计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000－1239.2019.20180723
56(9):1927－1938,2019

微博热门话题关联商品品类挖掘

左笑晨[1] 窦志成[2] 黄 真[1] 卢淑祺[1] 文继荣[2]
[1](中国人民大学信息学院 北京 100872)
[2](大数据管理与分析方法研究北京市重点实验室(中国人民大学) 北京 100872)
(zuoxc@ruc.edu.cn)

Product Category Mining Associated with Weibo Hot Topics

Zuo Xiaochen[1], Dou Zhicheng[2], Huang Zhen[1], Lu Shuqi[1], and Wen Jirong[2]
[1](*School of Information, Renmin University of China, Beijing* 100872)
[2](*Beijing Key Laboratory of Big Data Management and Analysis Methods* (*Renmin University of China*), *Beijing* 100872)

Abstract Weibo is one of the widely used social media platforms for online sharing and communication. Some widely-received topics have been formed into Weibo hot topics by being forwarded, reviewed, and searched by a large number of users in Weibo. And the widespread dissemination of these hot topics may further stimulate and promote users' offline behaviors. As a typical representative of it, some hot topics on Weibo may stimulate sales of products related to the topics under the e-commerce platform. Mining out the relevant product categories of Weibo's hot topics in advance can help e-commerce platforms and sellers to do a good job of commodity operation and inventory deployment as well as promote the search conversion rate of users and bring about an increase in the sales of corresponding products. This paper proposes a method of mining potential shopping categories associated with hot topics of Weibo. First, the method builds a product knowledge map, and then uses a variety of in-depth network models to perform textual matching between the information of the associated knowledge of product categories and the content of the Weibo topics. The strength of association of each hot topic and product category is identified. Experiments show that the method can effectively identify the relationship between hot topics and shopping categories, and most of the hot topics of Weibo can be associated with at least one product category in the e-commerce platform.

Key words knowledge graph; textual match; Weibo hotspot; entity recognition; deep learning

摘 要 微博是目前人们广泛使用的在线分享和交流的社交媒体平台之一.某些被广泛关注的话题因为在微博中被大量网友转发、评论和搜索而形成微博热门话题,而这些热门话题的广泛传播则可能进一步刺激和推动用户的线下行为.作为其中的典型代表,某些微博热门话题可能会刺激电商平台中和该话题相关的商品的热销.提前挖掘出与微博热门话题相关联的商品品类,可帮助电商平台和卖家提前做好商品运维以及库存的调配,提高用户搜索的购物转化率,带来相应商品销量的提升.提出了一种微博热门话题所关联的潜在购物品类的挖掘方法.首先构建商品知识图谱,然后采用多种深度网络模型对商品品类的关联知识图谱信息与微博话题内容进行文本匹配,识别出每个热门话题和商品品类的关联强度.

收稿日期:2018－11－01;修回日期:2019－04－02
基金项目:国家重点研发计划项目(2018YFC0830703);国家自然科学基金项目(61872370);中央高校基本科研业务费专项资金(2112018391)
This work was supported by the National Key Research and Development Plan of China (2018YFC0830703), the National Natural Science Foundation of China (61872370), and the Fundamental Research Funds for the Central Universities (2112018391).
通信作者:窦志成(dou@ruc.edu.cn)

实验表明,该方法能够有效识别出热门话题和购物品类的关联关系,大部分的微博热门话题都可以关联到电商平台中至少一个商品品类.

关键词 知识图谱;文本匹配;微博热点;实体识别;深度学习

中图法分类号 TP183

随着移动互联网的高速发展,在线购物和社交媒体成为中国网民频繁使用的2类互联网应用.在在线购物方面,2017年双十一期间京东全球好物节历时10天累计交易额达1 271亿元,天猫更是以1 682亿的交易额创下历史新高.人们在日常生活中对于电商平台的依赖程度也越来越高,在线购物成为人们生活中的常态.在社交媒体方面,包括微博微信在内的社交平台已经成为了人们沟通交流和获取信息的主要手段.很多大家广泛关注的话题一方面在微博上产生大量的浏览和转发行为,同时可能进一步刺激和推动用户的其他行为,包括在线购物.例如,一位明星发微博晒照抒发一天的心情,可能会引起网友购买照片中同款衣物或者饰品的热情.微博上有关冬日保暖养生的博文,则可能会引导用户购买保暖用品、养生茶甚至茶具.微博上"吃鸡"游戏话题的广泛传播,会激发网友购买相应游戏鼠标、键盘、显示器等相应外设的购买意愿.针对这一问题,本文重点研究如何挖掘微博热门话题和购物品类的关联关系.

及时有效地挖掘微博热门话题所对应的电商购物品类是非常有价值的.首先,提前知晓热门话题可能会带来某些商品的热销,可帮助商城运维人员提前做好相应商品的库存调配,避免出现缺货或者断货的状态,实现用户购物意图的高转化率.其次,可帮助商家或者商城运维人员及时进行商品标题运营,解决用户查询词与商品名称失配的问题.在现有的电商平台中,大部分商品的标题与描述都仅仅与商品本身的特性相关,着重突出商品样式与功能,比如商品的类别、规格、适用人群等.这些描述都是商品本身固有的,并不会随着时间变化.而社交媒体上的热门话题是会随着时间的迁移而变化的.用户受社交媒体上的热门话题驱动在购物引擎中检索相关商品信息,所使用的查询经常是和话题想关联的.例如,在热门话题"吃鸡"的驱动下,用户可能会在购物引擎中搜索"吃鸡耳机".某些满足用户购买需求的游戏耳机的商品名称中因为不包含"吃鸡"字样,而无法出现在搜索结果中.及时挖掘出商品和热门话题的关联关系,可帮助卖家及时在商品标题中增加热门话题相关关键词,一方面提升搜索转化率,同时提升用户满意度.最后,商城对应品类的运维人员可根据挖掘出的热门话题进行促销活动或者设计专门的购买入口.例如,在电商电脑外设频道或者首页上发布"吃鸡"外设专属促销页面,可进一步吸引用户(包括非微博用户)购买商品.

在如今的一些电商平台中,一些商品描述中已包含与热点有关的词汇,在搜索一些热门话题时,有可能得到一些满意的结果.但在大部分情况下,这部分商品是在热门话题产生之后一段时间出现的,或者是这些店家根据自己对部分时事热点的了解,在商品描述上面进行的修改.在电商平台中,商户间普遍存在竞争,最先捕捉到用户需求的商户往往会占据先机.当一个社会热门话题产生之后,相应的消费需求也随之产生.如果店家通过自己对于热点的发现来更新商品的描述信息,很可能会产生消息的滞后,原因就在于店家无法时刻关注热搜话题,而要做到尽快地更新信息,通过人工的方式会消耗大量成本.对于电商平台也同样需要运维部门及时快速地对热门话题进行响应.定期对微博热门话题进行扫描,挖掘出话题对应的商品品类,同时反过来为商品推送相关热门话题,对电商平台和商户都是具有重要价值的.

针对这一实际需求,本文提出了一种微博热门话题所关联的潜在购物品类的挖掘方法.首先根据已知的商品实体信息,构建出商品品类知识图谱.然后根据采集到的微博热门话题,获取相关的微博文本,对微博文本进行分词与命名实体识别,提取出与商品存在潜在关联的实体,将这些实体在之前构建好的知识图谱上进行检索,通过设计规则对检索结果进行评估,从而得到该话题与商品品类的关联性.进一步,为了考虑微博文本的语义信息,本文引入用于商品标题与微博文本的匹配模型——基于核函数的神经网络排序模型(kernel based neural ranking model),将匹配模型的结果与之前知识图谱的检索结果结合,得到最终的匹配模型K-KCM(KNRM-knowledge graph Weibo content matching model).实验显示通过知识图谱检索的方式可以发现许多与

热点关联的商品品类,但召回率较低,仍有许多应有的相关商品未被发掘.在加入了商品标题与微博文本的匹配结果之后,召回率得到了显著的提升,大部分显著相关的商品都能够被发现.

1 相关工作

1.1 基于社交媒体的产品推荐

目前有一系列工作致力于研究社交媒体和在线购物之间的关系,其中具有代表性的研究是基于社交媒体的商品推荐.在电子商务迅猛发展的背景之下,帮助用户在海量信息中找到合适的商品变得愈加重要.社交媒体上拥有的海量数据对于商品个性化推荐的作用不容忽视.其作用体现在很多方面:首先体现在冷启动方面.对于已使用电商平台的顾客,在推荐时可以参考该用户的访问记录与反馈记录[1],然而对于初次使用电商平台的用户不具备消费历史,有研究通过从与该用户关联的社交网络中提取知识用于跨站点冷启动[2].具体地,可以采用社交媒体平台中提取的人口统计信息[3]进行产品推荐,以及使用产品图像与用户评论作为推荐系统的依据[4],或者使用产品采用者的在线评论信息做产品推荐[5],还有研究采用卷积神经网络学习每个用户和商品的特征表示进而完成推荐[6].其他方面,有的研究通过对于早期在线评论的预测来对顾客的购买行为进行指导[7].还有研究根据消费者购物时更倾向朋友意见的现象,从多维性和动态性2点出发,提出了基于社交网络信任模型的商品推荐系统[8].进一步地,还有研究从网络信任角度探讨了消费者的认知能力、关系强度和交互作用对社交媒体网络中消费者网络购物决策的影响,并通过微博数据进行回归分析[9].另外,还可以通过结合用户吸引力相似度和用户交互相似度来获得多属性综合相似度,结合多属性相似度采用加强协同过滤的算法完成推荐[10].

然而,文献[1-10]主要针对用户的个性化推荐问题,并没有考虑到社交媒体中引起广泛关注的热点话题潜在促进商品热销的作用.事实上,除了个性化推荐以外,微博数据能够发现新的热门趋势带来的商品销售机会.有关研究[11]通过学习微博中的商业意图来识别与热门趋势相关的产品.本文同样基于微博热点话题与内容,挖掘其与商品品类存在的潜在关联,从而对电商平台的运营起到辅助作用.

1.2 微博文本分析

对微博文本的分析问题中,比较重要的是实体与事件的抽取.对微博内容进行实体抽取的研究很多:姜仁会等人[12]提出了一种基于统计与规则相结合的命名实体识别的方法.李刚等人[13]提出了一种基于条件随机场模型改进的方法.李治国等人[14]针对大量存在于网络信息中不规则书写的命名实体和商务领域中系列类型的命名实体,利用它们的特点提出了在篇章中使用词与词之间的互信息来识别命名实体类的办法.刘玉娇等人[15]提出了一种基于深度学习的微博命名实体识别方法.陈箫箫等人[16]针对微博中的开放域事件抽取问题进行了深入研究.主要通过序列标记方法提取微博语句中的命名实体和事件短语表征事件,利用非监督分类方法对事件进行分类.其中运用条件随机场模型完成事件抽取中的序列标记任务,非监督分类方法使用的是LDA主题模型.基于微博文本进行实体抽取是本文模型中的一个步骤,但不是本文的研究重点.

2 商品品类与微博热点的匹配

本节主要介绍商品品类与微博热点话题的匹配模型K-KCM.在此之前,商品与热点话题的关联关系并没有引起足够的重视,然而这种关联关系实际上是存在且十分重要的,热点话题一定程度上影响着社会潮流,因此也对人们的消费行为产生着不可忽略的影响.然而将这种关联关系挖掘出来并不容易,因为商品品类与微博话题属于不同领域,有着不同的表达结构.为了将这2种不同结构的事物相关联,本文提出一种商品品类与微博热点话题的匹配算法,将商品抽象为3级品类结构,并使用商品品类知识图谱与微博热门话题匹配以及商品标题与微博内容匹配2种方法综合得出商品与微博匹配得分,整体流程如图1所示.

2.1 商品3级品类结构

在大部分电商平台中,由于商品种类繁多,以及许多不同品类的商品之间差异并不明显,导致商品品类划分粒度极其细微,给商品管理造成了极大不便.因此,大部分的购物引擎对于商品品类使用分级管理.通常分为3级,从商品的一级品类到三级品类,商品所属的范畴被不断地压缩.商家通过将自己的产品准确划分到所属的三级品类,可以提高用户搜索该产品的准确度.家用电器和电脑/办公品类下的3级品类结构如图2所示.

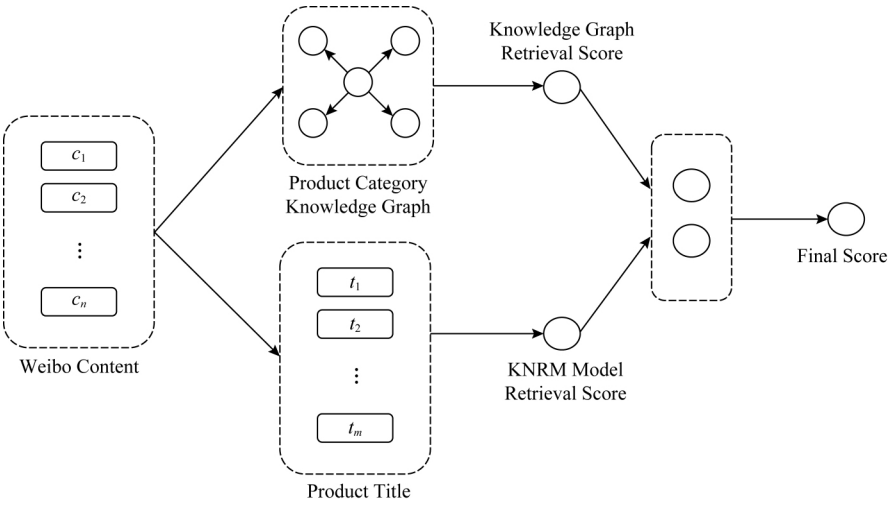

Fig.1 The process of K-KCM algorithm

图 1 K-KCM 匹配算法流程

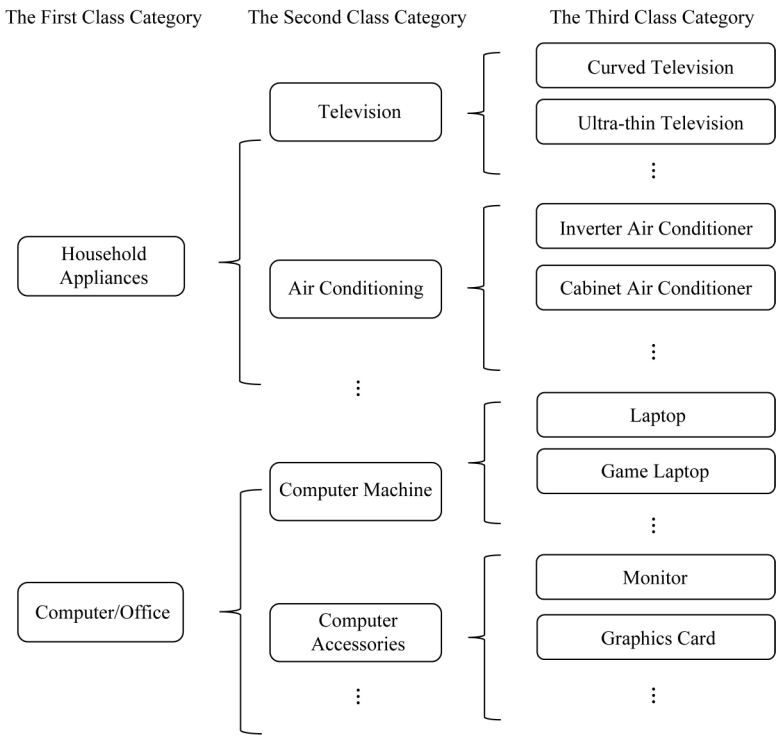

Fig. 2 Examples of the three-level structure of category

图 2 3 级品类结构示例

本文在商品匹配过程中使用的是三级品类，同时从某电商平台爬取了各级品类名称.虽然图 2 中展示的级品类名称在匹配时差异不大，但实际上许多属于同一个二级品类下的三级品类商品仍有不小差异.比如同属于手机/运营商/数码一级品类下、电子教育二级品类下的早教益智和电子词典 2 个三级品类，在匹配过程中并不能当成含义相近的品类，因而使用三级品类.

2.2 商品品类知识图谱与微博热门话题匹配(KCM)

2.2.1 商品品类知识图谱的构建

挖掘商品品类与微博话题的关联，首先需要让计算机对于商品品类有一定的认知.比如对于一个三级商品品类中央空调，仅仅知道这个名字对于关联的挖掘来说是远远不够的，需要知道这个名字的含义.这个含义的表达方式有很多，比如它的形状(长方形)、颜色(白色)、用途(制冷)等.当掌握了

这些信息之后,计算机便对某个商品品类的含义有了真正的认识,这样才能与之后分析得到的微博文本语义做关联挖掘.因此,本文构建了商品品类知识图谱,其结构如图 3 所示:

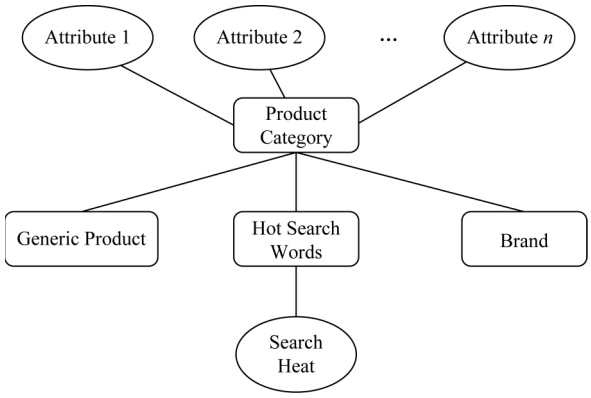

Fig. 3 Knowledge graph of product category

图 3 电商品类知识图谱

从图 3 可知,每一个商品品类与其他 3 个实体关联,分别是泛产品品类、热搜词和品牌,其中商品品类本身还具有一些品类本身特有的属性,热搜词还有词频属性.具体的实体含义为

1) 泛产品品类.原始数据给出了所有的商品品类,但是在这些商品品类中有很多品类对于顾客的需求没有太大的差异,如表 1 中品类名称.表 1 中呈现了 3 组泛产品品类名称,它们分别属于休闲娱乐、VR 设备、保温壶这 3 个商品品类.泛产品品类存在的意义在于将多个相近的品类集中起来,作为匹配过程中的一个整体,可以减少匹配品类总数,提高话题匹配成功的概率.

Table 1 Examples of Generic Product Category

表 1 泛产品品类示例

Category	Generic Product Category
Leisure and Entertainment	Leisure Shopping, Leisure Fitness, Leisure Vacation, Catering and Leisure
VR Device	VR Glasses, VR Helmet, VR Device, VR Head Display Device
Insulation,Pot	Insulation Pot, Hot Water Bottle, Thermal Flask, Thermos

2) 品牌.对于每一个商品品类,都拥有许多商品品牌.比如卫衣品类下有诸如丹杰仕、乔丹、朵比妮等品牌名称.在微博文本中,许多商家的官方微博内容中经常会涉及到许多品牌名,例如 Dior 官博发布的微博:“青年演员身着 Dior 迪奥 2018 早秋系列精彩演绎时尚街拍……”中提到的品牌名 Dior.对于品牌名的匹配可以准确找到关联的商品品类.

3) 热搜词.用户在搜索指定商品时输入的搜索词.比如对于中央空调品类下有关的热搜词有家用中央空调、美的中央空调、吸顶空调等.热搜词和微博内容类似,都有口语化现象,因此也更容易在微博文本中匹配成功.加入热搜词之后,大部分的热门话题都与部分商品关联成功.

热搜词具有词频的属性,不同的热搜词被使用的次数不同,使用次数高的热搜词更能够代表对应的品类,在匹配过程中匹配成功之后贡献的得分也相应更高.

4) 商品品类属性.除了几个与商品品类相关的实体之外,商品品类本身也有若干属性.比如品类 T 恤下拥有属性衬衫领形、袖长等属性;品类珍珠胸针下拥有属性镶嵌材质等.例如,戒指品类下知识图谱结构具体实例如表 2 所示:

Table 2 Examples of Knowledge Graph Structure

表 2 知识图谱结构示例

Category	Ring
Generic Product	K Gold Ring, Diamond Ring, Wedding Ring, Silver Ring
Brand	AFN, LOQI, CGC
Hot Search Words	DO Ring, Ring Female, Ring Male Domineering, Couple Ring
Attributes	Mosaic Material, Main Stone Weight, Cleaness, Bye Stone Weight

在实际匹配过程中发现,一些出现频率比较低的热搜词实际上对于匹配结果的影响却很大.原因在于虽然这些词在商品搜索过程中出现频率较低,理论上对于匹配结果的贡献值也不太高,但这些词往往都是人们日常生活中常用却对搜索结果没有什么意义的词,比如:男士、女士……人们一般不会在搜索栏中输入这样的词语,因为这种描述过于模糊,并不能够代表该类商品的特点.虽然这类词很少出现,但是在微博文本中却大量出现,累计的贡献值要远远超出想象,最终得到的匹配结果也受到影响.因此实际上删除了热搜词中出现频率低于某一阈值的词,该阈值与实际日志数据的长短有关.

2.2.2 微博热门话题内容的获取

微博数据通过网络爬取,抓取最新的热搜微博内容,这些微博内容围绕同一个微博热搜榜话题,不仅包括话题发起者的微博,同时也包括微博用户对于该话题的相关评论,以及引用该话题的其他微博.将这些微博整理为文本,对其进行除噪过滤,作为语料文本进行匹配.过滤方法有 3 种:

1) 去除所有的标点符号以及表情等非常用文本符号.发微博或者评论微博的用户用语具有口语化以及随意性等特点,甚至有时整篇内容都是没有意义的符号.比如表示震惊的情绪时,可能会使用大量的感叹号,以及表达一些丰富的情感时,常使用一些特殊的表情符号,这些加强情感的符号对于商品品类的匹配没有较多的帮助,属于文本噪音,需要删去.

2) 去除所有以"@"开头以及冒号结尾的字符串.微博内容中一个非常鲜明的特点就是当微博涉及到其他用户或者是想让其他用户看到这篇微博时,会使用@加上该用户的昵称.除了一些官方微博以外,大部分用户的昵称对于商品的匹配过程是没有贡献的,甚至会产生极大的误导,因此用正则表达式匹配的方法将这些昵称删除.

3) 去除以"#"开头与结尾的字符串.与前文提到的昵称问题类似,以"#"开头结尾的往往表示一个话题的名称.正常情况下,在一个话题中使用这样的符号引用另一个与之相似的话题并不会有不良影响,但通过观察数据发现,许多微博用户并不遵循这种相似性规则,甚至有的人喜欢在某话题下面引用与之毫不相关的话题,这便对不同话题之间的比对造成干扰,所以删除类似这样的话题引用.

将过滤之后的所有微博内容连接在一起,作为此话题对应的用于分析的微博内容.

2.2.3 知识图谱与热门话题内容的匹配

对于每一个实时产生的热门话题,通过 2.2.2 节方法获得该话题对应的微博内容.对于其中涉及到的知识图谱中涉及的不同实体,采用不同的分析方法.将分析结果在已构建好的知识图谱上进行检索,计算流程如图 4 所示:

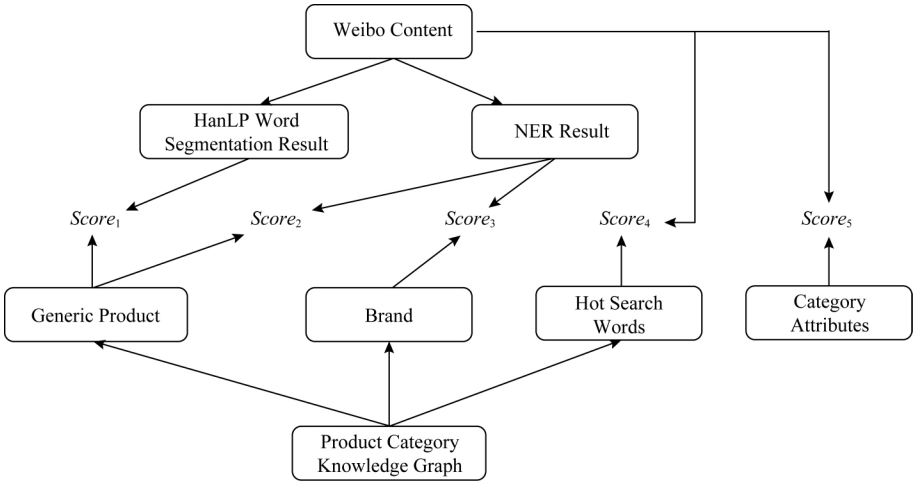

Fig. 4 The process of knowledge graph retrieval

图 4 知识图谱检索流程

2.2.3.1 泛产品名称识别

1) 使用 HanLP 汉语言处理包对微博内容进行分词,并将所有的泛产品品类名称作为词典对分词结果进行过滤.过滤后统计出现次数最多的前 10 个词,去除其中出现次数不超过 1 次的词.用这些词在知识图谱中的泛产品品类名称部分进行检索,即与每一个商品品类下相关的所有泛产品品类名称进行精确匹配.每匹配成功,便为该品类累计得分 $Score_{g_1}$,其计算为

$$Score_{g_1}=\sum_g w_{g_1}\ \mathrm{lb}(Freq_g^{\mathrm{gen}}),$$

其中的 $Freq_g^{\mathrm{gen}}$ 代表第 g 个泛商品类型名称词在微博内荣中出现的次数,w_{g_1} 表示泛产品品类识别对于匹配结果的贡献权重.

2) 对微博内容进行命名实体识别(named entity recognition, NER),这里采用的是双向长短期记忆网络结合条件随机场(bi-direction long-short term memory-conditional random field, biLSTM-CRF)模型识别出微博文本中所有类型为泛产品品类的实体.将得到的实体在知识图谱中的泛产品品类名称部分进行检索.为了避免重复,如果识别出的实体在之前 HanLP 分词结果中出现,则不再重复计算.实体识别结果在知识图谱中检索的累计得分 $Score_{g_2}$ 计算为

$$Score_{g_{21}}=\sum_g w_{g_{21}}\ \mathrm{lb}(Freq\_ner_g^{\mathrm{en\text{-}gen}}),$$

$$Score_{g_{22}}=\sum_g w_{g_{22}}\ \mathrm{lb}(Freq\_ner_g^{\mathrm{gen\text{-}en}}),$$

$$Score_{g_2}=Score_{g_{21}}+Score_{g_{22}},$$

其中 $Freq\_ner_g^{\text{en-gen}}$ 和 $Freq\_ner_g^{\text{gen-en}}$ 都表示识别出的实体在微博内容中出现的次数.区别在于前者表示的是包含某泛产品品类的实体,比如实体名称为纯牛奶,包含名为牛奶的泛产品名称;后者表示的是泛产品品类名称中包含的实体,例如某泛产品品类名称为游戏周边,包含实体游戏和周边.其中的 $w_{g_{21}}$ 和 $w_{g_{22}}$ 分别表示这 2 种实体对于最终匹配结果的贡献权重.

2.2.3.2 品牌名称识别

品牌名称识别部分直接使用 2.2.3.1 节所述实体识别结果,识别出所有类型为品牌的实体.将这些实体在知识图谱中进行检索,与每个商品品类下相关的所有品牌进行比对,累计得分 $Score_{b_1}$ 计算为

$$Score_{b_1}=\sum_b w_{b_1}\ \text{lb}(Freq_b^{\text{brand}}),$$

其中的 $Freq_b^{\text{brand}}$ 代表商品的第 b 个品牌在微博内容中出现的次数,w_{b_1} 表示品牌名称匹配结果对于最终匹配结果的贡献权重.

2.2.3.3 热搜词识别

热搜词不同于泛产品品类名称与品牌名称,它的内容往往很随意,比如对于品类项链,有热搜词迪士尼黄金苹果吊坠、SOINLOVE 钻石旗舰店,这样的热搜词里面不仅可能包含泛产品品类名称和品牌名称,还可能包含其他的实体,例如上述热搜词中的迪士尼和旗舰店.因此无法使用简单的分词技术或者命名实体识别方法得到满意的结果.因此这一部分与之前采用的方法不同,对于所有的商品品类,找到该品类下相关的所有热搜词,将它们在微博内容中进行检索,检索结果累计得分 $Score_{h_1}$ 计算为

$$Score_{h_1}=\sum_h w_{h_1}\ \text{lb}\left(\frac{Value_h}{\sqrt{sl}}\right),$$

其中,$Value_h$ 代表第 h 个在微博内容中出现热搜词的词频,w_{h_1} 表示热搜词匹配对于匹配结果的贡献权重,sl 表示该品类具有的热搜词数量.由于热搜词数据中不同品类下拥有的热搜词数量不同,热搜词数量多的品类在匹配中有可能会得到更高的分数,但实际上热搜词数量多的品类并不代表与话题有更多的关联,而是代表该品类在用户搜索过程中的表述形式更多样.因此,为了降低热搜词数量过多或过少对匹配得分造成的偏差,在原匹配分数上除以 $\sqrt{sl}$ 来消除影响.

2.2.3.4 商品属性识别

由于不同商品品类属性种类各异,属性值在表达方式上也不规范,因此匹配过程与热搜词的匹配过程类似.对于所有的商品品类,找到品类具有的属性值,将它们在微博内容中进行检索,检索结果累计得分 $Score_{a_1}$ 可计算为

$$Score_{a_1}=\sum_a w_{a_1}\ \text{lb}(Freq_a^{\text{attr}}),$$

其中 $Freq_a^{\text{attr}}$ 表示商品第 a 个商品属性值在微博内容中出现的次数,w_{a_1} 表示商品属性值匹配对于匹配结果的贡献权重.

最终匹配得分 $Score$ 可计算为

$$Score=Score_{g_1}+Score_{g_2}+Score_{b_1}+Score_{h_1}+Score_{a_1}.$$

2.3 商品标题与微博内容匹配

通过设计规则得到的电商品类知识图谱与微博热门话题内容的匹配结果有一定的局限.首先,在各个匹配过程中都采用精确匹配,比如对于泛产品品类热水袋,当微博内容中出现“暖宝”、“热水囊”类似的实体时并不会对匹配结果产生贡献,在实匹配中只有准确匹配的词语才会对结果产生贡献,这在处理微博这种语言规范性极低的文本过程中并不合理.另外,知识图谱检索的匹配方式并没有考虑到微博文本的语义信息.比如当微博文本中出现“笔记本”时,电脑品类和记事本品类下都存在笔记本这个泛产品品类名称,而实际上微博内容很可能只表达其中的一个实体,要判断这里的笔记本指的是电脑还是纸质本,还需要结合文本的上下文语义进行判断,这在知识图谱检索的算法中是难以实现的.

为了解决该问题,本文采用文本匹配的思路,使用机器学习的方法.采用文本匹配的思路首先要确定待匹配的文本,微博热门话题采用整理好的微博正文内容文本即可,而在商品品类方面,我们使用的是商品标题文本,因为大部分的商品标题都是由商家书写,同时没有绝对规范的格式,与微博内容中常见的日常用语风格相近.对于某一商品品类,将该品类下的若干条商品标题连接起来形成待匹配的文本.为解决此局限,本文使用 KNRM[17] 模型,相比于传统的基于交互的匹配模型 DRMM[18],KNRM 通过引入核函数机制,在多个不同相似度下统计每个词的贡献值,其模型结构如图 5 所示.

将商品标题与微博内容文本的词向量矩阵计算相似度得到相似度矩阵.并在相似度矩阵上使用多个不同的核函数,在多种相似度水平上,分别计算微博文本内各个词的软词频(Soft-TF),之后将各词的软词频加和得到用于排序的特征,通过多层感知机得到最终的匹配分数.

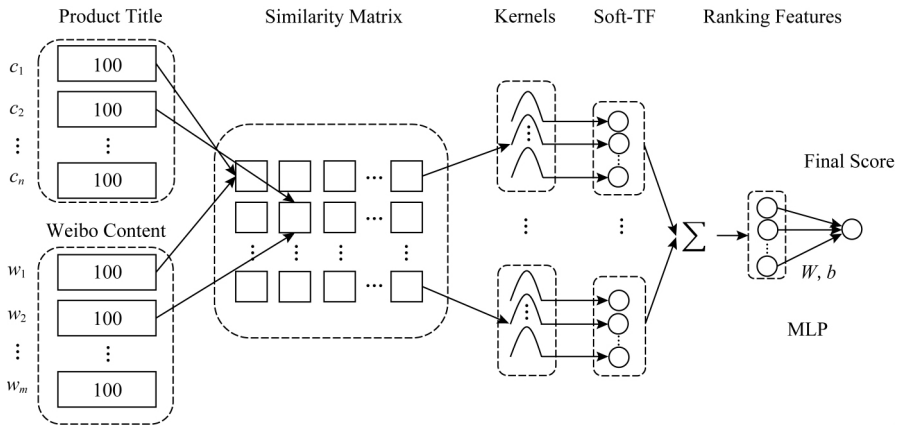

Fig. 5 The structure of KNRM model

图 5 KNRM 模型结构

另外,为了与本文提出的 K-KCM 模型对比,本文参照文献[19]使用了多个深度网络模型:ARC-I 模型[20]、ARC-II 模型[20]、Matchpyramid 模型[21]和 MVLSTM 模型[22].

3 实 验

3.1 实验数据集

本文使用某电商平台提供的商品数据集,其中包括商品三级品类名称、商品属性、商品品牌和商品相关热搜词.其中,有 751 个三级商品品类,平均每个商品包括 1～10 个商品属性、100～2 000 个商品品牌,经过滤后每个商品包括 0～500 个热搜词.在此基础上,将三级品类中类型相近的商品品类综合在一起作为同一个品类,该品类具体包含的所有品类作为泛产品品类.通过热搜词词频对热搜词进行过滤时,使用长度为 1 年的日志数据,根据经验将阈值设置为 100.此外,还需要对商品的属性做筛选,去除一部分品类间区分度不大的属性.比如价格、规格、省份以及颜色等,保留诸如自由度、机身系统、像素、净化技术等具有一定区分度的属性.最终再通过这些数据构建商品品类知识图谱.

使用计算机爬取微博数据集,通过每个小时访问微博热搜榜,获取话题集与相关的内容集,对其中的话题内容做索引,存储在 Solr 搜索引擎中.每当从热搜榜单上获取新的热搜话题,就到搜索引擎中查找,返回所有相关的微博正文,将这些正文作为微博文本数据.数据中包括话题 500 个,微博约 3 000 条.实验训练集、测试集、验证集划分比例为 10∶1∶1,并进行人工标注约 2 500 例匹配数据.

3.2 实验设置

在实验过程中不断根据匹配结果调整各匹配部分权值,调整过程中发现泛产品品类名称和品牌名称识别结果的准确度要高于热搜词与商品属性识别结果的准确度,同时,泛产品品类名包含实体名称时结果的准确度要高于实体名称包括泛产品品类名称时的结果.因此,最终泛产品品类识别过程中使用分词方法获得分数的权值 w_{g_1}、泛产品品类识别过程中使用实体识别方法获得分数的权值 $w_{g_{21}}$ 和 $w_{g_{22}}$、品牌识别获得分数的权值 w_{b_1}、热搜词识别获得分数的权值 w_{h_1} 的最优权值以及商品属性值识别获得分数的权值 w_{a_1} 分别为 3,3,2,3,1,1.

商品标题与微博内容的匹配部分涉及的参数主要是模型中结构中的参数和一些超参数.实验中发现,使用 Word2Vec 训练出的词向量为 100 维时,训练的效果会更好.模型内部的超参数设置如表 3 所示.

Table 3 Model Parameters

表 3 模型参数

Model	Hyperparameter	Value
KNRM	Sigma	0.1
	Exact_Sigma	0.001
	Kernel Number	21
ARC-I	Convolution Kernel Number	32
	Convolution Kernel Size	3×3
ARC-II	1D Convolution Kernel Number	4
	1D Convolution Kernel Size	3×3
	2D Convolution Kernel Number	[4,3]
	2D Convolution Kernel Size	[3×3,2×2]
Matchpyramid	Convolution Kernel Number	32
	Convolution Kernel Size	3×3

在汇总各个深度模型的结果以及知识图谱匹配结果时,使用排序学习(learning to rank)中的LambdaMart模型,模型中回归树的总数设置为1 000,每棵回归树的叶子节点数量值设置为10.

3.3 对比模型

1) ARC-I+KCM.使用商品品类知识图谱匹配结果,与ARC-I模型得到的商品标题和微博文本匹配结果相结合.其中ARC-I模型用于匹配商品品类标题与热门话题内容的模型.使用卷积神经网络,首先在2段文本各自的词向量矩阵上使用多个相同尺寸的卷积核进行1维卷积操作,将多次卷积的结果经过池化层之后拼接成各自的特征向量,将2个特征向量连接起来放入多层感知机中训练得到最终的匹配得分.

2) ARC-II+KCM.使用商品品类知识图谱匹配结果,与ARC-II模型得到的商品标题和微博文本匹配结果相结合.其中ARC-II模型用于匹配商品品类标题与热门话题内容的模型.使用卷积神经网络,同时对2段文本的词向量矩阵做1维卷积操作并对卷积结果进行池化操作,得到匹配2段文本的特征矩阵,并对该矩阵使用2维卷积操作并池化,将获得的矩阵铺平(flatten)得到匹配向量,将该匹配向量放入多层感知机中训练得到最终的匹配得分.

3) Matchpyramid+KCM.使用商品品类知识图谱匹配结果,与Matchpyramid模型得到的商品标题和微博文本匹配结果相结合.Matchpyramid模型用于匹配商品品类标题与热门话题内容的模型.使用卷积神经网络,将2段文本的词向量矩阵交互得到相似度矩阵,并在该相似度矩阵上做卷积与池化操作,得到的结果作为2段文本的匹配特征向量.将该向量放入多层感知机中训练得到最终的匹配得分.该模型与ARC-I和ARC-II模型均采用卷积神经网络,不同之处在于ARC-I和ARC-II是基于文本的表示,该模型是基于文本矩阵的交互.

4) MVLSTM+KCM.使用商品品类知识图谱匹配结果,与MVLSTM模型得到的商品标题和微博文本匹配结果相结合.MVLSTM模型用于匹配商品品类标题与热门话题内容的模型,使用循环神经网络.对2段文本分别使用双向LSTM网络训练得到标题文本与话题内容特征向量,将2个向量结合起来放入多层感知机中训练得到最终的匹配得分.

5) Learning to rank.使用商品品类知识图谱匹配结果与另外5个商品标题和微博内容匹配模型的得分结合起来,作为6维的特征,通过Learning to rank[23],使用gradient boosted regression tree模型[24]得到综合匹配结果.

6) KNRM+KCM(K-KCM).本文使用的模型,使用商品品类知识图谱匹配结果,与KNRM模型得到的商品标题和微博文本匹配结果相结合.

对比模型中除了Learning to rank模型之外,其他的模型均需要将2个独立的模型结果结合.在实验过程中,首先采用无监督的方法得到KCM模型的实验结果,然后在其他5个深度模型输出层之后添加一个全连接层,通过训练分别得到KCM模型与其他5个深度模型结合的权重,得到最终的匹配结果.

3.4 实验结果分析

在得到标注数据之后,对3.3节提出的各个对比模型进行实验,实验结果如表4所示.实验过程中各个模型采用分类的思路,评测指标为Accuracy值(Acc)、$F1$值、Precision值(Pre)和Recall值(Rec).

Table 4 Overall Result

表4 实验结果

Model	*Acc*	*F*1	*Pre*	*Rec*
ARC-I+KCM	0.682	0.343	0.270	0.521
ARC-II+KCM	0.671	0.341	0.274	0.509
MVLSTM+KCM	**0.729**	0.300	0.264	0.375
Matchpyramid+KCM	0.507	0.336	0.222	0.798
Learning to rank	0.597	0.359	0.230	**0.813**
KNRM+KCM(K-KCM)	0.689	**0.386**	**0.290**	0.580

Note: Bold figures represent the best-performing results under the corresponding metric.

通过表4可以发现,本文提出的KNRM+KCM对商品品类标题与微博热门话题内容文本进行匹配得到的Accuracy值、$F1$值和Precision值相对较高.综合所有结果采用Learning to rank得到的结果召回率最高.此外,所有匹配模型得到的结果普遍比采用商品品类知识图谱检索得到的结果要高(除了MVLSTM+KCM在召回率方面比较低).分析其原因,一方面是由于微博内容文本中蕴含着一定的语义信息,在单纯进行知识图谱检索时难以挖掘;另一方面有可能是商品标题文本中蕴含了很多知识图谱以外的词语.比如在吊坠这个品类中,在一些商品标题中包含了“送女友”、“惊喜”这一类很可能在微博

中出现的却并不与商品有直接关系的词语.

另外,结果表明采用深度网络的 ARC-I, ARC-II, MVLSTM, Matchpyramid 这 4 个模型效果不如 KNRM,甚至在最后 Learning to rank 综合排序中拖低了结果.原因可能在于商品标题文本有一定的特殊性,文本中很多词语并没有很强的词序关系,甚至颠倒顺序仍然通顺,比如把容量、商标、适用人群这些属性任意排列仍可以看作同一个商品的标题.因而基于计数原理忽略词序的 KNRM 模型有可能会有更好的结果.

针对 K-KCM 模型,综合知识图谱匹配结果与 KNRM 模型匹配结果时使用的最佳权重是通过训练得到的,为了验证训练效果,使用不同的权值计算得到相应结果如图 6 所示:

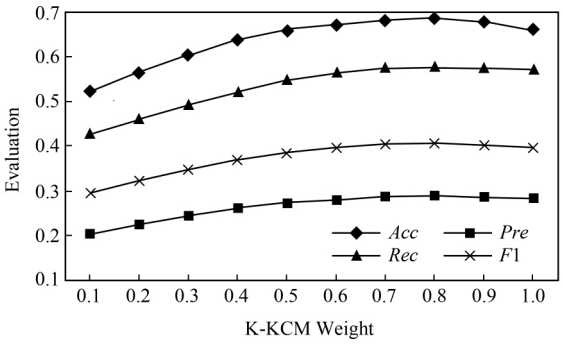

Fig. 6 Influence of K-KCM weight on model effect

图 6 K-KCM 权重对模型效果影响

实验发现,当 KNRM 权值为 0.813 左右时综合结果最佳,可见 KNRM 模型对于最终结果的影响更大.但 KCM 模型也对综合结果有一定的贡献,体现在 KNRM 模型权值大于 0.813 之后模型综合效果会下降.

进一步地,将知识图谱各部分识别结果的重要程度做对比,可以得到 4 个对比模型.

1) BH-KCM.只计算商品品类相关的品牌、热搜词、商品属性值的匹配分数和,不考虑泛产品品类名称部分匹配分数.

2) GH-KCM.只计算商品品类相关的泛产品品类名称、热搜词、商品属性值的匹配分数和,不考虑品类相关品牌部分匹配分数.

3) GB-KCM.只计算商品品类相关的品牌、泛产品品类名称、商品属性值的匹配分数和,不考虑热搜词部分匹配分数.

4) KCM.考虑知识图谱中所有实体部分的匹配得分,通过表 5 中知识图谱部分对比模型的结果,可以发现知识图谱各部分实体的匹配结果均对最终的匹配结果有贡献.其中热搜词部分对于模型的贡献值最大,当去掉热搜词相关匹配结果之后模型 GB-KCM 的效果显著下降.原因在于热搜词的语言风格与微博文本的语言风格更为相近,精确匹配成功的可能性更高,而泛产品品类名称和品牌名均是官方提供,与口语习惯不符.比如对于水杯商品品类,微博内容中更可能会出现"杯子",这种情况下热搜词更可能会匹配成功.

Table 5 Comparison Result of Knowledge Graph

表 5 知识图谱部分对比结果

Model	*Acc*	*F*1	*Pre*	*Rec*
BH-KCM	0.493	0.255	0.187	0.402
GH-KCM	0.452	0.229	0.165	0.372
GB-KCM	0.326	0.176	0.154	0.206
KCM	**0.502**	**0.264**	**0.195**	**0.409**

Note: Bold figures represent the best-performing results under the corresponding metric.

通过使用商品品类知识图谱与微博热点内容匹配的方法不仅可以获得匹配得分,还可以获得与匹配相关的匹配词,可以由此对匹配结果进行定性分析,匹配结果如表 6 所示.对于"甜馨公主裙"这个话题,匹配结果中得分比较高的 4 个品类分别是早教启智、芭比娃娃、裙子和儿童配饰.其中早教启智得分最高,因为它有 3 个关键词与话题相关,这 3 个关键词都来自于热搜词,由此也不难发现在知识图谱匹配过程中热搜词起了很大的作用.对于芭比娃娃和裙子这 2 个品类,都只有一个"公主"的热搜词与话题相关,但是芭比娃娃品类的得分却比裙子品类的高,原因在于芭比娃娃品类的热搜词数量比裙子品类的热搜词数量少,可以认为芭比娃娃这个品类具有更强的识别度.

Table 6 Knowledge Graph Matching Details

表 6 知识图谱匹配详情

Topic	Category	Matching Keywords	Score
Tianxin Princess Dress	早教启智	5 Years Old Princess Piano	15.98
	芭比娃娃	Princess	5.944
	裙子	Princess	5.105
	儿童配饰	Lovely	2.962

4 结 论

本文针对商品品类与微博热门话题的关联问题进行了深入探究,提出了 K-KCM 匹配模型,在电商品类知识图谱检索的基础上添加了文本匹配的方法,采用 KNRM 匹配模型对商品标题与微博热点内容文本进行了匹配,并通过实验证明模型的有效性,可以挖掘出微博话题与商品品类的关联.

在本文中电商品类知识图谱的检索结果并不高,下一阶段我们希望对知识图谱进行填充,从而提高这一部分的结果,同时在商品标题与微博内容匹配的部分,采用更多的模型进行尝试,提高综合排序的结果.

参 考 文 献

[1] Yin Zhimin, Yu Xiangzhan, Zhang Hongli. Commodity recommendation algorithm based on social network [C] // Advances in Computer Science and Its Applications. Berlin: Springer, 2014: 27-33

[2] Zhao Xin, Li Sui, He Yulan, et al. Connecting social media to e-commerce: Cold-start product recommendation using microblogging information [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(5): 1147-1159

[3] Zhao Xin, Li Sui, He Yulan, et al. Exploring demographic information in social media for product recommendation [J]. Knowledge and Information Systems, 2016, 49(1): 61-89

[4] Ye Wenwen, Zhang Yongfeng, Zhao Xin, et al. A collaborative neural model for rating prediction by leveraging user reviews and product images [C] //LNCS 10648: Proc of Asia Information Retrieval Symp. New York: Springer, 2017: 99-111

[5] Wang Jinpeng, Zhao Xin, He Yulan, et al. Leveraging product adopter information from online reviews for product recommendation [C] //Proc of the 9th Int AAAI Conf on Web and Social Media. Menlo Park, CA: AAAI, 2015: 464-472

[6] Rani A S, Bharathi K F. Product recommendation using convolution neural network in social media [J]. International Journal of Emerging Technology in Computer Science & Electronics, 2017, 24(4): 78-83

[7] Bai Ting, Zhao Xin, He Yulan, et al. Characterizing and predicting early reviewers for effective product marketing on e-commerce [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2271-2284

[8] Zeng Sai. Commodity recommendation system based on social network trust model [D]. Guangzhou: South China University of Technology, 2012 (in Chinese)
(曾赛. 基于社交网络信任模型的商品推荐系统[D]. 广州: 华南理工大学, 2012)

[9] Liang Linmeng, Qin Xiaohong. Research on consumers online shopping decision-making and recommendation of commodity based on social media network [C] //Proc of Cluster Computing. Berlin: Springer, 2018: 1-11

[10] Jian Yi, Xiao Yunpeng, Liu Yanbing. Incorporating multiple attributes in social networks to enhance the collaborative filtering recommendation algorithm [J]. International Journal of Advanced Computer Science and Applications, 2016, 7(4): 60-67

[11] Wang Jinpeng, Zhao Xin, Wei Haitian, et al. Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs [C] //Proc of the 2013 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1337-1347

[12] Jiang Renhui, Wang Ting, Tang Jintao. Named entity recognition for micro-blog [J]. Computer and Digital Engineering, 2014, 42(4): 647-651

[13] Li Gang, Huang Yongfeng. An approach to named entity recognition towards Micro-blog [J]. Application of Electronic Technique, 2018, 44(1): 118-124

[14] Li Zhiguo, Cai Dongfeng, Zhou Qiaoli, et al. Mutual information method used to recognize name entity [J]. Journal of Shenyang Institute of Aeronautical Industry, 2007, 24(1): 35-37 (in Chinese)
(李治国, 蔡东风, 周俏丽, 等. 在篇章中利用互信息识别命名实体的研究[J]. 沈阳航空工业学院学报, 2007, 24(1): 35-37)

[15] Liu Yujiao, Ju Shenggen, Li Ruochen, et al. Chinese Weibo named entity recognition based on deep learning [J]. Journal of Sichuan University, 2016, 48(2): 145-149 (in Chinese)
(刘玉娇, 琚生根, 李若晨, 等. 基于深度学习的中文微博命名实体识别[J]. 四川大学学报, 2016, 48(2): 145-149)

[16] Chen Xiaoxiao, Liu Bo. Extracting open domain events in Microblogs [D]. Beijing: Beijing University of Technology, 2016 (in Chinese)
(陈箫箫, 刘波. 微博中的开放域事件抽取[D]. 北京: 北京工业大学, 2016)

[17] Xiong Chenyan, Dai Zhuyun, Jamie C, et al. End-to-end neural ad-hoc ranking with kernel pooling [C] //Proc of SIGIR'17. New York: ACM, 2017: 55-64

[18] Guo Jiafeng, Fan Yixing, Ai Qingyao, et al. A deep relevance matching model for ad-hoc retrieval [C] //Proc of CIKM'16. New York: ACM, 2016: 55-64

[19] Pang Liang, Lan Yanyan, Xu Jun, et al. Depth text matching review [J]. Chinese Journal of Computers, 2017, 40(4): 985-1003 (in Chinese)
(庞亮, 兰艳艳, 徐军, 等. 深度文本匹配综述[J]. 计算机学报, 2017, 40(4): 985-1003)

[20] Hu Baotian, Lu Zhengdong, Li Hang, et al. Convolutional neural network architectures for matching natural language sentences [C] //Proc of NIPS'14. Cambridge, MA: MIT Press, 2014: 2042-3060

[21] Pang Liang, Lan Yanyan, Guo Jiafeng, et al. Text matching as image recognition [C] //Proc of AAAI'16. Menlo Park, CA: AAAI, 2016: 2793-2799

[22] Wan Shengxian, Lan Yanyan, Guo Jiafeng, et al. A deep architecture for semantic matching with multiple positional sentence representations [C] //Proc of AAAI'16. Menlo Park, CA: AAAI, 2016: 2835-2841

[23] Cao Zhe, Qin Tao, Liu Tieyan, et al. Learning to rank: From pairwise approach to listwise approach [C] //Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 129-136

[24] Li Ping, Burges Christopher J C, Wu Qiang. McRank: Learning to rank using multiple classification and gradient boosting [C] //Proc of NIPS'07. New York: Curran Associates Inc, 2007: 897-904

Zuo Xiaochen, born in 1996. BSc. His main research interests include information retrieval, text matching and named entity recognition.

Dou Zhicheng, born in 1980. PhD. His main research interests include information retrieval, natural language processing, Web search, data mining, information extraction, and big data analysis.

Huang Zhen, born in 1995. Master. His main research interests include named entity recognition, sentiment analysis, text matching and knowledge graph.

Lu Shuqi, born in 1997. BSc. Her main research interests include data mining, natural language processing and information retrieval.

Wen Jirong, born in 1972. PhD. His main research interests include Internet big data management, information retrieval, data mining, and machine learning.