用户评论中产品特征的抽取及聚类①

韩雪婷, 李 炜, 沈奇威

(北京邮电大学 网络与交换技术国家重点实验室, 北京 100876)
(东信北邮信息技术有限公司, 北京 100191)

摘 要: 在用户评论中蕴含了大量的产品特征和用户对这些特征的观点和态度. 本研究提出了基于 Apriori 关联规则算法的产品特征抽取方法, 利用与种子特征集合的互信息和与观点词的共现度对候选特征进行过滤; 并提出了一种特征自动聚类方法, 以特征词间的字符串相似度和语义相似度以及特征所对应的观点词作为衡量产品特征之间关联程度的特征, 采用 K-means 聚类算法对产品特征进行聚类. 本研究采用大众点评网对美食店铺的评论语料, 对该方法进行了数据实验,实验结果初步验证了该方法有效性.

关键词: 用户评论; 产品特征; 特征抽取; 聚类; 观点词

Extracting and Clustering Product Features from User Reviews

HAN Xue-Ting, LI Wei, SHEN Qi-Wei

(State Key Lab, Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)
(EBUPT Information Technology Co. Ltd, Beijing 100191, China)

Abstract: User Reviews contains a large number of product features and user's opinions towards these features. This paper proposed an approach to extract product features, which is based on Apriori algorithm, and using PMI with the seed set and co-occurrence degree with opinion words to filter features. And then an approach to group product features based on K-means algorithm is proposed, in which sharing words, lexical similarity and opinion words are chosen as the tokens to represent the association of product features. With the Chinese reviews of restaurants from the Internet, experimental results demonstrate the validity of the proposed method.

Key words: user reviews; product features; feature extraction; clustering; opinion words

1 引言

随着 web2.0 的飞速发展, 在电子商务网站、专业论坛等网络站点中出现了大量的针对产品或服务的用户评论, 这些用户评论中蕴含了丰富的信息, 对消费者的行为模式产生了深刻的影响, 越来越多的证据表明, 评论信息影响到消费者的购买决定[1,2].

数据挖掘技术越来越多的被用在去发现海量互联网信息中有用的信息, 包括聚类、分类和关联规则等技术[3]. 通过对产品的评论进行分析, 可以挖掘出这些产品的主要特征, 进一步发现用户对这些特征的意见和态度, 让商家更好的了解顾客反馈, 以便更好地调整经营策略, 让用户无需翻阅大量评论就找到自己关注的特征信息, 以帮助做出可靠的决策. 所以, 以获取产品评论中有用信息为目标的非结构化数据挖掘技术—“评论挖掘”, 已经成为一项重要的研究课题[4,5].

评论中的产品特征过于繁多, 同类的特征可以有多种描述, 如“口味”和“味道”描述的就是同一类特征, 对提取出的众多特征进行自动聚类, 可以方便用户和商家进行浏览和总结, 也利于后续的情感分析. 本研究制订了以下的用户评论挖掘任务:

a) 提取评论中的产品特征, 并找到对每个特征的评论观点;

b) 依据产品特征和评论观点的对应关系对产品特征进行聚类.

① 基金项目:国家自然科学基金(61072057,61101119,61121001,60902051);长江学者和创新团队发展计划(IRT1049);国家科技重大专项(2011ZX03002-001-01)
收稿时间:2012-10-26;收到修改稿时间:2012-11-26

2 研究背景

2.1 产品特征抽取的相关方法

已有不少学者对英文评论中产品特征的自动抽取进行了研究, 具有代表性的工作有 Hu M 和 Liu B[6,7]等人提出的应用关联规则提取特征词, 根据候选特征词的共现识别高频特征词, 然后利用修剪规则来提高准确率和覆盖率; Popescu A M 等人[5]抽取评论中频繁出现的名词和名词短语作为候选产品特征, 借助搜索引擎计算点互信息值(Point-Wise Mutual Information, 即 PMI) 来对候选特征进行评估, 利用贝叶斯分类提取产品特征, 提高了 Hu 的准确率, 但是覆盖率却有所下降, 同时由于利用搜索引擎计算 PMI 值效率较慢.

近些年在中文评论中特征提取方面也有了很大的进展, 如李实, 叶强等将 Hu[7]的方法针对中文产品评论的特点做了一些修改, 在挖掘中文产品特征时也取得了较好的效果[4]; 黄永文[8]通过定义一些常见的产品特征和观点词, 并生成特征和观点的表达模式, 采取 Boot-strapping 方法迭代抽取新的产品特征和观点词以及新的表达模式.

2.2 产品特征归类的相关方法

在产品特征词归类方面, 也有一些学者做出了研究. Guo 等人[9]提出了 mLSA 无监督算法, 根据上下文信息建立潜在语义关联模型, 进行产品特征归类. Zhai 等人[10]提出了一种半监督的 SC-EM 算法进行特征归类, 并通过对比实验证明了该算法的可行性和优异性. 杨源[11]等人在 Zhai[10]提出的 SC-EM 算法上进行改进, 抽取评论中特征词和观点词的搭配关系, 形成一个二部图, 然后用权重标准化 SimRank 计算不同特征之间的相似度, 并把所得的结果与贝叶斯分类器进行融合, 得到了更好的分类结果. 张姝等人[12]以语素和观点词作为特征, 利用 K-Means 方法对特征进行自动聚类.

3 用户评论中产品特征的抽取及过滤

在实际应用中, 用户关注和评论的对象往往为产品的某些特征, 将产品的信息按照产品特征进行组织是十分有必要的, 而这当中一个必不可少的工作就是产品特征的抽取. 产品特征包括产品的属性或功能、产品的部件、产品部件的属性或功能、产品的相关概念等, 一个特征词需要满足以下三个条件之一: 1)是给定主题的一部分; 2)是给定主题的一个属性; 3)是给定主题的部分的一个属性[13]. 举例: 如“这里的服务很好”中“服务”就是一个产品特征.

本文选择名词或名词短语作为产品特征, 按照图 1 所示步骤进行特征的抽取及过滤.

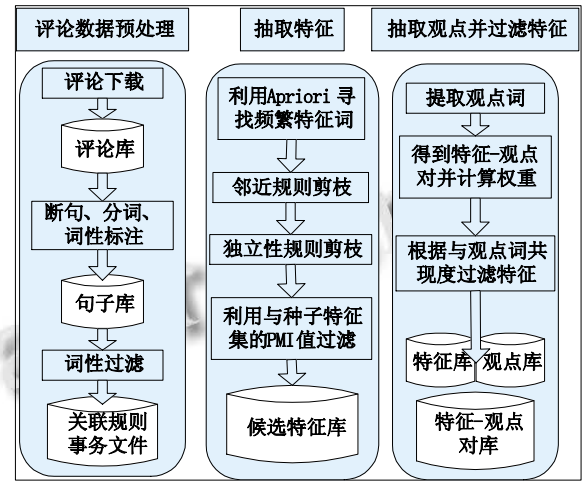

图 1 产品特征抽取的框架

3.1 评论数据预处理

本文以评论中的句子为单位提取产品特征, 首先对爬虫得到的评论语料进行断句处理, 按照评论中出现的标点符号(句号、逗号等)、空格符等进行断句. 应用中科院分词器 ICTCLAS2011[14]对评论中的句子进行分词、二级词性标注得到句子库, 并选择名词或名词短语(包括动名词、形容词性名词)作为产品特征, 通过词性过滤得到关联规则的事务文件.

在分词过程中, 进行了停用词处理, 将出现较为频繁而非产品特征的常见的人称名词(如“家人”、“客户等”)加入停用表中, 同时将用户评论中出现的专有名词, 如菜名、餐厅名等加入到用户词典中, 以提高分词准确率和后续计算的效率.

3.2 利用 Apriori 提取特征词以及特征词的修剪

本文利用 Apriori 关联规则算法, 利用上一步生成的事务文件, 提取满足最小支持度的频繁项集作为产品的候选特征. 只考虑 3 项及其以下的频繁项集, 对于频繁一项集采用的最小支持读为 0.5%, 2、3 项集采用的最小支持度较一项集的最小支持度更小, 本文采用 0.2%.

得到频繁特征词之后, 需要利用邻近规则和独立性规则进行剪枝, 过滤掉一些非特征词.

定义 1: 邻近规则

令 f 是一个频繁特征词且 f 含有 n 个单词, 假定 f

在句子 s 中按(w1 ，w2 ，…，wn)顺序出现且任意两个相临的词之间的距离不超过 3 个单词，则称 f 在 s 中是邻近的. 假如 f 在 m 个句子出现，在 n 个句子中邻近，若 n/m>α 且 n>2，则我们称 f 是一个邻近的特征短语，否则过滤之. 其中阈值 α 根据实验选取.

定义 2：独立性规则

在评论中一个特征词 f 的独立支持度(pSupport)是包含 f 的且不包 f 的父集作为频繁特征项的句子数量. 若 f 在 m 个句子中出现，在 n 个句子中独立出现，若 n/m>β 且 n>3,则 f 符合独立性规则，否则过滤之. 其中阈值 β 根据实验选取.

3.3 利用与种子特征集合的 PMI 过滤特征

在评论句子中存在一些名词或名词短语频繁出现，却与美食商铺不太相关，不是真正的美食特征，例如：“话”、“关系”、“状元”等，这些词需要过滤掉. 本文通过互信息 PMI 去衡量一个词与美食领域的相关性，首先人工从频繁出现的特征词中选出 6 个具有代表性的特征词组成种子特征集合，包括：味道、价格、环境、服务员、菜、餐厅, PMI 的计算公式如下:

$$PMI(w1)=\sum_{w\in Seeds} log_2 \frac{hits(w1,w)}{hits(w1)*hits(w)}$$

其中 *Seeds*={味道，价格，环境，服务员，菜，餐厅}，*hits*(*w1*,*w*)为候选特征词 *w1* 和种子特征词 *w* 在评论语料库中共同出现的次数，*hits*(*w1*)、*hits*(*w*)为词 *w1*、*w* 单独出现的次数. 高的互信息值意味着强关联关系，设定一个阈值 γ，将 PMI 值与 γ 比较，若大于等于 γ 则该词为特征，否则过滤之. γ 的取值由实验确定.

在有限的语料中候选词与种子特征词的共现可能是不均匀的，为了弥补语料库的不全面，提高准确度，可以借助搜索引擎 API 来检索词条，以返回的页面数作为词条出现的次数.

3.4 提取观点词，并利用特征与观点词的共现度过滤特征

在用户评论中，出现产品特征的句子中往往会伴随着对该特征的评价性的观点词，观点的抽取可以按照一定的规则从抽取特征词附近的形容词(包括形容词性动词和形容词性的名词，如“推荐”、“爱吃”、“麻烦”等)入手. 由于两者的共现关系，我们将同时出现的特征和观点叫做“特征-观点对”，用<f,o>表示. 在评论语句中除了特征和观点以为，还有一些因素也很重要，即表示程度的词(d)和表示否定语义的词(n)，故可以用 pair<f,o,d,n,weight>表示一个具体的特征-观点对，其中 weight 表示<f,o>匹配的权重，只有在 weight 大于阈值 0.25 的情况才认为<f,o>提取正确. 之后对所有提取<f,o>进行频次统计，用 freq 表示.

由于特征和观点经常是成对出现的，对于候选特征词 f，在其上下文出现观点词的比例越高，那么它为特征的可能性越大. 设定：f 与观点词同现的次数即由 f 提取出的 pair 数为 n，含 f 的句子数即 f 的独立支持度为 m，将 n/m 与阈值 δ 进行比较，若大于阈值 δ 则该候选特征是产品特征，否则过滤之. δ 的取值由实验确定. 具体的特征-观点对提取及特征过滤见图 2.

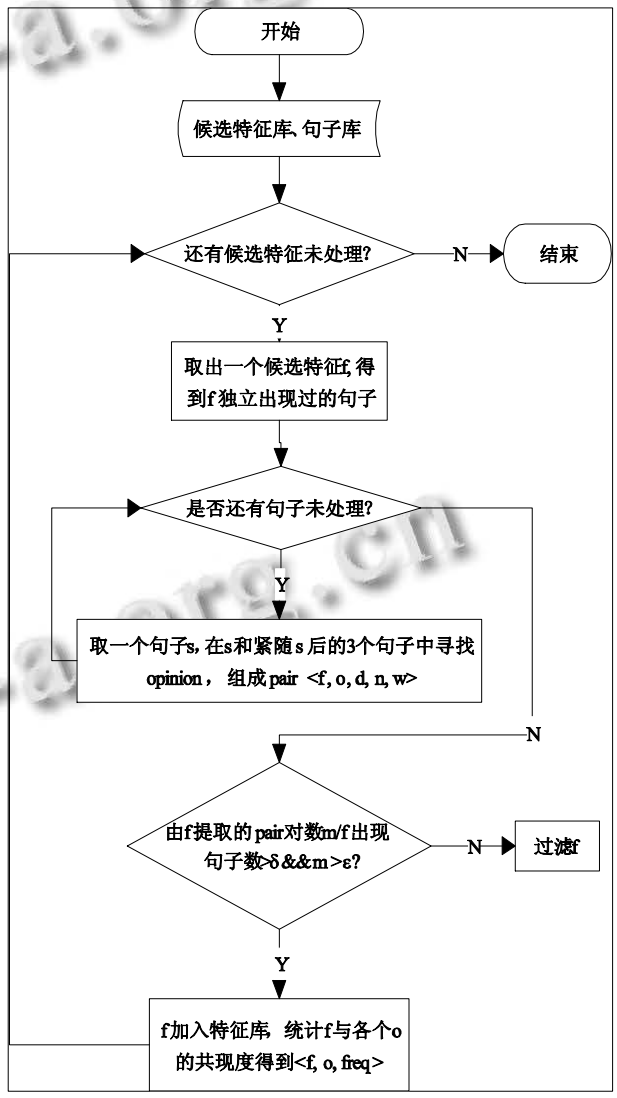

图 2　特征-观点对提取及特征过滤算法框架

具体的观点词匹配规则为:

(1) 若一个特征词已经与一个观点词匹配上，则

匹配上其他观点词的概率将减少, 故每匹配上一个观点词, weight 减去 0.3. 但如果存在并列的观点词, 则并列观点词匹配的概率相同. 如“装修古朴大方”中“古朴”和“大方”以相同的概率匹配.

(2) 对于前面有“的”字出现的特征, 优先向前匹配观点词, 若匹配到则 weight 减去 0.5, 否则优先向后匹配观点词. 如“悠闲的去处”, 向前匹配得到观点词“悠闲”.

(3) 若句子结尾是句号、问号等表示句子结束的符合, 则继续匹配到观点的概率将很低, 故 weight 减去 0.6, 若是空格、逗号等则 weight 减去 0.3.

(4) 观点词有时候可能出现在该句的后续几个句子中, 如“今天去吃了那的汽锅鸡, 很不错,我喜欢!”, 特征为“汽锅鸡”, 匹配的观点词有“不错”、“喜欢”. 故本文考虑了特征出现的句子后续的 3 个句子来提取可能观点. 当后续的句子中有名词出现时, 则该句子中的观点匹配的概率将降低, 故 weight 减去 0.2.

4 产品特征的聚类

在评论中, 特征词和观点词经常是成对出现的, 对于同类特征, 其对应的观点词往往是相同或相近的, 所以我们选择修饰特征词的观点词作为聚类特征. 含有相同字的或者是同义词的两个属性往往是同类属性, 所以在计算两个特征相似度时, 我们还考虑了特征词的字符串相似度和语义相似度, 本文借助 HowNet 和《同义词词林》[15]计算语义相似度.

本文采用 K-means 聚类算法对产品特征进行聚类, 运用向量空间模型(VSM)表示产品特征信息, 把产品特征向量化为 feature (f1,f2,…fn,o1,o2,…om), 其中 f1 到 fn 表示提取出的所有特征词, o1 到 om 表示提取出的所有观点词, f1-fn 的权重由 feature 与 fi(i∈[1,n])的字符串相似度和语义相似度联合计算, o1-om 的权重由 feature 和 oj(i∈[1,m])的点互信息表示, 计算公式如下:

$$PMI(f,o)=\log_2\frac{p(f,o)}{p(f)*p(o)}$$

其中, $p(f,o)$是观点词 f 和特征词 o 在语料中的联合概率, $p(f)$和 $p(o)$是特征词 f 和观点词 o 分别在语料中出现的概率.

5 实验分析

5.1 实验数据及评价方法

实验语料是来自大众点评网站[16]的 20710 句对美食店铺的评论, 从中选择 5 家店铺的评论数据进行分析, 每家店铺含有 100-300 条评论. 针对每家店铺的全部评论,人工提取和标注评论中所提到的产品特征.

按照前面提出的方法采用 JAVA 语言构造实验系统. 为了评估方法的性能, 利用了以下指标: 准确率(*Precision*), 覆盖率(*Recall*), *F-measure*, 公式如下:

$$Precision=\frac{\text{正确识别的产品特征数目}}{\text{识别的产品特征数目}},$$

$$Recall=\frac{\text{正确识别的产品特征数目}}{\text{产品特征总数}},$$

$$F\text{-}measure=\frac{2*Precision*Recall}{Precision+Recall}$$

5.2 实验结果

通过对 5 个店铺(分别为石屏会馆、福照楼汽锅鸡饭北大门店、嘉华饼屋、桥香园过桥米线文化街店、一颗印东风西路店)的评论数据进行实验, 我们得到产品特征提取的平均准确率 88.0%和平均覆盖率 66.3%, 证明了本实验方法的有效性. 且从表 1 中可以看出, 利用与种子特征集合的 PMI 过滤特征以及利用与观点词的同现过滤特征有助于提高特征提取的准确率.

表 1 特征提取及过滤实验结果

商户	人工标注特征数	关联规则及剪枝后特征提取准确率	PMI 过滤后准确率	观点词过滤特征后准确率	覆盖率	F 值
1	106	83.7%	85.7%	88.9%	67.0%	76.4%
2	102	82.6%	83.5%	88.5%	66.7%	76.1%
3	94	77.6%	79.5%	84.3%	60.6%	70.5%
4	98	80%	81.1%	88.7%	65.3%	75.2%
5	107	81.3%	82.1%	89.5%	72.0%	79.8%

表 2 中列出了几个有代表性类别的特征聚类结果, 证明了特征聚类方法的有效性.

表 2 特征聚类结果

类别	代表性特征词
1	甜味 味道 滋味 口味 味 香味 味,盐 味道,饭 口味,总体 味道, 烧烤 云南,口味 味道,牛肉 味,奶
2	装修 气氛 环境 楼,环境 地方,环境 店面,环境 卫生,环境
3	份量 分量 量
4	肉质 肉 鱼肉 猪,肉 肉,酱 肉,骨 肉,鸭 肉,香 肉,虾肉,鱼 皮,肉
5	材 材料 食材 材,食

6 结论

用户评论中蕴含了大量有价值的信息, 识别出用户关注的产品特征并将产品信息按照特征进行组织, 可以满足不同用户的信息需求. 但是现有的大部分特征提取方法得到的特征之间缺乏逻辑关联, 致使挖掘结果可读性较差, 所以对产品特征进行聚类对于产品评论挖掘结果的汇总和展示十分重要.

本文专注于解决用户评论中产品特征的提取及聚类问题, 并将研究结果应用在大众点评网评论挖掘中, 取得了不错的效果, 且系统具有良好的可移植性. 接下来, 我们将致力于提高产品特征识别的准确率和召回率, 并将研究观点的情感倾向判断等问题.

参考文献

1 郁亚辉,张明,袁方,王煜.产品评论挖掘研究综述.山东大学学报(理学版),2011,46(5):16–23,38.

2 Senecala S, Nantela J. The influence of online product recommendations on consumers' online choices. Journal of Retailing,2004,80(2):159–169.

3 Ni P, Liao JX, Wang C, Ren KY. Web information recommendation based on user behaviors. 2009 WRI World Congress on Computer Science and Information Engineering, 2009,4:426–430.

4 李实,叶强,李一军,Law R.中文网络客户评论的产品特征挖掘方法研究.管理科学学报,2009,12(2):142–152.

5 Popescu AM, Etzioni O. Extracting product features and opinions from reviews.Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Proc.Stroudsburg, DA, USA: Association for Computational Linguistics,2005:339–346.

6 Hu MQ, Liu B. Mining and summarizing customer reviews. Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM Press,2004:168–177.

7 Hu MQ, Liu B. Mining opinion features in customer reviews. Proc. of 9th National Conference on Artificial Intelligence. Men lo Park, CA, USA: American Association for Artificial Intelligence, 2004:755–760.

8 黄永文.中文产品评论挖掘关键技术研究[博士学位论文].重庆:重庆大学,2009.

9 Guo H, Zhu H, Guo Z, Zhang X, Su Z. Product feature categorization with multilevel latent semantic association.Proc. of CIKM.2009:1087–1096.

10 Zhai ZW, Liu B, Xu H, Jia PF. Grouping product features using semi-supervised learning with soft-constraints. Proc. of the 23rd International Conference on Computational Linguistics (COLING-2010). 2010: 1272–1280.

11 杨源,马云龙,林鸿飞.评论挖掘中产品属性归类问题研究.中文信息学报,2012,26(3):104–108.

12 张姝,贾文杰,夏迎炬,孟遥,于浩.产品属性归类技术研究.第六届全国信息检索学术会议论文集,2010.

13 Yi J, Niblack W. Sentiment mining in Web Fountain. Proc. of the 21st International Conference on Data Engineering (ICDE 2005). Washing,DC,USA:IEEE Computer Society Press,2005:1073–1083.

14 http://www.ictclas.org/

15 HIT-IRLab-同义词词林(扩展版).哈尔滨工业大学信息检索研究生: http://ir.hit.edu.cn/

16 www.dianping.com

(上接第 132 页)

国防科技大学学报,2010,32(4):105–109.

23 杨亚让.基于 X3D 的虚拟现实全景技术设计.绵阳师范学院学报,2009,28(2):82–85.

24 顾大权等.基于纹理技术生成立方体表面全景图的算法.系统仿真学报,2009,21(19):6140–6143.

25 祝剑锋等.基于 CT 断层数据的仿 X 线全景图生成技术.生物医学工程学杂志,2011,28(6):1189–1193.

26 王海颍,秦开怀.一种全景图构造与全局调整的新方法.系统仿真学报,2010,22(8):1908–1911.

27 Foote J,Kimber D.FlyCam:Practical panoramic video and automatic camera control[C].Proc. of International Conference on Multimedia and Expo.New York,USA:IEEE Press, 2000:1419–1422.

28 周金广等.PTZ 自主跟踪中的全景视频生成.中国图象图形学报,2011,16(1):110–117.