文章编号:2095-2163(2019)05-0167-04 中图分类号:TP301.6 文献标志码:A

基于 Trie 树的关键词匹配算法在电子政务领域的应用

陈有伟,康 磊
(西安石油大学 计算机学院,西安 710065)

摘 要: 传统的行政管理方式随着互联网的高速发展,其效率低下的弊端已经逐渐显露。各级部门在依托互联网快速发展的基础上积极引进现代互联网技术,结合现有行政管理的基本方式形成了符合当代环境的电子政务行政管理方式。民生诉求是电子政务的一个重要组成部分,保障和妥善解决民生问题是职能部门的重要职责,是反映其办事效率的一个窗口。然而由于民生诉求涉及到的投诉信息范围广、数量多、情况错综复杂,这给职能部门快速处理民生诉求带来了挑战。本文通过在电子政务系统中引入基于 Trie 树的关键词匹配算法,对市民提交的信息进行分析、匹配,从而快速分派到相应部门处理、极大地提升了各部门处理事务的效率。

关键词: 电子政务; Trie 树; 模糊匹配; 关键词匹配

Application of keyword matching algorithm based on Trie tree in e-government

CHEN Youwei, KANG Lei
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 With the rapid development of the Internet, the dilemma of the low efficiency of traditional administrative management methods has gradually emerged. On the basis of the rapid development of the Internet, departments at all levels actively introduce modern Internet technologies, and in combination with the basic methods of government administration, form an e-government administrative approach that conforms to the contemporary environment. People´s livelihood appeal is an important part of e-government. Safeguarding and properly solving people´s livelihood issues is an important duty of the department and a window reflecting the efficiency of department affairs. However, due to the wide range of complaints and the large number of complaints involved in the people´s livelihood appeals, this has brought challenges to the department´s rapid handling of people´s livelihood demands. This paper introduces Trie tree based on keyword matching algorithm in the e-government system to analyze the information submitted by the citizens, and then quickly dispatch them to the corresponding departments for processing, which greatly increases the efficiency of the department´s handling of affairs.

【Key words】 e-government; Trie tree; fuzzy matching; keyword matching

1 国内外研究现状

随着经济社会的快速发展,民众的诉求呈现出多样化的趋势,涵盖着从医疗、就业、教育等大的方面,直至寻物、家政等小的方面在内的众多议题观点[1]。研究可知,信息匹配技术在电子政务系统中是处理民生诉求的一项核心技术。如今,信息匹配技术在世界各国都取得了长足进步,依靠国家力量的支持,以信息匹配技术为核心的应用系统也得以广泛的发展[2]。斯坦福大学的特克和郝克特开发了一种基于内容的关键词匹配系统 SIFT(Standford Information Filtering T001)[3]。用户凭借这个系统,能够单独创建属于自己的词汇库,并通过使用相关关键字和空间模型来完成用户的诉求和网络信息内容间的相互匹配。美国国家安全局为了应对恐怖活动、军事威胁,建设了“Echelon”通信监视网络[4],可以通过卫星拦截大量包含个人信息的传真、电话和电子邮件等,Echelon 也是一个通过关键字匹配来获取通信的电子通信系统[5-6]。在英国,一个专门收集情报机构“英国政府技术援助中心”,在英国政府的主导下也随之成立,这个援助中心可以获取进出英国网络的所有信息[7]。

在国内,由于信息匹配技术和文本处理技术革新的相继问世,相关科研机构、高等院校以及公司,也设计了大量结合系统化技术的优秀产品[8]。例如中科天玑公司与中国科学院联合设计研发的“天机网络网页关键字监测系统”[9]。2009 年 1 月国内首个网络关键字安全研究机构在北京交通大学成立,如今该机构正在全方位地推进网络关键字的产生、传播和导控等方向的研究以及网络舆论安全关

作者简介: 陈有伟(1994-),男,硕士研究生,主要研究方向:计算机系统结构、管理信息系统; 康 磊(1968-),女,副教授,主要研究方向:嵌入式系统、计算机体系结构。

收稿日期: 2019-06-03

键技术的研发[10]。北京大学方正技术研究院设计推出了“方正智思网页关键字预警辅助决策支持系统”[11],该系统依靠对网页中的离线数据的自动解析和预报,合理分析并规划网页关键字的监控内容,产生了一种具有生命周期特征的社情民意反馈系统[2]。

随着国内外对于网络信息关键字的分析技术逐步成熟,关于信息匹配的软件产品得到了大量推广,国内电子政务领域的处理流程得到了部分改善。但是在处理专用信息上,关键词匹配技术还不够完善。特别是,对于市民提交的民生诉求信息的识别技术也仍表现出一定不足,难以满足智能化的要求,其准确率和时效性也有待提高,存在许多问题亟待解决。

2 基于关键字的布尔模型匹配算法

布尔模型因为实现方式简单、匹配速度快、检索方式易于用户理解[12]等特点,在诸多领域得到了应用,成为了网站搜索引擎使用的首选方案。布尔模型是结合集合论和布尔代数思想的简单数学模型,这种模型把文本信息中的关键词从文本信息中提取出来,作为文本的特征值[13]。匹配过程也很简单,把匹配词用“与”、“或”、“非”进行连接就可以组成相应的正则表达式,而后利用正则表达式与模型关键词进行对比得出匹配到的内容是否存在于该文档中。

设文档 $d_i(i=1,2,3,\cdots,n)$ 为文本集 $D=(d_1,d_2,\cdots,d_n)$ 中任意一个文档, $T_i=(t_1,t_2,\cdots,t_m)$ 为文档 d_i 标引词集,对于某检索,形如 $Q=W_1\wedge W_2\wedge\cdots\wedge W_n$,如果存在 $W_1\in T_i, W_2\in T_i, W_i\in T_i$,则称文档 d_i 存在于检索结果当中,这里 d_i 为命中文档,反之 d_i 为不命中文档;对于检索形式为 $Q=W_1\vee W_2\vee\cdots\vee W_n$ 的检索式,如若存在一个或多个 $W_k\in T_i$,$(k=1,2,\cdots,n)$,则 d_i 为命中文档,反之若不存在满足条件的 $W_k\in T_i$,$(k=1,2,\cdots,n)$,则 d_i 为不命中文档[14]。

布尔模型的优势表现在其匹配速度快、实现方式简单等方面,但是这种模型的不足也十分明显。对此可做阐释分析如下。

(1)布尔模型对满足其前提条件的文档进行匹配时容易造成遗漏。由于布尔模型拥有严格的匹配规则,关键字的选取如果稍有偏差就有可能会被过滤,例如当使用“与”作为连接词进行匹配时,系统匹配仅仅只命中与匹配词一致的文档,但是那些和匹配词不一致、内容却一致的文档通常会被遗漏,所以如何选取合适的匹配词就变得十分困难[15]。

(2)无法匹配重点结果。由于布尔模型匹配到的结果是一个大致的范围,对于数据量小的情况比较适用,但是对于在电子政务领域逐步增长的海量数据信息,布尔检索在处理能力上的不足就显得尤为突出。

(3)容易造成匹配结果的冗余。

(4)因为布尔匹配的实现方式过于简单、往往不能完全反映出想要的结果。

正是由于民生诉求包含的社会信息十分复杂、庞大,为了能快速地处理这些信息,引入了一种高效的数据结构—Trie 树。

3 基于 Trie 树的匹配算法

3.1 Trie 树

Trie 树也叫作字典树,是对一组词进行结构化处理后的组织[16]。

其中,字典树对含有相同前缀的词进行压缩处理,使其所占用的空间得到了极大优化。同时由于将相同公共前缀的词放在了一起,则使得通过前缀进行匹配也变得十分迅速。研究中构建的一颗字典树即如图 1 所示。

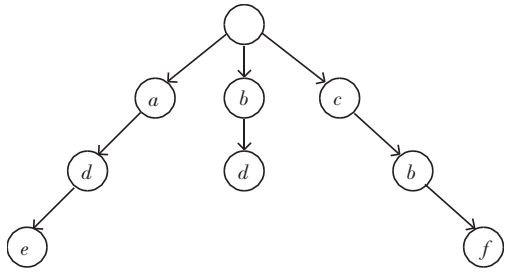

图 1 Trie 树示意图

Fig. 1 Schematic of Trie tree

字典树通过从根节点到子节点的路径来表达一个词,图 1 中 e,f 节点为一个词的最后一个节点,也就是说图 1 字典树代表的单词有 ade、ad、bd、cbf、cb。字典树的根节点不表示任何字符。字典树不仅节省了存储空间,同时为模糊匹配技术的发展提供了坚实的基础。

3.2 构造基于中文的 Trie 树

英文 Trie 树的结点是由 26 个英文字母组成的,所以英文 Trie 树的一个节点最多拥有 26 个子节点。但是中文却不一样,生活中常用的汉字就高达 7 000 多个,如果按照英文 Trie 树的构建法则来构建中文 Trie 树,将会极大地降低匹配的效率。因此如何构造基于中文的 Trie 树结构就有着至关重要的研究

意义。

比如，在向教育局投诉的信息中，根据教育局的相关关键词构建属于教育局的 Trie 树结构，以关键词“教育局”为例：

首先，基于拆词的思想，利用正则表达式将关键词“教育局”拆分为教、育、局三个字。

接着，检查根节点是否已经有字符“教”节点，如果已经有这个节点，依次重复检验并添加“育”、“局”两个节点。如果没有，则将“教”添加在根加点下。

最后，当插入了每个关键词时，在其末尾增加一个标志符，使用这个字符作为此关键词的结束标志（如图 2 中的灰色三角），利用这个字符来标记查找到了这个关键词。

循环插入所有关键词。构造出的中文 Trie 树如图 2 所示。

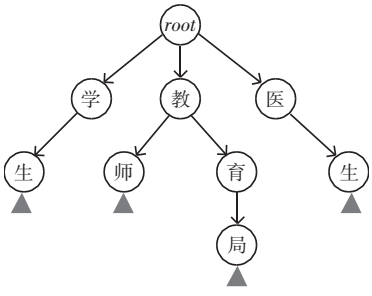

图 2　中文 Trie 树

Fig. 2　The Chinese Trie tree

3.3　利用中文 Trie 树解决中文匹配

以一则民生投诉为例：“我是 X 中初四学生家长，听孩子说上体育课跑操时老师大声骂学生，有时还用脚踢学生，学生真害怕，3、4 班的。请求帮助。”利用图 2 已经构造好的中文 Trie 树来开始匹配。

首先，将投诉内容利用正则表达式拆成单个字符“我”、“是”、…；从根节点处查找第一个字符“我”，并没有查找到以“我”为首字符的关键词，然后继续移动字符指针，直到查找到符合条件的字符节点“学”；接着在“学”这个字符节点下查找字符节点值为“生”的节点，成功找到时计算子树的深度为 2，关键词的长度是 2，此时字符指针继续移动，如果发现结束标志，就意味着匹配成功，将匹配到的关键词返回，如果未碰到结束标志则继续向后移动指针结点寻找下一个字符。

循环遍历完毕，返回所有匹配到的关键词。

3.4　Trie 树的数据结构设计

Trie 树的数据结构设计采用 PHP 语言，结合了 PHP 数组的 hash 特性，代码如下：

```
Private $root = array(
    'depth' => $depth,
    // 深度，用来判断命中的字数
    'next' => array(
    $val => $node,  // 使用 PHP 数组的 hash 结构，增加子节点的查找速率
    …
    )
)
```

4　实验结果及分析

实验环境为 MacBook Pro(Retina，15-inch，Mid 2015)，处理器为 2.2 GHz Intel Core i7，内存 16 GB 1600 MHz DDR3，使用 PHP 语言实现。实验中的给定文本内容来源于某市民心网 1 000 个市民提交的诉求问题。

将 1 000 个市民提交的问题内容分成 4 个小组，每组 250 篇，并计算其查全率、查准率以及所耗时间。基于 Trie 树结构的关键词匹配结果，见表 1。

基于正则表达式的关键词匹配结果，见表 2。

表 1　基于 Trie 树的关键词匹配结果

Tab. 1　Key word matching results based on Trie

实验组号	敏感词总个数	正确过滤个数	查全率/%	查准率/%	所用时间/s
第一组	256	242	94.53	93.07	1.25
第二组	265	257	96.98	94.48	1.11
第三组	234	227	97.00	95.69	0.86
第四组	274	255	93.06	92.72	1.37

表 2　基于正则表达式的关键词匹配结果

Tab. 2　Keyword matching results based on regular expressions

实验组号	敏感词总个数	正确过滤个数	查全率/%	查准率/%	所用时间/s
第一组	256	235	91.80	94.00	3.37
第二组	265	244	92.08	93.12	3.42
第三组	234	217	92.73	94.34	3.08
第四组	274	248	90.51	93.28	3.63

要应用在电子政务领域，至关重要的就是效率与准确率。通过以上实验结果可以发现，与在电子政务系统中单纯使用正则表达式相比，使用 Trie 树结构处理 250 条数据基本耗时在 1 s 左右，并且根据关键词匹配到的结果，将其分发到命中的部门，准确率基本都高达 93% 以上，明显改善了处理民生诉求问题的效率，符合电子政务领域的基本要求。

5 结束语

本文通过在电子政务系统中引入 Trie 树这种效率极高的数据结构结合正则表达式,极大地提高了匹配效率,使得职能部门在处理民众诉求时,能够及时将民众反映的相关问题分派到相应的部门去办理,优化部门办事效率,提升了民众对职能部门的工作满意度。

参考文献

[1] 麦范金, 李东普, 岳晓光. 基于双向匹配法和特征选择算法的中文分词技术研究[J]. 昆明理工大学学报(自然科学版), 2011,36(1):47-51.

[2] 靳瑞敏. 网页关键字过滤研究及改进[D]. 呼和浩特:内蒙古大学,2012.

[3] http://zjnustdl.blogdriver.com/zjnustdl//1196699.html.

[4] 俞文洋,张连堂,段淑敏. KMP 模式匹配算法的研究[J]. 郑州轻工业学院学报(自然科学版),2007,22(5):64-66.

[5] HARALICK R M. Statistical and structural approaches to texture [J]. Proceedings of the IEEE, 1979,67(5):786-804.

[6] TAMURA H, MORI S, YAMAWAKI T. Textural features corresponding to visual perception [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1978,8(6):460-473.

[7] CHEN Yixin, WANG J Z, KROVETZ R. Clue:Cluster-based retrieval of images by unsupervised learning [J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2005,14(8):1187-1201.

[8] FLECK M, FORSYTH D, BREGLER C. Finding naked people [C]// 1996 European Conference on Computer Vision. Berlin, Germany:Springer-Verlag, 1996,2:592-602.

[9] WU S, MANBER U. A fast algorithm for multi-pattern searching [R]. Tucson:University of Arizona, 1994.

[10] SAGE D, NEUMANN F R, HEDIGER F, et al. Automatic tracking of individual particles: Application to the study of chromosome dynamics [J]. IEEE Transactions on Image Processing, 2005,14(9):1372-1383.

[11] http://www.ekany.corn/wd998/cg/tutorialapter8/lesson8 - 6.html.

[12] 李静. 基于概念匹配度模型的文献检索系统[D]. 成都:西南交通大学,2009.

[13] 段立娟,崔国勤,高文,等. 多层次特定类型图像过滤方法[J]. 计算机辅助设计与图形学学报,2002,14(5): 404-409.

[14] 范晓,申铉京. 基于 IE 浏览器的色情图片过滤器[J]. 吉林大学学报(信息科学版),2004,22(6): 631-637.

[15] 冯军红,刘桂林,高立新,等. 基于小样本训练集的肤色模型建立方法[J]. 计算机工程与应用,2003(28):67-71.

[16] 赵晓晖. 基于内容的敏感图片过滤技术的研究及其在 IE 浏览器中的实现[D]. 长春:吉林大学,2005.

(上接第 166 页)

原则,分别是计划性原则、时效性原则、针对性原则及连贯性原则,这有利于保障搜集到的经济信息是全面、有效,且符合实际要求的。此外,还要规范经济信息搜集的程序,严格按照步骤来开展工作,提高信息搜集的效率。

4.2 构建宏观经济信息管理系统

在搜集到相关的经济信息后,就需要对经济信息进行整合,从中挖掘出有用的经济信息,并应用于宏观经济管理的工作中,这就亟需要构建宏观经济信息管理系统。首先,在系统的构建过程中要严格遵循的一系列基本原则,对此可分述如下。

(1)要在结合具体实际需要的情形下进行系统构建。

(2)构建的系统要符合独立性、完整性、经济性等方面的要求。

(3)构建的经济信息管理系统要为宏观经济管理服务。

其次,在构建经济信息管理系统时,还需要提高信息化程度,建立健全网络化信息系统,从而提高宏观经济信息管理系统的应用率。

4.3 加大人才的培养

对于宏观经济管理中高素质人才稀缺的问题,要从人才的质量与数量两方面入手,来满足人才市场的需求。一方面,高校应该适当开设经济信息管理专业,扩充经济信息管理类的人才队伍;另一方面,要加大人才培养的投资力度,加强培训,特别是对于经济管理的专业与信息技术方面的知识培训,并且要提供多种机会,增加此类学员的实践经验,提高实践能力。这样,才能够尽快达到市场对经济信息管理类的人才需求培养目标,进而提升经济信息在宏观经济管理中的应用水平。

5 结束语

随着中国改革开放的深入与经济的发展,提高经济信息在宏观经济管理中的应用具有重要的意义。而提高对经济信息的认识,转变观念,深入研究如何进一步提高经济信息在宏观经济中的应用,有利于推动经济的稳定可持续性发展。

参考文献

[1] 李学成. 经济信息在宏观经济管理中的应用[J]. 现代经济信息,2012(4):1.

[2] 谢芳芳. 经济信息在宏观经济管理中的应用研究[J]. 长春金融高等专科学校学报,2015(1):86-89.

[3] 李金霞,杨丽媛. 典型相关分析在我国财政政策与宏观经济发展分析中的应用[J]. 市场经济与价格,2014(1):39-42.