

文章编号:1001-9081(2015)10-2727-06

doi:10.11772/j.issn.1001-9081.2015.10.2727

基于 top-k 显露模式的商品对比评论分析

刘璐¹, 王怡宁¹, 段磊^{1,2*}, Jyrki Nummenmaa³, 晏力¹, 唐常杰¹

(1. 四川大学 计算机学院, 成都 610065; 2. 四川大学 华西公共卫生学院, 成都 610041;

3. 坦佩雷大学 信息科学学院, 芬兰 坦佩雷 FI-33014)

(*通信作者电子邮箱 leiduan@scu.edu.cn)

摘要:随着电子商务的发展,许多购物网站都提供商品评论作为用户购物的决策参考。由于商品评论具有海量、冗余、不规范的特点,用户难以在短时间内浏览所有商品评论,更难以基于评论内容发现商品对比特征。对此,设计了 top-k 显露模式挖掘算法,并将此算法应用于商品评论对比分析,实现了用户购物决策支持系统—ReviewScope。ReviewScope 能够从不同商品的评论中发现特定商品的对比评论,并以此作为购物决策可视化地提供给用户。基于京东商城真实商品评论数据的实验结果表明 ReviewScope 具有效、灵活、用户友好的特点。

关键词:商品评论;购物决策支持;模式可视化;显露模式挖掘;对比评论

中图分类号: TP311.13 **文献标志码:**A

Analysis on distinguishing product reviews based on top-k emerging patterns

LIU Lu¹, WANG Yining¹, DUAN Lei^{1,2*}, NUMMENMAA Jyrki³, YAN Li¹, TANG Changjie¹

(1. School of Computer Science, Sichuan University, Chengdu Sichuan 610065, China;

2. West China School of Public Health, Sichuan University, Chengdu Sichuan 610041, China;

3. School of Information Sciences, University of Tampere, Tampere FI-33014, Finland)

Abstract: With the development of e-commerce, online shopping Web sites provide reviews for helping a customer to make the best choice. However, the number of reviews is huge, and the content of reviews is typically redundant and non-standard. Thus, it is difficult for users to go through all reviews in a short time and find the distinguishing characteristics of a product from the reviews. To resolve this problem, a method to mine top-k emerging patterns was proposed and applied to mining reviews of different products. Based on the proposed method, a prototype, called ReviewScope, was designed and implemented. ReviewScope can find significant comments of certain goods as decision basis, and provide visualization results. The case study on real world data set of JD.com demonstrates that ReviewScope is effective, flexible and user-friendly.

Key words: product review; shopping decision support; pattern visualization; emerging pattern mining; contrastive review

0 引言

随着电子商务的快速发展,网络购物在人们的生活中扮演着越来越重要的作用。据中国互联网络信息中心发表的《第 35 次互联网报告》,截至 2014 年 12 月,我国网络购物用户规模达到 3.61 亿,电子商务市场交易规模达到 12.3 万亿元。

由于在网络购物过程中,用户难以直观地感受商品特点,加之可选商品品目繁多,许多用户通常面临相似商品的选购困惑。对此,如何让用户更清晰地了解商品的特点,既是用户的需求,也是网络购物网站提高销量的关键。据艾瑞咨询发布的 2013 年度电子商务市场核心数据中的调查显示:64.4% 的网购消费者使用商品评论作为参考信息,商品评论在用户购物决策中占有重要的地位。目前,常见的网络购物网站,例如:淘宝、京东商城、Amazon 和 Best Buy 等,都提供了已购物消费者对某商品或者相关服务的体验评论,来为潜在用户提

供购物决策。

例 1 苹果 iPhone6 和三星 Note4 是两款硬件配置相似、功能相当、价格接近的手机,不少潜在消费者在购买前都会看看已购用户的相关评价作为购物参考。表 1 例举了苹果 iPhone6 和三星 Note4 分别在京东、淘宝、苏宁、Amazon 及 Best Buy 的商品评论数量(数据统计截止时间为 2015 年 1 月,商品评论数量来源于网站提供的统计)。

表 1 不同网站关于苹果 iPhone6 和三星 Note4 商品评论数量

商品	京东	淘宝	苏宁	Amazon	Best Buy
iPhone6	3 461	5 897	9 326	182	555
Note4	1 798	588	2 283	185	421

通过观察实际网络购物网站提供的商品评论,可以发现商品评论信息具有以下特点:

1) 数量大。如表 1 所示,评论数量最少的 Amazon 也有 182 条评论。由于时间关系,鲜有用户阅读所有评论。

收稿日期:2015-06-15;修回日期:2015-06-26。 **基金项目:**国家自然科学基金资助项目(61103042);中国博士后科学基金资助项目(2014M552371);软件工程国家重点实验室开放研究基金资助项目(SKLSE2012-09-32)。

作者简介:刘璐(1991-),女,内蒙古五原人,硕士研究生,主要研究方向:数据挖掘、序列数据; 王怡宁(1993-),女,内蒙古鄂尔多斯人,主要研究方向:数据挖掘; 段磊(1981-),男,四川成都人,副教授,博士,主要研究方向:数据挖掘、知识发现; Jyrki Nummenmaa(1961-),男,芬兰人,教授,博士生导师,博士,主要研究方向:数据库、大数据处理; 晏力(1988-),男,四川泸州人,硕士研究生,主要研究方向:数据挖掘; 唐常杰(1946-),男,重庆人,教授,博士生导师,博士,主要研究方向:数据库、知识工程。

2)冗余。很多评论表述的内容相似,如:有关 iPhone6 的“蛮好速度快”“运行流畅”,这样的评论表达的意义相似。

3)不规范。很多评论内容实际意义偏低,如:有关 Note4 的“送人的,还没用过”“快递太慢”,这样的评论没有描述商品特征,此类商品评论参考性较低。

通过对商品评论进行比较,过滤掉商品评论中的重复、冗余信息,可以获得独有的商品评论信息。如:虽然苹果 iPhone6 和三星 Note4 都有不少“音量大”的商品评论,但是三星手机“屏幕大”的商品评论在苹果手机的商品评论中并不常见。获得这样的评论信息,可以帮助偏好大屏幕的用户作出更好的购物决策。由此给出如下定义。

定义 1 给定查询商品和参照商品的商品评论集,在查询商品的评论集中频繁出现而在参照商品的评论集中非频繁出现的评论短语,称为对比评论。

对比学习泛指以比较为主要途径学习不同类对象间的差异特征、模型的一系列方法。限于篇幅,本文仅给出对比学习的基本概念,更多细节可以参考文献[1]。本文的研究目标在于针对指定的待比较商品,利用对比学习方法,从大量评论中找到最显著的商品评论,并可视化地提供给用户作为购物决策支持。

据我们所知,尚没有使用对比学习方法分析商品评论的研究工作。高效、准确地比较商品评论并找出评论间的差异面临如下挑战:1)评论数量大,人工获取难度高,且 AJAX 技术使信息通过异步更新,导致传统爬虫无法有效获取动态网页内容,完全、准确地从网站上获取全部的商品评论非常困难。2)直观上讲,虽然能实现显露模式挖掘算法^[2]来区分获取的评论,但是传统的显露模式挖掘算法要求用户设定支持度阈值,然而设置合理的支持度阈值又非常困难。需要考虑如何将显露模式挖掘算法运用到系统中又方便用户使用。3)对比评论挖掘结果为多个描述短语的集合,当集合中评论数量增加时,用户难以在众多相似评论中找到关键性的差异化评论,因而需要考虑如何友好地展示挖掘结果。

综上,为帮助用户更好地作出购物决策,本文设计并开发了一个原型系统——ReviewScope。本文主要工作包括:

1)设计了 top-k 显露模式挖掘算法。传统显露模式挖掘算法^[2]需要用户显示地指定支持度阈值,但是在不具备先验知识的情况下,用户难以设定恰当的阈值,随意地设定阈值则可能导致挖掘结果不合理。

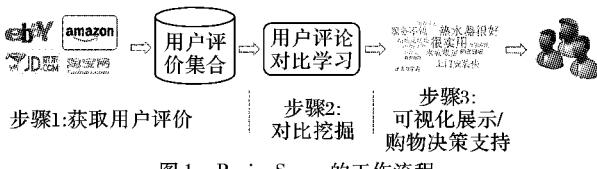


图 1 ReviewScope 的工作流程

2)将 top-k 显露模式挖掘算法应用于商品评论对比分析。利用商品评论的标签,将每条商品评论转换为由一组商品评论标签组成的项集。对不同的商品之间运用 top-k 显露模式挖掘算法可以发现特定商品的对比评论模式。

3)实现了用户购物决策支持系统 ReviewScope。在之前工作^[3]的基础上,实现了动态网页中的商品评论标签的自动获取。并运用可视化技术,以文字云的方式动态地展示特定商品的显著评论模式,作为用户的购物决策支持。

ReviewScope 的工作流程:首先,ReviewScope 通过爬虫获取动态网页商品评论,生成商品评论集;其次,基于对比学习

对商品评论进行 top-k 显露模式挖掘,发现 top-k 对比评论模式,产生对比评论;然后,运用可视化技术,结合动态文字云展示对比评论,以此为用户提供购物决策支持。

1 相关工作

1.1 动态网页信息获取

获取商品评论内容是进行评论分析研究的前提和基础。但评论数量巨大,手工获取难度大,且由于 AJAX 技术的广泛应用,有效获取这类异步更新信息网页是开展评论分析的新挑战。文献[4]提出了一种新颖的挖掘商品评论的方法。该方法引入了新概念——浅依赖文法,通过利用大多数商品评论特征是短语的事实,构造浅依赖树(shallow dependency tree),由此提取商品评论。文献[5]运用关联规则来挖掘商品评论。首先,获取商品评论,找出该商品评论中描述的频繁和非频繁的特征;对这些特征进行词性标注,然后运用关联规则挖掘找到所有频繁项集;在确定评论是正面的还是负面的之后,对所有的特征进行总结。文献[6]将文本挖掘和计量经济学结合在一起,将评论分词后来评估一个产品的个体特征。

1.2 显露模式挖掘

对比挖掘是近年来数据挖掘领域的热点之一。文献[7]提出的对比挖掘泛指发现在不同类别或条件下,数据集中具有“对比”特性模式的方法。文献[2]定义了显露模式,显露模式挖掘作为对比挖掘的一个分支,是发现从一个数据集到另一个数据集支持度发生显著变化的项集。根据显露模式的定义,其发现的目标是通过对比的方法,找出支持度变化大的项集,因此具有广泛的应用。文献[8]提出了一种基于基本显露模式的电子邮件分类与过滤算法,该算法将基本显露模式的分类算法与新的特征提取方法相结合,是一种十分高效的方法。文献[9]提出一种应用显露模式的分类器——PCL,用以分析急性淋巴白血病 6 种亚型以上的基因表达谱。文献[10]还提出一种基于对比模式的聚类算法来识别动态逻辑博客社区中的主题和趋势。

1.3 模式可视化

可视化是利用计算机图形学和图像处理技术,将数据转换成图形或图像展示在屏幕上,并进行交互处理的理论、方法和技术。信息可视化利用图像等形式可以更有效地向用户传达信息。近年来,关于数据挖掘及可视化相结合的研究越来越广泛,文献[11]设计了一个系统——FeatureLens,将文本中挖掘的频繁模式和可视化功能集成在该系统中,用以创造新的见解和发现。文献[12]提出了一个称为 SocialViz 的观察仪,应用频繁模式可视化的方法,为用户提供网络中多实体之间社交关系上频繁的信息。SocialViz 将频繁模式在二维空间中用坐标系表示出来,但这种可视化表示方法无法表达文本信息的关系。文献[13]将空间数据挖掘技术与可视化技术结合起来,运用关联规则挖掘和聚类的方法,通过地理信息系统(Geographic Information System, GIS)和 Google Earth 设备进行可视化,用以分析斯洛文尼亚的不同地区不同时间段的交通安全状况。

2 ReviewScope 设计与实现

2.1 评论自动获取

为对比不同产品商品评论之间的差异性,ReviewScope 首先从网页上爬取商品评论。ReviewScope 接收查询商品及参

照商品的起始 URL。根据给定商品 URL 找到相关页面,再将网页解析为 DOM 树,找到包含商品评论的节点抽取内容,遍历树中的节点,动态获取评论模块中所有子页面的商品评论。该设计使得用户在购物商品展示页面获得起始链接后,可将该链接直接输入 ReviewScope,从而获得所有子页面的商品评论,不需要复制每页评论的子页面链接。

如何获取商品评论遇到两个较大的挑战:1)为改善传统 B/S 模式下用户浏览网页的体验,绝大多数购物网站将 AJAX 和 JavaScript 等技术应用于动态网页中,使信息通过异步方式更新,同时网站在评论模块进行分页显示,导致传统的静态爬虫技术无法完全获取商品评论;2)某款商品对应的页面中,除商品评论外,还存在许多商品介绍的信息,因此在设计评论获取算法时,爬虫必须在商品的整个网页中筛选无关信息并且准确快速地定位商品评论模块。

在之前的工作中,设计了一种基于对比学习的动态网页商品评论的获取方法——ReviewCrawler^[3],该方法将网页 p 解析为一棵 DOM 树 T , T 中每个节点代表 p 中的标记(tag)。由于 p 中的标记包含页面信息,因此 ReviewCrawler 的目标在于找到 T 中包含商品评论的节点并将其内容抽取出来。本文将 ReviewCrawler 应用于 ReviewScope 中来获取商品评论标签。

2.2 评论对比学习

获得不同商品的评论标签后,ReviewScope 可以发现特定商品所有的对比评论。基于此特性,将显露模式挖掘方法应用于该系统。回顾显露模式的定义,对于一个项集,当它在不同数据集中的支持度差异大于设定阈值时,这样的项集就是显露模式。为了区别,用“+”代表查询商品的评论集,“-”代表参照商品的评论集。形式化定义显露模式:给定数据集 D_+ 和 D_- 及模式 X ,当模式 X 在 D_+ 中的支持度大于 α 而在 D_- 中的支持度小于 β (其中 α 和 β 都是用户指定的参数),这样的模式 X 就是显露模式。可看出,显露模式挖掘就是通过对比,找出支持度差异最大的项集。然而在大部分的实际应用中,用户很难根据已有的经验设定合适的阈值。因此本文先定义商品评论对比度,依据商品评论对比度定义 top- k 对比评论模式,定义如下:

定义 2 给定项集 ts , $Sup_+(ts)$ 为查询商品的评论集中包含 ts 的支持度, $Sup_-(ts)$ 为参照商品的评论集中包含 ts 的支持度,则 ts 的商品评论对比度为:

$$CR(ts) = Sup_+(ts) - Sup_-(ts) \quad (1)$$

例 2 对表 2 中的商品评论数据集,令 $ts = \{\text{外观漂亮,系统流畅}\}$,根据式(1), $Sup_+(ts) = 0.75$, $Sup_-(ts) = 0.25$, $CR(ts) = 0.75 - 0.25 = 0.5$ 。

表 2 商品评论数据集

评论编号	评论集合	类标
1	{外观漂亮,系统流畅,功能齐全}	
2	{外观漂亮,分辨率高,功能齐全,音效好}	+
3	{系统流畅,功能齐全,外观漂亮}	+
4	{外观漂亮,系统流畅}	
5	{屏幕大}	
6	{反应快,屏幕大,照相不错}	-
7	{外观漂亮,系统流畅,屏幕大,功能齐全}	-
8	{照相不错,屏幕大,反应快}	

定义 3 对于查询商品的评论集 D_+ 和参照商品的评论

集 D_- ,在所有挖掘出的对比评论项集中,商品评论对比度最大的前 k 个项集为 top- k 对比评论模式。

例 3 对表 2 中的商品评论数据集,令 $k = 5$,类标为“+”的序列作为查询商品的评论数据集,类标为“-”的序列作为参照商品的评论数据集,那么 top- k 对比评论模式如表 3 所示。

表 3 top-5 对比评论模式

项集 ts	$Sup_+(ts)$	$Sup_-(ts)$	$CR(ts)$
{外观漂亮}	1.00	0.25	0.75
{系统流畅}	0.75	0.25	0.50
{功能齐全}	0.75	0.25	0.50
{外观漂亮,系统流畅}	0.75	0.25	0.50
{外观漂亮,功能齐全}	0.75	0.25	0.50

如定义 3,设定 k 值比设定支持度阈值更易于用户理解。当出现多个 CR 值相等的情况下,top- k 对比评论模式挖掘算法将会选择最长评论模式,这样选择是因为较长的评论有助于用户对评论内容的理解。同时,由于 k 值受限于网购网站展示给用户的标签数量,所以如果 k 值过大,获得的标签较多,用户很难理解挖掘的结果。本文在第 3 章展示不同 k 值下挖掘结果的可视化效果及 k 值对系统效率的影响。在 ReviewScope 中用户可以根据自己的喜好设定不同的 k 值改变挖掘的结果,以获得需要的对比特征。

2.3 对比评论可视化

据我们所知,目前还没有实际的可视化效果应用于文本显露模式的展示,然而,用户更喜欢图形化的结果展示方式而非冗长的文字表达。为了更好地展示挖掘结果,将文献[14]中的文字云形式的可视化效果应用于对比评论的展示中,同时改进其显示效果使其更好的适应对比评论模式。

对比评论可视化主要由两部分构成:第一部分是动态文字云,挖掘出的评论标签在文字云中用不同大小的字体进行展示,字体大一些的标签代表出现频繁的商品评论,动态效果体现在当鼠标移至文字云中的任意一个评论标签时,文字云中该标签的 top- k 对比评论模式中包含的所有标签会高亮展示,而其他无关标签会隐藏,以此方便用户的查看及交互。同时文字云中标签展示为相应 top- k 对比评论模式中所有模式包含标签的并集。第二部分是一个文本框,文本框中会列出选中标签的所有 top- k 对比评论模式。当在文字云中选中并点击一个标签,就可以在右边的文本框里看到包含该标签 top- k 对比评论模式。该项设计是为方便用户的查看。展示这样的 top- k 对比评论模式可以清楚地让用户知道商品的哪些特性在商品评论中是同时频繁出现的。

例 4 对表 3 中得出的 top-5 对比评论模式进行可视化展示,其中 $k=5$,那么文字云中所展示标签为模式中所包含标签的并集,即 $\{\text{外观漂亮}\} \cup \{\text{系统流畅}\} \cup \{\text{功能齐全}\} \cup \{\text{外观漂亮,系统流畅}\} \cup \{\text{外观漂亮,功能齐全}\} = \{\text{外观漂亮,系统流畅,功能齐全}\}$,所以文字云中展示的标签个数为 3,当鼠标移至文字云中的 {功能齐全} 标签时,文字云中相应的 {外观漂亮} 标签会高亮展示,因为模式中有项集 {外观漂亮,功能齐全}。而当鼠标选中 {功能齐全} 标签时,文本框部分中会列出的对比评论包括: {功能齐全} 和 {外观漂亮,功能齐全}。

2.4 用户界面设计

本文使用 Java 实现基于 J2EE 架构的 ReviewScope 系统,用以帮助用户挖掘 top- k 对比评论。图 2 展示了 ReviewScope 的用户界面。图 2 中 A 部分文本框表示系统输入,在上面两个文本框中分别输入查询商品及参照商品的网页链接,下面的文本框由用户输入指定 k 值大小,表示希望找到的前 k 个对比评论模式, k 值应小于对比评论模式结果集的大小,如果 k 值超过该上限,系统会自动提示错误输入。输入 URL 链接及 k 值后,点击“Compare”按钮,系统即可开始运行,“Reset”按钮可以帮助用户重置图 2 中 B 部分文字云的形状,方便用户的查看。



图 2 ReviewScope 用户界面

图 2 中 B 部分即可可视化的动态文字云,文字云中的标签包含 top- k 对比评论模式中包含的所有标签,每个标签的大小表示该标签在数据集中支持度的大小,字体大的表示其在评论集中支持度较大。鼠标移至其中的任一标签,文字云会高亮显示所有包含该标签且满足 k 值设定的差异化标签集合,并且会隐藏其他无关标签。

图 2 中 C 部分即可可视化的文本框,文本框中会列出选中标签在两个商品评论中支持度差值最大的前 k 项评论项集,即 top- k 对比评论模式,ReviewScope 通过这样的方式为用户提供海量评论中具有参考价值的频繁模式。

3 实验分析

本文使用 Java 程序实现 ReviewScope 系统。实验环境配置如下:JDK 1.8;Tomcat 8.0;Intel i7 4790 3.60 GHz;8 GB 内存;Windows 7 操作系统。实验过程显示出了 ReviewScope 系统友好、高效、灵活的特点。

3.1 实验数据集

为验证 ReviewScope 的有效性和性能,本文分别选取了京东商城多组相同类型且价格相近的商品进行实验。实验数据集信息如表 4 所示。

表 4 实验数据集信息

商品类别	商品名称	商品编号	评论数量	类标
音箱	惠威多媒体音箱	110664	12741	+
	哈曼卡顿水晶音箱	567895	3642	-
手机	华为荣耀 6	1185017	6002	+
	魅族 MX4	1209645	6110	-
加湿器	小熊加湿器	1010243	7054	+
	德尔玛加湿器	671473	4088	-

注:“+”为查询商品的评论集,表示用户指定的商品评论;
“-”为参照商品的评论集,表示用于对比的商品评论。

表 4 所示 6 种商品在各自品类中都属于热销商品,价格相近且配置相似,因此用户在选购时难以抉择。表中数据统

计时间截止至 2015 年 2 月,商品评论数量来源于网站提供的统计。

3.2 有效性分析

ReviewScope 广泛应用于各种类型的商品评论数据集,帮助用户在选购不同类型商品时作出决策。实验选取三类共 6 种商品,每组中的两个商品在价格、配置等参数上具有可比性。

如图 3 所示的实验结果表示的是在每组数据集中满足 k 值的所有对比评论标签的集合。

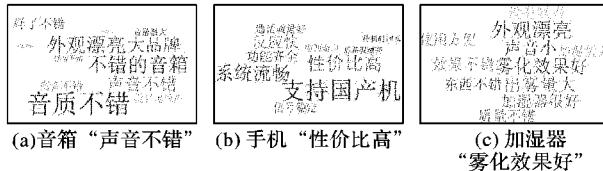


图 3 不同类型商品评论集对比结果

点击文字云上的标签,可以看到所有包含点击标签且支持度变化最大的 top- k 对比评论模式。图 4 展现了在图 3 的 3 组结果中选择音箱的“声音不错”、手机的“性价比高”、加湿器的“雾化效果好”这 3 个标签后相对应的对比评论模式列表。

图 4 中的文本框列出选中标签在两个商品评论中支持度差值最大且满足 k 值的 top- k 对比评论模式,例如,当点击“声音不错”时,包含声音不错的标签集合如图 4(a)所示,表明许多认为音箱“声音不错”的人,也同时认为其具有“外观漂亮”“高音不错”等特点。

(a) 音箱 “声音不错”	(b) 手机 “性价比高”
声音不错 声音不错, 音质不错 声音不错, 不错的音箱 外观漂亮, 声音不错 大品牌, 声音不错 高音不错, 声音不错 没有电流声, 声音不错 样子不错, 声音不错 音量很大, 声音不错	性价比高 性价比高, 系统流畅 性价比高, 支持国产机 性价比高, 信号稳定 性价比高, 待机时间长 性价比高, 电池耐用 性价比高, 反应快 性价比高, 后盖很漂亮 通话质量好, 性价比高

(c) 加湿器 “雾化效果好”
雾化效果好 没有噪音, 雾化效果好 雾化效果好, 声音小 雾化效果好, 外观漂亮 质量不错, 雾化效果好 出雾量大, 雾化效果好 加湿器很好, 雾化效果好 效果不错, 雾化效果好 使用方便, 雾化效果好 东西不错, 雾化效果好

图 4 包含选中标签的对比评论

图 5~7 展示了音箱、手机和加湿器商品评论数据集中,当 $k=5,10,15$ 时,挖掘结果生成标签文字云的视觉效果。由图观察得知:如果 k 值设置太大,获得的标签数量较多,不利于用户理解挖掘结果;反之当 k 值很小时,有价值的评论标签很可能被从结果中过滤掉。为了得到理想的效果,ReviewScope 允许用户可以根据需要自定义 k 值。



图 5 不同 k 值下音箱数据集文字云展示结果

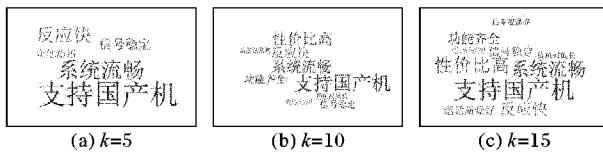


图6 不同k值下手机数据集文字云展示结果

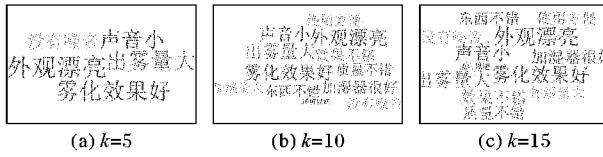


图7 不同k值下加湿器数据集文字云展示结果

3.3 性能分析

ReviewScope系统的响应时间主要包括商品评论获取时间和对比评论挖掘时间。由于获取商品评论的时间受限于用户网速及服务器响应时间,因此,将每次用户获取的评论缓存在服务器的本地文件中,以此减少系统联网的响应时间。

3.3.1 k值对ReviewScope执行效率的影响

k 值决定了商品评论对比度最显著的前 k 个对比评论模式,同时也会影响展示在可视化结果集中的对比评论个数。 k 值作为top- k 显露模式挖掘算法的唯一参数,测试当 k 值变化时,挖掘算法运行时间的变化,本文测试了3组数据集中 k 值和运行时间变化的关系,结果如图8所示。从图8可看出,当 k 值增加时,运行时间随 k 值变大而增加,但运行时间保持在1000 ms以内,系统响应时间在用户可接受范围内。

3.3.2 数据集规模对系统性能的影响

为研究数据集规模对ReviewScope执行效率的影响,选择3组商品评论作为实验数据集(见表4),设查询商品的评论集为“+”,参照商品的评论集为“-”。数据集大小用 $|D|$ 表示,参数 α 控制数据集大小的变化,测定当其中一个数据集规模固定 $|D|$,另一个数据集规模 $\alpha \cdot |D|$ ($\alpha = 0.2, 0.4, 0.6, 0.8, 1.0$)变化时系统运行时间的变化,结果如图9所示。

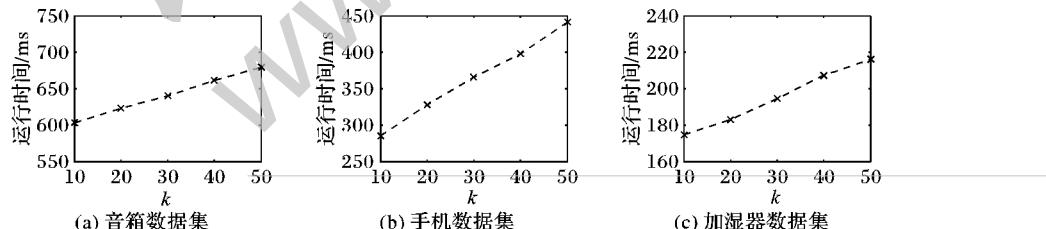


图8 不同k值下系统运行时间

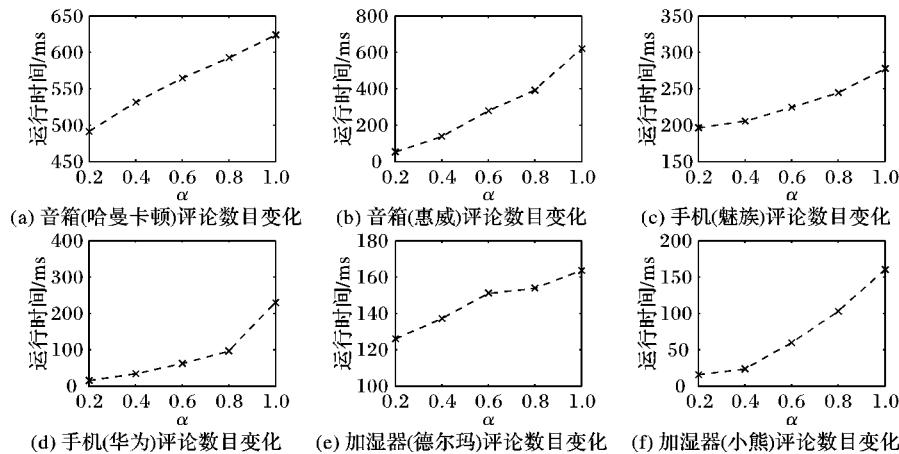


图9 ReviewScope伸缩性测试

图9(a)表示当惠威音箱商品评论数据集大小($|D_+|$)固定,哈曼卡顿音箱商品评论数据集大小($|D_-|$)随参数 α 变化时,ReviewScope运行时间的变化。图9(c)表示当华为手机商品评论数据集大小($|D_+|$)固定,魅族手机商品评论数据集大小($|D_-|$)随参数 α 变化时,ReviewScope运行时间的变化。图9(e)表示当小熊加湿器商品评论数据集大小($|D_+|$)固定,德尔玛加湿器商品评论数据集大小($|D_-|$)随参数 α 变化时,ReviewScope运行时间的变化。相应地,图9(b)、(d)、(f)分别表示当哈曼卡顿音箱、魅族手机及德尔玛加湿器的商品评论数据集大小($|D_-|$)固定时,惠威音箱、华为手机和小熊加湿器商品评论数据集大小($|D_+|$)随参数 α 变化时,系统运行时间的变化。

观察得知,当被参照数据集规模保持不变时,ReviewScope运行时间随参照数据集的增大而增加;并且,图9中运行时间变化表明,查询商品的评论集规模变化比参照商品的评论集规模变化对系统运行时间产生的影响更大。这是因为对比挖掘算法中要发现的目标是在查询商品的评论集中频繁而参照商品的评论集中非频繁的项集,当查询商品评论集的规模增大时,需要对比的项集数目同时增加,因而对运行时间影响更大;反之,如果参照商品评论集的大小增加,其中的项集增加,但是不满足对比挖掘的定义,算法对其进行剪枝,故其大小变化对ReviewScope运行时间影响较小。

4 结语

随着以用户为中心的B2C电子商务网站的普及,线上商品评论的数量快速地增长,同时商品评论指导购买决策的作用开始突显,但现有研究未提供商品评论之间的对比挖掘。这样的对比挖掘可以使用户从产品属性、产品使用体验和产品售后等多维度出发来指导购物决策。然而,由于商品评论存在海量性、重复性等特点,需要一种适用于分析这些商品评论的工具,帮助用户作出合适的购物决策。本文提出基于挖

掘商品评论的差异性来指导购物决策的方法,设计为用户提供购物决策系统——ReviewScope。ReviewScope 帮助用户发现在一个商品中频繁出现而在另一个商品中很少出现的评论以此作为区别两个商品特征的依据,并为用户提供友好的挖掘结果展示。该系统自动挖掘前 k 个支持度差值最大的项集,免除用户设定参数的困扰,提高系统的实用性。最后,通过京东商城大量真实商品评论数据集验证了系统具有有效、灵活且友好等特点。

未来的工作将引入自然语言处理的方法^[15]到原型系统中,使其提取商品评论的同时进行去噪声处理,商品评论标签更加规范化,去除冗余评论,提高分辨商品评论有效性等指标,提高系统的可行性;并且将增加多个商品评论的对比,根据实验用户使用情况进行系统的调整。同时为了免除用户设定阈值的困扰,本文采用 top- k 定义对比度最大的前 k 个对比评论模式,根据图 4 结果观察得知,此种设定未满足最小化的需求,对比评论模式存在一定的冗余,我们会在未来的工作中针对模式最小化进行改进。

参考文献:

- [1] DONG G, JAMES B. Contrast data mining: concepts, algorithms, and applications [M]. Boca Raton: CRC Press, 2013: 3 – 12.
- [2] DONG G, LI J. Efficient mining of emerging patterns: discovering trends and differences [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 43 – 52.
- [3] RAN X, DUAN L, LYU G. A contrast learning based approach for customer reviews crawling from dynamic Web pages [C]// Proceedings of the 29th National Database Conference of China. Beijing: Science Press, 2012: 52 – 57. (冉熙璐, 段磊, 吕广奕, 等. 基于对比学习的动态网页商品评论获取方法 [C]// 第 29 届中国数据库学术会议论文集. 北京: 科学出版社, 2012: 52 – 57.)
- [4] ZHANG Q, WU Y, LI T, et al. Mining product reviews based on shallow dependency parsing [C]// SIGIR 2009: Proceedings of the 2009 International Conference on Research on Development in Information Retrieval. New York: ACM Press, 2009: 726 – 727.
- [5] HU M, LIU B. Mining and summarizing customer reviews [C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 168 – 177.
- [6] ARCHAK N, GHOSE A, IPEIROTIS P G. Show me the money!: deriving the pricing power of product features by mining consumer reviews [C]// KDD 2007: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 56 – 65.
- [7] DUAN L, TANG C, DONG G, et al. Survey on emerging pattern based contrast mining and applications [J]. Journal of Computer Applications, 2012, 32(2): 304 – 308. (段磊, 唐常杰, DONG G, 等. 基于显露模式的对比挖掘研究及应用进展 [J]. 计算机应用, 2012, 32(2): 304 – 308.)
- [8] LI Y, FAN M. E-mail categorization and filtering technology based on essential emerging pattern [J]. Journal of Nanjing University: Natural Science, 2008, 44(5): 544 – 550. (李艳, 范明. 基于基本显露模式的电子邮件分类与过滤技术 [J]. 南京大学学报: 自然科学版, 2008, 44(5): 544 – 550.)
- [9] LI J, LIU H, JAMES R D, et al. Simple rules underlying gene expression profiles of more than six subtypes of Acute Lymphoblastic Leukemia (ALL) patients [J]. Bioinformatics, 2003, 19(1): 71 – 78.
- [10] DONG G, FORE N. Discovering dynamic logical blog communities based on their distinct interest profiles [EB/OL]. [2014-12-10]. http://www.thinkmind.org/download.php?articleid=sotics_2011_2_10_30018.
- [11] DON A, ZHELEVA E, GREGORY M, et al. Discovering interesting usage patterns in text collections: integrating text mining with visualization [C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2007: 213 – 222.
- [12] LEUNG C, CARMICHAEL C. Exploring social networks: a frequent pattern visualization approach [C]// Proceedings of the 2010 IEEE Second International Conference on Social Computing. Piscataway: IEEE Press, 2010: 419 – 424.
- [13] LAVRAC N, JESENOVE D, TRDIN N. Mining spatio-temporal data of traffic accidents and spatial pattern visualization [J]. Metodoloski Zvezki, 2008, 5(1): 45 – 63.
- [14] KASER O, LEMIRE D. Tag-cloud drawing: algorithms for cloud visualization [EB/OL]. [2015-01-20]. <http://arxiv.org/abs/cs/0703109>.
- [15] ZHOU L, ZHANG D. NLPIR: a theoretical framework for applying natural language processing to information retrieval [J]. Journal of the American Society for Information Science and Technology, 2003, 54(2): 115 – 123.

(上接第 2720 页)

- [17] CAO L, LI F F. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes [C]// ICCV 2007: Proceedings of the IEEE 11th International Conference on Computer Vision. Piscataway: IEEE Press, 2007: 1 – 8.
- [18] WANG X, GRIMSON E. Spatial latent Dirichlet allocation [EB/OL]. [2014-10-10]. <http://www.ee.cuhk.edu.hk/~xgwang/papers/wangG07nips.pdf>.
- [19] NIU Z, HUA G, GAO X, et al. Spatial-DiscLDA for visual recognition [C]// Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2011: 1769 – 1776.
- [20] MACKEY L. Latent Dirichlet Markov random fields for semi-supervised image segmentation and object recognition [EB/OL]. [2014-10-10]. <http://stanford.edu/~lmackey/papers/lmdmf-cs281a07.pdf>.
- [21] GUO Q, LI N, YANG Y, et al. LDA-CRF: object detection based on graphical model [J]. Journal of Computer Research and Development, 2012, 49(11): 2296 – 2304. (郭乔进, 李宁, 杨育彬, 等. LDA-CRF: 一种基于概率图模型的目标检测方法 [J]. 计算机研究与发展, 2013, 49(11): 2296 – 2304.)
- [22] WELLING M, TEH Y W, KAPPEN H. Hybrid variational/Gibbs collapsed inference in topic models [EB/OL]. [2014-10-10]. <http://arxiv.org/ftp/arxiv/papers/1206/1206.3297.pdf>.
- [23] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C]// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2006: 2169 – 2178
- [24] FAN R E, CHANG K W, HSIEH C J, et al. LIBLINEAR: a library for large linear classification [J]. Journal of Machine Learning Research, 2008, 9(12): 1871 – 1874.