北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.047

# 利用 URL-Key 进行查询分类

李雪伟[1] 吕学强[1,†] 董志安[1] 刘克会[2,3]

1. 北京信息科技大学网络文化与数字传播北京市重点实验室, 北京 100101; 2. 北京理工大学管理与经济学院, 北京 100081; 3. 北京城市系统工程研究中心, 北京 100035; † 通信作者, E-mail: lxq@bistu.edu.cn

**摘要** 针对查询分类问题, 借助互联网中人工组织的分类网站领域 URL, 利用 URL-key 在各个类别中使用的频度, 提出基于方差的领域 URL-key 识别方法, 利用机器翻译、拼音翻译和搜索结果反馈等技术对 URL-key 进行过滤, 构建领域 URL-key。然后结合伪相关反馈技术, 选取 URL-key 为特征, 构建 URL-key 向量, 利用 SVM 对查询串进行分类。实验结果表明, 该方法不仅 *F* 值比对比方法提高 7%, 而且资源的使用也远远小于对比方法, 提高了系统的时效性。

**关键词** 查询分类; URL; URL-key; 伪相关反馈

**中图分类号** TP391

# Query Classification by Using URL-Key

LI Xuewei[1], LÜ Xueqiang[1,†], DONG Zhian[1], LIU Kehui[2,3]

1. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101; 2. School of Management and Economics, Beijing Institute of Technology, Beijing 100081; 3. Beijing Research Center of Urban Systems Engineering, Beijing 100035; † Corresponding author, E-mail: lxq@bistu.edu.cn

**Abstract** For the problem of query classification, a variance based method is proposed to identify domain URL-key by using domain URL organized manually from aggregator sites and the use frequency of URL-key in each category. Then, the URL-key is filtered by using machine translation, pinyin and search results feedback technology. Finally, coupled with relevance feedback, we classify the query by selecting the URL-key as feature and establishing the URL-key vector with a SVM multi - class classifier. Experimental results show that our method not only uses less resources, but also the *F*-value is 7% higher than contrast method.

**Key words** query classification; URL; URL-key; feedback

随着社会的发展, 搜索引擎已经成为人们生活的工具。然而, 搜索引擎反馈结果的准确性还不能满足人们的需求, 因此需要对查询串分类进行研究。

查询分类是将用户提交的查询按照查询的意图分配到已定义好的类别中, 对于用户输入的查询串, 根据历史用户查询日志分类来推断当前查询的类别。查询分类被广泛应用于搜索广告投放[1]、搜索结果改进[2]、用户行为分析[3]以及知识挖掘[4]等领域。

查询分类与文本分类的不同之处在于处理的对象不同, 它处理的是用户输入的查询串, 其主要特点有以下两方面: 1) 查询串比较短, 通常 4 个字左右, 缺少上下文语境; 2) 查询串容易出现歧义现象[3]。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[5]。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。此类方法通常要对在线搜索结果进行复杂处理, 由

国家自然科学基金(61271304)、北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目(KZ201311232037)和北京市科学技术研究院创新工程(PXM2013-178215-000004)资助

收稿日期: 2014–07–27; 修回日期: 2014–10–13; 网络出版时间: 2014-11-28 15:20

此导致的延时将影响实际应用。张宇等[6]提出基于URL匹配的查询分类方法，在一定程度上解决了查询分类的问题，但同时也存在不足，如时效性差、URL的收集困难等。由于领域URL收集的数量有限，而互联网中网站的增长速度非常快，有限的领域URL无法满足爆炸式增长的网站，所以应该找到一个更好的方法来取代URL。

URL作为Internet上用来描述信息资源的字符串，主要由三部分组成：协议、域名和路径。域名通常由两部分构成：域名主体和域名后缀。例如，对于“match.sports.sina.com.cn”这个域名来说，“match.sports.sina”为该域名的主体词，“com.cn”为该域名的后缀。域名的主体词通常与信息资源的类别有关。路径指信息资源具体存放的位置。人们在构建路径时，为了方便信息资源归类，通常会考虑将不同的资源放在不同的路径下，且为路径起的名称与资源的主题相关，如football的路径下应该放与足球相关的信息资源。

通过上述对URL的分析发现，人们在申请域名或者创建路径时，通常会思考与信息资源的相关性。虽然各网站的域名不同，但是相同主题下的域名主体词及路径引导词可能会相同。通常，会集中使用一些领域性强的词语，如体育类中的“sport”，经济类中的“finance”，汽车类中的“auto”，等等。也就说，互联网中的域名主体词及路径引导词被赋予人们智慧的结晶。本文将URL中含有类别信息的域名词主体词及路径引导词统称为URL-key。

本文放弃传统的搜索引擎结果文本扩展的方法，利用反馈的URL作为桥梁，利用URL-key类别来判断URL的类别，从而判断查询串的类别。例如：用户输入的查询串为“凯美瑞”，搜索引擎返回的URL={“http://www.52car.net/”，“http://www.yicars.com/”，…}，由于car在汽车类中容易出现，在其他类别中很难出现，所以很容易得出“凯美瑞”为汽车类的查询串。本文根据URL-key在不同领域中使用的差异性，对文献[6]中收集的领域URL进行领域URL-key提取，构建领域URL-key词表。然后，根据待分类的查询串反馈的URL结果集，构建领域URL-key向量，最终利用SVM对查询串进行分类。

## 1 领域URL-key构建

为了得到更多的领域URL-key，需要找到更多的领域URL。在互联网中很多资源都凝聚了人类的群体智慧，如分类网站，通常是经过人工整理分类后呈现给用户的，具有很高的可信度。张宇等[6]采集了Yaboo，Google和Baidu的中文分类网页目录，以Yahoo网页目录的前两层作为分类的标准，并将Google和Baidu网页目录全部映射到Yahoo的类别体系中，最终得到10类领域，40895个URL(表1)。

文献[6]采用URL匹配的方法进行查询分类，而网络中存在着海量的URL，且每天都在不断的增长，因此使用URL匹配的方法已经不能满足当前的需要。本文通过对文献[6]收集的URL进行分析发现，在构建URL时词语的使用存在差异性，分布不均衡，特定词语使用的频次较高。例如：“sport”在体育类URL中大量出现，“car”在汽车类URL中大量出现，“finance”在金融类URL中大量出现。因此，本文尝试用领域URL-key代替领域URL进行查询分类，这样不仅可以解决URL增长带来的问题，同时也可使用较少的资源，提升计算速度。

本文利用分割符$T$={:, /, ., -}对文献[6]收集的URL进行切分，构成一系列的字符串，删除长度为1的字串、纯数字串和一些无意义的特殊字符序列(如：http, www, com等)，最终构建候选字符串集合为CandidatekeySet。如http://www.tennis.com/player/539/na-li/_经过切分后得到Candidatekey为tennis, player, na, li。本文将从CandidatekeySet中提取领域URL-key。

根据Candidatekey在每个类别中分布的差异

**表1 URL分布**

Table 1 Distribution of URL

| 领域 | URL |
|---|---|
| Economy | 15205 |
| IT | 5312 |
| Health | 2353 |
| Sports | 1502 |
| Travel | 1856 |
| Education | 7560 |
| Job | 276 |
| Art | 6162 |
| Military | 255 |
| Auto | 414 |

性, 本文提出基于方差的领域 URL-key 识别方法, 并利用机器翻译、拼音翻译和搜索结果 3 种方法对领域 URL-key 进行过滤。

## 1.1 基于方差的领域 URL-key 提取

李素建等[7]利用方差的思想实现领域术语的提取。本文借鉴此思想解决领域 URL-key 的领域性问题。本文利用 Candidatekey 在各个领域中使用差异性, 衡量 Candidatekey 的领域性。使用差异性越小则领域性低, 反之则领域性越强。

首先, 计算词语在不同领域中使用概率的总方差。设共有 $N$ 个词语, 每个词语为 $W_i$ (0<$i$<$N$), 有 $M$ 个不同的类别, 词语 $W_i$ 在领域 $j$ 中使用的概率为 $P_j(W_i)(1 \leqslant j \leqslant M)$, 词语 $W_i$ 在 $M$ 个领域中使用概率的平均值用 $R(W_i)$表示, 由此可以得到 $W_i$ 的使用性差异 Var($W_i$), 如式(1)所示。

$$\mathrm{Var}(W_i)=\frac{\sum_j (P_j(W_i)-R(W_i))^2}{M}, \tag{1}$$

$$P_j(W_i)=\frac{\sum_j W_i}{\sum_j W}, \tag{2}$$

$$R(W_i)=\frac{\sum_j P_j(W_i)}{M}。 \tag{3}$$

在 CandidatekeySet 集合中, 若某个 Candidatekey 使用概率差异性较大, 则说明该 Candidatekey 可能为领域 URL-key。但是, 该 URL-key 所属领域还需要进一步计算。本文认为一个 URL-key 在其他类别中使用概率值小, 而在特定类别中使用概率大, 则该 URL-key 为该领域的领域 URL-key。为了得到领域 URL-key 的所属领域, 本文主要从两方面进行考虑: 1) URL-key 的方差贡献度; 2) URL-key 在各领域中的使用情况, 即 URL-key 的 IDF($W$)。

1) 方差贡献度: 由于 Var($W_i$)的值是根据 $W_i$ 在 $j$ 中的使用概率来确定的, 也就是说 $W_i$ 在 $j$ 领域中的 $P_j(W_i)$越大, 则对 Var($W_i$)贡献越大, 因此词语 $W_i$ 的方差贡献度为

$$\mathrm{Con}_j(W_i)=\frac{P_j(W_i)}{\sum_j P_j(W_i)}。 \tag{4}$$

2) URL-key 使用度: 通常一个 URL-key 在其他领域中没有出现, 只在某个特定领域中出现, 则该 URL-key 的领域性强。词语 $W_i$ 的使用度为

$$\mathrm{Use}_j(W_i)=\log 10\left(\frac{L}{M}+\lambda\right), \tag{5}$$

其中 $L$ 为类别总个数, $L$=10; $\lambda$ 为平滑因子, $\lambda$=0.001。

由上可知, 某领域 URL-key 的方差值不仅受其方差贡献度影响, 还受 URL-key 使用度影响, 因此 URL-key 的领域度公式如下:

$$\mathrm{Domain}_j(W_i)=\mathrm{Var}(W_i)\ \mathrm{Con}_j(W_i)\ \mathrm{Use}_j(W_i)。 \tag{6}$$

## 1.2 领域 URL-key 过滤

本文通过对领域 URL-key 的观察发现, URL-key 多为一些领域常用的词汇, 可明显地与其他类别词进行区分。领域 URL-key 可以分为 3 类, 如表 2 所示。

**表 2 URL-key 类型**

Table 2 URL-key type

| 类别 | 实例 |
|---|---|
| 英文单词或缩写 | sport, hospital, finance, nba, oly, … |
| 汉语拼音或缩写 | tiyu, rencai, junshi, caijing, ty, rc, … |
| 特殊网站域名 | zhibo, teixue, atohome, toyota |

由表 2 可知, 人们一般会使用简单、方便、领域性强的词语来构建网站 URL。本文根据领域 URL-key 的类型特点提出以下 3 种过滤方式。

1) 基于机器翻译的过滤。从表 2 可知, 领域 URL-key 中词语多为英文单词或缩写, 将领域词串 $W$ 放入 Google 翻译系统中, 得到$W \xrightarrow{\text{机器翻译}} \mathrm{G}$。如果 $W$ 能被翻译, 且 $G(W)$有意义, 则 $G$ 为 1, 否则 $G$ 为 0。

$$G=\begin{cases}1, & \text{if 机器翻译结果中}\exists \mathrm{G}(W),\\ 0, & \text{otherwise}。\end{cases} \tag{7}$$

2) 基于汉语拼音翻译的过滤。从表 2 可知, 领域 URL-key 中存在汉语拼音。由于搜狗输入法占据市场的很大份额, 同时也积累了很多汉语拼音书写的习惯, 因此本文选择 sogo 翻译进行拼音翻译。将 $G$ 过滤后的无法翻译的 $W$ 放入 sogo 翻译系统中, 选择自动检测语言, 得到$W \xrightarrow{\text{拼音翻译}} P$。如果 $W$ 能被翻译, 且 $P(W)$有意义, 则 $P$ 为 1, 否则 $P$ 为 0。

$$P=\begin{cases}1, & \text{if 拼音翻译结果中}\exists P(W),\\ 0, & \text{otherwise}。\end{cases} \tag{9}$$

3) 基于搜索引擎的过滤。将 $G$ 和 $P$ 过滤后无法翻译的词语放入 sogo 搜索引擎中搜索,得到

$W \xrightarrow{\text{搜索引擎}} S$。如果 $W$ 被检索，切 top10 中 $S(W)$ 有意义，则 $S$ 为 1，否则 $S$ 为 0。

$$S=\begin{cases}1, & \text{if 搜索结果top10中}\exists S(W),\\ 0, & \text{otherwise}。\end{cases} \tag{9}$$

根据 Candidatekey$_j$ 的 $G$, $P$, $S$ 值构建一个 $j$ 领域的 $\boldsymbol{R}_j$ 矩阵，如式(10)所示，由 $\boldsymbol{R}_j$ 矩阵的值来判别 Candidatekey$_j$ 是否为领域 URL-key，如果 Candidatekey$_j$ 的 $G \cup P \cup S$=1，则 Candidatekey$_j$ 为领域 URL-key，最终构建 URL-keySet$_j$ 集合。

$$\boldsymbol{R}_{\mathrm{j}}=\begin{bmatrix} G_i & P_i & S_i \\ G_{i+1} & P_{i+1} & S_{i+1} \\ & \cdots & \\ G_{i+n} & P_{i+n} & P_{i+n} \end{bmatrix}。 \tag{10}$$

## 2 基于 URL-key 的查询分类

本文在已构建的领域 URL-key 语料的基础上，利用搜索引擎伪反馈技术，得到相关的 URL 反馈信息。然后，利用 URL 中各类别 URL-key 的分数来构建特征向量，最后利用 SVM 分类器判断查询串的类别。

伪相关反馈是假设搜索引擎系统查询反馈的搜索结果排名越靠前与查询越相关[8]。本文基于以上假设，构建待分类查询串 query 的 URL-key 特征向量。首先利用 $T$ 集合对伪反馈 URL={URL$_1$, URL$_2$, …, URL$_l$} 进行切分，构建一个 bag$_i$={Candidatekey$_{i1}$, Candidatekey$_{i2}$, …, Candidatekey$_{ij}$}，目标类别为 $C$={$c_1$, $c_2$, …, $c_n$}，其中 $l$ 为搜索引擎返回的前 $l$ 条结果，$j$ 为第 $i$ 个 URL 切分后获取 Candidatekey 的个数，$n$ 为类别个数。为了构建 query 的领域 URL-key 向量，需要计算 query 伪反馈 URL 结果中每一类的 URL-key 分数 Score($c_n$|query)。

本文利用式(11)计算 query 在各个类别中含有的个数。由于反馈结果中排名越靠前，与查询串越相关，则反馈的 URL 与 query 的相关性与位置有一定的关系。也就是说，URL 中含有 URL-key 且其位置越靠前，Score($c_n$|query)越高。本文将反馈结果的前 10 个查询结果看成权重相同的，后面结果随着排名的增加权重逐渐降低。具体分布如式(12)所示。

$$\mathrm{Count}(\mathrm{Url-key}_{\mathrm{n}})=\begin{cases}1, & \text{if } bag\text{中含有url-key}\\ 0, & \text{if } bag\text{中不含有url-key}。\end{cases} \tag{11}$$

$$\mathrm{Pos}(i)=\begin{cases}1 & (1\leqslant i\leqslant 10),\\ \dfrac{1}{\log(i+1)} & (i>10)。\end{cases} \tag{12}$$

综上所述，查询串的 Score($c_n$|query)与反馈结果中含有的领域 URL-key 有关，同时也与位置信息有关。计算公式如下:

$$\mathrm{Score}(c_n|\mathrm{query})=\sum_{i=1}^{l}\mathrm{count}(\mathrm{Url-key})\mathrm{pos}(i)。 \tag{13}$$

根据每一类含有 URL-key 的分数值，最终构建出 query 的特征向量{Score ($c_1$|query), Score ($c_2$|query), …, Score($c_n$|query)}。最后，利用 SVM 分类器对查询串进行分类。

## 3 实验结果与分析

### 3.1 实验数据

本文使用的领域 URL 数据是文献[6]所收集整理的语料，分为 10 个类，共 40895 个 URL。为了与文献[6]进行对比，本实验中用到的实验语料与其相同，共 2073 个查询，分布如表 3 所示。在每一个类查询串中选取 20%作为测试语料，共计 411 条数据，其余 80%作为训练语料，共计 1662 条数据。测试查询串的分布如表 4 所示。

### 3.2 结果与分析

#### 3.2.1 评价指标

借用文本分类评价中采用的准确率($P$)、召回率($R$)、$F$ 值和精确率($A$)进行评价。

$$P=\frac{a}{a+b}\times 100\%, \tag{4}$$

**表 3 查询串分布**

Table 3 Distribution of query

| 领域 | 查询串数量 |
|---|---|
| Economy | 217 |
| IT | 152 |
| Health | 248 |
| Sports | 187 |
| Travel | 240 |
| Education | 230 |
| Job | 101 |
| Art | 249 |
| Military | 200 |
| Auto | 249 |

**表 4　测试查询串分布**
Table 4　Distribution of test query

| 领域 | 查询串数量 |
|---|---|
| Economy | 43 |
| IT | 30 |
| Health | 49 |
| Sports | 37 |
| Travel | 48 |
| Education | 46 |
| Job | 20 |
| Art | 49 |
| Military | 40 |
| Auto | 49 |

$$R=\frac{a}{a+c}\times 100\%, \tag{5}$$

$$F=\frac{2\times P\times R}{P+R}\times 100\%, \tag{6}$$

$$A=\frac{\text{正确分类的查询数}}{\text{总查询数}}。 \tag{7}$$

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 $c$。

### 3.2.2　结果分析

通过对领域 URL 的切分, 利用方差计算公式计算每一个候选领域 URL-key 的领域度, 再根据领域度的大小进行排序, 取 top100 作为候选领域 URL-key, 结果如表 5 所示。

通过对领域 URL-key 的观察, 发现领域 URL-key 主要有以下特点: 1) 由英文单词或缩写组成, 大约占总 URL-key 的 35%, 如 sport, auto, car, bond, nba, vw 等; 2) 为汉语拼音或缩写, 大约占总 URL-key 的 26%, 如 tiyu, che, ccb, icbc 等; 3) 为特殊网站域名, 大约占总 URL-key 的 14%, 如 autohome, zhibo 等。

本文根据领域 URL-key 的上述特点, 利用机器翻译技术、拼音翻译技术和搜索技术构建 $\boldsymbol{R}$ 矩阵。遍历 $\boldsymbol{R}$ 矩阵, 若 $\boldsymbol{R}$ 矩阵中 $\boldsymbol{R}_j$ 的 $G$, $P$, $S$ 的值都为零, 则过滤此候选 URL-key。最终得到 717 个领域 URL-key, 分布如表 6 所示。

本文利用 Query 伪反馈的领域 URL-key 分数构建特征向量, 使用 SVM 对查询串进行分类。图 1 展示 URL 数目变化时 $P$, $R$, $F$, $A$ 的表现, 也就是含有 URL-key 数目的变化。从图 1 可以看出, 随着 URL 数目增加 $P$, $R$, $F$ 总体上呈现先下降再上升的趋势。当 URL 个数为 100 时, 实验结果最好。分析其原因为: 搜索引擎返回的结果排名越靠前与查询串越相关, 但随着 URL 数量的增加会受到一定的噪音影响。实验结果随 URL 个数的增加呈现下降的趋势, 当 URL 数量为 50 个的时候 $P$, $R$, $F$ 值最低, 主要是因为增加的搜索结果与查询串的主题

**表 6　领域 URL-key 数量分布**
Table 6　Distribution of domain URL-key

| 领域 | URL-key 数量 |
|---|---|
| Economy | 72 |
| IT | 85 |
| Health | 74 |
| Sports | 69 |
| Travel | 76 |
| Education | 72 |
| Job | 87 |
| Art | 53 |
| Military | 55 |
| Auto | 74 |

**表 5　部分领域 URL-key**
Table 5　Examples of URL-key

| 领域 | URL-key |
|---|---|
| Sport | sports, sport, nba, espnstar, pingpang, guojizuqiu, basketaball, snooker, soccer, tennis, badminton, olympics, olympic, tiyu, tiyv, ty, zhongchao, cbachina, weiqi, guoneizuqiu, sportsbl, chess, saichang, nba, wangqiu, cba, volleyball |
| Auto | auto, car, xcar, vw, toyota, chinacars, dongfeng, nissan, vehicles, audi, bitauto, new_cars, chery, newcar, peugeot, webcars, qiche, autohome, che |
| Economy | stock, bond, eastmoney, soufun, forex, ccb, cmbchina, stockstar, icbc, invest, boc, insurance, money |

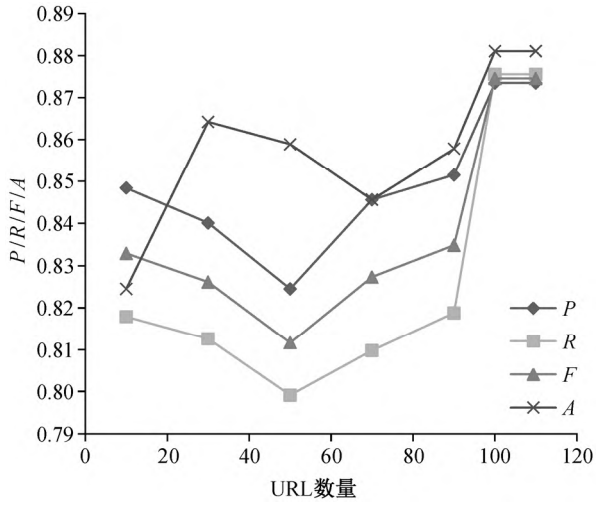


**图 1　实验结果随 URL 数目的变化**

Fig. 1　Performance of experiment results with different number of URL

无关。由此猜想，搜索引擎在第 3~5 页时检索的结果与查询串的主题相关性可能比较小。但是，当 URL 大于 50 个时，随着 URL 个数的增加，*P*，*R*，*F* 值呈增加的趋势。当 URL 达到 100 个时结果最好，且随着 URL 的增加趋于平稳状态。这是因为随着 URL 的增长，可被分类的 URL-key 数量增多，可以利用的资源更多，结果更好。但是用较大数目的 URL 有可能引入非相关文档，增加匹配时间，损害分类的效果。

表 7 给出本文方法与文献[6]方法的对比，均使用前 100 条 URL 作为实验语料。实验结果表明，本文的方法要好于对比方法，准确率为 87.35%，召回率为 87.56%，F 值为 87.45%，*F* 值提高近 7%，精确率提高近 6%。这是由于文献[6]采用的是基于 URL 匹配的方法，搜索结果反馈的 URL 中不能准确地与人工收集的 URL 进行完全的匹配。据统计，每返回 100 条 URL 时，有大约 20%的 URL 能够进行 URL 匹配，大约 40%的 URL 能够进行领域 URL-key 匹配。本文使用的基于领域 URL-key 的方法可以解决这个问题。本文不仅在实验结果上好于对比实验，而且使用的资源较少。本文只使用 717 个 URL-key 进行匹配，而文献[6]使用 40895 个 URL 进行循环剪枝匹配，大大影响了实验的时间效率。

**表 7　实验结果**

Table 7　Experimental result

| 方法 | *P*/% | *R*/% | *F*/% | *A*/% |
|---|---|---|---|---|
| 本文 | 87.35 | 87.56 | 87.45 | 88.07 |
| 文献[6] | 82.84 | 78.83 | 80.78 | 82.26 |

表 8 给出本文方法在不同类别中的各个指标参数，可以看出，对于健康类、旅游类、军事类和汽车类，本文方法取得较好的准确率和召回率，这是由于搜索引擎反馈的结果与主题查询比较明确。然而对于 IT 类，取得了较好的召回率，但准确率相对比较差。对实验结果分析发现，一些娱乐类的查询串被误分到 IT 类。分析其原因，娱乐查询中主要是小说查询，在返回的结果中含有“下载”的链接较多，而“下载”为 IT 类的 URL-key，所以 IT 类中娱乐类的数据会较多。此外，还有一些查询串本身就可以被分到多个类别中，如“j10”被标注为军事类，实际上被分配到 IT 类。对于教育类和求职类的查询串，搜索引擎反馈结果的类别 URL-key 不好区分，经常被相互误分，所以准确率、召回率都比较低。

**表 8　SVM 分类结果**

Table 8　Result of classification by SVM

| 领域 | *P* | *R* | *F* |
|---|---|---|---|
| Economy | 0.869565 | 0.930233 | 0.898876 |
| IT | 0.743590 | 0.966667 | 0.840580 |
| Health | 0.957447 | 0.918367 | 0.937500 |
| Sports | 0.939394 | 0.837838 | 0.885714 |
| Travel | 0.936170 | 0.916667 | 0.926316 |
| Education | 0.826087 | 0.826087 | 0.826087 |
| Job | 0.714286 | 0.750000 | 0.731707 |
| Art | 0.820000 | 0.836735 | 0.828283 |
| Military | 0.972222 | 0.875000 | 0.921053 |
| Auto | 0.956522 | 0.897959 | 0.926316 |
| 平均值 | 0.873528 | 0.875555 | 0.874541 |

## 4　总结

本文利用互联网中凝聚人们智慧的 URL 为资源，提取领域 URL-key，首次提出利用领域 URL-key 进行查询分类的方法，有效地解决了由网站爆炸式增长带来的领域 URL 不匹配问题，同时也解决了人工收集领域 URL 带来的时间效率问题。首先利用方差提取领域候选 URL-key，再利用机器翻

译技术、拼音翻译技术和搜索反馈技术对领域URL-key进行过滤, 构建领域URL-key词表。利用伪相关反馈技术, 构建query各个类别的URL-key得分向量, 使用SVM对query进行分类, 最终得到查询串的类别。本文方法的准确率达到87.35%, *F*值为87.45%, 在一定的程度上解决了部分查询分类的问题, 为查询分类提供了新的思路。

本文提出的方法仍有需要改进的地方, 例如: 领域URL-key的构建应该更加精确, 分类应该更细化。这些将在未来工作中进一步研究。

## 参考文献

[1] Broder A Z, Fontoura M, Gabrilovich E, et al. Robust classification of rare queries using web knowledge // Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 231–238

[2] Wang X, Zhai C X. Learn from web search logs to organize search results // Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 87–94

[3] 董志安, 吕学强. 基于百度搜索日志的用户行为分析. 计算机应用与软件, 2013, 30(7): 17–20

[4] [Taneva B, Cheng T, Chakrabarti K, et al. Mining acronym expansions and their meanings using query click log // Proceedings of the 22nd International Conference on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee, 2013: 1261–1272

[5] Shen D, Pan R, Sun J T, et al. Query enrichment for web-query classification. ACM Transactions on Information Systems (TOIS), 2006, 24(3): 320–352

[6] 张宇, 宋巍, 刘挺, 等. 基于URL主题的查询分类方法. 计算机研究与发展, 2012, 49(6): 1298–1305

[7] 李素建, 宋涛, 高杰, 等. 一种基于使用差异的词语领域性分析方法. 中文信息学报, 2009, 23(6): 72–78

[8] Yu S, Cai D, Wen J R, et al. Improving pseudo-relevance feedback in web information retrieval using web page segmentation // Proceedings of the 12th International Conference on World Wide Web. New York: ACM, 2003: 11–18