

一种新型的中文产品评论微摘要方法研究

赵福青¹, 刘柏嵩^{2*}, 葛文镇¹

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 宁波大学 图书馆, 浙江 宁波 315211)

摘要: 提出一种新的无监督的方法, 对网络上存在的大量中文产品评论信息进行处理, 生成简洁的非结构化的可读性强且具有代表性、简洁性的理解式评论微摘要。用 N-gram 语言模型来衡量可读性, 用改进的点间互信息函数来衡量代表性, 用同义词词林来计算词语相似度; 将这种产品评论微摘要问题归结为优化问题, 试图寻找具有可读性和代表性的简洁、低冗余的词组, 并提出了一个启发式算法来解决这个优化问题。

关键词: 产品评论挖掘; 无监督学习; 微摘要; web N-gram

中图分类号: TP391.1

文献标志码: A

文章编号: 1001-5132 (2014) 04-0038-04

随着互联网技术和电子商务的蓬勃发展, 消费者的行为模式发生了巨大的改变, 越来越多的消费者开始选择网上购物。互联网在很大程度上也改变了消费者反馈产品评价的途径。区别于以往的购物模式, 消费者可以凭借网络上大量用户对产品的评价信息对产品的特征、性能有一个全面的了解。随着电子商务网站上产品的不断多样化及消费者数量的不断增长, 网络上各种产品评论数量也在飞速的增长。

对网络上存在的大量产品评论信息进行挖掘的意义主要体现在以下两个方面: (1)可以帮助消费者更加全面、综合地了解该产品的评价情况, 从而能全面地了解产品的性能评价, 对产品有更加详细的认识和了解, 更好地满足自己的需求; (2)生产厂商可以通过分析产品评论信息进一步了解消费者的需求, 明确自己的优势及不足, 便于改善产品的设计与服务, 提升竞争优势^[1]。

网络上产品评论信息的数量日益增多, 有些热门商品评论信息更是达到了上千万条, 然而其中有用的信息可能只是一部分。如果仅凭人工对产品评论进行分析处理, 将是一件非常繁琐和耗时的任务, 而且很难获取全部的有效评论信息。因

此, 需要提供一种有效的方法能够自动或者半自动地对评论信息进行分析 and 处理, 从而挖掘出对用户和生产厂商有用的信息, 由此, 评论挖掘研究应运而生。评论挖掘是一种以获取产品评论中的有用信息为目标的非结构化数据挖掘技术。

1 产品评论微摘要

产品评论微摘要是指对大量的产品评论信息进行处理, 生成简短的具有代表性和可读性的简洁摘要。产品评论微摘要在帮助用户消化网络上众多的评论信息时是非常关键的。

目前研究产品评论主要集中在生成高度结构化的摘要。主要将评论挖掘分为 4 个子任务——产品特征抽取、评论观点抽取、评论观点的极性以及强度判读、评论挖掘结果的汇总按用户观点排序^[1]。这类似于简单的对评论信息情感分类^[2-3]或者是一个多方面的等级评定^[4-5]。虽然结构化的摘要可以提供产品的情感信息, 但是这种摘要缺乏非结构化摘要所提供的详细信息, 这就使得用户经常需要返回原文获取更多的评论信息。这样非结构化的微摘要就显得比较重要。

然而, 生成理解式的文本摘要是非常困难的。

收稿日期: 2013-12-31.

宁波大学学报(理工版) 网址: <http://journalg.nbu.edu.cn/>

基金项目: 浙江省部级实验室开放基金 (B2014) .

第一作者: 赵福青 (1989-), 女, 山东临沂人, 在读硕士研究生, 主要研究方向: 数据挖掘与评论挖掘. E-mail: zhaofuqing90@126.com

*通信作者: 刘柏嵩 (1971-), 男, 安徽安庆人, 博士/研究员, 主要研究方向: 人工智能及计算机网络. E-mail: lbs@nbu.edu.cn

首先, 摘要需要能够代表原文的主要观点. 其次, 生成的摘要可读性要强, 这样才能更好地被用户理解. 再次, 随着人们对信息获取效率的提高, 摘要的简洁性也占据了非常重要的位置. 在本文中, 我们试图完成的任务是生成一组非常简洁的词组, 每个词组都是原文中的主要观点意见, 并且研究的重点是生成简洁的理解式摘要, 而不是抽取式摘要, 这就使得我们的摘要问题比较独特, 而以往此方面的研究比较少. Ganesan 等^[6](Opinosis)也试图生成理解式的简洁摘要, 他们利用依赖文本结构冗余的图数据结构来发现翔实的短语.

下面是某个预想的产品评论微摘要的例子:

产品: iPhone 5s

微摘要:

优点: 系统流畅, 运行速度快, 外观漂亮, 手感舒适, 摄像头表现出色, 系统功能强大.

缺点: 电池续航能力一般, 价格高, 屏幕尺寸小, 发热量大温度高.

2 微摘要优化表达式

我们的目标是使用一组微观点生成一个紧凑的、内容丰富的微摘要. 微观点是一个简短的词组, 用来总结文本中的关键评论.

问题定义: 汇总产品评论信息, 得到评论信息文本, 从而建立 1 组句子 $Z = \{z_i\}_{i=1}^n$, 目标是生成 1 个评论微摘要, $M = \{m_i\}_{i=1}^k$, 每个 $m_i \in M$ 表达了 Z 中的 1 个关键观点. 这里, 我们要求 m_i 中的词语至少在 Z 中出现 1 次, 但是并不要求 m_i 是 Z 中的原有序列. 这使得我们的研究工作更大程度上是理解式摘要的问题, 与抽取式摘要有着本质的区别.

直观上, 想要生成的评论微摘要 $M = \{m_i\}_{i=1}^k$ 需满足以下几个条件: (1) 代表性. 每个 m_i 应该能够反映原文本的主要观点意见; (2) 可读性. 生成的每个 m_i 应该具有良好的语法规则; (3) 简洁性. M 应该使用尽可能少的词语来表达主要的评论观点.

理论上, 我们可以列举所有可能的摘要, 并且评估每个摘要是否能满足这三大条件, 因此可将这个微摘要问题归结为如下的优化问题:

$$M^* = \arg \max_{M=\{m_1, \dots, m_k\}} \sum_{i=1}^k [S_{\text{rep}}(m_i) + S_{\text{read}}(m_i)], \quad (1)$$

使满足:

$$\begin{aligned} \sum_{i=1}^k |m_i| &\leq \sigma_{\text{ss}}, \\ S_{\text{rep}}(M) &\geq \sigma_{\text{rep}}, \\ S_{\text{read}}(M) &\geq \sigma_{\text{read}}, \\ S_{\text{sim}}(m_i, m_j) &\leq \sigma_{\text{sim}}, \forall i, j \in [1, k], \end{aligned} \quad (2)$$

其中, $S_{\text{rep}}(m_i)$ 是计算 m_i 代表性的函数; $S_{\text{read}}(m_i)$ 是计算 m_i 可读性的得分函数; $S_{\text{sim}}(m_i, m_j)$ 是计算 m_i 和 m_j 相似性的相似函数; σ_{ss} 和 σ_{sim} 是 2 个用户调节参数, 它们可以控制摘要的最大长度和摘要中能接受的冗余度; σ_{rep} 和 σ_{read} 是代表性和可读性的最小阈值, 能够改善算法的效率.

这种优化方案的基本原理如下:

首先, 目标函数的目的是获取最优的代表性和可读性, 这样能确保生成的微摘要能反映原始文本的主要观点和意见, 也确保了生成的理解式微摘要的可读性.

其次, 通过设置摘要最大长度的阈值确保了摘要的简洁性, 任意 2 个词语之间的相似性阈值的设置保证了摘要的低冗余性.

最后, 为了确保得到具有代表性和可读性的最小词组, 我们设置了 2 个阈值, 通过剪枝确保了算法的效率. 与此同时, 这也保证了生成的摘要具有良好的代表性和可读性.

显然, 现在主要问题是确定代表函数 $S_{\text{rep}}(m_i)$ 、可读性函数 $S_{\text{read}}(m_i)$ 和相似函数 $S_{\text{sim}}(m_i, m_j)$.

2.1 代表性

一般来说, 通常情况下的评论观点是相当冗余的, 并且会包含着很多相互矛盾的观点信息. 因此, 生成一些具有高度代表性的词组是具有挑战性的. 定义高度代表性词组的 2 个关键属性: (1) m_i 中的词组之间, 必须在原始文本的狭窄的窗口内有很强的关联性; (2) m_i 中的单词, 在原始文本中出现次数应该足够频繁.

第一属性确保生成候选词组的词语是相关的. 上下文强关联属性可以通过计算 m_i 中词语的点互信息(PMI)得到. PMI 是目前用来衡量单词间关联度的最好指标^[8]. 如果 1 组词语被认为是具有代表性的, 那么它们在原文中也应该是具有代表性的. 这样第二属性会增大包含那些频繁在原始文本中出现单词的 N 元组. 形式上, 假定 $m = w_1 \cdots w_n$

是 1 个候选词组.

定义 $S_{\text{rep}}(m_i)$ 如下:

$$S_{\text{rep}}(w_1, \dots, w_n) = \frac{1}{n} \sum_{i=1}^n P_{\text{pmi, local}}(w_i), \quad (3)$$

其中, $P_{\text{pmi, local}}(w_i)$ 是局部点间互信息函数, 且为:

$$P_{\text{pmi, local}}(w_i) = \frac{1}{2C} \sum_{j=i-C}^{i+C} P'_{\text{pmi}}(w_i, w_j), i \neq j, \quad (4)$$

其中, C 是上下文窗口尺寸大小. $P_{\text{pmi, local}}(w_i)$ 测量了单词 w_i 与它 C 个所有的相邻单词的平均关联度. 为得到第二属性, 称之为 P'_{pmi} 的改进 PMI, 其中, 在 w_i 和 w_j 间的 P'_{pmi} 定义为:

$$P'_{\text{pmi}}(w_i, w_j) = \log_2 \frac{p(w_i, w_j) \cdot c(w_i, w_j)}{p(w_i) \cdot p(w_j)}, \quad (5)$$

其中, $c(w_i, w_j)$ 是 2 个词语同时出现在原始文本中文本窗口大小为 C 的 1 句话中的次数; $p(w_i, w_j)$ 是相应的联合概率. 同时出现频率 $c(w_i, w_j)$ 是为了增大原始文本中频繁出现的词语得分值而增加的项. 原始 PMI 函数的问题是它对于低频词产生了高分值. 通过加入 $c(w_i, w_j)$ 项到 PMI 函数中, 能确保低频词不再占优势, 且使得高共现率关联词组获得相对高的分值.

2.2 可读性

可读性方面是任何摘要工作的重要要求, 因为这可以让读者更好地理解摘要信息. 我们借助公开可用的微软 N-gram 服务^[9]来解决这个问题. 其假设是如果生成的词组频繁出现在网络上, 那么此词组更容易可读.

具体来说, 对文件的正文文本使用微软的三元语言模型进行训练, 以获得候选词组的条件概率. 在计算每个词组时, 首先获得词组中不同三元组的条件概率. 这些分数相结合、平均, 并产生最后的可读性分数. 假定 $m = w_1 \cdots w_n$ 是候选词组, 那么 $S_{\text{read}}(m)$ 可以定义如下:

$$S_{\text{read}}(w_1, \dots, w_n) = \frac{1}{K} \log_2 \prod_{k=q}^n P(w_k | w_{k-q+1} \cdots w_{k-1}), \quad (6)$$

其中, q 表示使用模型的 N-gram 的阶数; K 表示被计算条件概率的数目.

2.3 词语相似度计算

词语相似度是个数值, 取值范围在 $[0, 1]$ 之间. 词语相似性和词语相关性是影响词语相似度的两

个主要指标. 词语相似性可以用词语间的词语(语义)距离来衡量. 词语距离是个 $[0, \infty)$ 的实数. 词语相关性反映的是 2 个词语互相关联的程度. 一般用 2 个词语在同 1 个语境中共现的可能性来衡量词语相关性. 词语相似性反映的是词语间的聚合特征, 而词语相关性反映的是词语间的组合特征^[7].

我们采用文献[7]中的词语相似度计算方法, 计算词语相似度时将 2 个词语的所有义项分别两两计算, 然后取最大值作为 2 个词语的相似度值.

3 摘要算法

因为我们要使用原文本中的单词来组成新词组, 所以解决优化问题的潜在方案空间是巨大的. 在实践中, 列举所有可能的候选词组并逐个记下它们是不可行的. 因此, 我们提出了贪心算法解决优化问题, 通过用启发式剪枝策略系统地探索解空间, 使得我们只处理最理想的候选词组.

在英文中单词的界限是很明显的, 但是对于中文来说, 1 个“词”的划分却要复杂许多, 因此构建 N-gram 模型也有 2 种不同的方法, 通常根据应用的需要可以分为以字为基础构建 N-gram 模型和以词为基础构建 N-gram 模型. 文中笔者选择基于词的 N-gram 模型构建方法.

摘要算法的输入是包含了产品评论的一组句子. 我们用 $Z = \{z_i\}_{i=1}^n$ 表示所有关于 iPhone 5s 评论的句子. 输出是微观点摘要, 是一组 N 元组词组 $M = \{m_i\}_{i=1}^k$, 其中微观点的数目由优化问题的参数限制所确定.

摘要算法包含以下 3 步:

步骤 1 产品评论信息分词. 使用中国科学院计算机研究所研发的 NLPIR 汉语分词系统对评论信息分词处理, 并去除停留词, 分出的词作为一元组, 并将高频一元组合并生成二元组.

步骤 2 深度搜索产生候选词组. 输入为二元组, 并通过进一步连接共有重叠的二元组发展为 1 组理想的 N 元组. 虽然产生了 N 元组, 还需要计算其代表性分值 S_{rep} 和可读性分值 S_{read} , 并剪掉这些任何分值低于其相应阈值的情况. 接着进一步检查生成候选词组的冗余度, 如果 2 个词组的相似度高于阈值 σ_{sim} 时, 则抛弃 $S_{\text{rep}} + S_{\text{read}}$ 的分值较低的那个候选词组.

步骤3 微摘要的生成. 基于目标函数值(即 $S_{rep} + S_{read}$)排序所有的候选 N 元组, 且通过逐渐地添加最高分值的词组到摘要中, 直至累加的摘要长度达到长度的阈值 σ_{ss} , 以此产生满足要求的产品评论微摘要.

4 总结

针对网络上存在的大量中文产品评论信息, 提出一种新的无监督的方法来处理评论信息, 预期生成简洁的非结构化的可读性强且具有代表性、简洁性的理解式评论微摘要, 并突出强调微摘要的三个特性——可读性、代表性及简洁性.

在未来的工作中, 我们将进一步实现提出的产品评论微摘要生成方法, 实验检测本方法的效果, 并与人工摘要及其他学者研究结果对比, 进一步改进遇到的问题与不足.

参考文献:

- [1] 郗亚辉, 张明, 袁方, 等. 产品评论挖掘研究综述[J]. 山东大学学报: 理工版, 2011, 46(5):16-23.
- [2] Li Zhuang, Jing Feng, Zhu Xiaoyan. Movie review mining and summarization[C]. Proceedings of the 15th ACM international conference on Information and knowledge management ACM, 2006:43-50.
- [3] Liu Bing, Hu Mingqing, Cheng Junsheng. Opinion observer: Analyzing and comparing opinions on the web[C]. Proceedings of the 14th international conference on World Wide Web, ACM, 2005:342-351.
- [4] Hu Mingqing, Liu Bing. Mining and summarizing customer reviews[C]. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004:168-177.
- [5] Lu Yue, Zhai ChengXiang, Sundaresan N. Rated aspect summarization of short comments[C]. Proceedings of the 18th international conference on World wide web, ACM, 2009:131-140.
- [6] Ganesan K, Zhai Chengxiang, Han J. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions[C]. Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010:340-348.
- [7] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6):602-608.
- [8] Terra E, Clarke C L. Frequency estimates for statistical word similarity measures[C]. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003:165-172.
- [9] Wang Kuansan, Christopher T, Evelyne V, et al. An overview of microsoft web N-gram corpus and applications [C]. Proceedings of the NAACL HLT 2010 Demonstration Session, Association for Computational Linguistics, 2010:45-48.

Micro-abstract of Chinese Product Reviews: A Novel Approach

ZHAO Fu-qing¹, LIU Bai-song^{2*}, GE Wen-zhen¹

(1.College of Information Science and Engineering, Ningbo University, Ningbo 315211, China;

2.Library, Ningbo University, Ningbo 315211, China)

Abstract: This paper presents a new unsupervised approach to generate micro-abstract of Chinese product reviews, mainly focusing on generating the readable, representative, concise and unstructured abstractive summaries. We model the readability with an N-gram language model, measure representative impact based on a modified mutual information function, and calculate words similarity with Tongyici Cilin. We also formulate the addressed issue as an optimization problem in a move to seek a set of concise and non-redundant phrases that are readable and can represent key opinions in text. In the end, we propose the heuristic algorithms to efficiently solve the formulated optimization problem.

Key words: product reviews mining; unsupervised approach; micro-abstract; web N-gram

(责任编辑 章践立)