时态文本数据流特征流行趋势模型及算法

孟志青 许微微
(浙江工业大学管理学院 杭州 310023)

摘 要 当今在电商和社交等平台上每天会产生大量的文本数据流。快速提取文本数据流的特征并将其用于发现一些事物的趋势变化来指导企业运营十分重要,比如服装企业必须尽可能快速而又准确地感知流行信息,服装特征的流行趋势对设计生产与经营起着至关重要的作用。以线上商品的文本数据流为研究对象,结合线上的销售文本实时数据流,定义了商品的时态文本数据流特征趋势模型,然后提出了一种文本数据流特征趋势发现的实时挖掘算法。将该算法应用到服装销售的文本描述以提取流行特征应用,可以获得有效的服装流行趋势,为企业制定生产计划、选择营销策略提供了决策支持。使用电商平台的真实销售数据进行实验,结果证明:该算法提取流行特征的准确率较高、速度较快,具有重要的理论与实际意义。

关键词 时态文本模型,文本数据流,特征快速提取,实时挖掘算法
中图法分类号 TP311 **文献标识码** A

Temporal Text Data Stream Feature Trend Model and Algorithm

MENG Zhi-qing XU Wei-wei
(School of Management, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Today, on the platform of e-commerce and social networking, there will be a lot of text data streams. It is very important to extract the characteristics of text data flow quickly to find some trend for guiding the operation of enterprises. For example, clothing enterprises must perceive popular information as quickly and accurately as possible. Fashion trends are of vital importance to the design, production and operation. Taken the text data flow of online goods as the research object, combining the online sales text real-time data flow, this paper defined a characteristic trend model of the temporal text data flow. Then, it proposed a real-time mining algorithm for text data stream feature trend finding. The algorithm was applied on the description of clothing sales text to extract popular feature applications. It can obtain an effective fashion trend and provide decision support for enterprises to formulate production plans and select marketing strategies. On the real sales data of the e-commerce platform, the experiment results prove that the algorithm has good accuracy and fast speed. Therefore, the proposed algorithm has important theoretical and practical significance.

Keywords Temporal text model, Text data stream, Feature extraction, Real-time mining algorithm

1 引言

每天在许多电商平台上产生了大量的文本数据流,这些数据包含了许多对企业有价值的信息,例如淘宝上服装销售时产生的大量的服装描述文本数据流,淘宝的服装种类上亿,其描述文本是一个大数据,可用于发现服装流行趋势,对于指导服装企业的设计、生产、销售具有重要的作用。从服装文本描述数据流中发现服装流行趋势,显示是一个非常困难的课题。服装作为时尚流行的代号,其流行趋势是时尚界的风向标,极大程度地引领着消费者的审美倾向。服装的流行趋势是指某个时期内服装流行的概念、特征和样式,是特定时间段内对人们物质生活和精神世界的反映。一件服装单品,从开发到售出,在整个过程中流行趋势都起着引导的作用,一件符合流行趋势的服装才能被市场和消费者接受[1]。流行趋势的重要性是显而易见的,将其带入到实际的产品开发与销售中去,必定可以产生巨大的经济效应,带动整体产品的推广,增加服装企业的自主研发能力和市场竞争力。服装的流行趋势是动态变化的,但并不是没有规律可循,流行趋势具有可预测性。流行趋势在很大程度上是由市场决定的,即由消费者引爆潮流。据悉,专业时尚研究机构的研究人员分布在全球各地,研究当地消费者们的喜好,再将各地的流行风潮汇总并加以提炼,最终确定下一季的流行趋势。

近年来,服装行业的各品牌、各高校甚至消费者对服装流行趋势的关注度越来越高,服装流行趋势预测的系统化、科学化研究俨然成为了我国培养和发展服装设计专业人才的关键问题[2]。当下的服装流行趋势预测,无论是在趋势形成和阐释方面,还是在信息的收集和提炼分析方面,在很大程度上完全依赖于人的判断力,即直观预测法。在服装设计领域,设计师及其他专家需要靠自己的时尚敏锐度和长年的经验来预测流行趋势,从而引领时尚。但是这种方法具有较高的主观性,

本文受浙江省自然科学基金项目(LY15G010007)资助。

孟志青(1962—),男,博士,教授,主要研究方向为数据挖掘、最优决策理论,E-mail:mengzhiqing@zjut.edu.cn(通信作者);**许微微**(1992—),女,硕士生,主要研究方向为数据挖掘、自然语言处理。

正确率得不到保证,且对普通消费者和小企业不友好,可操作性差,并不适合大范围的应用,从而导致一些企业只能被动地跟随时尚的脚步。除了直观预测法之外,部分学者致力于市场统计预测法和数学模型预测法的研究。常丽霞等[3]探讨了时间序列的长短对灰色 GM(1,1)模型预测服装流行色时色相预测性能的影响,研究表明,6 年长度的时间序列建立的灰色模型的性能最佳,精度高且绝对误差低。Yu 等[4]用实证的方式系统地比较了自回归移动平均模型(ARIMA)、灰色模型(GM)、人工神经网络(ANN)在流行色预测中的表现,实验证明人工神经网络(ANN)对流行色预测的精度优于其他模型。目前很多关于服装流行预测的方法都是利用历史数据进行的,而不是利用实时数据进行预测,这些数据多多少少地存在滞后的问题,且数据量也有限,这极大地限制了预测精度的提高,预测结果滞后性较强,对于掌握服装企业流行趋势的指导作用较差。

随着互联网的发展,我国早已进入大数据时代,人们获取信息的方式也由传统渠道向网络渠道转移。在大数据背景下,不论是流行趋势的形成还是关于流行趋势的信息收集等都会受到大数据技术的影响而发生改变,拥抱大数据,进行服装流行趋势的预测方式的转型势在必行。陈云依澜[5]将服装流行趋势的预测与大数据中的云计算技术相结合,提出了基于云计算的服装流行趋势预测机制,讨论了利用大数据云计算技术代替传统预测机制的可能性,得出了基于云计算的预测机制优于大部分的传统预测机制的结论。Liu 等[6]利用图片识别相关技术,生成了一个多功能的服装图片分类器,可以从图片中直接识别并提取流行元素,并创建了一个新的流行图片数据库。Nogueira 等[7]从社交网站 Facebook 和 Instagram 的信息中捕捉新的服装流行趋势,为服装图片进行更加精准的自动注解添加。

面对线上线下日益激烈的竞争形式,服装企业必须在尽可能短的时间内获知产品的流行趋势,随着"新零售"时代的到来,纯电商的时代即将被打破,线上线下将深度融合,各大电商平台产生的大量交易信息,对服装的流行趋势预测有着巨大的价值。这些真实的销售信息数据流反映了当前时期消费者内心对服装产品的外在需求,从这些数据中可以挖掘出消费者当前的关注重点和审美喜好,是流行特征最直观的反映。以国内最大的电子交易平台——天猫为例,一个商品的详情页面中包含着大量的信息,例如店铺评分、宝贝描述等,其中商品标题、关键词、属性描述、月销量、累积评价、收藏人气等重要信息是消费者关注的焦点。那么,是否可以利用文本挖掘及其他相关的数据分析技术,从这些信息中提取流行特征,从而预测流行趋势呢? 电商平台上的销售信息数据量大,易获取,具有时效性,并且与消费者密切相关,非常适合作为流行趋势预测的数据源。

本文通过文本分析,将服装线上的销售信息与其对应的商品描述相结合,构成商品时态文本数据流,提出了一种基于文本数据流运算的服装流行特征算法,为服装流行趋势的预测提供了一种全新的思路。该算法易操作且适用性广。通过对大量电商数据的实验分析,证明该方法的正确率较高。本文的研究有助于服装原材料供应商、成衣商以及零售商在内的服装行业从业者进行更为科学的产品研发设计和推广销售,减少了不必要的库存堆积,加快了整条供应链的协调运作,推动了整个服装行业的前进。

2 文本数据流

2.1 文本表示

在对文本数据流进行挖掘之前,必须对文本数据进行一定的预处理,包括中分文词以及停用词过滤。

词是指语言中有意义的能单独使用或用来造句的最小单位[8],中文分词是一切中文文本分析的基础,做好中文分词之后才能进行后续的研究。为了去除无意义的词语或标点,初步分词之后需对数据进行停用词过滤,本文的研究对象是文本数据流特征趋势发现,下面以服装数据文本为对象进行讨论,服装文本具有特殊性,不能直接套用现有的停用词库,因此在数据采集的同时需制作停用词表。

在对文本进行预处理后,需要将其转化为适用于计算机理解的数据化文本特征表示,这里采用 Salton 等[9]提出的向量空间模型(Vector Space Model,VSM)进行商品文本表示。一件商品 P 可以用一系列的词项 C_i 来表示,$p=(C_1,C_2,C_3,\cdots,C_n)$,其中 C_i 表示商品 P 中所包含的某一词项,$1\leqslant i\leqslant n$,n 为该商品的词条总数。在进行文本的特征提取时,每个词项 C_i 对应一个特征权重 W_i,采用 TF-IDF 算法计算这个权值,TF-IDF 理论的核心思想是:一个词在一篇文档中出现的次数越多,即词频 TF(Term Frequency)越大,则该词区分这篇文档的能力就就越高;在所有文档中包含该词的文档数越多,则该词区分某一文档的能力越弱[10]。基于此,我们可以做出如下定义。

定义 1(商品词项词频(Pruduct Term Frequency,PTF))指一个词项在对应的商品文本集中出现的频次,一个商品记一次。那么在商品文本集 d 中,词项 C_i 的词频可记为 $PTF(C_i,d)$。

定义 2(词项逆商品率(Inverse Pruduct Frequency,IPF))其描述的是主题关键词在商品文本中的区分度,某词项在所有商品中出现的次数越少,则在文档中的区分就度越高。假设对应的囊括了 N 个商品的语料库数据集为 D,词项 C_i 的 IPF 的计算公式如下:

$$IPF(C_i,D)=\log\frac{N}{PTF(C_i,D)+1} \tag{1}$$

其中,$PTF(C_i,D)$指的是词项 C_i 在语料库 D 中的商品词项词频。综上,若要计算某类商品集 d 中词项的权重,d 的商品数应选择小于语料库 D 的商品数 N,完整的计算公式为:

$$\begin{aligned}W_i&=PTF(C_i,d)*IPF(C_i,D)\\&=PTF(C_i,d)*\log\frac{N}{PTF(C_i,D)+1}\end{aligned} \tag{2}$$

2.2 时态文本数据流

Henzinger 等[11]于 1998 年提出了数据流的概念,并将其确立为一种新的数据模型来进行处理,从此数据流的研究便成为热点,每年在各大顶级数据挖掘领域的会议上(如 SIGMOD,SIGKDD,ICDM 等)都有相关文章出现。在日常生活中,Web 日志数据、电信呼叫数据、股票交易数据、电商零售交易数据等就是数据流的典型例子。数据流是实时的、连续的、有序的数据序列,元素的出现顺序、输入速率均不可控。数据流与传统的关系模型的不同点在于,数据项是在线到达的,其数据规模在理论上是无限多个,若不将处理过的数据存于内存中,则不可二次访问[12]。为方便后文算法的阐述,给

出下述简单定义。

定义 3（文本数据流，Data Stream）　是由文本数据项构成的无限集合，其中 $x_n=\{a_1^n,a_2^n,\cdots,a_d^n\}$，$d$ 代表了文本样本数据的特征维度。文本数据流中的数据项只能依次按下标有序地到达，基于此，一般的文本数据流可以形式化为 $DS=\{\cdots,x_{n-1},x_n,x_{n+1},\cdots\}$。

对于线上文本数据流来说，数据项具有时间属性，并且该属性对相关领域的研究来说十分重要。由于文本数据流的实时性，采用时态数据流进行分析时，不同的情况可能需要考虑不同的时间粒度[13]。在时态文本模型中，文本数据流中的每个数据项都代表了一个独立的特征信息。

定义 4[22]　一个时态型 μ 是绝对时刻到绝对时间的一个映射 $\mu:R\to 2^R$，设对于任意绝对时刻 $t\in R$，有：1)（非空性）$t\in\mu(t)$；2)（单调性）若 $t_1<t_2$，$\mu(t_1)\bigcap\mu(t_2)=\varnothing$，对任意的 $t'\in\mu(t_1)$ 和 $t''\in\mu(t_2)$，有 $t'<t''$，记 $\mu(t_1)<\mu(t_2)$；3)（同一性）对任意 $t'\in\mu(t)$，有 $\mu(t')=\mu(t)$；4)（有界性）对任意 $t'\in\mu(t)$，$|t'|<+\infty$，则 μ 称为一个时态型，$\mu(t)$ 称为时态型 μ 的时态因子（Temporal Factor）。

时态型是对时间数轴 R 的一个划分，每个时态因子是一个区间（一般为半开半闭或开或闭），我们可以用秒、分、小时、日、周、月和年来划分时间数轴 R，因此它们又都是时态型。若 $\mu(t)(\forall t\in R)$ 为单点集，则称 μ 为原子时态型，若 $\mu(t)(\forall t\in R)$ 为非单点集，则称 μ 为非原子时态型。

性质 1　对任何 $t\in R$，$\mu(t)$ 存在上确界 $\sup\mu(t)$ 和下确界 $\inf\mu(t)$。

由时态型的单调性和有界性可知性质 1 成立。将绝对时间长度定义为：

$$Length(\mu(t))=\sup\mu(t)-\inf\mu(t)$$

定义 5[22]　设 μ,ν 是两个时态型，如果 $Length(\nu(t))<Length(\mu(t))(\forall t\in R)$，则称时态型 ν 小于时态型 μ。若时态型 ν 小于时态型 μ，且对 ν 的任何时态因子 $\nu(t)$，存在唯一 $\mu(t')$ 使得 $\nu(t)\subset\mu(t')$，则称 ν 是 μ 的一个基时态型。

性质 2[22]　设 μ,ν 是两个时态型，ν 是 μ 的一个基时态型，则对于 μ 的任何一个时态因子 $\mu(t)$，一定存在 n 个 $t_1,t_2,\cdots,t_n$，使得 $\mu(t)=\bigcup_{i=1}^{n}v(t_i)$，其中 $t_1<t_2<\cdots<t_n$，$v(t_i)\bigcap v(t_j)=\varnothing$，$i\neq j$，$i,j=1,2,\cdots,n$。

定义 6[22]　如果一个非原子时态型 μ 具有等长的绝对时间长度，则称 μ 为一个时间粒度。

定义 7　设时间粒度 ν 是时间粒度 μ 的一个基时态型，定义一个时态顺序流 $v(t_i)$，$i=1,2,\cdots,n$，$t_1<t_2<\cdots<t_n$，其中时态因子 $\mu(t)=\bigcup_{i=1}^{n}v(t_i)$，时态数据流（Temporal Data Stream）是一个多维数据集 $X(\mu)=\{X(v(t_i))|i=1,2,\cdots,n\}$，其中文本数据流 $DS=\{X(v(t_i))=(a_1(v(t_i)),a_2(v(t_i)),\cdots,a_d(v(t_i)))\}$ 为时刻的样本向量，d 是文本特征维度，$a_k(v(t_i))$ 表示在时间粒度 $v(t_i)$ 上的第 k 个特征。

以天为基时间粒度 ν，以月为 μ，考虑服装流行特征数据流，取描述服装文本的特征：颜色、款式、花样、板式和风格等，这些特征具有很强的季节性，不论是企业家还是消费者，最关心的是最近一段时间的数据信息，且内存空间有限，无法全部保存所有数据。Datar 等提出的滑动窗口数据模型很好地兼顾了内存的消耗问题以及数据流实时、快速的特点，被广泛应用于时序型数据流的分析和管理中。

定义 8　数据流窗口模型（Data Stream Window Model）　在数据流上设置一个可滑动的窗口，实时更新和维护当前处理的数据，假设滑动窗口的大小为 W，则在时间点 t_n，查询范围为 $\{X_{\max(0,t_n-W+1)},\cdots,X_{t_n}\}$，除此之外，在时间点 $\max(0,t_n-W+1)$ 之前的数据将不予保存。

滑动窗口模型如图 1 所示，这里假设数据流自右向左流入。$X(v(t_i))$ 包含在 t_i 时刻窗口中服装数据出现得越早，它对当前时刻的流行趋势预测的贡献无疑会越弱，此时应在数据前添加衰减因子。假设在 t_m 时刻流入的数据项为 $X(v(t_m))$，为其添加权重为 $(1-\gamma)^{n-m}$，γ 为 (0,1) 内较小的一个常数，通常赋值为 10^{-6} 或 10^{-9}，这样离当前时刻越远的数据的权重越小，而当天的数据的权重则为 1。

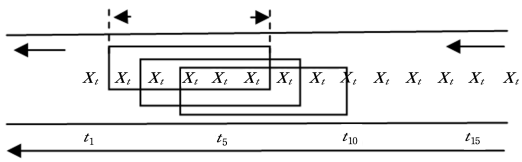

图 1　滑动窗口模型

2.3　时态文本数据流模型

本文在 2.1 节和 2.2 节中已经阐述了文本表示数据流的相关概念，线上的服装商品销售信息的特殊点在于，除了时间、商品 ID 等基本属性之外，还带有由消费者或潜在消费者产生的销售信息。众所周知，线上消费者在购物或浏览商品时，常常直接或间接地对销量、收藏人气、评论数等数据造成影响，我们将这几样数据统称为销售信息。对一件商品 P 来说，它所附带的销售信息 S 可以直接附加在对应的文本对象 A 上。

定义 9　商品时态文本数据流（Pruduct Temporal Text Stream）　对于属于商品 P 的文本对象 A，它的时态文本数据流模型表示为 $\{P,A,S,X(\mu),T\}$，$T=(t_1,t_2,t_3,\cdots)$ 是时间属性的集合，交易数据中通常包含多个粒度的世界属性，比如年、月、日、小时和分等，设 v 是时间粒度 μ 的基时态型 $T:=\mu$ 表示一个时间特征区间。$S=(a_1,a_2,\cdots,a_d)$ 和 $X(v(t_i))=(a_1(v(t_i)),a_2(v(t_i)),\cdots,a_d(v(t_i)))^{\mathrm{T}}$ 是商品 P 在时间 T 的各种特征信息，如颜色、款式、板式、品牌、位置等，特征刻画值为销量 $Sell$、收藏人气 $Favor$ 和评论人数 $Rate$ 等。文本的对象可以是一段话、一个句子、一个词项或一个字，特征集合是文本对象提取产生的。

结合本文的研究要求，服装文本特征对象定为词项，如设一种词特征项：$a_j(v(t_i))=(P,Sell,Favor,Rate,t_i,\mu)$，$P$ 为对应的商品唯一标识，通常为 SKU 或 ID。

性质 3　表达式中各项都非负且有界，$0\leqslant a_j(v(t_i))<+\infty$。

性质 4　各数值具有可加性，当遇上另一件商品 P_j，使得 $E(item,P_j,v(t))=1$ 成立，那么原特征词项带有的销售信息可直接与商品 P_j 所带的销售信息 $S(P_j)$ 相加。

性质 5　具有时效性，只有特征项 $item$ 所界定的时态因子 $v(t_i)$ 与商品的销售信息的时态因子 $v(t_i)$ 一致时才能继承商品的销售信息。

3　时态文本数据流在线特征算法

本文提出时态文本（服装）数据流（流行）特征提取算法，

其主要含数据采集、数据预处理以及特征提取 3 部分。

3.1 流行趋势的相对稳定性

服装的流行趋势虽然是动态变化的，但是并不是没有规律可循，具有可预测性[14-22]。此外，时尚流行的特征还有的区域性和滞后性，有的消费者已经开始追赶新的风潮，而有的消费者还没有意识到流行趋势的变化，这使得服装的流行趋势具有相对稳定性。即在较长的一段时间周期内，流行趋势是变化的，而在较短的一个周期内，比如一个月或一个季度，流行趋势几乎没有太大的变化，是相对稳定的。为了证明这一点，我们收集了女装 T 恤类目下 2017 年 3 月 1 日至 2017 年 3 月 31 日淘宝热搜词的前 100 位，淘宝相对其他电商平台来说，服装类商品的交易量最大，它的热搜词排名一定程度地反映了消费者关注的焦点，从侧面推测出流行趋势的变化幅度。以 3 月 1 日的 100 个热搜词为基准，记录每天的变化情况，结果如图 2 所示。由图 2 可以看出，即使将 3 月 31 日的数据与 3 月 1 日的数据进行对比，其每日热搜词的变化率也不到 20%。这说明流行趋势在一定时间内是相对稳定的，我们利用当前的数据提取出的流行特征至少在近期内仍具有时效性。

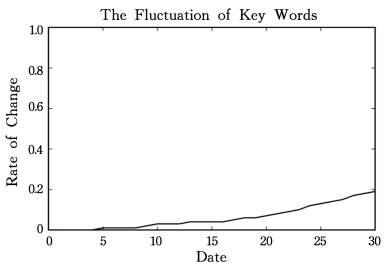

图 2 淘宝 T 恤热搜词的变化率

3.2 数据准备

为了为后续工作做充分准备，在数据流采集阶段需要进行 3 方面的工作：首先选取服装类目下某一品类的商品，一方面按需采集该品类下商品的销售信息，通过对各大电商的商品详细页面的初步分析，可以获得累计销售量、累计收藏人气以及累计评论数等信息；另一方面采集该商品对应的文本信息，同样地，经过分析可以采集到商品标题、商品属性描述及其他一些细节描述。另外，在这两部分工作之外，还必须收集语料库的数据。本文将抓取的大量电商服装商品的文本信息作为语料库，值得注意的是制作语料库的文本描述最好来自于抓取销售信息品类的上一级类目。比如，若一开始选择“牛仔裤”这一类目进行分析，那么语料库的信息最好来自于“裤子”或“下装”这一类目。这是由于制作语料库需要的商品数量较为庞大，若类目太细则数量不足且范围太窄，从而影响逆商品频率的有效性。

3.3 数据预处理

首先需要删除不符合标准的数据，包括文本描述乱码的商品、销售信息抓取不全的商品，接着采用开源项目 Jieba 分词对文本信息进行处理，并在数据采集的同时，阅读相关文献自行制作服装行业分词词库，补充词条 2000 多条，停用词 100 多个。将其补充到 Jieba 分词的现有词典中，以提高分词准确率。整个分词操作分为 3 步。

1)初步分词：将文本信息以单个商品为单位进行分词，词与词之间以空格隔开，每个商品的文本描述之间以换行符隔开；

2)去重：去除单个商品描述中的重复词语，保证商品与词的对应性；

3)去停用词：去除对研究无意义的词汇和符号，例如“加入购物车”“的”“立即购买”等。

中文分词之后便可得到商品的文本表示以及初始商品词频等信息，为提高在线运算时的速率，同时减少 CPU 和内存的消耗，必须在离线时利用采集到的数据先行进行计算。利用滑动窗口数据模型前，需利用离线信息设立 3 个概要表：表 V、表 P 和表 S，以保存与利用原有窗口的部分信息。表 S 保存当前窗口内的商品销售数据，包括销量、收藏数、评论数等，并按照到达的时间顺序排列，商品用 ID 进行唯一标识。表 P 保存的是预处理后所有商品的文本信息，商品 ID 与词项呈一对多的关系，即将商品的文本表示以另一种形式提前存入内存。表 V 保存了已知词项的相关信息，包括商品词频(PTF)、逆商品率(IPF)、窗口时间内的累积销量、累积收藏数、累积评论数以及最终计算得到的流行指数得分，以 TermID 作唯一标识并将每个词项按其流行指数得分值由大到小排列。在离线阶段，我们可以得到初始商品词频(PTF)和逆商品率(IPF)，在无新商品上架的情况下，这两个值不会发生变化，先行计算以降低工作量。3 个表的逻辑结构图 3 所示。

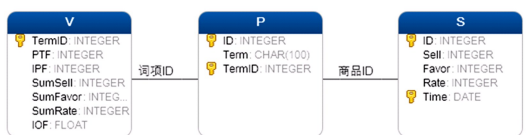

图 3 概要表结构图

3.4 时态文本数据流特征趋势模型及算法

本文根据时态文本数据流模型提出了一种提取特征趋势的算法，为了描述更加具体清晰，下面对服装文本数据流行特征提取趋势发现问题进行描述，提出权重流行指数 IOF(Index Of Fashion)的计算方法。在现实生活中，线上服装的交易数据是一种典型的数据流数据，选取有效的信息使得输入形式为(这里仅取特征描述的 3 个数值，也可取更多特征刻画描述)：

$$a_j(v(t_i))=(P.Sell,P.Favor,P.Rate;v(t_i))$$

其表示商品对象 P 在时态因子 $v(t_i)$ 时商品的针对特征 a_j 的销售量、收藏数和评论数(点击率、成交率、浏览量等的处理方式类似，此处略)。得到一个文本数据流特征向量序列：

$$X(v(t_i))=(a_1(v(t_i)),a_2(v(t_i)),\cdots,a_d(v(t_i)))^{\mathrm{T}},$$

$$i=1,2,\cdots$$

由于 d 非常大，时态文本数据流$\{X(v(t_i))\}$是一个 $d\times 3$ 高维矩阵序列集合，那么针对这个集合可以进行特征时间序列回归分析、特征聚类分析、特征相关分析、特征周期提取、特征趋势发现等。时态文本数据流特征趋势模型的描述如下：

(1)计算 $a_j(v(t_i))$对应值$(P.Sell,P.Favor,P.Rate)$的权重$(\omega_1,\omega_2,\omega_3)$，然后对其进行主成分分析。通过对$\{X(v(t_i))\}$进行相关分析来获得特征相关图，如品牌、位置、款式与流行特征等。

2)发现流行特征集合，计算流行特征强度：

$$E(a_j,T)=\sum_{i=1}^{T}d_{ji}^{+}+\sum_{i=1}^{T}d_{ji}^{-}$$

其中，$\varphi_j(v(t_i))+d_{ji}^{-}-d_{ji}^{+}=\bar{a}_j$ 是偏离方程，T 是在一个时间区间粒度 v 上的时间窗口宽度，其中取特征 a_j 的时间变换：

对应的 $\varphi_j(v(t_i))=\omega_1 P.Sell+\omega_2 P.Favor+\omega_3 P.Rate$，设 $\bar{a}=(\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_d)^T$ 表示最低流行阈值，d_{ji}^- 和 d_{ji}^+ 分别是特征 a_j 的负偏差和正偏差向量，刻画了实际特征与最低流行阈值的差距，$E(a_j,T)$越大，特征 a_j 就越流行，记时态因子 $v(t_i)$ 的最大流行强度值为：

$$S(a_j(v(t_i)))=\max\{d_{ji}^-+d_{ji}^+ \mid j=1,2,\cdots,T\}。$$

3) 建立流行特征时间趋势综合得分排序。

4)建立流行特征趋势回归方程，取特征 a_j 时间轴变换对应的 $v(t_i)$，对应的流行强度值为：

$$y_i^j=\frac{d_{ji}^-+d_{ji}^+}{S(a_j(v(t_i)))},j=1,2,\cdots,T$$

特征值 a_j 一个流行趋势回归公式为 $f(a_j(v(t_i)))=w\cdot v(t_i)+b$，判定特征 a_j 随时态变化的流行趋势预测，模型如下：

$$\mathrm{FFSVM}(a_j)\ \min_{w,b,\xi_i,\xi_i^*}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$

$$\text{s.t. } ((w\cdot v(t_i))+b)-y_i^j\leqslant\varepsilon+\xi_i,i=1,2,\cdots,T$$

$$y_i^j-(w\cdot v(t_i)+b)\leqslant\varepsilon+\xi_i^*,i=1,2,\cdots,T$$

$$\xi_i,\xi_i^*>0,i=1,2,\cdots,T$$

上述模型为二次规划。选取不同的时态型 v 和时间窗口 T，得到不同的流行集合 $S(T)$从而得到不同流行趋势方程。

每个词项初始商品词频(PTF)和逆商品率(IPF)在准备期间利用离线数据进行计算，在算法运行过程中，若非遇到新上线的商品则不需要变动。这样节省了大量的运行时间和内存消耗。时态文本数据流的流行特征提取算法如算法 1 所示。

算法 1 （服装）流行特征趋势算法

输入：时态文本数据流{X(v(t_i))}，窗口大小 W，特征提取数量 N，衰减因子 r，时态型 v，时间窗口 T

输出：N 个流行特征

1. While(新元组 X(v(t_i))到达)：
2. if ID is new：
3. Get text(ID_i)
4. Segment(text(ID_i))
5. insert value to table P
6. update table V
7. elseif ID_i is exsit in table p：
8. delete X_{t-w+2}^i
9. $P.Sell(Item_j\rightarrow a_j)=\sum_{i=0}^{i=Max(i)}\sum_{T=t-w+1}^{t}(1-\gamma)^{t-T}Sell_T(ID_i,Item_j\rightarrow a_j)$，$P.Rate(Item_j\rightarrow a_j)=\sum_{i=0}^{i=Max(i)}\sum_{T=t-w+1}^{t}(1-\gamma)^{t-T}P.Rate_T(ID_i,Item_j\rightarrow a_j)$，$P.Favor(Item_j\rightarrow a_j)=\sum_{i=0}^{i=Max(i)}\sum_{T=t-w+1}^{t}(1-\gamma)^{t-T}Favor_T(ID_i,Item_j\rightarrow a_j)$//所有包含该词项的商品在窗口内的汇总
10. $(\omega_1,\omega_2,\omega_3)=PCA(P.Sell,P.Rate,P.Favor)$ //权值系数做主成分分析
11. 计算流行强度 $E(a_j,T)=\sum_{i=1}^{T}d_{ji}^+ +\sum_{i=1}^{T}d_{ji}^-$
12. 获得 $S(a_j(v(t_i)))=\max\{d_{ji}^-+d_{ji}^+ \mid j=1,2,\cdots,T\}$
13. $IOF(Item_j)=IDF(Item_j)*S\_TF(Item_j)$
14. Sort $IOF(Item_j)$ //将词项按流行指数得分排序
15. 解 $FFSVM(a_j)$，得分类特征预测方程

如果除去求解 FFSVM(a_j)的时间，假设该品类有 m 个初始特征项，目标品类的商品数量为 n，则算法复杂度为 O($n+m*n$)。

在算法的第 13 步中将文本分析中词频的概念与各销售信息相结合，从而得到一个流行特征集合，运算前各变量需进行标准化。在电商平台，判断一个产品是否受欢迎，最佳的衡量指标是其销售情况。我们咨询了大量的淘宝卖家后发现，一个商品的标题及文字描述并不是随便制定的，在描述一件服饰基本特征的前提下，店家会参考服装资讯和网络流行信息，筛选合适的字眼，尽量使自己的商品在描述上符合流行趋势。而且，若极大比例的服装商品都包含某一特征，则进一步印证了这个特征的流行性，因此本文决定将词频与销售信息以一比一的权重进行计算。而对于销售指标，除了销售量之外，收藏人气和评论数量也必须加以考虑。收藏人气表明了消费者的购买意愿，从侧面显示出了该商品受关注的程度，而评论是消费者对购买商品的反馈，在一定程度上影响着销售量[15]。但显而易见的是，这 3 个变量必定是高度相关的，任选 150000 条数据进行相关性检验，结果如表 1 所列。

表 1　销售量、收藏数和评论数的相关矩阵

		sellCount	rateCount	favorCount
sellCount	Pearson	1	0.426**	0.403**
	P		0	0
rateCount	Pearson	0.426**	1	0.445**
	P	0		0
favorCount	Pearson	0.403**	0.445**	1
	P	0	0	
	N	150000	150000	150000

注：** 表示在 0.01 水平(双侧)上显著相关

由表 1 可知，3 个数据量之间存在着显著的正相关关系，因此不能简单地将 3 个变量以相同的权重处理，为了能准确地表示销售指标，在运算时进一步对数据进行主成分分析，以得到权重系数 $\omega_1,\omega_2,\omega_2$，详细过程不再进行赘述。

4　实验结果与分析

4.1　实验数据

本文选用天猫作为数据采集平台，用 python 编写爬虫代码，自 2016 年 11 月 1 日起开始收集销售信息，持续收集了 T 恤、连衣裙、针织衫、小西装等多个类目的页面信息，并将其保存至本地数据库备用。本文实验选取 T 恤、针织衫、小西装、衬衫 4 个品类进行分析，每个类目的商品数量在 12 万到 17 万之间，其中大部分销售信息没有变化。中文分词后的全部商品对应的不重复词条为 1 万到 2 万之间，单个商品对应了 10 到 20 个有意义词项。同时收集女装上衣各类目共 60 万商品的文本描述，对其进行处理后将其制作成为语料库备用。

实验在 CPU 为 PIV3.4 GHz、内存为 2 GB、操作系统为 WIN2007 的 PC 机上运行，实验程序均用 Python 实现。

4.2　实验结果及分析

实验将数据按天为时间粒度分块，数据窗口大小为 30 天，即一个月，这个时间周期对于流行趋势的预测来说较为合理。为了统一，每天均选取 T 恤、针织衫、小西装 3 个品类中销售信息不为 0 的前 50000 条数据项进行实验，商品语料库由 60 万女装上衣的商品文本经处理后组成。得到其中一个特征结果：

S＝{薄款/瑜伽/运动服/百搭/圆领/纯色/基础/打底衫/修身/显瘦/莫代尔/广东广州/均码/西瓜红/黄色/白色/肉色/玫红/彩蓝/荧光绿/黑色/大红/聚氨酯/弹性纤维/氨纶/贴布/长袖/电商/拼色/文化/经典/通勤/秋季/…}。

选用 2016 年 11 月 30 日、2016 年 12 月 30 日、2017 年 1

月 30 日、2017 年 2 月 30 日、2017 年 3 月 30 日为当前时间节点 t，为了减小数据计算规模，选取 1 天作为时态因子进行文本数据流统计汇总，同时将当时由业内专家、服装企业主提供的当月流行特征与当时的淘宝热搜词相结合，整理出服装流行特征的基准关键词并由专家对其进行排序，以此为标准验证每次流行特征提取的准确率、特征提取数量、衰减因子，最终结果取 5 次实验的平均值。特征提取对比结果如图 4 所示，由图 4 可知，本文通过时态文本数据流特征提取算法的计算结果与人工判定值具有较好的相似度，特别是有非常好的趋势拟合效果。

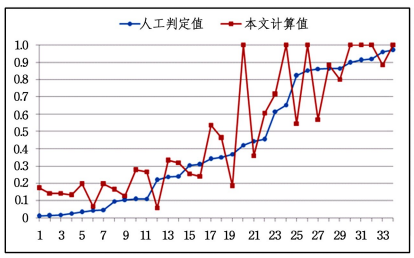

图 4 特征提取比对结果

针对 4 类产品的流行特征趋势算法的准确率的实验结果如图 5 所示。由图 5 可以看出，对各品类来说，流行特征提取的准确率较好，准确率大多数超过 70%，大致能反映当前的流行趋势。总体来说，随着 N 取值的不同，准确率有所变化，效果最好的 N 值在 7 到 10 之间，之后便略微下降并逐渐趋于稳定。这可能是由于有些商品的部分描述是以图片的形式展现在页面上的，从而无法准确地识别一些文本数据，影响了本文对一些流行特征的准确率的提高。

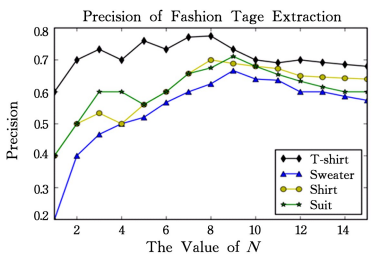

图 5 流行特征提取实验

结束语 随着服装产业全球化，如何准确又快速地把握流行信息，是服装企业提高自主研发能力、制定精准营销策略、增强市场竞争能力的关键。本文创造性地将线上销售数据用于流行趋势的预测，将销量、收藏数、评论数等在线销售数据流与商品标题、关键词、详细描述等文本信息相结合，提取服装流行特征，从而了解线上服装的流行趋势。本文的研究可以使普通消费者迅速了解当前的时尚风向，购买自己喜欢又不失潮流的商品；对零售商来说，有助于这些信息来源可以帮助他们制定更合适的营销策略，可以为制定下一步的库存补货策略提供决策支持；对服装原料供应商来说，有助于了解市场行情，也对调整生产作业计划以及下一步的开发打样有较大帮助，从而加快整个供应链的协调运作。

本文的流行特征提取方法可以类推到其他类型的商品中，后续也可以考虑以并行计算的方式运行，运用神经网络技术提取流行生命周期和流行特征分布。

参 考 文 献

[1] RU H Y. Application research of fashion trend in product design [D]. Shanghai: Donghua University, 2015.

[2] TENG Z Y. Exploration of China's teaching in prediction of fashion vogue[J]. Journal of Textile Research, 2011(5): 112-117.

[3] CHANG L X, et al. Hue prediction on Intercolor for women's spring/summer using GM(1,1) models[J]. Journal of Textile Research, 2015, 36(4): 128-133.

[4] YU Y, HUI C L, CHOI T M. An empirical study of intelligent expert systems on forecasting of fashion color trend[J]. Expert Systems with Applications, 2012, 39(4): 4383-4389.

[5] CHEN Y Y L. Study of cloud comfuting based clothing fashion trend forecasting mechanism[D]. Shanghai: Donghua University, 2016.

[6] LIU S, et al. Fashion Parsing With Weak Color-Category Labels [J]. IEEE Transactions on Multimedia, 2014, 16(1): 253-265.

[7] NOGUEIRA K, VELOSO A A, SANTOS J A D. Pointwise and pairwise clothing annotation: combining features from social media[J]. Multimedia Tools and Applications, 2016, 75(7): 4083-4113.

[8] HUANG C N, ZHAO H. Chinese word segmentation: a decade review[J]. Journal Of Chinese Information Processing, 2007(3): 8-19.

[9] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the Acm, 1975, 18(11): 613-620.

[10] ERRA U, SENATORE S, MINNEUA F, et al. Approximate TF-IDF based on topic extraction from massive message stream using the GPU[J]. Information Sciences, 2015, 292: 143-161.

[11] HENZINGER M R, RAGHAVAN P, RAJAGOPALAN S. Computing on Data Streams[OL]. http://citeseerx.ist.psu.edu/viewdoc/download? doi = 10.1.1.63.7109&rep = rep1&type=pdf.

[12] BABCOCK B, BABU S, DATAR M, et al. Models and issues in data stream systems[C]// ACM Sigact-Sigmod-Sigart Symposium on Principles of Database Systems. Madison, Wisconsin, USA, 2002: 3-5.

[13] 孟志青，蒋敏，姜华. 时态数据挖掘算法[M]. 北京：经济科学出版社，2014.

[14] JONES K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972(1): 11-21.

[15] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24(5): 513-523.

[16] YANG M, QI W, YAN X B, et al. A study on the effectiveness of online commodity reviews[J]. Journal Of Management Sciences In China, 2012, 15(5): 65-75.

[17] 汝海洋. 流行趋势在产品设计上的运用研究[D]. 上海：东华大学，2015.

[18] 滕兆媛. 基于实践的中国服装流行趋势预测教育探索[J]. 纺织学报，2011(5): 112-117.

[19] 常丽霞. 灰色 GM(1,1)模型在国际春夏女装流行色色相预测中的应用[J]. 纺织学报，2015(4): 128-133.

[20] 陈于依澜. 基于云计算的服装流行趋势预测机制研究[D]. 上海：东华大学，2016.

[21] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007(3): 8-19.

[22] 杨铭，祁巍，闫相斌，等. 在线商品评论的效用分析研究[J]. 管理科学学报，2012, 15(5): 65-75.