基于最大频繁模式的在线评论标签抽取 [1]

李良强　徐华林　袁华*　邵培基

（电子科技大学 经济与管理学院，四川 成都 610054）

摘　要　随着电子商务的蓬勃发展，日益增加的海量在线评论数据影响了在线消费者获取有用信息的效率。本文提出了一种从海量在线评论数据中自动抽取评论标签的文本处理技术。该方法综合了语法搭配（词性搭配）和最大频繁模式，从海量的在线评论数据自动抽取出用户对该产品/服务的主要评论内容。实验表明基于词性搭配的最大频繁模式的过滤技术能灵活有效地从在线评论中抽取核心的用户评论。并且，该方法在不同的数据集和不同方法的比较中，都表现出较好的抽取效果。

关键词　文本挖掘，最大频繁模式，评论标签，模式过滤

中图分类号

1. 引言

随着互联网和Web 2.0 技术的发展，在线用户评论对电商和消费者影响重大。以往研究表明多数消费者在做出购买决策前会在网上收集产品/服务相关信息，并且购买后会在网上分享消费体验和购买评价[1]。另一方面，大量网络用户在购买产品/服务前会阅读用户评论，并且受到评论内容的影响[2]。因此，海量的在线用户评论是帮助消费者发现产品质量和做出相应购买决策的重要信息源。然而，随着在线评论数量的增加，不可避免地出现了信息过载的问题。例如，部分畅销产品拥有成千上万条评论数据，过多的数据让消费者难以从中获得准确的核心产品评价信息。因此，合理地抽取核心信息并制作成信息标签（Tag），对潜在用户的信息检索行为起到高效的归纳和引领作用（如图1中“买家印象”）。

图 1 JD.com 用户在线评论标签示例

一般地，信息标签在内容上有三方面的要求：用户生成（意见可信性），大多数人的意见（避免意见的过度有偏性）以及保留足够的语义（易于理解）[3]。从用户评论的文本数据中抽取出核心信息满足了“用户生成”要求。将抽取过程实现自动化，无疑会增加这一方

[1] 基金项目：国家自然科学基金（71271044/U1233118）
通信作者：袁华，电子科技大学经济与管理学院，副教授，E-mail:yuanhua@uestc.edu.cn。

法的效率，因而许多人工智能方面的方法被应用到这个领域[2][4][5]。自动文本信息抽取方法的共同点是把用户的评论内容表现为一堆自然语言句子或者语言字符的集合，然后用机器学习方法来形成信息标签[6][7][8][9][10]。由于其较高的自动化能力，并且增加了对集合中词（项）的位序关系信息，逐渐得到研究者和应用领域的重视。机器学习方法的缺点，需要大量“优质”数据进行训练，显然，在网络中由海量背景各异的用户生成的评论文本质量难以保证。因此，在机器学习的基础上，需要再加以专家修正（通常是标注）以增强结果可用性。另外，时间消耗高也是机器学习方法的弱点。为了在海量在线文本中形成“大多数人的意见”，频繁集挖掘方法表现出很高的效率[11]。但是，简单地应用频繁集挖掘方法在生成信息标签时面临两个问题：一是该方法主要考查词汇的相关性（Correlation），亦即共现（Co-occurrence）频率；而在语义理解中非常重要的词汇之间的位序关系却被忽略。二是，频繁集方法仅以频率的高低来评价生成模式质量的好坏。这样，项数少的集合具有优势。但是自然语言的理解上，显然项数多的集合能够提供更多的语义信息。

为了生成具有语义的信息标签，需要尽可能多地保留频繁集中的词汇并且能够梳理频繁集中词汇的语义关系。因此，本文提出一种基于最大频繁模式的在线用户评论标签抽取方法。该方法包括两个步骤：(1)利用最大频繁模式（Maximal Frequent Pattern, MFP）挖掘评论中的频繁词集。(2)利用词性搭配规则过滤出拥有语义信息的频繁词集合，尤其代表评论的客体和评论者情绪的词性搭配组合。实验结果表明该方法能灵活有效地从用户在线评论中抽取出关键信息标签。本文结构组织如下：第二部分为介绍相关工作；第三部分是阐明方法的结构框架和相关细节；第四部分为实验结果展示和相关分析；第五部分是结论。

2. 相关工作

从海量在线评论中抽取标签的相关工作主要集中在评论中的“评价对象+情感倾向”的特征提取、观点摘要和标签生成三个方面。

在特征提取方面，除了使用人工定义抽取特征外[12]，主要是使用各种统计方法和机器学习的方法自动从大量评论中提取出评价对象和情感倾向的特征。一类工作是从抽取用户评价对象及其评价词出发。Hu 和 Liu 抽取评论中的名词和名词短语作为产品特征，并利用关联规则抽取与其相近的形容词，从而得到评论中的用户观点[5]。Popescu 等通过计算名词短语与所要抽取评价对象的分类的点间互信息（Point Mutual Information，PMI）来评价名词短语[13]。Zhu 等通过评价对象种子集出发，计算每个候选评价对象中的词的共现频率，接着不断应用 Bootstrapping 方法挑选候选评价对象[14]。王洪伟等选取词性、词性组合、N-gram 作为情感文本的潜在特征项对中文网络评论特征进行选择[15]。Jin 和 Ho 等使用词汇化的 HMM 模型来学习抽取评价对象和评价词的模式[16]。另一类工作是通过解析句子的依存关系以确定评价词修饰的对象[17][18][19]。Chen 等在比较了基于规则的和基于统计的方法之后发现，条件随机场在挖掘客户评论时具有较高的精度[22]。Mukherjee 和 Liu 从用户提供的评价对象种子集开始，应用半监督联合模型不断迭代，产生贴近用户需要的评价对象[23]。Brody 和 Elhadad 则先使用主题模型识别出评价对象，再考虑相关的形容词作为评价词[24]。

观点摘要的任务就是通过用户评论对用户感兴趣的产品特征进行情感总结[5][13]。在情感分析领域，其中一些研究已经完成对评论文本的情绪分析[4]和主观分析[25]。Li 等将产品特征提取、评论萃取、极性侦测综合成一个连续的标注问题来摘要产品评论特征[26]。

Zhang 利用条件随机场来识别产品的特征，并利用语法方面的关系来识别和获取用户对产品特征的情感[27]。Miao 等人将高词频的名词作为产品特征，利用语法规则将与之对应的高频的动词、副词、形容词并结合人工的方式识别观点[28]。

在标签生成方面，Gupta 等从社会标签的角度阐述了标签的产生、属性、模型、语义和应用方面的问题[29]。李丕绩等提出了为针对<主题词（n），（ADVs），修饰语（a）>三元组句子中的依存关系，抽取关键标签的方法[30]。吕海燕等从微博内容中选取名词作为候选关键词，进行自底向上的层次聚类，然后选取簇代表词扩展生成用户兴趣标签[31]。Lappas 提取出的代表性评论集必须涵盖所有评论中涉及到的商品属性[32]。Tsaparas 等提出取出的代表性评论，不但涵盖所有被提及的属性，而且包含了各个属性的所有正、负面的评价[33]。Nazi 等提出了关于提取用户标签的三项指标：相关性(Relevance))、覆盖度(Coverage)和情感极性(Polarity)，根据这三项指标对生成的标签进行优化[34]。

3. 研究方法

本文的抽取系统主要包括数据预处理、最大频繁模式挖掘、词性搭配模式标注和用户评论标签的生成。下面图 2 说明了整个研究的流程。其中，数据预处理是文本挖掘相关的经典方法，主要是去除噪音并把有语义的文本转化成一个事务数据记录。预处理生成的事务数据将被用于挖掘最大频繁集。从原始的评论数据中标注出词性搭配模式可以从最大频繁集中抽取出有意义的用户评论标签。

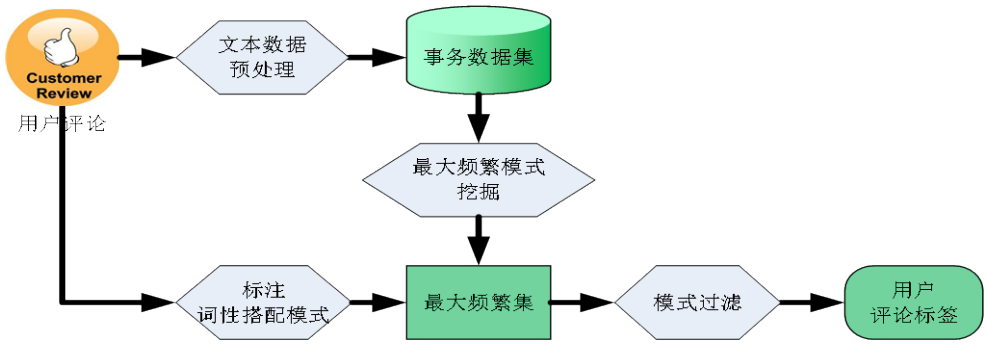


图 2 在线用户评论抽取技术框架.

3.1 文本数据预处理

文本数据预处理主要包括分词和词性处理。

假定 C 代表所有用户评论的句子集合，给定一个用户评论 $c \in C$,（c 是由句子的形式表达）。分词就是运用工具软件将句子 c 中的基本词元素标注并拆分。最后用词项 $\boldsymbol{x}_c=\{x_1, x_2,\ldots\}$表示拆分结果。其中的词元素满足关系$x_i \neq x_j$，如果$i \neq j$。例如，有评论句 c=“嗨，酒店很好!”，可以被拆分成词项集 $\boldsymbol{x}_c$={“嗨”，“酒店”，“很”，“好”}。值得说明的是，一个句子的词项集生成与分词软件算法相关。

分词之后，需要进行词性处理来删除没有意义的词，让词项集 $\boldsymbol{x}_c$ 中只包括一些重要的观点词。例如，语气词“嗨”可以从前面的例子项集 $\boldsymbol{x}_c$ 中删除，因为它不能给评论阅读者带来有价值的信息。为了高效率地从在线用户评论中抽取出用户观点，我们通过只保留观点词的方式来降低计算复杂度。其中观点词包括名词、动名词、副词、形容词和实体名，因为这些词更可能传递出评论者的观点[35][36]。另外，为了从观点词中获得丰富的语义信息，必须先区分出观点词的相关词性。人们可以利用 Monty Lingua Python library 对用户评论句

子进行分词和词性标注[37]。经过词性处理和词性标注的用户评论句子 c 进一步可以被表示成为“词＋词性”形式的复合项集，如 $\boldsymbol{x}_c$={“酒店/n”，“很/d”，“好/a”}。其中，所有的词性组合称为词项集 $\boldsymbol{x}_c$ 的词性模式，表示为 $M(\boldsymbol{x}_c)$＝｛n，d，a｝。

分词是数据预处理最基础的工作，处理结果将产生一个事务数据集（Transactional data set）$T = \cup\{\boldsymbol{x}_{c_i}\}$，其形式见表 1：

表 1　评论文本转化成事务数据记录

ID 评论句子	词项集
c_1	$\boldsymbol{x}_{c1}$
c_2	$\boldsymbol{x}_{c2}$
…	…
$c_\#$	$\boldsymbol{x}_{c\#}$

3.2 最大频繁模式挖掘

最大频繁模式指的是一个频繁项集，并且其所有的超集都不是频繁的[38]。事实上，在文本相关数据挖掘中，最大频繁模式能找出事务数据集中那些较长的频繁项集，并且这些较长的频繁项集可能拥有更多的语义信息。

本研究中，我们旨在挖掘出由观点词构成的事务数据集中的最大频繁项集。因此，挖掘出的最大频繁集 MFP 中的任一个模式 mfp_i 也是一个基于词组合的项集。另外，挖掘算法不是本研究的关注点，因而通常的 MFP 挖掘算法均可适用。

3.3 信息标签的词性搭配模式要求

从集合 T 中挖掘出的最大频繁集可以被看做是一组词汇的组合搭配。但是，并非所有的组合搭配都是有意义的或者有益于评论标签的抽取。为此，我们需要根据在线评论文本中的语义表达，标注一些基本的词性搭配模式。

首先，如果最大频繁集只包含一个词项，无论它是任何词性，都不会包含评论标签抽取需要的目标语义信息。例如一个最大频繁集是{“酒店”}，这个最大频繁模式只能说明用户评论的目标是“酒店”，而没有任何与“酒店”有关的语义信息。所以，我们删除所有只包含一个词项的最大频繁集。

其次，最大频繁集中词项间至少有一种搭配关系，这样的最大频繁集才包含有价值的语义信息。因此，包含语义信息的最大频繁集中必须拥有代表评论客体的词性，比如名词。并且还需要包含能与名词进行合理搭配并且能表达用户评论意思的词性。信息标签在评价系统中的形式通常如图 1 所示，研究者将其内容构成定义为“对象特征＋评价”[24]，其中的“对象特征”指用户评价的对象，一般为名词。而“评价”指用户对评价对象的观点看法等，一般为修饰作用的副词和相容词。

最后，由于汉语是少形态或非形态的语言，词组装成短语，一靠词序，二靠虚词。并且由词与词搭配起来产生各种关系，但词与词的搭配不是任意的，它既要受词性的制约，也要受词义的制约[39]。所以在挖掘出的最大频繁集中既要考虑有意义的评论抽取也要考虑词汇搭配之间的合理性。通过对评论数据的观察并结合了《现代汉语实词搭配词典》抽取出了，评论中常用的搭配模式（RULE={ R_1, R_2, R_3}）如下：

- R_1：名词+（副词）+形容词，即 R_1＝｛n, d, a｝；
- R_2：（副词）+形容词+名词，即 R_2＝｛d, a, n｝；
- R_3：动词+名词，即 R_3＝｛“有 / 送 / 提供”, n｝。

最大频繁模式 mfp_i，作为潜在的评论标签其词性模式需满足：$\bigcup_{j=1,2,3} mfp_i \cap R_j \neq \emptyset$。

3.4 基于词性搭配的评论标签过滤算法

基于词性搭配的 MFP 过滤算法见表 2。

表 2 基于词性搭配的评论标签过滤算法

	算法 1 评论标签过滤
1	**Input**：最大频繁集 MFP，搭配模式集 RULE；
2	**Output**：潜在评论标签集 *TAG*。
3	*CustomerOpinion=Φ*;
4	**For** each *mfp*∈MFP **do**
5	**For** each *R*∈RULE **do**
6	**If** $mfp \cap R \neq \emptyset$ **then**
7	*CustomerOpinion*←*mfp*;
8	**End If**
9	**End For**
10	**End For**
11	**For** each *mfp*∈ *CustomerOpinion* **do**
12	*tag*←对 *mfp* 按照{n, d, a}或者{v, n}模式排序;
13	**End For**
14	**Return** ∪ $\{tag\}$.

算法主要步骤包括潜在最大频繁模式过滤（第 3-10 行）以及评论标签（第 11-13 行）的生成。考虑到 R_1 和 R_2 的结构在中文中表达的语义差别（尤其是语义对象）不大，例如“房间”＋“大”和“大”＋“房间”。最后，为了自动生成信息标签，我们都将其排列成 R_1 或者 R_3 模式（第 12 行）。

4. 实验结果

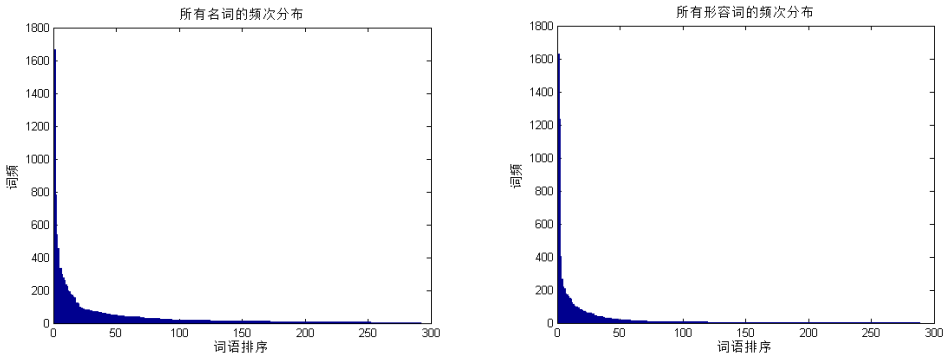
我们收集了来自携程网成都西藏饭店的用户在线评论数据。数据集包括 3119 条用户评论，评论时间介于 2012 年 7 月 02 日——2015 年 9 月 30 日之间，大部分评论数据由中国消费者用中文撰写。收集的数据包括用户 ID、评分、评价时间、评价（文本）内容等相关数据特征。最终的信息标签由评价内容生成，数据集中评价内容的分布情况如下（表 3）。

表 3 评价文本内容的分布

统计项目	统计值
文本数	3119
最大长度	565 字
最小长度	1 字
平均长度	33.6 字
标准差	42.7

4.1 分词

中文是一门表意语言，而且中文句子中词与词之间没有分隔符。因此分词是中文文本处理的第一步。实验中，我们选取了中国科学院计算技术研究所的分词软件 ICTCLAS 对评论数据进行分词和词性标注[40]。



(a) 名词的词频分布　　(b) 形容词的词频分布

图 3 数据集中名词和形容词的词频分布情况

图 3-(a)和(b)分别展示了其中名词和形容词的词频分布情况。可以看出极少数词占据了非常高的词频，剩下的词有一个典型的长尾分布。这说明用户在评价对象（名词）和评价（形容词）上用词都非常稀疏，在进行频繁集挖掘时需要设定较低的支持度阈值。

4. 2 最大频繁模式挖掘及用户评论抽取

由于 MFP 需要保留尽量多的语义元素，我们将评论数据集分三类做了对比实验：

(1) 直接分词后的评论数据集（data1）；

(2) 基于处理停用词后的评论数据集（data2）；

(3) 基于词性处理后的评论数据集（data3），

其中，data1 主要用于考察保留最多的原始数据参与运算，结果会不会有大的提升。消除停用词是自然语言处理中一种重要的数据预处理的方法，data2 用于考察停用词对计算的影响。实验结果比较见图 4。在最小支持度为 3%时，获得 7.9%带有语义信息的最大频繁模式，而对基于词性处理后的评论数据集进行 MFP 过滤算法挖掘则获得 70.2%带有语义信息的最大频繁模式。基于停用词处理后的评论数据集获得 19.6%带有语义信息的最大频繁模式。然而，算法效率问题，直接分词后的评论数据集的运算成本最高，而基于停用词处理后的评论数据集中保留的词数量也远远超过了基于词性处理后的数据集。

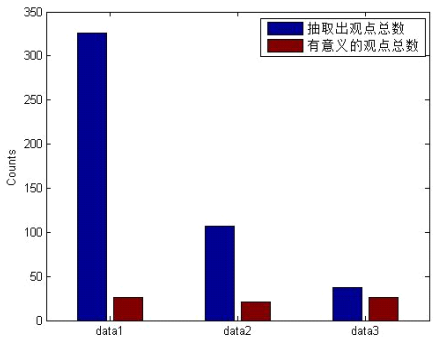


图 4 不同数据集上的结果比较

进一步实验比较了不同数据集的挖掘内容。基于词性搭配的 MFP 过滤算法挖掘出的带有语义信息的最大频繁模式和直接分词后进行 MFP 挖掘出模式差异不大。这是因为用户评

论用词的稀疏性，导致高频词在文本中的使用比较集中，因而保留更多语义元素的 data1 在相同的支持度阈值情况下，显得并无多少优势。而数据集 data2，由于去除了“有／没有”这两个在 R_3 中常用的动词相关组合，所以挖掘的带有语义信息的最大频繁模式略少于前两者。但是在酒店相关的评论中，“有早餐”，“有无线网络”成了人们决策的重要信息。与去除停用词的数据集的比较实验表明，MFP 加上词性过滤方法可以不用预先去除停用词。因为很多停用词是高频词，在 MFP 中可能增加语义，有利于对挖掘模式的理解。表 4 列出了基于词性搭配的方法从 135 条 MFP 中过滤出的 20 个信息标签。

表 4 基于词性搭配信息标签生成

MFP（支持度 ）	信息标签
便利/an 交通/n (0.85882)	交通便利
周到/a 服务/vn (0.587153)	服务周到
丰富/a 早餐/n (0.692314)	早餐丰富
方便/an 交通/n (0.753659)	交通方便/
有/tyc 特色/n (0.508281)	有特色
干净/a 房间/n (0.788713)	房间干净
老/a 酒店/n (0.578389)	酒店老
位置/n 不错/a (0.569626)	位置不错
位置/n 好/a (0.674787)	位置好
性价比/n 高/a 很/d (0.508281)	性价比很高
环境/n 不错/a (0.587153)	环境不错
环境/n 好/a (0.648497)	环境好
大/a 房间/n (0.701078)	房间大
小/a 房间/n (2.02436)	房间小
早餐/n 不错/a (0.83253)	早餐不错
早餐/n 好/a (0.63097)	早餐好
不错/a 服务/vn (1.41092)	服务不错
不错/a 酒店/n (2.2785)	酒店不错
服务/vn 好/a 很/d (1.35834)	服务很好
好/a 酒店/n 很/d (1.11296)	酒店很好

4.3 方法比较

为了进一步检验本方法抽取效果，分别选取了具有代表性的点互信息 PMI[41]和文档主题生成模型 LDA[42]与本文方法进行比较。三种方法均在词性处理后的评论数据集上进行的实验，其中 MFP 最小支持度设为 3%。

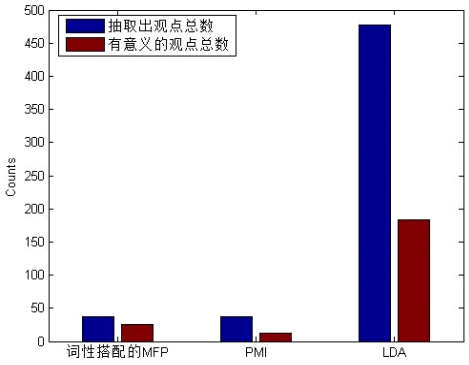


图 5 基于词性搭配的 MFP 与 PMI 和 LDA 的结果比较

从图 5 中，我们发现 LDA 方法找到的潜在评论最多（评论对象词＋评价词）。而 PMI 在相同数据集上找到的潜在评论数最少。LDA 方法虽然可以找出最多的用户潜在关注话题

和观点，但是其中的噪音非常多，需要人工进行总结筛选。PMI 由于主要集中找出观点对（pair），所以倾向找出两项词之间的关系，而基于词性搭配的 MFP 过滤算法不仅可以发现两项词之间的关系还可以发现两项以上词之间的关系。因此，从效率和结果上看，MFP 方法有较好的实用性。

事实上，MFP 过滤算法获得的结果与用户评价系统抽取的信息标签内容（示例见图 6）能相互对应。并且 MFP 过滤算法还发掘出如“酒店有特色”和“交通便利”等极有决策参考价值的信息标签。换言之，利用观点词词性处理过滤的方法除了获得相似的信息外，还能对实时系统进行补足，获得更加丰富和有用的结果。

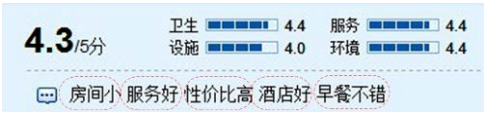


图 6 携程网的用户评价信息标签示例

5. 结论

多数在线商务平台都建立了用户评价系统，用户可以自由的分享自己对产品/服务的消费体验。因此在线用户评论会积极的影响潜在购买者的购买决策，在线用户评论对电商和消费者都非常重要。但与此同时，由于大量的用户评论之间背景差异大，导致消费者非常的迷茫，从而无法从用户评价中获得正确的产品/服务信息。

本文提出了基于观点词词性处理、词性搭配与最大频繁模式挖掘方法相结合的在线用户评论抽取系统，从海量的在线用户评论中抽取出核心的产品/服务的评价信息。实验结果表明基于观点词词性处理、词性搭配与最大频繁模式挖掘方法相结合的在线用户评论抽取系统能灵活有效地抽取出用户的核心评价。并且该方法在不同的数据集和不同方法的比较中，都表现出较好的抽取效果。

参 考 文 献

[1] Butler P, Peppard J. Consumer purchasing on the Internet:: Processes and prospects[J]. European Management Journal, 1998, 16(5): 600-610.

[2] Ye Q, Law R, Gu B. The impact of online user reviews on hotel room sales[J]. International Journal of Hospitality Management, 2009, 28(1): 180-182.

[3] Ames M, Naaman M. Why we tag: motivations for annotation in mobile and online media[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,ACM, 2007: 971-980.

[4] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1-135.

[5] Hu M, Liu B. Mining and summarizing customer reviews[C]. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004: 168-177.

[6] Eck D, Lamere P, Bertin-Mahieux T, et al. Automatic generation of social tags for music recommendation[C]. Advances in Neural Information Processing Systems,2008: 385-392.

[7] Mirizzi R, Ragone A, Di Noia T, et al. Semantic tags generation and retrieval for online advertising[C]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM,

2010: 1089-1098.

[8] Gupta M, Li R, Yin Z, et al. Survey on social tagging techniques [J]. ACM SIGKDD Explorations Newsletter, 2010, 12(1): 58-72.

[9] Yatani K, Novati M, Trusty A, et al. Analysis of adjective-noun word pair extraction methods for online review summarization[C]. IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 2011, 22(3): 2771.

[10] Liu B, Zhang L. A Survey of Opinion Mining and Sentiment Analysis [M]. Mining Text Data, Springer US, 2012: 415-463.

[11] Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions [J]. Data Mining and Knowledge Discovery, 2007, 15(1): 55-86.

[12] Kobayashi N, Inui K, Matsumoto Y, et al. Collecting Evaluative Expressions for Opinion Extraction[M]. Natural Language Processing–IJCNLP 2004, Springer Berlin Heidelberg, 2005: 596-605.

[13] Popescu A M, Etzioni O. Extracting Product Features and Opinions From Reviews [M]. Natural Language Processing and Text Mining, Springer London, 2007: 9-28.

[14] Zhu J, Wang H, Tsou B K, et al. Multi-aspect opinion polling from textual reviews[C]. Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009: 1799-1802.

[15] 王洪伟, 郑丽娟, 刘仲英, 霍佳震. 中文网络评论的情感特征项选择研究[J]. 信息系统学报, 2012, 01:76-86.

[16] Jin W, Ho H H, Srihari R K. A novel lexicalized HMM-based learning framework for web opinion mining[C]. Proceedings of the 26th Annual International Conference on Machine Learning, 2009: 465-472.

[17] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 2006: 43-50.

[18] Somasundaran S, Wiebe J. Recognizing stances in online debates[C]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics, 2009: 226-234.

[19] Kessler J S, Nicolov N. Targeting sentiment expressions through supervised ranking of linguistic configurations[C]. ICWSM, 2009.

[20] Zhu J, Wang H, Tsou B K, et al. Multi-aspect opinion polling from textual reviews[C]. Proceedings of the 18th ACM Conference on Information and Knowledge Management,. ACM, 2009: 1799-1802.

[21] Kessler J S, Nicolov N. Targeting sentiment expressions through supervised ranking of linguistic configurations[C]. ICWSM, 2009.

[22] Chen L, Qi L, Wang F. Comparison of feature-level learning methods for mining online consumer reviews[J]. Expert Systems with Applications, 2012, 39(10): 9588-9601.

[23] Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012: 339-348.

[24] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews[C]. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010: 804-812.

[25] Ghose A, Ipeirotis P G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics [J]. Knowledge and Data Engineering, IEEE Transactions On, 2011, 23(10): 1498-1512.

[26] Li F, Han C, Huang M, et al. Structure-aware review mining and summarization[C]. Proceedings of the 23rd

International Conference on Computational Linguistics,Association for Computational Linguistics, 2010: 653-661.

[27] Zhang S. Extracting product features and sentiments from Chinese customer reviews [C]. Proceedings of the 7th Int'l Conf. on Language Resources and Evaluation, 2010: 1142-1145.

[28] Miao Q, Li Q, Zeng D. Fine grained opinion mining by integrating multiple review sources [J]. Journal of the American Society for Information Science and Technology, 2010, 61(11): 2288-2299.

[29] Gupta M, Li R, Yin Z, et al. Survey on social tagging techniques [J]. ACM SIGKDD Explorations Newsletter, 2010, 12(1): 58-72.

[30] 李丕绩, 马军, 张冬梅, 等. 用户评论中的标签抽取以及排序[J]. 中文信息学报, 2012, 26(5): 14-19.

[31] 吕海燕, 张杰, 王丽娜. 基于聚类分析的微博用户标签自动生成[J]. 电子设计工程, 2015 (7): 67-69.

[32] Lappas T, Gunopulos D. Efficient Confident Search in Large Review Corpora [M]. Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, 2010: 195-210.

[33] Tsaparas P, Ntoulas A, Terzi E. Selecting a comprehensive set of reviews[C]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,ACM, 2011: 168-176.

[34] Nazi A, Das M, Das G. The tagAdvisor: luring the lurkers to review web items[C]. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015: 531-543.

[35] Li B, Chen J, Chen X, et al. Sentimental direction analysis: A framework for chinese sentiment computation and resource construction [J]. International Journal of Knowledge and Language Processing, 2011: 19-39.

[36] Benson, M. Benson, E., & Ilson, R. The BBI Dictionary of English Word Combinations [M]. Amsterdam: John Benjamins, 1997.

[37] Haizhou L, Baosheng Y. Chinese word segmentation [J].In Proceeding of Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 212-217.

[38] Pang-Ning T, Steinbach M, Kumar V, Introduction to Data Mining [M]. Addison Wesley, 2006

[39] 张寿康,林杏光.现代汉语实词搭配词典[M].北京:商务印书馆,1996.

[40] Gao J, Li M, Wu A, et al. Chinese word segmentation and named entity recognition: A pragmatic approach [J]. Computational Linguistics, 2005, 31(4): 531-574.

[41] 代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26-32.

[42] Blei, D.M., Ng, A.Y., Jordan, M.I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3), 993-1022.

Maximal Frequent Pattern Based Opinion Tag Extraction from Online Customer Reviews

Li Liangqiang, Xu Hualin, Yuan Hua*, Shao Peiji

(School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China, 610054)

Abstract With the development of online business, too many reviews published online by various people will make the review readers confused in getting right information. In this work, we propose a text processing techniques to extract useful customer opinions from massive online reviews, which is a combination method of linguistic (word collocation) with traditional maximal frequent pattern mining. It can be used to identify the common concerns about a product/service from mass customer review contents automatically. The experimental results show that the presented opinion words and word-collocation-based MFPs filtering methods can extract customer opinions from online reviews flexibly and efficiently.

Key words Text mining, Maximal frequent pattern, Opinion tag, Pattern filtering

作者简介:

李良强（1981-），男，四川广元人，电子科技大学经济与管理学院博士生，研究方向：电子商务、信息管理、商务智能等.E-mail：langmalee@gmail.com。

徐华林（1978-），男，四川双流人，电子科技大学经济与管理学院博士生，讲师，研究方向：电子商务，商务智能.E-mail：bruce123.xu@gmail.com。

袁　华（1973-），男，四川达州人，博士，电子科技大学经济与管理学院副教授，研究方向：电子商务，商务智能.E-mail：yuanhua@uestc.edu.cn。

邵培基（1946-），男，江苏南京人，电子科技大学经济与管理学院教授、博士生导师，研究方向：商务智能、信息管理与电子商务、营销创新.E-mail：shaopj@uestc.edu.cn。