文章编号：2095-2163(2020)07-0061-04 中图分类号：TP311.5 文献标志码：A

# 服装个性化定制需求提取问题的研究

唐豪杰，刘国华
（东华大学 计算机科学与技术学院，上海 201620）

**摘 要**：提供个性化定制服务是“工业 4.0”时代服装生产的一大特色，准确理解顾客的个性化需求是保证个性化定制服务质量的前提。但是由于个性化需求有提供形式多样化（语音、文字、图片等）、内容复杂且具有歧义性等特点，给个性化需求提取工作带来了困难，从而制约了对个性化需求的理解。本文针对服装个性化定制需求提取的问题，根据服装设计过程和需求提取流程的特性，为服装个性化定制需求提取问题提供了一套解决方法。该方法首先对需求文本进行分词和词性标注的预处理工作，然后应用有限状态自动机以及模式匹配等理论，通过构建有限自动机，识别由关键词组成的正则语言，从而提取相关的服装属性。

**关键词**：工业 4.0；有限状态自动机；需求提取；服装个性化定制

# Research on the demand extraction of personalized clothing customization

TANG Haojie, LIU Guohua
（School of Computer Science and Technology, Donghua University, Shanghai 201620, China）

**【Abstract】** Providing personalized customized service is a major feature of garment production in the era of industry 4.0. Accurate understanding of customer´s personalized needs is the premise to ensure the quality of personalized customized service. However, due to the diversity of forms (voice, text, picture, etc.), complexity and ambiguity of the content, personalized needs extraction work is difficult, which restricts the understanding of personalized needs. According to the characteristics of the process of clothing design and demand extraction, this paper provides a set of solutions for the problem of demand extraction of clothing personalized customization. This method first preprocesses the word segmentation and part of speech tagging of the demand text, then applies the theory of finite state automata and pattern matching to construct the finite automata, recognize the regular language composed of keywords, and extract the related clothing attributes.

**【Key words】** Industry 4.0; Finite-State Machine; Demand extraction; Clothing customization

## 0 引 言

互联网的出现和信息技术水平的不断提高，使得互联网技术与其他行业的融合更加丰富和深入，传统制造业为了取得发展，正不断优化自身生产模式，向智能化的先进制造业转型升级。2013 年，德国在汉诺威工业博览会上提出“工业 4.0”战略，其核心是“智能+网络化”，即通过虚拟-实体系统（CPS），构建智能工厂，实现智能制造的目的[1]。国内很多传统制造企业借鉴智能制造的思路，希望抢得先机建立核心竞争力，完成企业转型升级。如何发展智能制造已经成为行业中最为热门的话题之一。与此同时，近年来新一轮科技革命和产业变革孕育兴起，《中国制造 2025》中数次提及发展“定制”[2]。

我国服装产业在发展过程中形成了大量生产，集中消费的模式。近几年，随着生活水平的提升，人们对服装个性化、差别化的要求越发明显，消费者选购衣物时融入了更深层次的自我意识。消费者的需求出现创新、个性化、短周期的趋势，这种趋势要求制造商必须迅速适应市场的变化，定制化的产品模式在这种背景下应运而生[3]。

目前，市场上有两大类服装定制服务，一种是高端私人定制，这种传统定制服务由于成本高，生产周期长等劣势，为服装企业带来了巨大压力；另一种是服装制造商承接的集体服装订单，如制服、工作服等，并不能满足个性化需求，不属于真正意义上的服装定制。结合网络信息技术的智能化服装个性化定制是未来服装行业定制服务的一个发展方向。本文对服装个性化需求提取问题进行了探究，利用有限状态自动机等计算机理论，提出解决方法。对从自

**基金项目**：科技部国家重点研发计划（2017YFB0309800）；上海市工业互联网创新发展专项项目（2019-GYHLW-004）。

**作者简介**：唐豪杰（1995-），男，硕士研究生，主要研究方向：计算理论、数据库理论；刘国华（1966-），男，博士，教授，博士生导师，主要研究方向：人工智能、大数据、关系数据库。

**通讯作者**：刘国华 Email：ghliu@dhu.edu.cn

**收稿日期**：2020-04-08

然语言描述中提取结构化的服装定制需求这一问题,做出论述。

## 1 问题分析与描述

服装制造产业专业性较强,普通消费者、设计师和生产厂家对同一件服装的关注点是不同的,设计师需要规范化和结构化的需求描述,而消费者的描述存在着数据信息不规范,无用信息过多等问题。智能化的服装个性化定制过程要节省沟通成本,需求提取的规范性和精确性十分重要。服装定制需求中属性众多,本文以服装风格为例,结合网络信息技术的智能化服装个性化定制流程描述如下:有定制需求的消费者使用语音或者文本描述对服装的个性化需求,利用信息技术对需求描述进行转化和预处理,对需求进行提取,将需求变为符合实际生产的结构化数据,如图 1 所示,得到的结构化需求可以为各种服装推荐系统提供数据,比如基于情感语义的推荐系统[4],还有基于层次分析法的推荐系统[5]。

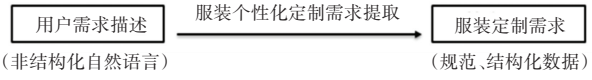


**图 1 服装个性化定制需求提取示意图**

**Fig. 1 Sketch map of demand extraction for personalized clothing customization**

### 1.1 问题分析

个性化需求提供形式是多样化的,除了文本描述还可能有语音和图像描述,目前已经有较成熟的技术可以把语音内容、图像内容转化为文本,提取出服装设计师设计服装时所需要的服装属性。提取用户需求数据,提供给服装个性化定制的下一阶段。由于个性化需求提供的形式多样化、内容复杂且具有歧义性等特点,给个性化需求提取工作带来了困难,从而制约了对个性化需求的理解。

服装的需求包含很多属性,如:风格、颜色、男女款式等,每种属性存在着多种属性值,如现代服装风格根据传播方式和分类依据的不同分为二十多类[6],每种分类下存在多个对应该风格分类的关键词,如:民族、传统文化、民俗、扎染、绣花、棉麻、唐装等关键词均对应着民族风格的服装[7]。这些指向服装风格的关键词是不统一的,非标准化的,而它们指向同一个服装分类属性的值(民族风格)是标准化的,表 1 是各个角色的语义关系,通过提取定制者需求描述中非标准化的关键词进行匹配,就可以得到标准化、结构化的服装需求属性。因此,根据一定规则和方法,匹配定制者描述中的关键词,是解决该问题的关键。

**表 1 各个角色的语义关系**

**Tab. 1 Semantic relationship of each role**

| 属性 | 属性值(标准) | 关键词(非标准) |
|---|---|---|
| 风格 | 民族风格 | 民族、传统文化、民俗、扎染、少数民族、棉麻、唐装…… |

### 1.2 问题描述

已知一个语言的集合:$Y = \{Y_{ij} \mid i,j = 1,2,\ldots,n\}$,需求文本经过预处理的词库集合:$C = \{C_i \mid i = 1,2,\ldots,m\}$,集合 $Y$ 中的 $i$ 代表服装定制需求的各项属性,集合 $Y$ 中的 $j$ 代表服装定制需求属性的值,$Y_{ij}$ 是该服装 $i$ 属性下属性值为 $j$ 的正则语言。求集合 $C$ 中属于集合 $Y$ 中语言的元素,并构成定制需求集合 $O = \{O_i \mid i = 1,2,\ldots,p\}$。词库集合 $C$ 中的元素中可能存在可以表达服装定制需求的关键词,也存在对于定制需求提取没有实际意义的词组,本文需要解决的问题就是通过集合 $Y$ 找出集合 $C$ 中有定制含义的关键词,再通过 $Y$ 集合中的正则语言的 $i$(定制需求属性) 和 $j$(对应属性值),即可得到结构化的服装定制需求数据。

## 2 相关技术

模式匹配是计算机科学中字符串的一种基本运算,定义如下:假设 $P$ 是给定的子串,$T$ 是待查找的字符串,要求从 $T$ 中找出与 $P$ 相同的子串,这个问题称为模式匹配问题。$P$ 称为关键词,也称为模式,$T$ 称为目标。如果 $T$ 中存在模式为 $P$ 的子串,则给出该子串在 $T$ 中的位置,称为匹配成功;否则匹配失败[8]。模式匹配在信息检索、入侵检测、DNA 测序等应用中扮演重要角色。

模式匹配算法有多种分类,根据一次匹配模式数目分为:单模式匹配和多模式匹配;根据匹配精确度可以分为:精确匹配和模糊匹配。在服装个性化定制需求提取问题中,需要提取多个标准化的需求属性值。

### 2.1 AC 多模式匹配算法

Aho-Corasick 算法是由 Alfred V. Aho 和 Margaret J.Corasick 提出的字符串搜索算法,用于在输入的一串字符串中匹配有限组“字典”中的子串。它与普通字符串匹配方法有所不同,AC 算法可以同时与所有字典串进行匹配。平均情况下,算法下具有近似于线性的时间复杂度,约为字符串的长度与匹配的数量的和。AC 算法基于有限状态自动机,在匹配前会对模式串的集合进行预处理,构建一棵树形有限状态自动机 FSA,根据被查找的目标字符串,即关键词,从树的根节点开始向叶子节点逐字匹配,在匹配过程中,如果发生失配,要根据失配跳转点跳转,如果找到匹配

的模式串则完成匹配。AC 算法在扫描文本时完全不需要回溯，如果只考虑匹配的过程，该算法的时间复杂度为 O(n)，也就是只跟待匹配文本的长度相关。

在预处理中会生成 3 个函数：goto(转移)函数、failure(失效)函数和 output(输出)函数。从树的初始状态出发，每次取文本串中的一个字符，根据 goto 函数或 failure 函数进入到下一个状态，当某个状态的 output 函数不空时，表明在该状态匹配到了相应的模式，输出值。例如：模式串集合：K = {he，she，his，hers}，根据 AC 算法构建的 goto 图如图 2 所示。

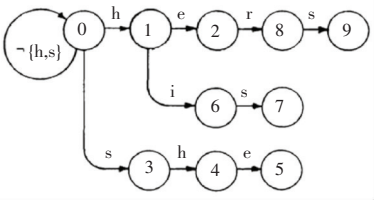


**图 2　根据 AC 算法构建的 goto 图**

**Fig. 2　Goto graph based on AC algorithm**

### 2.2　有限状态自动机

有限状态自动机(FSM "finite state machine" 或者 FSA "finite state automaton" )是为研究有限内存的计算过程和某些语言类而抽象出的一种计算模型，是具有离散输入和输出的系统的一种数学模型，可以表示为一个有向图。有限状态自动机具有有限个状态，不同的状态代表不同的意义，系统在任何一个状态下，从输入字符串中读入一个字符，根据当前状态和读入的这个字符转到新的状态[9]。有限状态自动机的形式化定义是一个五元组 $M = (Q, \Sigma, \delta, q_0, F)$，$Q$ 是状态的有限集合，$\forall q \in Q$，$q$ 称为 $M$ 的一个状态，$\Sigma$ 是输入符号的有限集合、$\delta$ 是转移函数 $Q \times \Sigma \rightarrow P(Q)$，$q_0$ 是初始状态或启动状态，$q_0 \in Q$，$F$ 是 $M$ 的接受状态的集合，$F \subseteq Q$，任给 $q \in F$，$q$ 称为 $M$ 的终止状态。

## 3　个性化需求提取流程

### 3.1　数据预处理

在定制者提供的定制需求描述中，因为难以制定填写规范，往往包含大量口语化的无用信息，如语气词、连词和标点等与服装定制无关的非实词描述。这些文本内容会影响需求提取的准确性和效率，对自然语言的描述进行预处理尤为重要。

使用自动机提取需求，需要从文本的各个位置作为起点进行识别，这种方法无效验证的次数多，即大部分起始位置都不能与正则表达式匹配。

由于服装行业专业性较强，目前没有相关的语义知识库参考。为了解决以上问题，选择使用北京理工大学张华平教授团队的中文分词系统 NLPIR 对原始数据进行中文分词和词性标注的处理。人工描述的语句“我想为下个月的校园草坪音乐节定制一款连衣裙，可以用红色或者其他亮色，裙摆长一些。”，经过分词处理后可以得到“我/rr 想/v 为/v 下/vf 个/q 月/n 的/ude1 校园/n 草坪/n 音乐节/n 定制/v 一/m 款/q 连衣裙/n，/wd 可以/v 用/p 红色/n 或者/c 其他/rzv 亮色/n，/wd 裙/ng 摆/v 长/a 一些/mq。/wj”的结果，去除影响识别的虚词，保留服装需求提取中需要识别的名词、形容词等，最后可以得到词库集合 $C$ {“校园”、“草坪”、“音乐节”、“连衣裙”、“红色”、“亮色”、“裙”、“长”}，对 $C$ 中的元素进行验证，可以提高提取的效率和精确性。

### 3.2　构建有限状态自动机

有限状态自动机可以分为确定有限自动机(DFA)和非确定有限自动机(NFA)。DFA 的起始状态是唯一的，NFA 的起始状态是一个集合，DFA 一个输入对应着一个状态转换，NFA 一个输入对应着一个状态集。基于 DFA 的匹配在速度上有优势，但对存储空间需求较大。

本文服装属性对应属性值下的一组关键词构成一个正则语言，根据语言集合 $Y$ 中的元素 $Y_{ij}$ 构建 DFA。不以关键词构建，以一组关键词构成的语言构建，一个服装需求属性对应一个较大的 DFA，如图 3 所示。可以减少检测次数，也更便于添加关键词，修改匹配模式。

### 3.3　个性化需求提取流程

本节将以服装风格为民族风格的需求为例，介绍需求提取流程。

集合 $Y$ 中的 $i$ 代表服装定制需求的各项属性，设 $i = 2$ 时代表服装风格，集合 $Y$ 中的 $j$ 代表服装定制需求属性的值，设 $j = 1$ 时为民族风格的服饰，则 $Y_{21}$ 是所有能够指向民族风格的关键词所构成的正则语言，设 $Y_{21}$ = {地方服饰，传统文化，民族，民族风，民俗，扎染}，由于自动机是根据标准化、结构化的服装属性构建的，所以能够确保抽取出的需求符合生产要求。构建确定有限状态自动机，如图 4 所示。

本方法中可以提取的合法需求为正则语言，需要解决的理论问题为定制者的需求文本经过分词处理后，所得到的词库中是否存在属于该语言的输入串。词库中的输入串进入自动机，从初始状态出发，每读入字符就根据当前状态和转移函数确定下一状态，直到到达接受状态，说明这台自动机可以识别该输入串，成功抽取出该自动机代表的服装需求属性。

(下转第 66 页)