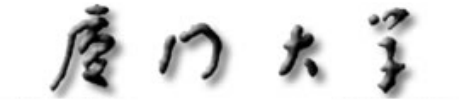
学校编码：10384　　分类号____密级____
学号：23020101153065　　UDC____



硕 士 学 位 论 文

基于在线评论的个性化多产品摘要算法的研究

Personalized Multi-product Summarization based on Online Reviews

王菁菁

指导教师姓名：林琛副教授
专 业 名 称：计算机科学与技术
论文提交日期：2016 年　　月
论文答辩时间：2016 年　　月
学位授予日期：2016 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：王菁菁

2016 年 5 月 17 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：王菁菁

2016年 5月17日

摘 要

近年来，观点摘要技术为世界各地的消费者带来了极大的便利。从大量的在线商品评论中，观点摘要技术自动为给定商品的大众观点生成摘要。然而，当前的观点摘要系统为每个商品所提供的摘要通常是静态、粗粒度的，这样的摘要在处理高度动态和个性化的用户偏好时具有很大的局限性。因此，在用户评估候选商品的阶段，这种摘要无法为其提供所需要的有效的指导意见。

在本文中，我们通过生成个性化的多商品摘要为消费者提供决策支持。本文的目标是生成简洁的商品动态摘要，它可以体现出用户所喜爱的特征的重要信息，同时能够兼顾不同商品之间的差异性。

首先，为了使得生成的摘要满足以下的三个特征：高度精简性、集中覆盖性、差异性，本文将个性化的多商品摘要问题建模为特征树上的带有可变覆盖半径最小代表特征集问题，树上每个被覆盖的区域都包含了商品各种各样特征的观点，从层次结构上来看，这些特征在语义上是相互关联的。为了获得最优的覆盖半径，我们会为层次结构上的每个特征赋予一个实时推导出来的偏好权重，并结合商品本身的差异性，以此来指导最优半径的选择。

除此之外，本文中使用了有监督的模型实现特征识别，同时在提供部分已标注语义层次关系的前提下自动构造出特征的层次结构。在特征识别和特征的排序学习中，我们都使用到了观点挖掘技术。同时，商品特征的层次结构中使用了满足最大召回率的贪心算法。

最后，从真实的数据集上的实验结果和用户案例分析的结果上来看，本文中提出的方法展示出了有效性和合理性。

关键词：个性化摘要；观点挖掘；特征层次结构

Abstract

Nowadays, opinion summarization technologies have brought conveniences to consumers worldwide by making an abstract of the public opinions on a given product automatically from massive online product reviews. However, state-of-the-art opinion summarization systems, which provide a static, coarse grained summary for each product, have their limits in handling highly dynamic and personalized user preferences; and thus fail to function fully when users need constant support in the pre-consumption session of evaluating candidate products.

In this paper we propose to facilitate consumption decision by personalized multi-product summarizations. The aim is to tailor brief summarizations that cover important information on users' preferred aspects and clear distinctions among products. In order to achieve high brevity, focused coverage on preferred aspects, and clarity on distinguished aspects, we first present to model the personalized multi-product summarization as a minimal representative feature set problem, with a set of mutable radiuses on separated regions. Each region consists of opinions on various product aspects, which are semantically related to each other in the form of a product aspect hierarchy. A preference weight inferred real-time and dictinctions between products are assigned to each product aspect in the hierarchy to obtain the optimal radiuses.

Then we present to automatically extract the product aspects and the hierarchy from online reviews, supervised by a few annotated aspects and semantic relationships. Opinion mining techniques are exploited to learn new aspects and the rankings between every pair of aspects. We present a greedy algorithm to generate the product aspect hierarchy with maximal recall on the rankings.

Experimental results on real data sets, as well as user studies, demonstrate the effectiveness of the proposed method.

Key words: Personalized Summarization; Opinion Mining; Aspect Hierarchy

目 录

Table of Contents

第一章 绪 论

1.1 研究背景及研究目的

当今社会中所提供的丰富的商品，并没有如所预期的那样，提高消费者的满意程度。相反地，由于这种丰富性而导致的紧张感、焦虑感、从大量的候选商品中选择合适的商品而产生的忙碌感都会不利于提升我们的幸福感。“选择的悖论”——更多的选择带来更少的乐趣——已经被各种心理学家不断地研究。虽然选择和情绪之间的关系很模糊，但是通过减少冗余的选择来降低选择恐惧症这种做法被证明是行之有效的。在心理学家 Barry Schwartz 的畅销书[1]中，他描述了做出最后的购买决定之前通常涉及到几个主要步骤的循环执行，这些步骤包括：明确目标，对选择进行排序，评估每种选择是否满足目标，挑选出最满意的选择以及在必要的情况下修改目标。

随着电子商务的快速发展，选择过载的问题变得越来越严重，因为用户更容易接触到更多的商品了。传统的做法是，通过专家（即通过询问售货员）来实现对选择的评估，但是这种做法经常因为个人的兴趣爱好而有失客观性。近些年来，通过在线商城或者反馈网站来分享个人的购买经历，网络口碑就可以被用来指导消费者评估不同的候选商品。比如，表 1- 1 所示的是大众点评[1]网上对“全聚德(朝阳北路店)”的一些评论，这些评论涉及到餐馆方方面面的特征，比如“菜量”、“服务”和“味道”等等。而且，很多真实的研究结果显示，消费者不管是在线上或者线下做消费决策之前，都高度依赖了在线评论。

大量的在线评论为用户（消费者）理解公众的观点造成了负担。如果能对这些评论进行统一的分析处理，从海量的数据中抽取出用户的主要观点，即对这些观点产生一个综述，得到的综述必然会同时为用户和商家产生巨大的方便和利益。一方面，用户可以通过该信息了解到其他用户对这个商品的实际评论，从而为他的购买决策提供支持。另一方面，商家可以通过该综述快速了解用户的反馈，及时地改进商品或者相关的服务，从而提高自身的竞争力。因此，为了减轻这种信息过载的问题，观点挖掘和多文档摘要技术近些年来受到了越来越多的关注。比如说，一家餐馆的评论中包含了对餐馆各个角度的评价，包括餐馆的环境、装修、

[1] www.dianping.com

性价比、菜相、菜的味道等方面，对这些评论使用观点摘要技术，最终生成的摘要包含三个方面的评论：“这家餐馆的性价比超高”、“菜的味道相当不错”、“环境相当舒适”。这样简洁有概括性的摘要，一方面可以提炼出这家餐馆的特色，从而给用户更明确的指导。另一方面，通过过滤大量的评论噪声，可以缓解用户面对过量信息的焦虑。从摘要的结果来看，相应的分析处理系统中需要包含两方面的功能，一方面是捕捉评论中的主要评论对象以及对该对象的态度，其中，所捕捉到的评论对象我们称之为商品的特征(aspect)，比如“环境”即是餐馆的一个特征，对这个特征的态度可以是赞赏或是不满，这主要通过观点挖掘技术来实现。另一方面，对观点挖掘的结果使用多文本摘要技术，得到在特定的特征上的摘要结果，从而更进一步提炼出评论中的重要信息。

表 1- 1 “大众点评”上对“全聚德(朝阳北路店)”的一些评论

量少了些，后来加了好多菜才够吃。
慕名而来，总体不算很失望，就是服务有些跟不上。
鱼不好吃，太腥，吃不惯。烤鸭还不错。
菜量挺大，服务也不错，地方略显拥挤。
服务很棒，虽然没有前门的和王府井的店有名，但是味道差不多，价格也实惠多了，就是距离远了点，但是坐地铁很方便，出了地铁很好找。11:30 到的，人不多，很开心不用等位子。份量很足，两个人点了半只烤鸭，两个素菜，吃的饱饱的。
味道还可以，名声很大，环境不错，服务周到。

在淘宝[2]等很多电商平台、百度音乐[3]等娱乐平台和美团[4]等 O2O 平台中，都存在一个类似的板块，即根据用户的历史浏览记录推断用户的偏好，从而为用户推荐一些商品或者娱乐电台等，这些为用户量身定做的推荐结果将大大改善用户的体验。为了更有效地生成对消费者有指导意义的摘要，如果能够在最终生成的摘要中体现出用户的偏好，那将更进一步减轻消费者的选择困难症。而在用户选

[2] www.taobao.com

[3] music.baidu.com

[4] www.meituan.com

择或者删除（即筛选）商品的过程中，往往体现出了用户的偏好。比如，有些消费者更喜欢选择性价比很高但是对于环境要求相对较低的餐馆，能体现用户这个偏好的摘要可以是“这家餐馆的性价比很高，但是环境稍微差了点”。

除了可以为每个商品生成为每个用户量身定做的摘要之外，如果可以同时考虑多个商品之间的差异性，并且将这种差异性在摘要中体现出来，这将减少用户对不同的商品进行横向对比的时间，为用户提供更直接的指导。例如，A、B 两个餐馆的评论中都包含了对“口味”的评价，其中 A 餐馆在“口味”上的评价要优于餐馆 B，最终为 A、B 两个餐馆生成的摘要中都包含了“口味”这个特征，这会减少用户在相似商品上的比对时间。现阶段已经出现了很多商品的摘要（如各大电商网站上出现的商品摘要），这些摘要都很简洁，能够体现大众的评论观点，但是它们都没有考虑用户的不同偏好和不同商品之间的差异性。

为了减轻信息负载问题，已经出现了很多观点挖掘的技术[2-9]。然而，很多已经存在的观点摘要系统有明显的缺点——摘要的粒度是固定的。以上所有观点摘要系统都是在商品的一些固定特征上提供了不同形式（比如，评分、极性、简短的文本片段）的摘要，它们展现出来的特征是静态的。比如，一个旅馆的静态摘要涉及到的特征可能包含旅馆的位置、清洁度、服务质量等等。或者，摘要中包含的特征数量是固定的，比如，规定摘要的结果中需要包含对 5 个特征的观点。然而这样的摘要无法让用户“缩小”地去看更多他们感兴趣的特征，或者也无法“放大”地去忽略一些特征的细节，也即无法根据实际情况动态地改变摘要中包含的特征，从而提供更灵活的结果。

在本文中，我们考虑的是粗粒度、静态的观点摘要无法为用户提供购买决策支持的场景。从下面的例子场景中可以看出，用户最初会受到基本的商品评论的引导，接着他们可能会通过点击不同类别的选择来改变他们的目标。自然地，摘要应该对于用户特别喜欢的特征有详细的描述，而只需要大致概括用户不感兴趣的特征。

例子：Alice 正在在线购买一个笔记本电脑。一开始，她无法明确自己的选择，所以她会随机选择 A 和 B 两种型号的笔记本电脑。接着，她发现大量的负面评价是关于 A 的电池，相反地，对于 B 的电池有很多正面评价。Alice 是一个经常需要出差的商人，所以她觉得电池的性能对她而言是很重要的。因此，她舍弃了 A 而选择了 C 型号的笔记本电脑（其电池具有很好的性能），此时她想要更