

一种基于双数组 Trie 的 B2B 规则串提取方法

李 慧 杨炳儒 潘丽芳 钱文彬

(北京科技大学计算机与通信工程学院知识工程研究所 北京 100083)

摘 要 针对 B2B 垂直搜索引擎中提取产品规格信息困难的问题,提出了一种基于双数组 Trie(Double-Array Trie)的规则串提取方法。该方法针对 B2B 系统中“参数名:参数值”字符串的规则特征构建规则串,生成双数组 Trie 树;并优先处理分支结点最多的子树,来提高存储效率。该方法对搜索文本进行一次扫描就能得到所有规则串;通过在规则串中加入约束条件,对候选串进行有效过滤,以提高规则串的提取准确率。实验表明,该方法能够降低传统规则串查找的算法复杂度,查找规则串的时间复杂度是 $O(n)$ 。

关键词 双数组 Trie,垂直搜索,规则串,B2B 系统

中图法分类号 TP311.12 **文献标识码** A

Rules String Extracting Method for B2B System Based on Double-array Trie

LI Hui YANG Bing-ru PAN Li-fang QIAN Wen-bin

(Knowledge Engineering Institute, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract To extract the data of product specification in B2B system, the ruled string extracting method based on double-array trie was proposed. The data feature is formed as "name:value" for the parameters of the product specification in B2B system. The method constructs the rule according to the data feature of specification parameters. The double-array trie is generated for the extracting processing according to the rules database. The optimization measures are adopted to improve the storing efficiency for the double-array trie. The measures include giving high priority to handle the sub tree with more child node. The method can extract all the ruled string by scanning the input text data once. The accuracy of the extracting results is improved via filtering according to the restrictions condition of the rules. Experimental results show that the extracting method can improve accuracy and decrease complexity comparing to the traditional method. The complexity of the extracting algorithm is $O(n)$.

Keywords Double-array Trie, Vertical search, Rules string, B2B system

1 引言

随着网络技术的发展,对于 B2B 供应商和采购商来说,一个快速、准确的网络平台进行产品信息交流成为一个急切的需求。垂直搜索引擎^[1,2]对网页信息进行了结构化的信息抽取,将网页的非结构化数据抽取成特定的结构化信息数据,以结构化为最小单位,并对其进行去重、分类等加工处理,然后进行分词、索引,最终向用户提供获取专业的、精确的、深入的网页信息,提高了信息获取效率。B2B 垂直搜索引擎技术因而成为这个网络平台的重要支撑,它能够帮助采购商快速找到产品的规格描述,从而使采购商迅速了解商品。

但在实际应用中,产品的规格信息往往只包含在产品的文本描述中,这使得采购商不得不面临巨大的浏览量,同时采购网站根据产品规格进行过滤也尤为困难。因此,如何有效

提取产品详细描述中的商品规格信息成为一个急需研究的问题。

商品规格信息规则化主要利用自然语言处理中的规则串提取手段来完成,双数组 Trie 算法是规则串提取的一种非常有效的手段。双数组 Trie 树算法最早由 Aoe^[3,4]提出并实现,其用两个一维数组表示一棵 Trie 树。Theppitak^[5]对双数组 Trie 进行了实现,并对其可操作性进行了研究。

王思力^[6]对双数组 Trie 树的构造过程提出了一种优化策略,其通过优先处理分支结点数更多的结点,在保证该算法数据查找效率不变的同时,减少数据稀疏,提高空间利用率。

赵欢^[7]在构造双数组 Trie 树的过程中,通过构造一个空状态序列,将冲突的结点放入 Hash 表中,以减少冲突。优化后的双数组 Trie 树插入速度和空间利用率得到了很大提高,且分词查询效率也得到了提高。

到稿日期:2012-07-18 返修日期:2012-11-22 本文受国家自然科学基金项目(61175048, 60875029),科技部创新方法工作专项项目(2010IM020900)资助。

李 慧(1983—),女,博士生,主要研究方向为数据挖掘、知识获取, E-mail: lihuimuizi@126.com; 杨炳儒(1943—),男,教授,博士生导师,主要研究方向为数据挖掘、人工智能; 潘丽芳(1980—),女,硕士,主要研究方向为数据挖掘; 钱文彬(1984—),男,博士生,主要研究方向为数据挖掘、粗糙集。

刘燕兵^[8]针对网络内容过滤系统中的多模式串匹配算法进行了存储优化,使用双数组 Trie 表示匹配串,显著地提高了串匹配的速度,使其适合于在线高速匹配的应用环境。

刘群^[9]提出一种基于层叠隐马模型的汉语词法分析方法,即引入角色 HMM Viterbi 算法标出角色序列,识别登录词,计算其可信度。

本文针对 B2B 垂直搜索引擎中提取产品规格信息困难的问题,提出了一种基于双数组 Trie 树的规则串提取方法。该方法针对系统中“参数名、参数值”形式的字符串规则特征生成规则串,以及参数名、参数值分别生成参数名 Trie 和参数值 Trie。经过约束条件过滤,最终得到需要的提取串。在构建双数组 Trie 树过程中,优先处理分支结点最多的子树,从而提高 Trie 树的运行效率。

2 规则串提取算法

本文提出的基于双数组 Trie 的规则串提取方法由 4 部分处理组成:

1)规则库管理模块。通过该模块将规则串导入规则库,在具体应用中规则库可以进行添加、删除等维护操作。同时在该模块中可添加规则间的约束条件。

2)双数组 Trie 生成模块。通过该模块系统将根据规则库中的规则串,生成双数组 Trie,在本方法中将针对“参数名”和“参数值”分别生成双数组 Trie。

3)规则串提取模块。该模块是进行商品规格信息提取的关键模块,在生成双数组 Trie 后,该模块将遍历所有需搜索的网页文本,并从中提取商品信息。因此该模块的效率是本方法的关键。

4)约束性判断模块。商品规格参数名与参数值存在一定的对应关系,且同一件商品的参数名与参数值在理论上也应具备一定的位置约束,因此本方法对提取的规则串进行约束性判断,过滤掉不符合约束条件的候选串,以提高搜索的准确度,得到准确性更好的规则串(参数名;参数值)。

本文提出的“基于双数组 Trie 的规则串提取方法”体系结构可表示为图 1。

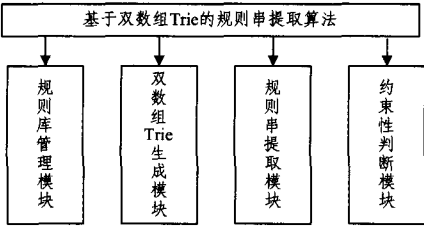


图 1 规则串提取方法模块图

2.1 规则库构建

从信息检索角度看,B2B 垂直搜索技术不但能够检索网页中的结构化信息,还能够提供结构化与非结构化信息相结合的检索方式。比如我们关注的商品信息:类别:服装;适合身高:165cm;颜色:橙色。

根据 B2B 行业产品的数据特点,本文提出的方法通过分析 B2B 商品的数据特征,首先将参数名、参数值及两者的约束关系添加至规则库。本文规则串提取方法中的规则包含 3 类元素:参数名、参数值以及两者的约束关系。

规则库中的规则由人工整理添加,也可批量导入,对“参

数名;参数值”形式的字符串,生成支持该规则特征的规则串。参数名表和参数值表分别对应下面的数据结构:

参数名表:

{参数名 1,参数名 2,...,参数名 n}

参数值表:

{参数值 1,参数值 2,...,参数值 n}

规则约束关系能够添加对规则的约束条件,比如参数名和参数值要符合一定的条件才能成为合法的规则串。参数名和参数值的约束关系,在实验中具体化成参数名和参数值的对应关系。

约束关系表 1:

{参数名 1;参数值 11,参数值 12,...,参数值 1m}

约束关系表 2:

{参数名 2;参数值 21,参数值 22,...,参数值 2m}

.....

规则串存储在规则库中,参数名和参数值对应关系存储在 hash 结构中。

2.2 双数组 Trie 规则树生成

在规则库创建完毕后,使用双数组 Trie 树生成算法生成本系统使用的参数名 Trie 树和参数值 Trie 树,参数名 Trie 和参数值 Trie 均被生成两个双数组 Trie,两 Trie 树的生成原理如图 2 所示。

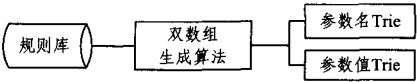


图 2 双数组 Trie 生成原理图

2.2.1 双数组 Trie 树存储结构

双数组 Trie 是 Trie 搜索树的一种变形,其本质上是一个确定的有限状态自动机,如图 3 所示。图 3 中有 8 个状态,每个节点代表一个状态,1 状态是初始状态,双线圈代表可终止状态(如 3、5、7、8 状态),弧代表转移条件。图 3 中有 4 个终止状态表示它可以识别 4 个字符串,可识别的字符串包括“北京”,“北京大学”,“北大”,“大学”。

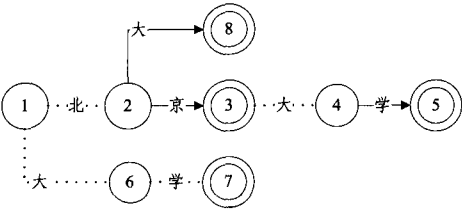


图 3 双数组 Trie 状态自动机

根据输入变量的不同,状态机进行状态转移。前一状态接收输入进入下一状态,随着输入结束,状态从开始状态转移到终止状态,从开始状态到终止状态的转移过程就是输入字符串的识别过程。

双数组 Trie(Double Array Trie)的存储结构是两个整数数组,一个是 base[],另一个是 check[],这两个数组拥有一致的下标。base 数组用于确定状态的转移,数组的每个元素相当于 Trie 树的一个节点,其值作状态转移的基值。check 数组用于检验转移的正确性,check 值相当于校验值,用于检测该状态是否存在。

设数组下标为 i,如果 base[i],check[i]均为 0,表示该位置为空,如果 base[i]为负值,表示状态为终止状态(即可识别

词语), $Check[i]$ 表示该状态的前一状态。

在图 4 所示的状态转换中, 初始状态为 s , 输入为 c , 输入后系统转为状态 t 。

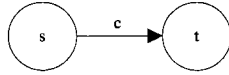


图 4 状态转换图

在双数组 Trie 的存储结构中, 从状态 s 输入 c 到状态 t 的一个转移必须满足式(1),

$$\begin{cases} base[s] + c = t \\ check[t] = s \end{cases} \quad (1)$$

根据上述条件, 双数组 Trie 的 $base$ 数组以及 $check$ 数组的数据存储结构及其对应关系如图 5 所示。

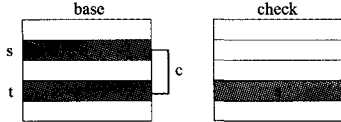


图 5 双数组 Trie 的存储结构

2.2.2 双数组 Trie 树生成

以图 3 的 Trie 树为例, 将树中的字符排序为“北—>1, 京—>2, 大—>3, 学—>4”, 根据双数组 Trie 树模型, $base$ 数组和 $check$ 数组构造过程如下:

- 1) 将第一个结点入队, 然后出队, 并将其直接孩子结点入队。初始化它们的序列码为下标的 $base$ 和 $check$ 的值为 0。
- 2) 当队列不为空时, 将第一个结点的所有孩子结点数最多的一个结点加入队列中, 孩子结点数相同的按照入队的顺序排列, 并将该结点作为当前结点。
- 3) 当前结点出队, 修改 $base$ 数组中的值, 确定各个直接子结点在数组中的位置, 令各直接子结点在 $check$ 数组中的值为当前结点的数组下标。
- 4) 为了解决访问中的冲突问题, 当前结点为可终止状态 (即可组成一个词语) 时, 修改其 $base$ 为原值的相反数。
- 5) 当前结点的直接孩子结点加入队列, 重复 2) 直至所有节点处理完毕。

基于图 3 的 Trie 树建立的 $base$ 数组和 $check$ 数组的存储内容如表 1 所列, 表 1 中空白项均为初始值 0, $Base[i] < 0$ 表示一个终止状态。通过 $base$ 数组的负值数可知, 该双数组 Trie 可识别“北京”、“北京大学”、“大学”、“北大”4 个字符串。

表 1 $base$ 和 $check$ 的存储内容

下标	base	check	识别字符串
0			
1	8		北
2			
3	8		大
4			
5			
6			
7			
8	-8	13	北京大学
9			
10	-10	1	北京
11	-11	1	北大
12	-12	3	大学
13	4	10	北京大

2.3 基于双数组 Trie 的规则串提取

根据双数组 Trie 中的字符排序, 将规则字符串中的字符量化为数字, 如北—>1、京—>2、大—>3、学—>4, 不在规

则字符串中的字符量化为无穷大值; 根据 $base$ 和 $check$ 数组的存储内容, 识别过程如下:

- 1) 初始状态为 $base[0]=0, check[0]=0$;
- 2) 输入“北”, “北”的量化数字为 1, 即 $c=1$, 根据公式 $base[0]+1=1$ 到达状态 1, 此时 $base[1]=8, check[1]=0$ 。
- 3) 输入“京”根据公式 $base[1]+2=10$, 到达状态 10, 此时 $base[10]=-10, check[10]=1$ 。由于 $base[10]<0$, 因此该节点为一个终结点, 已识别出“北京”字符串。
- 4) 由于“北京”字符串可以成为一个独立词语, 也可作为“北京大学”词语的一部分, 因此, 该字符串的判断取决于下一输入造成的状态转换是否为合法转换。
- 5) 输入“大”, 根据公式, 下一状态是 $base[10]+3$, 在查找下一状态时, 规定如果 $base[i]<0, base[i]=0-base[i]$, 那么下一状态是 $0-base[10]+3=13$, 到达状态 13, 此时 $base[13]=4, check[13]=10$ 。
- 6) 输入“学”, 根据公式 $base[13]+4=8$, 到达状态 8, 此时 $base[8]=-8, check[8]=13$ 。由于 $base[10]<0$, 因此该节点为一个终结点, 识别出“北京大学”字符串。

在上述识别过程中, 假如输入为非识别字符, 将使该状态变换至非法状态, 从而跳出当前词语的识别过程。

2.4 提取串约束性判断

通过规则串提取算法, 可以调用参数名 Trie 提取出所有的参数名字符串, 并记下参数名字符串在待提取字符串中的位置; 然后调用参数值 Trie 提取出所有的参数值字符串, 并记下参数值字符串在待提取字符串中的位置。

在所有提取到的参数名规则串和参数值规则串中, 候选提取串要满足参数名和参数值的约束条件, 才能组合成有效的“参数名: 参数值”规则字符串。

本方法已拥有的约束性条件包含如下几方面:

- 1) 所有的参数值必须在参数名的值域范围之内;
- 2) 有效的参数值字符串必须和其前的参数名字符串位置距离保持在一定阈值范围内;
- 3) 参数名字符串和其后的参数值字符串组成候选提取串, 如果参数名字符串之后还是参数名字符串, 则前一参数名字符串丢弃。

对于不满足约束条件的提取串, 将自动过滤掉。经过约束条件过滤, 最终得到需要的提取串。

3 算法实现及实验分析

本文提出的基于双数组 Trie 的 B2B 商品规格数据规则串提取方法, 已经在 PC 机上实现, 其算法流程图可描述为图 6。

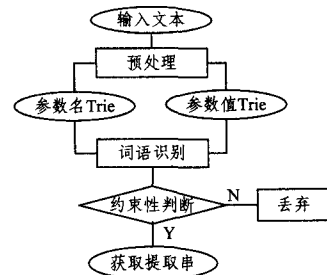


图 6 规则串提取算法实现

其具体算法可描述为:

(下转第 223 页)

- [23] 赫然,王永吉,王青,等.一种改进的自适应逃逸微粒群算法及实验分析[J]. 软件学报,2005,16(12):2036-2044
- [24] 邹晓峰,陆建江,宋自林.基于模糊分类关联规则的分类系统[J]. 计算机研究与发展,2003,40(5):651-656
- [25] 李锦泽,叶晓俊.关联规则挖掘算法研究现状[C]//第18届计算机技术与应用学术会议(CACIS). 2007:216-220
- [26] Robinson J, Rahmat-Samii Y. Particle swarm optimization in electromagnetics[J]. IEEE Transactions on Knowledge and

Data Engineering, 2004, 52(2):397-407

- [27] 陈翔,顾庆,王子元,等.一种基于粒子群优化的成对组合测试算法框架[J]. 软件学报,2011,22(12):2879-2893
- [28] Ratnaweera A, Halgamuge SK, Watson HC. Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients[J]. IEEE Trans Evol Comput, 2004, 8(3):240-255
- [29] 蔡昭权,黄翰.自适应变异综合学习粒子群优化算法[J]. 计算机工程,2009,35(7):170-171,202

(上接第208页)

1)在建立参数名 Trie 和参数值 Trie 后,首先对输入的网页文本数据预处理,根据逗号、句号等分隔符将文本分成长字符串;

2)通过参数名 Trie 和参数值 Trie 分别对输入的长字符串串进行词语识别;

3)循环处理每个长字符串,对提取获得的规则字符串进行约束性判断,如果不符合判断条件,则丢弃;否则作为提取串保存至数据库作进一步分析。

系统可以很方便地扩展成过滤系统。例如过滤垃圾邮件,只要将垃圾邮件中经常出现的字符串作为规则即可。

通过设计实验,针对本文的提取方法进行验证,实验1采用的两段输入文本如下:

输入文本1:单相控制隔离变压器符合 EN61558 标准接触保护(VBG4)绝缘等级 T40 变压器 MCT 输入电压 230/400V AC 输出电压 24V 绝缘等级 T 40/B

提取串:输入电压:230/400V

输出电压:24V

绝缘等级:T 40/B

输入文本2:此款上衣胸围 86cm 肩宽 38cm 适合平时穿 160/84 的 mm,颜色黑色。

提取串:胸围:86cm

肩宽:38cm

颜色:黑色

实验1中根据所需提取串的特征,所建立的参数名表、参数值表以及约束条件构建如下:

参数名表(输入电压,输出电压,绝缘等级,胸围,肩宽,颜色)

参数值表(v,b,cm,黑色)

约束条件(输入电压:v),(输出电压:v),(绝缘等级:B),(胸围:cm),(肩宽:cm),(颜色:黑色)

实验1的运行结果表明,本方法能够正确、准确地提取所需要的提取串。

实验2通过加大输入文本的数据量,来测试提取串的正确率和召回率,所获取的实验结果数据如表2所列。

表2 提取串的正确率

类型	测试数目	正确率	召回率
品牌词	1000	95.7%	93.1%
类别词	10000	96.3%	94.6%
服务项目	10000	97.1%	95.4%

通过表2的数据可以看出,本方法能够达到的正确率和召回率均满足 B2B 商务搜索引擎的需求。

结束语 本文提出了一种基于双数组 Trie 树的规则串提取方法,用于提取 B2B 垂直搜索引擎中产品规格的信息。通过对 B2B 系统中的参数名、参数值等规则串构建规则库,并生成双数组 Trie 树,实现在一次扫描过程中获取规则串;通过规则库中的约束条件判断,对候选提取串进行过滤,以提高提取准确率;在双数组 Trie 存储方面,优先处理分支结点最多的子树,以提高存储效率。实验表明,该方法能够降低传统规则串查找的算法复杂度,查找规则串的时间复杂度是 $O(n)$ 。提取算法已经应用到实践中,能够提高网站的数据质量。本文的进一步研究内容包括:对参数值的进一步约束以及约束性自动判断方面。

参考文献

- [1] Curran K, Glinchey J M. Vertical Search Engines [J]. ITB Journal, 2008, 16:22-28
- [2] 雷育生.基于垂直网站的网络信息支持系统研究[J]. 计算机应用研究, 2005, 7:105-107
- [3] Aoe J. An Efficient Digital Search Algorithm by Using a Double-Array Structure [J]. IEEE Transactions on Software Engineering, 1989, 15(9):1066-1077
- [4] Aoe J, Morimoto K, Sato T. An Efficient Implementation of Trie Structures [J]. Software Practice and Experience, 1992, 22(9):695-721
- [5] Karoonboonyanan T. An Implementation of Double-Array Trie [OL]. <http://linux.thai.net/theop/datrie/datrie.html>, 2003
- [6] 王思力,张华平,王斌.双数组 Trie 数算法优化及应用研究[J]. 中文信息学报:人工智能及识别技术, 2006, 20(5):24-30
- [7] 赵欢,朱红权.基于双数组 Trie 数中文分词研究[J]. 湖南大学学报, 2009, 36(5):77-80
- [8] 刘燕兵,刘萍,谭建龙,等.基于存储优化的多模式串匹配算法[J]. 计算机研究与发展, 2009, 46(10):1768-1776
- [9] 刘群,张华平,俞鸿魁,等.基于层次隐马模型的汉语语法分析[J]. 计算机研究与发展, 2004, 41(8):1421-1429
- [10] Dorji T C, Atlam E-S. New methods for compression of MP double array by compact management of suffixes[J]. Information Processing & Management, 2010, 46(5):502-513
- [11] Schubert P, Legner C. CAKES-NEGO: Causal knowledge-based expert system for B2B negotiation[J]. Expert Systems with Applications, 2011, 35(1):459-471
- [12] Rosenzweig E D, Timothy M. Through the service operations strategy looking glass: Influence of industrial sector, ownership, and service offerings on B2B e-marketplace failures[J]. Journal of Operations Management, 2011, 29(1):33-48