基于大数据的
商业智能在电商数据分析中的应用

文 / 钱丹丹　周金海

摘要：为了将大数据与传统商业智能相结合，重新设计了商业智能的架构平台，着重探讨了数据获取方式，以中药饮片企业电商数据为例，用聚类分析中的K-Means算法对消费者进行分群，以此实现对不同消费者进行个性化营销的目的。

关键词：大数据；商业智能；数据挖掘；聚类分析

引言

商业智能（BI）概念由Gartner Group提出，涉及信息搜索、管理和分析，目的是使企业决策者获得知识，促使他们做出对企业更加有力的决策。商业智能不是一种独立的技术，而是一套完整的解决方案。它将数据仓库，联机分析（OLAP），数据挖掘和可视化等技术结合应用于业务活动，使企业的复杂信息转化为可供辅助的知识，最后将知识呈现给用户，以支持企业决策[1]。

随着Internet应用程序规模的不断扩大，需要处理的数据量呈指数级增长，数据结构变得越来越复杂。业务运营压力急剧增大，从而直接推动了大数据处理技术的发展[2]。随着电子商务、云计算、移动社交媒体等新一代IT技术的快速发展，传统的BI系统逐渐不能满足企业数据分析的需求。个性化、数据化、科学的数据分析技术逐渐使传统的BI系统需要与大数据技术相结合，实现一种满足大数据分析的新平台架构。

1、基于传统BI体系的大数据应用设计

在大数据时代，传统BI的数据存储能力、数据分析能力、实时数据处理能力不能胜任非结构化的复杂数据源的应用分析。因此，如何综合利用现有的BI和大数据技术是新平台架构设计的关键。传统的BI数据主要来自内部操作系统和管理系统；大数据的主要来源是互联网，如微博，网页和其他数据交换。在数据源、数据收集、数据处理、数据存储和以后的数据应用程序方面，这两者都有本质上的不同。基于以上考虑，设计了新的架构平台如图1所示。

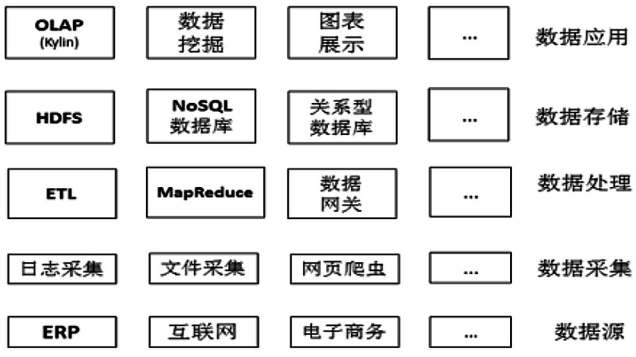

图1 BI与大数据结合的数据平台

数据源主要包括企业的内部数据和外部数据，内部数据由OA系统、ERP系统、财务报表系统等相关结构化数据组成；外部数据包括互联网上的非结构化数据，如超文本，图像和视频。数据采集在原有采集方式中新增了互联网网页爬虫的采集方式。针对结构化和非结构化的数据采用不同的处理方法。非结构化数据整理成结构化数据存储在分布式结构化数据库中；传统数据仍存储在关系型数据库中。大数据主要以分布式文件系统（HDFS）和NoSQL数据库的形式存储。最终数据主要用于联机分析处理，数据挖掘，数据可视化等方面。

2、数据采集方式

大数据背景下的数据收集方法主要包括三类：系统日志收集，网络数据收集和数据接口收集。日志数据的采集是通过设备中的日志记录子系统实现的，这个子系统能够在必要的时候生成日志消息。常用的商用数据API都支持REST API的方式获取数据信息。网络数据采集主要采用网络爬虫技术，其核心原则是：使用超文本传输协议HTTP仿真浏览器通过统一资源定位器URL地址访问Web服务器，获取Web服务器的权限，返回到原始页面并解析数据[3]。

传统的网络爬虫技术可能存在问题，因此为爬取web资源而设计的聚焦爬虫技术应运而生。聚焦爬虫有选择地访问因特网上的与网页相关的链接，以基于已建立的爬行目标（使用某电商销售主题）获得他们所需的信息。聚焦爬虫并不追求网页的全面覆盖，相反，它针对与特定主题相关的网页，并为面向主题的用户查询准备数据资源。

3、中药饮片企业电商数据应用案例

3.1 中药饮片企业发展状况

传统中药饮片在生产销售过程中比较混乱，没有统一的质量标准，因此，质量监督管理难度较大。由于中药饮片生产企业已经逐渐全面实施药品GMP认证，其生产已从纯手工加工独立出来成为中药行业的一项产业。也因此中药饮片、中药材、中成药并称为中药的三大组成部分。随着GMP认证的实施，中药饮片生产企业也发生了本质的变化，中药饮片的质量得到了提高，同时取得了良好的社会效益。然而，中药饮片的来源，加工方法和用途均有其传统特征。这一目标特性与GMP要求之间存在很大差异。因此，在实施过程中存在很多问题，特别是

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2010版的GMP和附录对中药饮片生产的要求达到了前所未有的高度，中药饮片企业的管理面临严峻挑战。

3.2 K-Means算法

K均值是一种广泛使用的聚类方法，它将D个实体划分为N个聚类。从而确保集群内的相似性尽可能高，集群之间的相似性尽可能低。K-means算法的过程如下：

(1) 随机选择N个数据点作为质心；

(2) 计算数据集中每个数据点到质心的距离，并将数据集中的所有数据点聚合为N个簇；

(3) 根据第2步计算得到的N组数据点，迭代计算出新的质心；

(4)重复步骤2-3，直到最终质心与前一个质心之间的距离很小（满足收敛）；

(5)最后读入所有的观察值，将每个观察值按照最接近质心的类别进行分类，分类结束。

质心和距离是K-MEANS算法的两个基本概念。质心可以被看做是一个样本，或者可以被认为是数据集中的某个数据点A，并规定它是具有相似性的一组数据的中心。质心的选择对聚类结果有很大影响，因为该算法是随机选择任何一个对象作为初始聚类的质心，并且最初表示聚类结果。当然，这个结果通常是不合理的，只是随机划分的数据集。质心的具体校正还需要多轮迭代计算才能逐渐逼近所需的聚类结果：具有相似性的对象被分组为一组，所有这些对象都具有共同的质心。另外，由于初始质心选择的随机性，最终结果不一定是预期的，因此需要多次迭代，在每次迭代时重新随机获得初始质心，直到最终聚类结果满足预期。

距离实际上是相似度的度量。常见的距离公式计算有：曼哈顿距离，欧几里德距离，闵可夫斯基距离，切比雪夫距离等。聚类分析中最常用的距离公式是欧氏距离，因为欧氏距离直观且容易计算，而且欧式距离对对象的点进行坐标偏移和变化旋转，最后，距离的值保持不变，因此仍然可以通过对象的原始相似性来判断对象相似性。设d（x，y）为对象a和b之间的距离，则d（x，y）应满足以下三个属性：

（1）非负性：即d(x，y)^{3}0恒成立；当且仅当x=y时，d(x，y)=0。

（2）对称性：即d(x，y)=d(y，x)。

（3）三角不等式：任意对象a，b，c恒有d(x，y)+d(y，z)^{3}d(x，z)。

3.3 中药饮片企业电商数据应用分析

在大数据时代，独立的数据本身价值不大，通过数据预测未来趋势以及利用数据发现隐藏的知识才是关键。 众多中药饮片企业紧跟时代发展，在电商网站都有相应的门店销售中药饮片，因此积累了大量顾客购买中药饮片的消费记录。对这些消费记录的分析可以对消费者进行分组，不同群体的消费者可以根据消费行为对营销进行个性化。 客户分类有利于中药饮片企业针对性的为不同群体客户提供差别化服务，也能够让企业及时察觉市场和客户的一些微小变化并针对其调整策略。

RFM模型是广泛应用的多因素客户分类方法，R(Recency)表示客户最近交易到当前时间的时间段。F（Frequency）代表在指定时间段内客户与企业合作的次数（即购买行为），M（Monetary）代表在指定时间段内客户与企业交易所产生的金额[4]，RFM是以客户创造的绝对金额来衡量客户价值的。

现从某中药饮片电商网站爬取相关数据，依据一定的数据处理原则对原始数据进行清洗采集，经过处理后得到消费者数据（3000条），R在这里表示最近一次购买中药饮片的时间间隔，F表示购买中药饮片频率，M表示在某平台上消费的总金额，截取部分有效数据见表1：

表1 消费者相关数据

客户序号	R（天）	F（次）	M（元）
1	27	6	334.21
2	3	5	759.19
3	4	16	1383.39
4	3	11	3280.77
5	14	7	154.65
6	19	6	501.38
7	5	2	1721.93
8	26	2	107.18
9	21	9	973.36
10	2	21	764.55
11	15	2	1251.4
12	26	3	923.28
13	17	11	1011.18
14	30	16	1847.61
15	5	7	1669.46

不同数据项之间存在着数值大小和数值单位的差异，因此不能直接用来参与运算。 比如，消费者购买的产品总金额M是一个很大的数值属性，单位一般在百以上，而在一定时间内购买产品的频率往往较小，且相对于消费金额来说没什么作用。为了让这些属性都能发挥作用，需要将属性与其自身对应的范围进行比较，保证单位和数值不存在差值性，以便后期直接使用这些标准数据进行运算。本文采用归一化处理方法对数据进行处理，以下表2是经过处理后的3000条数据中的部分数据。

表2 归一化后消费者相关数据

客户序号	R	F	M
1	0.000889182	0.000329254	0.000168387
2	9.8798E-05	0.000274379	0.000382507
3	0.000131731	0.000878011	0.000697002
4	9.8798E-05	0.000603633	0.001652971
5	0.000461057	0.00038413	7.79183E-05
6	0.00062572	0.000329254	0.000252613
7	0.000164663	0.000109751	0.000867571
8	0.000856249	0.000109751	5.40012E-05
9	0.000691586	0.000493881	0.000490414
10	6.58653E-05	0.00115239	0.000385208
11	0.00049399	0.000109751	0.000630501
12	0.000856249	0.000164627	0.000465182
13	0.000559855	0.000603633	0.000509469
14	0.00098798	0.000878011	0.000930893
15	0.000164663	0.00038413	0.000841134

使用K-Means算法设置簇的数量为3，最大迭代次数为3，距离函数使用欧几里德距离。由于初始质心是随机的，因此每个簇的结果可能不同。经过多次重复实验后，检测聚类结果基

（下转第96页）

(C)1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

然后根据品牌定位找出其不同于其他同类品牌的差异化特征，以此提高品牌辨识度。在这两个步骤中，意见领袖可以作为一个连接企业与消费者的纽带。企业通过意见领袖获取的消费者实时反馈，有助于企业对品牌的定位与以及对差异化特征的寻找。

在品牌形象构建完成之后，不能忽视接下来的维护与推广。一方面，在选择意见领袖时，要考虑到该博主的整体风格是否与品牌形象一致，例如微博内容、粉丝人群的定位等。另一方面，企业可借助意见领袖博主及时处理微博平台中不利于品牌形象的信息，同时积极扩散有利于品牌形象的信息。通过意见领袖来进行的长期的品牌推广与维护活动，可以有效的将大学生用户培养成为企业的潜在消费者。

5、总结及启示

面对大学生这一极具开发潜力的消费市场，微博这个开放度高且逾趋成熟的平台给网络营销业界提供了一个新的思路。借助微博中的意见领袖博主，企业可以低成本高效率的开展营销活动。但如果企业想保证微博营销活动的效果，则需要加大对微博意见领袖的运用力度，以及注重对该群体进行培养与挖掘。除此之外，在开展营销活动时，企业需要根据大学生群体的特点以及具体的情境来扩充目前已有的销售渠道，以此来提升产品的销售转化率。最后，企业若想有长远的发展，企业品牌价值的开发与维护是一定不能忽视的。只有兼顾以上四点，才能将企业的网络营销活动效果达到最大化，使企业走上可持续发展之路。

参考文献：
[1] 张晓霞.新浪微博受众的“使用与满足”研究[D].北京：北京邮电大学,2013.
[2] 彭兰.网络传播概论[M].北京:中国人民大学出版社,2017.
[3] 涂凌波.草根、公知与网红:中国网络意见领袖二十年变迁阐释[J].当代传播，2016(05):84-88.
[4] 王艳.民意表达与公共参与:微博意见领袖研究[D].北京：中国社会科学院研究生院,2014.
[5] 林建煌.消费者行为[M].北京:北京大学出版社,2016.
[6] 连环.网络意见领袖的营销价值及实现[J].电子商务,2009(07):35-37.

作者简介：
曹博文，湖南人文科技学院2015级本科生，研究方向：电子商务；
李李（通讯作者），硕士研究生，讲师，现任教于湖南人文科技学院，研究方向：电子商务。

（上接第30页）

本相同，因此可以采用此聚类结果，对聚类用户进行群体特征分析，并进行群体个性化营销。以下是K-Means算法聚类生成的群体一、二、三的图片，如图3消费群体所示：

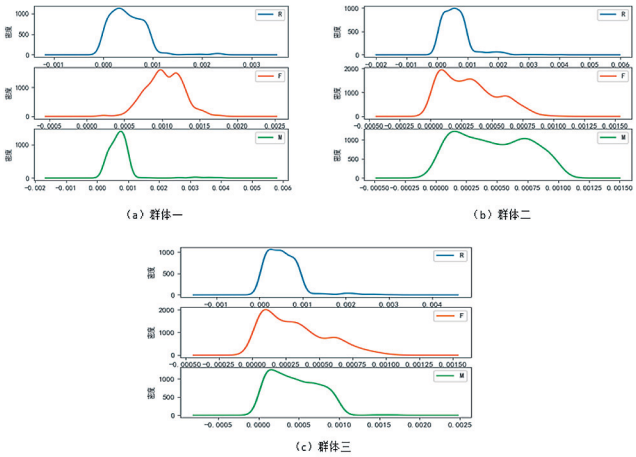

图2 消费群体

群体一：这些客户最近一次在电商网站消费间隔天数（R）较短，消费总金额（M）较多。他们是企业最理想的客户类型，同时也是潜在客户，对公司贡献大，但所占比例很小。 企业应优先考虑将资源投放到他们身上，以此实现差异化管理和一对一营销，从而提高此类客户的忠诚度和满意度，并最大限度地提高此类客户的高消费水平。

群体二：这些客户的购买频率（F）一般，最后一次在电子商务网站上消费的时间间隔（R）较短，并且消费总量（M）是适中的。他们客户价值变化的不确定性很高，消费下降的原因各不相同，因此及时了解客户信息并与客户保持互动尤为重要。企业可以根据近期消费间隔时间和消费频次来推测顾客消费行为的变化，重点关注这些客户并采用特定的营销方案来延长这类客户的生命周期。

群体三： 这类客户的购买频率（F）一般，最近一次在电商网站消费间隔天数（R）适中，消费总金额（M）较少。他们是中药饮片企业的一般用户与低价值客户，可能只有中药饮片打折促销时才会购买。

4、总结

在大数据的背景下，充分利用数据挖掘信息可以抓住市场机遇。 众多企业除了线下实体销售外也开展了具有独特优势的线上交易，从电商大数据中挖掘隐藏的信息，根据这些信息，针对不同的客户群体进行个性化营销，从而提高企业的客户满意度和经济效益。 本文主要研究了大数据与传统商业智能在电商企业（中药饮片电商网站）数据分析中的应用，重点描述聚类分析的 K- Means算法并应用于电子商务网站中客户消费数据的挖掘。 通过聚类分析将客户分为3个群体，根据不同客户群体的特征有助于企业识别客户，从而实现差异化的营销目标。

参考文献：
[1] 陈荣鑫,付永钢,陈维斌. 基于Pentaho的商业智能系统[J].计算机工程与设计,2008,09：2407-2409.
[2] 杨超. 基于大数据技术的BI系统关键技术研究[D].华南理工大学,2016.
[3] 卞伟玮,王永超,崔立真,郭伟,李晖,周苗,薛付忠,刘静.基于网络爬虫技术的健康医疗大数据采集整理系统[J].山东大学学报(医学版),2017,55(06)：47-55.
[4] 李品睿,许守任,许晖.基于 RFM 模型的核心客户识别与关系管理研究——以保险业为例[J].现代管理科学,2015,(6):24-26.

作者简介：
钱丹丹，硕士研究生，南京中医药大学，主要从事医药信息工程、人工智能研究工作；
周金海，教授，南京中医药大学信息技术学院硕士研究生导师，南京审计大学金审学院特聘教授，主要从事医药信息及人工智能研究工作。

(C)1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net