客,再判断其倾向性. 这也是把话题检索作为倾向性检索基础的原因.

在 2009 和 2010 年的话题检索任务中,笔者使用的方法基本相同,都是将其看作 Learning to Rank 问题,即通过学习博文的排序,利用一定的算法来获得博客的排序. 针对这一问题,采用 Voting 模型[10],即一个博客里的博文被看作是这个博客的支持者,该博客里的博文对于话题的相关性就越大,同时相关的博文数量越多,该博客的相关性就越大,排序越靠前.

具体的方法如下:将所有的数据以博文为单位输入 Indri 建立索引,用话题 Q 在 Indri 里进行查询,得到博文的相关性分数和排序. 通过此排序来获得博客排序,如式(3):

$$S_{\text{core}}(B,Q) = \sum S_{\text{core}}(p,Q)/|B|. \quad (3)$$

式中:B 表示一个博客,博客 B 中的一篇博文用 p 表示,$S_{\text{core}}(B,Q)$ 表示一个博客的相关性得分,$S_{\text{core}}(p,Q)$ 表示从 Indri 中获得的博文的相关性分数,$|B|$ 表示一个博客下博文的数量. 将获得的相关博客的分数排序,排在前 100 的被认为是与话题最相关的博客.

1.3.2　个人与官方(personal vs. official)

博客的兴起使个人和组织的言论表达变得更加便利,然而因特网用户可能不大喜欢宣传性、商业性的博客,更加喜欢以个人的名义发表的文章,这样就使得个人、组织搜索的研究变得具有现实意义.

博客的个人、组织检索,是 TREC 评测 2009 年新增加的一项子任务,被安排在话题检索之后. 在话题检索中,得到与话题相关的博客,再对其进行个人、组织检索. 最近 2 年分别采用了 2 种不同的方法来进行个人、组织检索.

2009 年主要采用了组织机构名的区分方法,因为官方/组织的博客的书写惯例,一般会将组织名称放在文章的开头位置,有种“开门见山”的感觉;所以根据相同的组织机构名称在文章中出现的频率和位置来给相关的博客进行打分,最后根据分数的高低来进行排序和检索,即可分别得到个人和组织的博客.

2010 年主要采用了基于机器学习的分类方法,将个人和组织的检索看作是一种分类的问题,在训练模型中,利用机器学习的方法来分别构建含有个人和组织信息的词典. 在构建词典前会做一个文本特征降维的处理,然后利用 VSM 模型用这 2 个词典对相关博客进行打分和排序[11],最后分别得到个人和组织的博客.

1.3.3　表达观点与描述事实(opinionated vs. factual)

博客的观点度与客观度排序评测旨在开发一种有效的检索系统,使其能根据博客中关于某话题所表达一种观点或陈述一个事实的强烈程度,来对这些博客进行排序.

笔者在 2008 和 2009 年都使用了同一种情感分析模型[12],对于博客的观点度打分如式(4):

$$S_o = \frac{|N_{\text{pos}} - N_{\text{neg}}|}{N_{\text{pos}} + N_{\text{neg}}}. \quad (4)$$

式中:N_{pos} 和 N_{neg} 分别代表主观和客观的博文数.

与前 2 年不同,2010 年的博客检索中使用了基于词典的方法,主要分为 3 个步骤:

1)利用信息增益与互信息自动生成“主观词词典”和“客观词词典”. 通过信息增益在训练集中挑选对观点型博客和客观型博客区分度高的词,作为词典的候选词. 由信息增益生成的候选词并没有被分类为“观点型”或“客观型”,为了生成最终的 2 种词典,利用互信息进一步将这些候选词分为“观点型”和“客观型”[13].

2)计算观点度得分和客观度得分. 对于每个查询 q 和词典中的词 t,在相关文档集中计算 TF-IDF 权重 $w_{\text{tfidf}}(t)$,同时用一种词权重模型[14]计算查询权重 $w_{\text{bol}}(q)$,然后将 2 个权重相加得到博客的观点度得分 S_{op} 和客观度得分 S_{fa}.

3)排序. 首先在相关文档集中找到每篇博客的相关性得分 $S_{\text{core}}(B,Q)$,然后将 $S_{\text{core}}(B,Q)\times S_{\text{op}}$ 和 $S_{\text{core}}(B,Q)\times S_{\text{fa}}$ 分别作为观点度排序和客观度排序的最终得分.

1.3.4　深入分析与浅层描述(in-depth vs. shallow)

2009 年首次提出博客的深浅度分析任务. 笔者提出了 L-Qtf 系数进行博文的深浅度分析[15]. 然后根据每一个博客下深度博文与浅度博文的数量,得到每一个博客的深度分析程度或浅度分析程度的排序. 最后将每一个博客深浅度的排序值与相应的博客精选的相关性值合并得到最终结果.

1)根据 L-Qtf 系数进行每一篇博文的深浅度分析:

$$k(\text{L-Qtf}) = \sum_{t\in Q\cap D}\frac{1+\ln(1+\ln(f_t))}{(1-s)+s\dfrac{l_d}{l_{\text{avg}}}}f_{qt}.$$

式中:f_t 和 f_{qt} 分别为查询中的单词在博文中的词频和在查询中的词频,在计算 f_t 和 f_{qt} 之前,进行词干化处理(stemming),其作用是将词的各个词形变化还原为同一词干,例如“selling”和“sells”是“sell”的不同词形,这样的处理可以提高查询词在博文中的覆盖率;l_d 为博文的长度;l_{avg} 为同一查询下全部相关博文的平均长度;在实验中参数 s 设置为 0.2.

2)根据博文的 L-Qtf 系数进行博客的深浅度分析. 在同一查询下,根据 L-Qtf 系数的值对博文进行排序,取该排序的前 45% 判定为深度表述的博文,

后45%判定为浅度表述的博文. 计算每一个博客下深度表述博文与浅度表述博文数量的差值,并对该博客下博文的数量进行归一化,得到该博客的深浅度分析结果 S_i.

$$S_i = S_{\text{core}}(b_x, Q) = \frac{\sum_{i=1}^{n} S_{\text{indepth}}(p_i, Q) - \sum_{i=1}^{n} S_{\text{shallow}}(p_i, Q)}{n}.$$

式中:$S_{\text{core}}(b_x, Q)$为深浅度分析结果,为了区分下面的合并方法,用 S_i 表示.

3)与博客的相关性结果合并得到最终排序. 一个博客深浅度分析的最终结果不能仅依赖于深浅度分析,还要考虑该博客对于查询词的相关性,所以提出了以下的合并模型:

$$S_j = \begin{cases} S_{\text{core}}(b_x, Q) \times S_{\text{norm}}(B, Q), S_{\text{core}}(b_x, Q) \geqslant 0; \\ 1 - S_{\text{core}}(b_x, Q) \times S_{\text{norm}}(B, Q), S_{\text{core}}(b_x, Q) < 0. \end{cases}$$

式中:$S_{\text{norm}}(B, Q)$为每个博客的相关性.

1.4 相关反馈

相关反馈是TREC在2008年发布的一项新任务,基本的任务是:对于一个给定的查询,对文档集索引中抽取相关文档,得到初始查询结果;然后再给定一些标注过的与查询相关或无关的文档,通过标记文档选择扩展词,对查询进行重构;最后重新查询得到反馈结果. 2008年采用了传统的Rocchio算法,即正负反馈的方法. 2009年相关反馈主要采用了文本分类、语言模型提取扩展词的方法[16],其效果较好. 2010年的相关反馈在2009年方法的基础之上加入了实体扩展、扩展词分类两部分.

1.4.1 结构流程

2010年相关反馈方法的流程如图1所示.

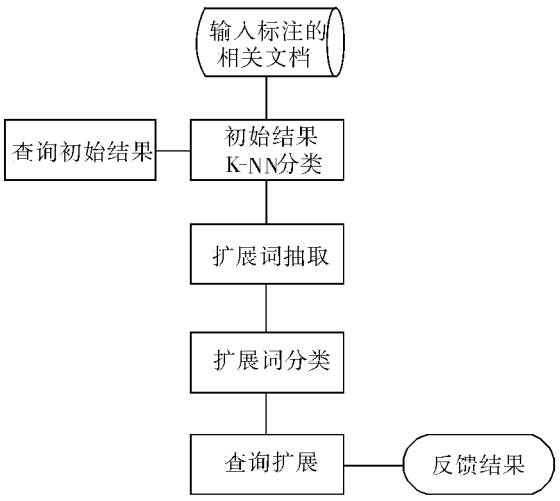


图1 相关反馈的流程

Fig. 1 The flow chart of relevance feedback

1.4.2 扩展词抽取

扩展词主要有2种:通过语言模型计算的权重排序得到的词[17]和通过相似性KL距离计算得到的命名实体. 扩展词的来源是初始查询结果通过标记文本分类得到的相关文档类.

语言模型进行扩展词抽取主要思想是将相关文档类看作一个模型[18],通过估计模型生成词的概率来对词进行排序. 词在相关文档类模型中的概率分布如式(5):

$$\hat{P}(t \mid M_d) = \begin{cases} P_{ml}(t,d)^{(1-\hat{R}(t,d))} \times P_{\text{avg}}(t)^{\hat{R}(t,d)}, f_t(t,d) > 0; \\ \frac{f_{ct}}{c_s}, f_t(t,d) \leqslant 0. \end{cases} \tag{5}$$

式中:$P_{ml}(t,d)$是词 t 在文档 d 中的归一化频率,$P_{\text{avg}}(t)$是词 t 的平均词频,$\hat{R}(t,d)$是一个风险函数,f_{ct}是 t 在文档类中的总词频,c_s 是相关文档集长度.

一些查询往往与特定的领域或主题相关,这些领域内部的人物、机构、地点等通常能有助于区分相关文档和不相关文档[19]. 因此,可以将这些命名实体(包括人名、地名、组织机构)作为扩展查询的一部分. 抽取的主要方法步骤是:1)对相关文档集进行命名实体标注,标注出人、组织和地名3类命名实体;2)基于命名实体的词频对实体进行排序,得到词频较高的前20个命名实体;3)去掉这20个命名实体中的噪声实体,噪声实体是指在相关文档集和不相关文档集中都经常出现的实体;4)计算去噪后每个实体和相关文档的KL距离[20],找到与相关文档距离最近的5个实体加入到扩展词集合中.

1.4.3 扩展词分类

通过语言模型提取出的扩展词,并不是都能改善原始查询的结果;因此采用对扩展词进行分类的方法,选择对原始查询改善效果比较好的扩展词,使得查询能够得到更好的优化. 在扩展词分类实验中,分类器采用LIBSVM,特征选取方面,主要考虑的是扩展词的分布特点、扩展词与查询词之间的共现频度和距离等特征,训练样本来源于2009年TERC相关反馈评测的数据.

根据扩展词对原始查询的不同影响,将扩展词分为好扩展和坏扩展2种,并进行扩展词标注. 好扩展是指当在扩展查询中该扩展词的权重为 w 时,返回的结果比原始查询好,即正反馈;当权重为 $-w$ 时,返回结果比原始查询差,即负反馈. 坏扩展与之相反. 实验中取 $w = 0.01$.

使用LIBSVM[21]进行SVM的训练和预测. 按照前面提到的标注方法,对2009年相关反馈提取的扩

展词进行了标注,为避免正负样本比例不协调的问题而影响分类效果,最后选定 191 个样本作为训练样本,其中 131 个负样本,60 个正样本. 在训练过程中,采取了交叉验证的方法,将数据平均分成 5 组,并保证每一组数据有 12 个正样本,最后达到的平均准确率为69.268 34%.

1.4.4　*查询扩展*

根据给定的原始查询和从相关文档集合中抽取的扩展词进行查询扩展. 扩展过程中查询的格式如下[22]:

combine (query). (title) #weight (1.0 #combine (query) 1.0 #uw (query) 0.2 terms 0.2 #combine (named entity)).

其中:"query"为原始查询,"terms"为语言模型抽取、SVM 分类过的扩展词,"named entity"为通过 KL 距离抽取的命名实体. 原始查询的权重设为 1.0,扩展词权重设为 0.2.

2　信息抽取

一般情况下,被用户认为有用的信息隐藏在大量文字中,或散乱分布在各种各样的网页中. 如何将这些符合特定需求的信息抽取出来,是当前文本搜索领域的热点问题. 著名的文本分析会议(text analysis conference, TAC)就将焦点放在信息的抽取和关联分析上. TAC 是由 IAD(information access division)组织的一个评测,该评测自 2008 年举办以来,已经进行了 3 届,最初是从 TREC 评测的 Question Answering Track 发展起来的[23]. 笔者自 2009 年已经连续 2 年参加了该评测的实体关联和实体填充[24] 2 项任务,并在评测中取得了较为优异的成绩.

2.1　实体关联任务及关键技术

实体关联(entity linking)的任务是根据每一个 query 的标题和支持文档找到 KB 中的惟一节点和它对应,或者返回空(表示该节点不和任何 KB 中的节点对应). 其中:KB(knowledge base)这个数据集中存放所有的 KB 节点;query 是评测开始时官方提供的数据,一个 query 包含 1 个 title(标题)和 1 篇支持文档.

1)系统总体框架. 系统主要包括以下几个模块:实体检索、命名实体识别、相似性判断、自动摘要,如图 2. 基本思想是,首先对每一个实体 query 进行实体检索,得到一批实体候选列表,然后针对每一个候选实体进行排序和相似度的打分,从而得到最终的结果.

2)实体检索. 在评测中,往往面对的是海量文本,如果对于每一个查询都去遍历 KB,那么其响应速度是不能接受的;因此,通常需要对 KB 建立索引,在 TAC 评测中,选用 Indri 作为建立索引的工具.

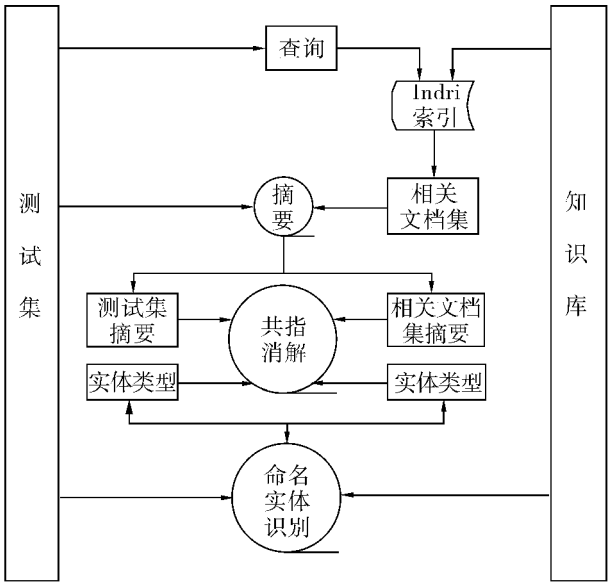


图 2　实体关联的流程

Fig. 2　The flow chart of entity linking

3)命名实体识别. TAC 评测中的 query 都是一个实体,并且该实体可能是以下 3 种类别之一:人名、地名、组织机构名. 首先需要判断该 query 是哪一种类别的实体,从而方便后续的处理,在 TAC 评测中,使用了斯坦福大学提供的命名实体识别开源工具包.

4)判定方法. 在评测中,需要对 1 个 query 和 1 个文档进行相似度的计算,采用了以下 2 种方法:

a)基于 VSM 模型的相似度判断:

$$S_{im}(S_1,S_2) = \frac{\boldsymbol{V}(S_1)\boldsymbol{V}(S_2)}{|\boldsymbol{V}(S_1)||\boldsymbol{V}(S_2)|} = \sum_{\text{common terms:}t_j} w_{1j} \times w_{2j}.$$

b)基于 KL 距离的相似度判断:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i (P(i) - Q(i)) \log \frac{P(i)}{Q(i)}.$$

5)实体关联的改进. 在 2010 年的 TAC 评测中,笔者加入了许多规则,这些规则的引入主要来自于对原始数据的观察,通过加入相关的这些规则,效果有了提高.

2.2　实体填充任务及关键技术

实体填充(slot filling)任务即在测试集中寻找与目标实体(查询)相关的信息,填充目标实体预先规定的一系列属性值. 目标实体分为 2 类:人名和组织机构,人名共有 26 种属性需要填充,组织机构共有 16 种属性需要填充. 属性有 single 和 list 的不同,其中 single 为只能有一个答案的属性,如人的生日;list 为可以有多个答案的属性,如人的子女.

1)系统总体框架. 实体填充系统的总体框架由

4个部分组成:实体检索模块、命名实体识别模块、关系抽取模块、结果决策模块,如图3.实体检索模块通过Indri检索平台,获取和查询实体最相关的前25篇相关文档及其相关度权值.命名实体识别模块使用斯坦福NER工具包识别人名、地名、组织机构名,使用时间规则模板匹配识别时间.关系抽取模块是实体填充系统的核心模块,把实体填充当作一个关系抽取任务,在这一模块中同时采用基于规则模板的方法与基于统计的方法.结果决策模块对关系抽取模块的结果进行优选得出最终结果.

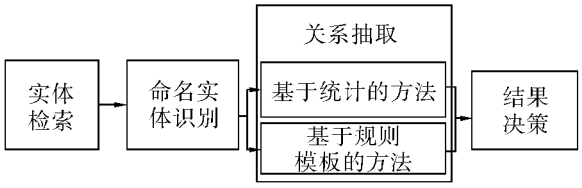


图3 实体填充的流程

Fig.3 The flow chart of slot filling

2010年实体填充的整体实现框架与2009年大体相同,但细节上有所改进,例如增加了URL的识别.采用基于规则方法识别为主、基于统计CRF识别方法做补充的实现方案.即当2种方法同时出现"single"的值,优选选择规则类方法;对于非"single"的值,综合考虑文档对于query的相关性值S_{EL}和填充结果的可信度值S_{SF},选择最优的若干个结果进行优选得出最终结果.

2)基于规则模板的方法.a)识别URL(网址)和LIST(title职称、charge罪名、cause of death死因、religion宗教等).其中URL识别采用正则表达式方法,LIST主要从训练语料中统计而来.b)根据规则模板输出实体填充结果.

3)基于统计的方法.基于统计的方法是一种半监督的机器学习方法,它将实体关系抽取看作一种多分类问题,从文本中抽取训练所需要的特征,然后利用条件随机场形成分类器.

利用9种特征来训练CRFs:词对、词特征、词性特征、顺序特征、动词位置特征、实体位置特征、二值特征、动词特征和类型特征.由于实体关系识别是一种多分类问题,而类别数越多,模型的准确率也会下降.为了尽可能降低类别数,根据目标实体的类型(人名或组织名)将初始的训练语料初步分为2份,然后再根据词对中的第2个词是否为命名实体,进一步将训练语料二次划分,最后用CRFs形成了4种分类器,这样做也提高了系统的整体效率.

4)结果合并.综合考虑文档对于query的相关性值S_{EL}和填充结果的可信度值S_{SF},选择最优的1个或若干个.选择策略如式(6)所示.

$$V_{\mathrm{alue}}(Q,s_{\mathrm{lot}},d_{\mathrm{oc}}) = \mu \times S_{\mathrm{EL}}(Q,d_{\mathrm{oc}}) + (1-\mu) \times S_{\mathrm{SF}}(Q,S_{\mathrm{lot}}). \quad (6)$$

式中:$V_{\mathrm{alue}}(Q,s_{\mathrm{lot}},d_{\mathrm{oc}})$即为综合考虑文档对于query的相关性值和填充结果的可信度值的权值.对于基于机器学习的方法,CRF++工具包[25]可以为识别结果提供可信度值,记为crfvalue,即该判别结果正确的概率,S_{SF}=crfvalue;对于基于规则的方法,优先选取基于规则方法的结果,设置填充结果可信度值为1,$S_{\mathrm{SF}}=1$.实体关联提供相关文档的同时提供该文档的相关度值,记为S_{EL}.其中参数μ设置为0.5.

3 信息过滤

近年来,随着互联网技术的迅速发展,垃圾信息的数量在网络上呈现上升趋势,信息过滤成为一个业内的难题和挑战.以垃圾邮件为例,TREC从2005—2007年组织了垃圾邮件过滤评测(spam track)[26-27],目的是尽可能找到一种好的垃圾邮件过滤模型,保证过滤的有效性和可重复性满足需求.主要任务包括即时反馈、延时反馈、主动学习和部分反馈等[28].笔者参加了其中的3届评测,2005年在参赛的国内队伍中成绩是最好的.

当前的垃圾邮件过滤技术可以大致划分为黑名单技术、人力驱动的启发式过滤以及基于机器学习的过滤[29].这些技术中,朴素贝叶斯方法受到广泛关注.

3.1 朴素贝叶斯分类器

朴素贝叶斯分类器简单有效,经常用于文本分类的应用和实验中.垃圾邮件过滤属于文本分类问题,因此该分类器被广泛使用于垃圾邮件过滤.朴素贝叶斯分类器是一种基于概率的方法,基本思想是通过观察一些词是否在邮件中出现来判断是垃圾还是非垃圾,如式(7):

$$C_{\mathrm{NB}} = \arg\max_{i\in L} P(C_i) \prod_k P(w_k \mid C_i). \quad (7)$$

式中:w_k是组成邮件的词,L是类别的集合.常用的朴素贝叶斯模型有multi-variate Bernoulli模型、Poisson Naïve Bayes模型以及multinomial模型.它们的不同之处主要在于如何计算$P(w_k|C_i)$.对于垃圾邮件过滤问题,只有2个类别:垃圾邮件C_+和非垃圾邮件C_-,那么一封邮件M的对数得分可写为

$$S_{\mathrm{core}}(M) = \log P(C_+) + \sum_k \log P(w_k \mid C_+) - \left(\log P(C_-) + \sum_k \log P(w_k \mid C_-)\right).$$

如果$S_{\mathrm{core}}(M)>0$,待分类邮件被标注为C_+类(垃圾邮件),反之被标注为C_-类(非垃圾邮件).过滤模型如图4所示.在有监督情况下,用户判断垃圾邮件过滤器的结果并反馈给过滤器,而过滤器依据反馈进行自动学习.系统开始运行时并不预设标准,即是一个无初始记忆的分类器,而后不断更新达到

最佳效果. 系统关于垃圾邮件的知识均是从理想用户的反馈中得到的.

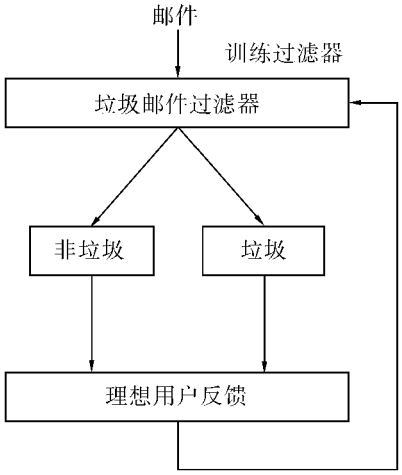


图 4　垃圾邮件过滤的流程

Fig. 4　The flow chart of spam filtering

3.2　加权朴素贝叶斯分类器

假设邮件的不同部分对过滤的贡献是不同的,某些部分对过滤的帮助更大. 若邮件分为 S 个部分,每个部分由 N_d 个词组成,$d=1,2,\cdots,S$. 那么朴素贝叶斯分类器的一个简单推广就是为邮件的不同部分赋予不同的权值 α. 式(7)可以更新成为

$$H_{\mathrm{NB}} = \arg\max_{i\in L}\{\log P(C_i) + \sum_{d=1}^{S}\left(\log \alpha_d^i + \sum_{k=1}^{N_d}\log P(w_k^d \mid C_i)\right)\}. \quad (8)$$

式中:α_d 为权值,$d=1,2,\cdots,S$. 式(8)用 N_d 和邮件长度正规化后可以写成

$$H_{\mathrm{NB}}^{\mathrm{normal}} = \arg\max_{i\in L}\{\log P(C_i) + \sum_{d=1}^{S}(N_d\log \alpha_d^i + \sum_{k=1}^{N_d}\log P(w_k^d \mid C_i))\}.$$

那么给定训练集后,参数集 α 就可以用最大似然准则求解了. 在实际中,划分的方法有很多. 可以按结构划分各部分,如标题、邮件头、正文、附件等,也可以按词的不同概率将邮件划分成不同的部分.

3.3　分类器集成

Bagging 是一种将一些弱分类器集成的技术. 弱分类器指的是准确率比 50% 高一点的分类器. 在分类过滤任务中,将弱分类器集成在一起,经过演进和变换达到最佳效果. 基于 Bagging 技术的朴素贝叶斯垃圾邮件过滤器,通过选择好的集成方法有助于提升过滤系统的性能. 常用的方法主要有嵌入决策树和分类错误加权等.

4　实际系统

4.1　互联网舆情系统

北京邮电大学模式识别与智能系统实验室的互联网舆情监控分析系统依托自主研发的文本搜索和文本挖掘技术,通过新闻、论坛、博客、微博、视频网站等内容源的自动采集与跟踪,进行敏感话题过滤分析、智能话题聚类分类、主题监测、专题聚焦和各类数据的统计分析,实现应用单位对相关网络舆情监督管理的需要,为决策层全面掌握舆情动态,做出正确舆论引导提供分析依据.

4.2　短信舆情系统

短信是人们日常生活中进行通信的重要手段,通过对短信文本的分析,可以掌握大众平时的舆论导向,并且可以帮助政府职能部门尽早地发现一些不良的、危及安全的不法短信. 但是短信有其自身的特点:短小、口语化等,这也给分析带来了很大的难度. 因此,基于短信进行舆情分析既有一定的学术价值,也有一定的现实意义.

短信舆情系统主要有以下一些模块:短信分类模块根据短信的内容将短信分到不同的类别,并且可以通过训练自动调整各类别下关键词的权重;敏感过滤模块可以过滤出涉及国家和人民生命财产安全的非法短信;发送方式分析模块可以判断出一条短信的发送方式,例如群发、转发、直发等,从而可以获知什么样的短信被大规模群发,并进行有针对性的跟踪;短信溯源和用户交际圈模块可以根据某一用户或某一短信进行全方位地分析,从而掌握某用户的动态.

通过短信舆情系统,可以更好地加强对短信数据的监控,掌握普通用户的舆情情况,为政府职能部门制定相关决策,追踪某些特殊的现象提供手段.

4.3　校园对象搜索引擎系统

校园对象搜索引擎(campus object search engine, COSE),是一款在校园网内工作,致力于帮助用户寻找人物、组织机构以及课程信息的垂直搜索引擎. 从 COSE 的名字就可以看出该系统所针对的服务对象是校园中的学生群体. COSE 的主要特点在于它融入了信息抽取中的命名实体识别和实体关系抽取这 2 项技术,可以自动识别网页中的人名、课程名以及机构组织名,建立实体(也称对象)数据库,并且根据对象名在网页中抽取其关系(也称相关属性),建立相关属性数据库,供用户查询检索时使用.

COSE 系统包含的模块有:网络爬虫与索引、中文分词、命名实体识别、实体关系抽取和查询重构. COSE 采用广度优先搜索策略,只抓取各个大学网站域名下的网页信息,建立网页文档库及索引. 这可以在很大一定程度上屏蔽掉大量无用的广告网页和新闻网页. 对网页文档建索引能加快查找和排序的速度,COSE 系统综合使用全文索引技术和动态文

档索引技术. 中文分词是命名实体识别和实体关系抽取的前提和基础,COSE 中的中文分词技术综合应用基于字符串匹配和基于统计的中文分词技术. 命名实体识别是 COSE 系统的关键技术之一,采用基于统计与基于规则相结合的识别方法. 实体关系抽取是 COSE 系统中的另一项关键技术,鉴于正则表达式的灵活性和强大的字符串匹配能力,COSE 系统借助成熟的 Python 字符处理规则,提出一种正则表达式方案抽取对象属性信息. COSE 中查询重构模块旨在解决以下 2 种形式的查询:1)复杂查询:查询的不是单纯实体;2)问题式查询:比如某某老师属于哪个学院. 在用户使用 COSE 进行检索时,系统会返回 2 类信息:一类是与通用搜索引擎相似的和查询相关的网页信息,另一类则是相关网页中包含的命名实体及其相关属性.

5 总结与展望

传统的文本搜索技术已经难以满足用户的需求,融合了信息检索、信息抽取和信息过滤等技术的智能文本搜索新技术是当前的研究热点.

信息检索技术不再是单纯的按相关度呈现各个网页,更多的是对网页内容的深度挖掘、组织并反馈,提高检索的准确性、完备性、个性化程度. 企业检索主要研究在企业内部数据中的用户检索行为,主要包含邮件检索、文档检索和专家检索任务,使用了二阶排序模型和专家经验模型. 实体检索主要关注查找相关实体,除了使用文档中心模型和实体中心模型外,还加入了实体抽取的关键技术和用来惟一标识实体的主页. 博客检索对博客中出现的观点及其与查询的相似性进行研究,在此基础上对倾向性作分析,主要分为 3 类:个人与官方、表达观点与描述事实、深入分析与浅层描述. 相关反馈利用给定的与查询相关或无关的标注文档,选择扩展词,对查询进行重构,通过重排序改善原有检索系统的性能.

信息抽取技术在文本分析会议评测中得到很好的体现. 该评测分为实体关联和实体填充 2 个任务,深度剖析文本信息,致力于识别、分析、整合文本中出现的实体. 信息抽取技术非常重要,为其他工作的顺利进行起到了基础性作用.

信息过滤的关键技术被应用在垃圾邮件过滤评测中. 该评测的目的是尽可能找到一种好的垃圾邮件过滤模型,保证过滤的有效性和可重复性,主要任务包括即时反馈、延时反馈、主动学习和部分反馈等. 其中加权朴素贝叶斯和分类器集成的方法表现出了良好的效果.

信息检索、抽取和过滤三大技术是相互联系的,经常融合在一起,发挥最大的作用. 例如:在检索之前要抽取有价值的信息,过滤掉垃圾信息;抽取和过滤中也可以使用检索的方法进行初步处理;抽取和过滤都有基于规则和基于统计的方法等. 这些都很好地在互联网舆情、短信舆情和校园对象搜索引擎等系统中得到了体现. 新的智能文本搜索技术将是未来热门的研究方向,并且具有巨大的发展前景.

参考文献:

[1]郭军. Web 搜索[M]. 北京:高等教育出版社, 2009: 1-3.

[2]方慧. TREC 发展历程及现状分析[J]. 新世纪图书馆, 2010(1): 57.
FANG Hui. On developing course and status analysis of TREC[J]. New Century Library, 2010(1): 57.

[3]BALOG K, SOBOROFF I, THOMAS P, et al. Overview of the TREC 2008 enterprise track[EB/OL]. [2010-12-15]. http://trec. nist. gov/pubs/trec17/papers/ENTERPRISE. OVERVIEW. pdf.

[4]RU Zhao, CHEN Yuehua, XU Weiran, et al. TREC2005 enterprise track experiments at BUPT[EB/OL]. [2010-12-15]. http://trec. nist. gov/pubs/trec14/papers/ beijingu-of-pt. ent. pdf.

[5]RU Zhao, LI Qian, XU Weiran, et al. BUPT at TREC 2006: enterprise track[EB/OL]. [2010-12-15]. http://trec. nist. gov/pubs/trec15/papers/beijing-upt. ent. final. pdf.

[6]BAILEY P, CRASWELL N. Overview of the TREC 2007 enterprise track[EB/OL]. [2010-12-15]. http://trec. nist. gov/pubs/trec16/papers/ENT. OVERVIEW16. pdf.

[7]WANG Zhanyi, LIU Dongxin, XU Weiran, et al. BUPT at TREC 2009: entity track[EB/OL]. [2010-12-15]. http://trec. nist. gov/pubs/trec18/papers/bupt. ENT. pdf.

[8]ZHANG Suxiang, WEN Juan, WANG Xiaojie, et al. Automatic entity relation extraction based on maximum entropy [C]//Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications. Ji'nan, China, 2006: 540-544.

[9]LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[10]MACDONALD C, OUNIS I. Voting for candidates: adapting data fusion techniques for an expert search task[C]// Proceedings of the 15th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2006: 387-396.

[11]MANNING C D, RAGHAVAN P, SCHUTZE H, An introduction to information retrieval[M]. Cambridge, UK: Cambridge University Press, 2008: 120-126.

[12]WILSON T, WIEBE J, HOFFMANN P, Recognizing contextual polarity in phrase-level sentiment analysis[C]//

Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2005: 347-354.

[13] MANNING C D, SCHTZE H. Foundations of statistical natural language processing[M]. Cambridge, USA: The MIT Press, 1999.

[14] AMATI G. Probabilistic models for information retrieval based on divergence from randomness[D]. Glasgow, UK: University of Glasgow, 2003.

[15] SINGHAL A, BUCKLEY C, MITRA M. Pivoted document length normalization[C]//Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1996: 21-29.

[16] LI Si, LI Xinsheng. PRIS at 2009 relevance feedback track: experiments in language model for relevance feedback[EB/OL]. [2010-12-15]. http://trec.nist.gov/pubs/ trec18/papers/pris.RF.pdf.

[17] LALMAS M, MACFARLANE A, RUGER S. Advances in information retrieval[M]. New York, USA: Springer-Verlag, 2002: 74-172.

[18] PONTE J M, CROFT W B. A language modeling approach to information retrieval[C]//Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1998: 275-281.

[19] WANG Bingqing, HUANG Xuanjing. Relevance feedback based on constrained clustering: FDU at TREC'09[EB/OL]. [2010-12-15]. http://trec.nist.gov/pubs/ trec18/papers/fudanu.RF.pdf.

[20] LAVRENKO V, CROFT W B. Relevance-based language models[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2001: 120-127.

[21] CHANG Chihchung, LIN Chihjen. LIBSVM: a library for support vector machines[EB/OL]. [2011-04-09]. http://www.csie.ntu.edu.tw/ ~cjlin/libsvm/index.html.

[22] The Lemur Project. INDRI: language modeling meets inference networks[EB/OL]. [2011-03-23]. http://www.lemurproject.org/indri/.

[23] TAC 2009. Knowledge base population track[EB/OL]. (2009-09-29)[2010-12-16]. http://apl.jhu.edu/ ~paulmac/kbp.html.

[24] TAC 2010. Knowledge base population (KBP2010) track[EB/OL]. (2010-09-12)[2010-12-16]. http://nlp.cs.qc.cuny.edu/kbp/2010/.

[25] CRF + +: yet another CRF toolkit[EB/OL]. [2010-12-16]. http://crfpp.sourceforge.net/.

[26] YANG Zhen, XU Weiran, CHEN Bo, et al. PRIS Kidult anti-SPAM solution at the TREC 2005 spam track: improving the performance of naive Bayes for spam detection[EB/OL]. [2010-12-15]. http://trec.nist.gov/pubs/trec14/papers/beijingu-of-pt.spam.pdf.

[27] YANG Zhen, XU Wei, CHEN Bo, et al. BUPT at TREC 2006: spam track[EB/OL]. [2010-12-15]. http://trec.nist.gov/pubs/trec15/papers/beijing-upt.spam.final.pdf.

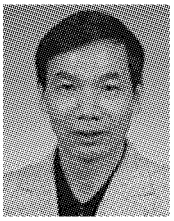
[28] CORMACK G V. TREC 2007 spam track overview[EB/OL]. [2010-12-15]. http://trec.nist.gov/pubs/trec16/papers/SPAM.OVERVIEW16.pdf.

[29] 杨震. 文本分类和聚类中若干问题的研究[D]. 北京:北京邮电大学, 2007: 10-86.
YANG Zhen. Research on key problems in text classification and clustering[D]. Beijing: Beijing University of Posts and Telecommunications, 2007: 10-86.

作者简介:

王占一,男,1984 年生,博士研究生,主要研究方向为信息过滤和信息检索等. 在国内外重要期刊和会议上发表学术论文 10 篇,获发明专利 2 项.

徐蔚然,男,1975 年生,副教授,主要研究方向为信息检索、模式识别和机器学习. 主持参加了 TREC、TAC、ACE 等国际著名检索评测,并且获得优异成绩,参与多项国家级科研项目,发表学术论文 20 余篇.



郭军,男,1959 年生,教授,博士生导师,主要研究方向为模式识别、网络管理、信息检索、基于内容的信息安全等. 主持多项"863"计划项目和国家自然科学基金项目,获省部级奖励多项,发表学术论文上百篇,获授权专利 5 项.