软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):262−277 [doi: 10.13328/j.cnki.jos.005018]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

无监督的中文商品属性结构化方法*

侯博议, 陈 群, 杨婧颖, 李战怀

(西北工业大学 计算机学院,陕西 西安 710129)

通讯作者: 侯博议, E-mail: nagisafurukawa@qq.com

摘 要: 从非结构化商品描述文本中抽取结构化属性信息,对于电子商务实现商品的对比与推荐及用户需求预测等功能具有重要意义.现有结构化方法大多采用监督或半监督的分类方法抽取属性值与属性名,通过文法分析器分析属性值与属性名之间的文法依存关系,并根据关联规则实现属性值与属性名的匹配.这些方法存在以下不足:(1) 需要人工标记部分属性值、属性名及它们之间的对应关系;(2) 属性值-属性名匹配的准确度受到语言习惯、句意逻辑、语料库及属性名候选集质量的严重制约.提出了一种无监督的中文商品属性结构化方法.该方法借助搜索引擎,基于小概率事件原理分析文法关系来抽取属性值与属性名.同时,提出相对不选取条件概率场,并使用 Page Rank 算法来计算属性值与属性名的配对概率.该方法无需人工标记的开销,且无论商品描述中是否显式地包含相应的属性名,该方法都能自动抽取到属性值并匹配相应的属性名.使用百度搜索引擎上的真实语料,针对 4 类商品的中文描述进行了实验.实验结果验证了对于候选属性名的自动生成,所提出的基于搜索引擎搜索属性值,并在包含属性值的搜索结果中抽取一般名词的候选属性名生成方法与只在描述句中抽取一般名词的候选属性名生成方法相比,查全率提高了 20%以上;对于非量化类属性,所提出的基于相对不选取条件概率场的属性值-属性名匹配方法与基于依存关联的方法相比,Rank-1 的准确率提高了 30%以上,平均 MRR 提高了 0.3 以上.

关键词: 结构化;相对不选取条件概率场;Page Rank;基于概率的文法分析;搜索引擎

中图法分类号: TP311

中文引用格式: 侯博议,陈群,杨婧颖,李战怀.无监督的中文商品属性结构化方法.软件学报,2017,28(2):262−277.http://www.jos.org.cn/1000-9825/5018.htm

英文引用格式: Hou BY, Chen Q, Yang JY, Li ZH. Unsupervised structralization method of merchandise attributes in Chinese. Ruan Jian Xue Bao/Journal of Software,2017,28(2):262−277 (in Chinese). http://www.jos.org.cn/1000-9825/5018.htm

Unsupervised Structralization Method of Merchandise Attributes in Chinese

HOU Bo-Yi, CHEN Qun, YANG Jing-Ying, LI Zhan-Huai

(School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Extracting attribute names and values from textual product descriptions is important for many e-business applications such as user demand forecasting and product comparison and recommendation. The existing approaches first use supervised or semi-supervised classification techniques to extract attribute names and values, and then match them by analyzing their grammatical dependency. However, those methods have following limitations: (1) They require human intervention to label some attributes, values and the matching relationship between them; (2) The matching accuracy may be greatly affected by language habits, semantic logic, and the quality of corpus and candidates sets. To address these issues, this paper proposes an unsupervised approach for attribute name and value extraction

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316203); 国家自然科学基金(61332006, 61472321); 西北工业大学基础研究基金(3102014JSJ0013, 3102014JSJ0005)

Foundation item: National Program on Key Basic Research Project of China (973) (2012CB316203); National Natural Science Foundation of China (61332006, 61472321); Northwestern Polytechnical University Foundation for Fundamental Research (3102014JSJ0013, 3102014JSJ0005)

收稿时间: 2015-08-15; 修改时间: 2015-12-02; 采用时间: 2015-12-25

and matching in Chinese textual merchandise descriptions. Taking advantage of search engine, it extracts the candidate set of attribute names with respect to a value by analyzing grammatical relation based on the principle of small probability event. A new algorithm for computing the matching probability between attribute names and values is also designed based on relative conditional deselect probability and Page Rank. The proposed approach can effectively extract attribute names and values from Chinese textual merchandise descriptions and match them without any human intervention, no matter whether the attribute name appears in the textual description or not. Finally, the performance of the proposed approach is evaluated on the textual descriptions of 4 types of merchandise using the search engine of Baidu. The experimental results show that the new approach for attribute name extraction can improve recall by 20%, compared with the approach of directly extracting attribute names from textual descriptions. Moreover, the new approach achieves considerably higher matching accuracy (above 30% if measured by the percentage of rank-1, above 0.3 if measured by MRR) than the existing techniques based on grammatical dependency analysis for non-quantization attributes.

Key words: structralization; relative conditional deselect probability field; Page Rank; grammatical relation analysis based on probability; search engine

商品的描述文本包含 1 个或多个结构化的属性.例如,“嘟麦依纯棉免烫处理修身时尚青年圆领长袖女款衬衫”这句描述对应着“品牌:嘟麦依,面料:纯棉,工艺:免烫,年龄段:青年,领型:圆领,款式:女,袖长:长袖,商品类型:衬衫/服装”的结构化信息.这些商品描述通常存在于商家对商品的介绍及用户的评论中.如何从非结构化商品描述文本中抽取结构化属性信息,对于电子商务实现商品的对比与推荐及用户需求预测等功能具有重要意义[1,2].

结构化属性抽取的关键在于对商品描述文本中属性值的识别,以及寻找该属性值所对应的属性名.因语言习惯、句意逻辑、语料库及属性名候选集质量不佳等诸多原因,目标属性名不一定出现在描述句及其上下文中.即使出现属性名,也未必是与该属性值对应的属性名.因此,为属性值寻找并匹配对应的属性名是结构化的难点.

现有针对商品信息抽取的研究大多关注如何基于统计学特征从商品描述及评论文本中抽取商品的属性名[3−8],而不涉及属性值与属性名的匹配.近年来有研究针对实体-类型的结构化及实体间关系的异构网络信息结构的构建[9−15],但只能用于结构化实体属性.有少部分研究[2]针对英文语料的属性信息结构化,依靠人工标记部分属性名、属性值及它们之间的对应关系,通过机器学习从语料库抽取属性值和属性名;再使用类似 Minipar 的文法分析器分析属性值与属性名语法上的依存关系,并根据关联规则进行配对[2]以实现结构化.还有部分研究使用监督或半监督的机器学习方法对隐性属性的抽取与结构化做了研究[2,16−18],可以抽取到结构化的隐性属性.针对中文语料的研究大多关注如何更有效地抽取商品属性[7,18,19],其原理与针对英文语料的属性抽取原理[2]或基于主题建模的关键短语抽取原理[14,15]类似.

然而,以上商品属性的抽取与结构化方法存在以下问题.(1) 需要人工标记部分属性值、属性名及对应关系,或人为指定候选属性名.(2) 没有结合商品类型判断词的类别,会导致分类错误.例如,CPU 既是电脑的一个属性名,同时其自身也是一种商品类型;CPU 的一个具体型号,如 FX9590,既是 CPU 作属性名时的一个属性值,同时也是一种商品.同样的例子也多见于材料和电子设备接口.(3) 因语言习惯、句意逻辑或语料库、候选词质量不佳等诸多原因,属性名并不总是与相应属性值关联最高的词,因此,基于依存关联的匹配方法的准确度会受到以上因素的严重制约;以“嘟麦依纯棉免烫处理修身时尚青年圆领长袖女款衬衫”为例,即使由人工过滤属性名候选集中非属性名类别的词,根据关联规则进行属性值与属性名的匹配效果依然欠佳:与“免烫”对应的最高关联词是“面料”,与“女款”所对应的最高关联词是“价格”、“参数”和“尺码”;当属性名候选集质量较差,如存在非属性名类别的干扰词时,基于依存关联的方法的准确度更低.

针对以上问题,本文提出了一种无监督的中文商品属性结构化方法.该方法通过信息熵原理确定商品类型,再结合商品类型,借助搜索引擎,基于小概率事件原理分析文法关系来抽取属性值与属性名;然后提出相对不选取条件概率场,并使用 Page Rank 算法进行属性值与属性名的配对.该方法无需人工标记的开销,无论相应的属性名是否出现在商品描述中,都能抽取到商品描述中的属性值及其属性名;并且,在语料库或属性名候选集质量较差,或是因语言习惯导致属性值与其相应属性名关联度较低时,与基于依存关联或词权重的匹配方法相比,本文提出的基于相对条件概率场的匹配方法的准确度显著提高,特别是对于非量化类属性值.

本文的主要贡献如下:

(1) 提出了一种借助搜索引擎的无监督商品属性结构化方法.无论商品描述中是否显式地包含相应的属性名,该方法都能自动抽取属性值并匹配相应的属性名.

(2) 提出了一种基于小概率事件原理分析文法关系的属性值、属性名类别判断方法,只要通过搜索引擎分析少量语料,就能判断出属性值与属性名类别的词.解决了自然语言处理方法分析效率较低,以及在商品描述或语料库质量欠佳时,特征学习的方法准确度较低的问题.

(3) 提出了基于相对条件概率场的属性值-属性名匹配方法.不同于单纯基于属性值-属性名依存关联的匹配方法,该方法不仅分析属性值与候选属性名之间的文法关系与关联规则的相互置信度,而且分析各候选属性名之间的文法关系进行属性值-属性名的匹配.解决了当语料库及属性名候选集质量较低,或因属性值-属性名依存度受到语言习惯与句意逻辑制约时,基于依存关联匹配方法的准确度较低的问题.

(4) 通过使用真实的语料进行实验,验证了本文方法的有效性.本文的基于搜索引擎搜索属性值,并抽取一般名词作为候选属性名的属性名候选集生成方法,与只在描述句中抽取一般名词作为候选属性名的属性名候选集生成方法相比,查全率提高了 20%以上.对于非量化类属性,本文提出的基于相对条件概率场的属性值-属性名匹配方法,与基于依存关联的匹配方法相比,Rank-1 的准确率提高了 30%以上,平均 MRR 提高了 0.3 以上.

本文第 1 节介绍目前对于商品属性的抽取及结构化的相关研究.第 2 节介绍属性值与候选属性名的自动抽取方法.第 3 节介绍相对条件概率场以及基于相对条件概率场的属性值-属性名匹配方法.第 4 节给出实验方法、实验结果及结论.

1 相关工作

现有的属性抽取研究大多针对在新闻或评论语料中抽取商品的属性,并不进行属性的属性值-属性名结构化.文献[3]通过制定 3 种启发式规则生成具有 BNP(base noun pharse),dBNP(definite base noun phrase),bBNP (beginning definite base noun phrase)结构的候选,再用类似于 TF-IDF(term frequency-inverse document frequency)原理的混合语言模型与似然比来检验两种算法的抽取属性.文献[4]针对日文使用特定的语义模式,基于新闻报纸题目中候选词与商品的关联规则来生成单名词属性.文献[5]从评论语料中挖掘事务集,并依据关联规则得出其中的频繁项作为商品属性.文献[16]还对多词属性,即名词短语形式的属性的识别做了研究.文献[20]使用朴素贝叶斯或 EM(expectation-maximization)方法自动抽取商品属性.当作为结构化目标语料源的各 Web 页面涉及不同的结构化格局(layout)时,文献[21]使用概率图(probabilistic graphical model)自动判断结构化属性格局对不同网页格局的依赖,以进行结构化属性抽取并对属性进行规范化.部分研究针对隐性属性的挖掘,能够挖掘出指定的没有显式出现于描述文本中的属性:文献[7,8]基于 DC(domain relevance)和 DR(domain consensus)两个统计特性挖掘隐性属性;文献[2,16,17]使用监督或半监督的机器学习方法对每类指定的隐性属性进行特征学习,并使用 EM 算法挖掘隐性属性;文献[18]提出了基于意见相似度的标准化主题建模,并通过与主题的关联分析来挖掘隐性属性.文献[22]基于 BNP 与 bBNP 进行商品属性的优缺点情感分析,并不直接针对商品属性进行结构化.

对于属性值-属性名的结构化,文献[6]提出的 OPINE 系统使用 KNOWITALL 知识库给出各类商品的属性名及相应属性值抽取规则,并结合点互信息(point-wise mutual information,简称 PMI)选出各属性名所对应的属性值,实现了远程监督的商品属性结构化;基于远程监督的方法的效果取决于知识库的完备性,而维护知识库完备性的人工开销甚至大于直接进行完全人工标注的属性结构化.文献[2]系统地提出了目前广泛使用的针对属性名-值对的抽取与结构化的方法:它对属性值、属性名进行少量的人工标记,并使用朴素贝叶斯公式或 Co-EM 算法将语料库中出现的词或短语分为属性值和属性名两类;再使用 Minipar 分析属性值与候选属性名的依存关联并作为配对依据.然而,以上方法都需要一定量的人力开销,而且基于属性值-属性名依存关联的匹配效果受语言习惯、句意逻辑与干扰词的严重制约.

对于中文语料的商品属性抽取,文献[7]使用 IDF(inverse document frequency),NDC(normalized domain consensus)与 NDR(normalized domain relevance)等统计学原理识别文本中的属性.文献[19]提出了一种语言依存分析和语料库统计相结合的未登录(out-of-vocabulary,简称 OOV)产品属性挖掘算法 OPINAX,其原理与文献[2]

提出的针对英文语料的商品属性抽取方法类似:该方法标注一个小规模的商品属性集,使用浅层语言分析工具分析这些属性的统计特征(如伴随情感词共同出现的频率)以进行机器学习,实现了从语料中抽取商品属性.然而,以上研究只针对从评论语料中如何抽取属性,而不进行属性值与属性名的匹配,且同样需要一定量的人工标注.

近年来有很多研究针对实体的抽取与实体-类型的结构化[9−15],实现了半监督或基于 Web 知识库的远程监督的实体-类型结构化.Han 等人[14,15]基于关键短语分析实体之间的联系,并基于实体之间的联系传播类型实现了更准确的异构网络信息结构化.然而,以上文献所研究的结构化与本文研究的无监督商品属性结构化存在以下不同.(1) 尽管实体包括一些属性值和属性名,但属性值和属性名并不一定是实体,如量化属性值和抽象概念属性名等.(2) 实体的候选类型必须预先给出(如使用 Web 知识库预定候选类型)并具有显著的区分度,且只有在区分属性名的粒度上才能用于属性值-属性名的匹配;而本文对属性结构化的研究目标是自动生成候选词,并在话题、候选类别及其粒度不尽相同的候选词中匹配相应的属性名.(3) 匹配实体类型需要半监督或远程监督实现,对人工标注量和人工维护的知识库完备性有一定的要求;而本文的目标是实现无监督的属性结构化.此外,由于以上方法针对较为严谨的学术文献、新闻语料和具有固定 html 格式的 Tweets,因此对语料质量要求极高.然而,大多数网络文本的质量都没有较为严格的保证.

2 属性值与候选属性名的自动抽取

本节依靠搜索引擎,根据文法关系抽取属性值与属性名.属性值和属性名的抽取都分为两个步骤:(1) 依靠 NLPIR 分词器与搜索引擎分别生成属性值与属性名候选集;(2) 基于文法关系确定候选集中词的类别,并过滤候选集以得到属性值与属性名.

第 2.1 节介绍属性值候选集的自动生成方法.第 2.2 节介绍属性名候选集的自动生成方法.第 2.3 节介绍如何根据文法关系判断词的类别,并过滤候选集以得到属性值与候选属性名.

2.1 属性值候选集的自动生成方法

本节使用 NLPIR 分词器提取出商品描述句中的专有名词、数量词、区别词和一般名词,将它们列入属性值的候选集.商品或其部件的品牌与型号是专有名词,包括数字与字母的组合.此外,根据区别词或数词与其后相邻的名词或量词组成名词和数量词的词法,区别词或数词需要结合后面相邻的名词或量词作为属性值候选词.

2.2 属性名候选集的自动生成方法

本节基于搜索引擎自动生成候选属性名,无论商品描述中是否显式地出现相应的属性名.

由于物品属性描述语料的表述语法、语意逻辑、语言习惯和上下文语境是随机的,因此相应属性名可能出现,也可能不出现,且同样是随机的,即服从 0-1 分布.由于属性值与相应属性名存在逻辑上的直接关联,在包含其中一词的语料及其上下文中,包含另一词的概率不为 0.因此,理论上只要与属性值相关的语料足够多,出现相应属性名的概率就足够大.

与属性值相关的语料通过在搜索引擎上对属性值搜索得到.如果属性值是一个量化属性值,则它在不同的商品类型及描述中可能对应不同的属性名,在数码产品中尤为居多.因此,对于量化属性值,需结合商品类型和相应描述句的关键词作为搜索内容检索相关语料.商品类型可通过在搜索引擎中搜索商品标题,并使用 NLPIR[23] 提供的基于统计学原理的关键词抽取算法 NLPIR_GetKeyWords,选取搜索结果中关键词权重最大的一般名词得出.NLPIR_GetKeyWords 是 NLPIR/ICTCLAS 汉语分词系统中采用信息熵自动抽取语料中关键词的算法[23],因此可用于抽取与特定商品或商品属性相关的语料中相应的商品名或属性名.

在实验中发现,只要取属性值搜索结果中的前 10 个 Snippet,相应属性名的召回率就接近饱和.因此,本文实验均选取前 10 个 Snippet 作为生成属性名候选集的语料.图 1 给出了使用百度搜索引擎生成候选属性名时,属性名召回率与所取 Snippet 数目的关系.

由于属性名只可能是一般名词或名词短语,因此本文使用分词器抽取语料中的一般名词,并使用 NLPIR_GetKeyWords 抽取语料中的名词短语作为属性名的候选集.

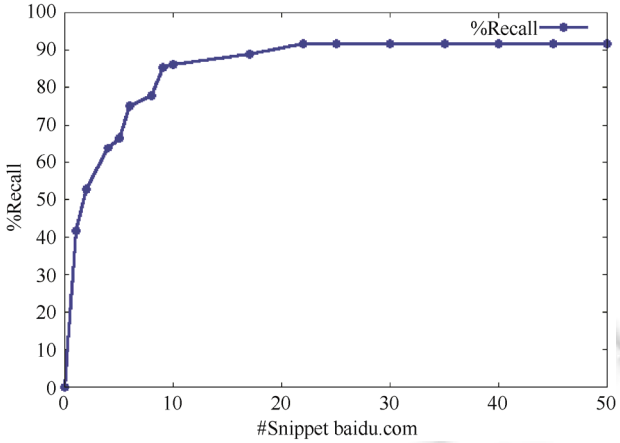


Fig.1 #Snippet effects %Recall of attribute (by baidu.com)

图 1 属性名召回率与所取 Snippet 数目的关系(使用百度搜索引擎)

2.3 基于文法关系的词类别判断及候选集过滤

本节基于判断关系和偏正关系判断一个词的类别,具体文法关系特征 $Q=\{Q_i\}$由表 1 给出.通过判断一个词 A 是否符合表 1 中目标类别 Y 的各个特征 $B(A;Q_i)$及定义相应的符合度 $h(A;Q_i)$,来计算目标类别 Y 的符合度 $h(A;Y)$并判断 A 是否符合目标类别 Y,以进行候选集的过滤.其中,偏正短语是指由修饰词和中心词组成,两词之间有修饰与被修饰关系的短语.为了便于区分偏正短语的修饰词与中心词,除指明外,本文中所有"A 和 B 组成偏正短语""A 和 B 具有偏正特征"的表述均指 A 作修饰词,B 作中心词.此外,为了增加可用的文法信息,本文还引入了并列属性值的概念:本文将语料中与属性值 V 具有相同量化单位、前后词缀以及通过并列特征标点符号或并列特征词与 V 组成并列短语的属性值类别的词记为 V',并赋予其与 V 的相似性度量为$\nabla_V V'$.如果它们是具有相同量化单位或前后词缀的属性值,则$\nabla_V V'$为它们的字符串相似度,否则,令$\nabla_V V'=1$.本文假定对于特定的商品类型,V'与 V 在文法关系上等价.因此,在文法关系考察中,∇可联络 V'使其近似地视为 V,增加了可用的文法信息.

Table 1 Grammatical relation features of value and attribute Q

表 1 属性值与属性名类别的文法特征 Q

类别	特征词	特征
商品属性值或商品属性	商品类型	组成偏正短语(特征词为中心词)
商品属性值	相应属性名	组成偏正短语(特征词为中心词)
商品品牌	"专卖店"	组成偏正短语(特征词为中心词)
商品属性名	相应属性值 商品类型或与商品类型组成偏正短语的商品部件	组成偏正短语(特征词为修饰词)或构成判断句 组成偏正短语(特征词为修饰词)

基于以上特征再定义相应类别的否定特征:在偏正短语中交换目标词与特征词的身份,该特征变为相应类别的否定特征.

对于 $B(A;Q_i)$的判断,由于现有的文法分析器大多只根据词性与形式判断文法,可能导致理论文法与实际文法不一致,因此本文基于搜索引擎及小概率事件原理判断 $B(A;Q_i)$:如果某词 A 在现实用语中不具有特征 Q_i,则 A 在同时包含 A 与相应特征词的语料中具有特征 Q_i 是小概率事件,进而在特定的考察中不会发生.其中,本文通过搜索引擎得到的包含 A 或 B 的语料 $R_{A,B}$ 来分析 A,B 的偏正特征.

(1) 在 $R_i\in R_{AB}:=\{R_k\in R_{A,B}|A\in R_k,B\in R_k\}$中,$A,B$ 直接相邻或 A,B 中只有偏正特征词"的",且 A,B 的两侧或者没有其他词,或者是动词或代词(偏正关系分界词).

(2) 在 $R_i\in R_{AB}$中分别距离 A,B 最近的动词或代词与 A,B 之间的词集合为 $L(R_i)$,则 $L(R_i)=\varnothing$或$\exists r\in R_{AB},L(R)\cap L(R_i)=\varnothing$,表示 R_i 中 A,B 的修饰关系不受 $L(R_i)$的影响,即 A,B 具有偏正特征.

(3) 当 $Card(\{R_i \in R_{AB} | F(A,B) \in F_{(1)} \cup F_{(2)}\})$时,即使距离 A,B 最近的动词或代词与 A,B 之间存在形容词、数词、量词、区别词或名词,也认为 A,B 具有偏正特征.其中,$Card(X)$表示集合 X 的势;$F(X)$表示词 X 所符合的表 1 中的文法特征.

本文将类别 Y 的符合度 $h(A;Y)$定义为 Y 的各文法特征的符合度 $h(A;Q_i)$之和.对于 $h(A;Q_i)$的计算,由于表 1 中所有特征都不是充要特征,因此需要定义并结合 Q_i 对类别 Y 的置信系数或否决系数.如果 Q_i 是 Y 类别的肯定特征,则本文定义 $c_Y(Q_i)$为 Q_i 的置信系数;如果 Q_i 是 Y 类别的否定特征,本文还定义 $u_Y(Q_i)$为 Q_i 对 Y 类别的否决系数:

$$u_Y(Q_i) := \frac{1 - c_Y(Q_i)}{c_Y(Q_i)}.$$

理论上,$c_Y(Q_i)$应与文法特征 Q_i 对类别 Y 的置信度 $P(A \in Y | F(A)=Q_i)$成正比.但是,文法特征的置信度会随商品类型及属性的不同而变化,并且置信度计算的目的是自动抽取结构化属性,因此不能通过标记每类商品类型的所有属性来计算置信度.因此,本文通过以下理论分析,得出属性名文法特征置信度的近似关系:当商品类型与某词具偏正关系时,该词为该商品类型属性名的置信度要远高于相应属性值与其具有偏正关系时的置信度.这是因为一个属性值修饰的对象可以是具有该属性的商品类型、参数类型或相应的属性名,当两者具有多重类别时情况更加复杂.而一个商品类型所修饰的对象大多是其部件等属性名,仅在两者具有多重类别时才会难以确定.因此,本文近似地认为:

$$\begin{aligned} c_{Attribute}(F_{Of-Modifier-Core-Phrase}(ProductType, A)) &= c_{Attribute}(F_{Of-Deter\min itive-Sentence}(A, CorrespondingValue)) \\ &= 8c_{Attribute}(F_{Of-Modifier-Core-Phrase}(A, CorrespondingValue)), \end{aligned}$$

并令基准置信系数: $c_{Attribute}(F_{Of-Modifier-Core-Phrase}(A, CorrespondingValue)) = 0.1$.

后文计算属性名候选词的匹配概率时也会涉及文法符合度的计算.由于所有候选词的符合度计算都使用相同的文法置信系数,因此理论上基准系数 $c_{Attribute}(F_{Of\text{-}Modifier\text{-}Core\text{-}Phrase}(A, CorrespondingValue))$的取值不影响匹配结果.

本文将至少符合目标类别 Y 的一个文法特征,且对目标类别 Y 的符合度 $h(A,Y)$大于显著性水平的词判断为 Y 类别.具体算法如下.

算法 1. 判断目标词是否属于目标类别.

输入:目标词 A,A 的待判断类别 Y,表 1 中的固定特征词及已确定类别的特征词的集合 C,表 1 中关于类别 Y 的文法特征集合 $Q(Y)$,文法特征 Q 的特征词集合 $K(Q)$,商品类型 T.

输出:A 是否属于 Y 类别 $B(A;Q_i)$.

1. FOREACH $Q_i \in Q(Y)$ AND $K(Q) \subseteq C$
2. $R=SearchEngine(A+K(Q_i))$
3. $P(A,Q_i) = \dfrac{Card\left(\{R_i \in R \mid F(A;R_i) = Q_i\}\right)}{Card\left(\{R_i \in R \mid A \in R_i, K(Q_i) \in R_i\}\right)}$
4. IF $P(A,Q_i)>\alpha$
5. $B(A;Q_i)$=TRUE
6. ELSE
7. $B(A;Q_i)$=FALSE
8. IF Q_i 是 Y 的肯定特征
9. $h(A;Q_i) := c_Y(Q_i)\dfrac{\sum_{R_j^k \in R_i} \nabla_v V' \cdot D(A;R_j^k)}{Card\left(\{R_i \in R \mid A \in R_i, K(Q_i) \in R_i\}\right)}$
10. ELSE (即 Q_i 是 Y 的否定特征)
11. $h(A;Q_i) := -u_Y(Q_i)\dfrac{\sum_{R_j^k \in R_i} \nabla_v V' \cdot D(A;R_j^k)}{Card\left(\{R_i \in R \mid A \in R_i, K(Q_i) \in R_i\}\right)}$

12. $D(A;R_j^k):=\begin{cases}1, & F(A;R_j^k)=Q_i\\ 0, & F(A;R_j^k)\neq Q_i\end{cases}$

13. $h(A;Y)=\sum_{Q_i\in Q(Y)}h(A;Q_i)$

14. IF $\exists B(A;Q_i)$=TRUE AND $h(A;Y)>\alpha$

15. $B(A;Y)$=TRUE

16. ELSE

17. $B(A;Y)$=FLASE

18. RETURN $B(A;Y)$

本文中的α都表示显著性水平,应取常用的 0.01~0.05.本文所有的显著性水平在实验中都指定为α=0.04.可以通过实验针对使用的语料库选择更优的显著性水平.

3 基于相对条件概率场的属性值-属性名匹配

在现有的监督或半监督的结构化方法中,属性值与属性名候选集的质量很高,较少存在干扰词,因此只根据依存关联进行匹配.然而,即使候选集质量很高,其匹配的准确度依然会受到语言习惯及语料库质量的制约.例如,在文献[2]给出的针对某种商品类型的结构化实验结果中,属性值-属性名对中至少一方正确的比例高达98.91%,这表明属性名候选集的质量很高;但属性对双方全部正确的比例仅为 54.90%,即在候选集质量很高时,匹配的准确率也只有55.5%.对于本文的无监督结构化方法,自动生成的属性名候选集的质量远低于监督生成的或经过人工筛选过的候选集的质量,并且所使用的语料库是普遍没有质量保证的网络语料.这将导致以下问题:

(1) 属性名候选集中存在大量与属性值关联性极强但却不是目标属性名的干扰词.

(2) 作为语料库的网络语料的语言习惯或句意逻辑有极大的随意性.

这些问题将导致基于依存关联或词权重等统计学原理的匹配方法准确度严重衰退.针对以上问题,本文提出了相对不选取条件概率场,并使用 Page Rank[24]算法进行属性值-属性名的匹配.Page Rank 是 Google 等搜索引擎根据网页之间相互的超链接引用,基于马尔可夫收敛过程计算网页权重与排名的算法.类比网页之间的相互影响,本文提出候选词间的相对条件概率场,并使用 Page Rank 算法计算各候选词的选取概率.与将属性值、商品类型与属性名的文法关系以及属性值与属性名的依存关联作为匹配依据的传统方法相比,本文的匹配方法仅根据它们得出各候选属性名的初始不匹配概率;然后,通过考察各候选属性名之间的文法关系得出相对条件概率场并标准化,再使用 Page Rank 算法迭代出最终的不匹配概率以作为匹配依据.

图 2 和图 3 对比了传统的属性值-属性名匹配方法与基于相对条件概率场的属性值-属性名匹配方法的不同.

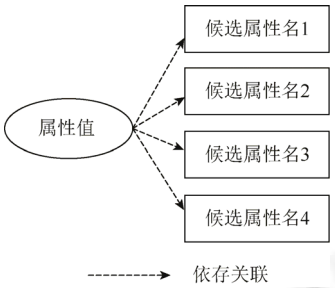


Fig.2 Conventional approach to match value and attribute

图 2 传统的属性值-属性名匹配方法

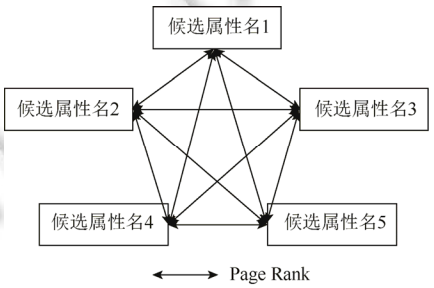


Fig.3 Approach based on relative conditional probability field to match value and attribute

图 3 基于相对条件概率场的属性值-属性名匹配方法

3.1 相对不选取条件概率场及属性值-属性名的匹配

本节提出了相对不选取条件概率场,并基于相对不选取条件概率场计算候选属性名的匹配概率,并得出匹配位序.记商品类型为 T,待匹配属性名的属性值为 V,V'代表 V 的并列属性值,相应属性名候选集 W 中的元素为 $W_1,W_2,W_3,...,W_n$, W_i 不与属性值相匹配的事件记为 A_i.

单纯考察属性值-属性名的文法关系或依存关联所得到的匹配概率受各种因素影响较大.因此,在属性值-属性名的匹配中,本文还考察了各属性名之间的文法关系,并得出各属性名间不匹配的相对条件概率:当 W_j 不是目标属性名时,W_i 也不是目标属性名的概率 $P(A_i|A_j)$.之所以将 A_i(即候选词 W_i 不与属性值相匹配)作为条件,而不是将候选词 W_i 与属性值相匹配作为条件,是因为多数属性值对应的正确或合适的属性名只有 1 个,并且在结构化时只为属性值匹配 1 个正确或合适的属性名即可;如果条件定为候选词 W_j 是与属性值相匹配的目标属性名,则其他候选词也同样作为目标属性名的条件概率恒为 0,不再具有表示相对关系的意义.本文称被作为假设条件的词为参考词,称在假设条件下被观测的词为目标词,并称相对条件概率矩阵$[a_{ij}=P(A_i|A_j)]$为相对条件概率场.

相对条件概率完全由目标词与参考词的文法关系及其逻辑得出,不涉及其他候选词.因此,通过不同的参考词所得出的相对条件概率间相互独立,受参考词自身词义等性质的影响,没有统一的度量标准.为了使相对于不同参考词的条件概率的度量具有统一标准,应将相同参考词下各目标词的相对条件概率标准化.标准化的相对条件概率的计算式为

$$\overline{\overline{P}}(A_i \mid A_j)=\frac{P(A_i \mid A_j)}{\sum_{k=1,k\neq j}^{n} P(A_k \mid A_j)}.$$

本文使用 Page Rank 算法计算各候选词的匹配概率:通过考察属性值、商品类型与属性名的文法关系,以及属性值与属性名的依存关联赋予各候选属性名初始的不匹配概率 $P_0(A_i)$;然后,将标准化的相对条件概率作为 Page Rank 的有向边权值并通过 Page Rank 算法迭代计算各候选词的概率 $P_k(A_i)$,直到 $P_0(A_i)$一致收敛,作为最终不匹配概率;最后,通过不匹配概率得出匹配概率,并得出匹配位序.其中,根据 Page Rank 算法,第 k 次迭代中 W_i 的不匹配概率计算式为

$$P_k(A_i)=dP_0(A_i)+(1-d)\sum_{j=1,j\neq i}^{n}\overline{\overline{P}}(A_i \mid A_j)P_{k-1}(A_j).$$

并且由于对$\forall j,A_j$的有向边权值都满足 Page Rank 算法的收敛条件:

$$\sum_{i=1,i\neq j}^{n}\overline{\overline{P}}(A_i \mid A_j)=1,$$

因此,算法一定收敛.

下面证明基于相对条件概率场使用 Page Rank 算法所得出不匹配概率的正确性.

定理 1. 以分析候选词两两相对文法所得到的标准化相对条件概率作为转移概率的马尔可夫收敛过程,一定收敛到各候选词的不匹配概率.

证明:为了准确地表示相对条件概率的意义,在定理证明过程中,$\forall W_i,W_j\in W$,相对条件概率 $P(A_i|A_j)$将完整地记为 $PW_i,W_j(A_i|A_j;T,V)$,即对于确定的商品类型 T 及属性值 V,通过在包含 W_i,W_j 的语料中进行 W_i,W_j 的文法分析,所得出的相对不匹配概率.

(1) 下面证明各候选词的匹配状态可以是以相对条件概率场为转移概率的马尔可夫过程[25].

由相对条件概率 $PW_i,W_j(A_i|A_j;T,V)$的计算方法可知:对于确定的商品类型 T 及属性值 V,W_i,W_j 的相对不匹配概率完全由包含 W_i,W_j 的语料中的 W_i,W_j 文法分析确定,与其他候选词 $C_W\{W_i,W_j\}$ 的匹配状态无关,且与所有候选词的历史匹配状态无关.因此,各候选词匹配状态的转移相互独立,匹配概率可由其他候选词的匹配-转移状态的多线性算子计算.此外,转移过程具有马尔可夫性质.

(2) 下面证明以相对条件概率场为转移概率的马尔可夫过程一定收敛,且收敛到各候选词的不匹配概率.

根据马尔可夫收敛定理[25],以标准化相对条件概率为转移概率的马尔可夫过程满足收敛所需的所有条件:

① 相对不匹配概率场具有有限的状态空间,因为候选词个数有限;

② 转移概率在转移过程中始终保持不变;

③ 标准化后的相对条件概率场满足转移矩阵的行列式值连乘收敛的条件;

④ 从任意一个状态能够变到任意其他一个状态,即对$\forall W_i,W_j\in W,\exists PW_i,W_j(A_i|A_j;T,V)$;

⑤ 转移过程不是简单循环.

因此,以相对条件概率为转移概率的马尔可夫过程一定收敛.

又由于转移概率的条件与目标都是相应候选词的不匹配概率,因此所收敛的各极限值一定是相应候选词的不匹配概率. □

基于相对条件概率场的属性值-属性名匹配方法如算法 2 所述.

算法 2. 基于相对条件概率场的属性值-属性名匹配方法.

输入:待匹配属性值 V,属性名候选集 W,商品类型 T.

输出:各 $W_i\in W$ 的匹配概率 $P(W_i)$及匹配位序 Rank(W_i).

1. FOREACH $W_i\in W$

2. $P_0(A_i)=InitialDeselectProbability(V,T,W_i)$

3. $P(A_i|A_j)=RelativeConditionalDeselectProbability(V,T,W_i,W_j)$

4. $$\bar{\bar{P}}(A_i\mid A_j)=\frac{P(A_i\mid A_j)}{\sum_{k=1,k\neq j}^{n}P(A_k\mid A_j)}$$

5. WHILE ($\exists W_i\in W,P_k(A_i)-P_{k-1}(A_i)>\delta$)

6. $$P_k(A_i)=dP_0(A_i)+(1-d)\sum_{j=1,j\neq i}^{n}\bar{\bar{P}}(A_i\mid A_j)P_{k-1}(A_j)$$

7. $P(W_i)=1-P_k(A_i)$

8. $Rank(W_i)=Rank_W(P(W_i))$

9. RETURN $Rank(W_i)$

其中,d 是 Page Rank 算法中的阻尼系数,取常用的 0.15.δ是 Page Rank 算法中的收敛阈值,取常用的 0.000 1. $InitialDeselectProbability(V,T,W_i)$是初始不匹配概率的计算函数,由第 3.2 节给出.$RelativeConditionalDeselect Probability(V,T,W_i,W_j)$是相对条件概率的计算函数,由第 3.3 节给出.

3.2 初始不匹配概率的计算

初始不匹配概率 $P_0(A_i)$由 $1-P_0(W_i)$得到,其中,$P_0(W_i)$是初始匹配概率.$P_0(W_i)$通过分析 V,T 与 W_i 的文法关系,并结合 V 与 W_i 的相互关联置信度得出:将 W_i 结合 V 与 T 作为搜索内容进行搜索,取定量搜索结果 R.记 $R_V:=\{R_i\in R|V\in R_i\}$,$R_{W_i}:=\{R_i\in R|W_i\in R_i\}$,$R_{VW_i}:=\{R_i\in R|V\in R_i,W_i\in R_i\}$,$R_{V'W_i}:=\{R_i\in R|V'\in R_i,W_i\in R_i\}$.考察以下特征.

(1) V 与 W_i 的判断或偏正特征符合度 $h(W_i;F_{(1)})$.

(2) T 与 W_i 的偏正特征符合度 $h(W_i;F_{(2)})$.

(3) W_i 与 T 的偏正特征否决度 $h(W_i;F_{(3)})$.

(4) V 对 W_i 的关联置信度 $P(W_i|V)$.

(5) W_i 对 V 的关联置信度 $P(V|W_i)$.

其中,

$$P(W_i\mid V)=\frac{Card(R_{VW_i})}{Card(R_V)},$$

$$P(V\mid W_i)=\frac{\sum_{R_i\in R_{V'}}\nabla_v V'}{Card(R_{W_i})}.$$

初始匹配概率应结合以上5个特征得出.为了避免陷入局部特征最优解,应视以上5个特征的度量都不能超过一定的阈值.本文定义概率量的上限为1,且视所有特征平权,因此赋予每个特征的度量上界 M=0.2.

此外,文法特征 $F_{(1)}$,$F_{(2)}$和 $F_{(3)}$并非为 W_i 是属性名类别的充分特征,其度量上界还应受到相应文法特征的置信度或否决度的约束.因此,$P_0(W_i)$应定义为

$$P_0(W_i)=\sum_{j=1}^{3}\min\left\{h(W_i;F_{(j)}),\mu_{(j)}M\right\}+\min\left\{P(W_i\mid V),M\right\}+\min\left\{P(V\mid W_i),M\right\},$$

其中,$\mu_{(j)}$为文法特征 $F_{(j)}$的置信系数($F_{(j)}$为肯定特征时)或否决系数($F_{(j)}$为否定特征时).

得到各候选词的初始匹配概率 $P_0(W_i)$后,由 $P_0(A_i)$=1−$P_0(W_i)$得出各候选词的初始不匹配概率.

3.3 相对条件概率的计算

本节介绍相对条件概率 $P(A_j|A_i)$的计算方法.

将各候选属性名两两组合,结合属性值及商品类型进行搜索,并在搜索结果中分析两候选词之间的文法关系;再通过依次假设各候选词不是目标属性名,并根据被假设词与其他候选词的文法关系及语意逻辑来确定其他词也不是目标属性名的概率,最终得出各候选词在相对于其他候选词的关系分析中的相对条件概率.其中,所考察的文法关系仍然是偏正关系与判断关系.由于在生成候选属性名时所考察的文法关系是偏正关系和判断关系,且修饰逻辑和判断逻辑可以传递,因此各候选属性名之间也会较多地产生偏正关系与判断关系.

如果搜索结果中第 k 句 R_k 同时包含 W_i 与 W_j,则计算该句中当 W_i 不是目标属性名时,W_j 也不是目标属性名的概率 $P^k(A_j|A_i)$.在被列入属性名候选集的词都具有属性名类别特征的前提下,具有属性名类别特征的词一般只能是目标属性名、其他属性名、相应属性值或相应商品的部件等子物品类型或技术等抽象子类.基于以上原因,本文对 $P^k(A_j|A_i)$定义以下4种取值.

(1) $P^k(A_j|A_i)$=1,包括3种情况.

① R_k 中出现短语 V 或 $V'+W_j+W_i$,并且 V 与 W_i 组成偏正短语,W_j 与 W_i 组成偏正短语,则 W_i 是被 V 或 W_j 修饰或限定的中心词.因此,当 W_i 不是 V 的属性名时,W_j 作为 W_i 的修饰或限定词,也不会是 V 的属性名.

② R_k 中出现短语 V 或 $V'+W_i+W_j$.如果 V 和 W_i,W_i 和 W_j 依次构成偏正短语,则当 W_i 不是 V 的属性名时,W_i 只能是具有属性值 V 的一种物体类型.按照语意逻辑,V,W_i 和 W_j 不会组成“V 或 V'+以 V 为属性值或修饰词的物品类型+V 的属性名”的迂回句式.如果 V,W_i 为 W_j 的并列修饰语,则 W_j 作 W_i 的中心词,因此,W_j 也不会是 V 的属性名.

③ R_k 中出现短语 W_j+V 或 $V'+W_i$.由于 W_j 在 V 或 V'的前面,因此,如果 W_j 是 V 的属性名,则 W_j+V 或 V'必须是构成省略判断动词的判断句,并修饰 W_i 或更后的名词.如果 W_i 后面的名词是目标属性名,则 W_j 一定不是,否则会语义重复;如果 W_i 后没有词或后面的词不是目标属性名,则 W_j+V 或 V'修饰着一个非目标属性名的名词,因此被修饰词只能是以 V 为属性值的物品类型.然而,当用属性值判断句来修饰物品类型时,为了强调判断句作修饰语成分,在判断句与修饰对象之间很少省略修饰特征词“的”;如果省略,则惯用的句式为“V 或 V 的修饰词+V 对应的属性名+物体类型”.因此 W_j 与 V 或 V'不具有判断关系,从而在 W_i 不是目标属性名的条件下,W_j 也不可能是目标属性名.

(2) $P^k(A_j|A_i)$=0,包括两种情况.

① 当 R_k 中出现短语 V 或 $V'+W_j+W_i$ 或 T 时,如果满足:(a) T 与 W_j 组成偏正短语,而不是并列短语;(b) W_i,W_j 作中心词与 V 或 V'组成偏正短语,或是与可修饰 V 的形容词或代词直接相连时,W_i 和 W_j 具有被 V 修饰且 W_j 有从属于 T 的特征.此时,在 W_i 不是目标属性名假设下,W_i 只能为可被修饰的物品类型.按照语意逻辑,它们一般符合“V 或 V 的修饰词+V 对应的属性名+物体类型”的句式.

② R_k 中出现短语 V 或 $V'+W_i+W_j$ 时,如果 V 和 W_j,W_i 和 W_j 构成偏正短语,则 W_j 是 V 与 W_i 修饰或限定的中心词,当 W_i 不是目标属性名时,W_i+W_j 在物品描述中应表现为某一限定词+目标属性名.

(3) $P^k(A_j|A_i)$=0.5.

在 R_k 中出现短语 W_j+V 或 V'+“的”+W_i 或 T.根据 $P^k(A_j|A_i)$=1 中第③种情况的分析,当 W_i 不是目标属性名时只能是以 V 为属性值的物品类型,则判断句与修饰对象间的“的”可能是强调判断句作修饰语成分,也可能是平凡的修饰特征词:如果是前者,则 W_j 一定是目标属性名;如果是后者,则 W_j 和 $P^k(A_j|A_i)$=1 中第③种情况一样,即一定不是目

标属性名.由于存在两种无法确定的情况与结果,因此将 W_j 是目标属性名的概率赋为两种情况随机发生时的概率.

(4) $P^k(A_j|A_i)=\dfrac{n-2}{n-1}$.

如果两词出现在 R_k 中,则它们一定具有某种逻辑关系.如果该逻辑关系无法确定,则将 $P^k(A_j|A_i)$赋予假设 W_i 不是目标属性名时随机匹配的条件概率.

本文令 W_j,W_i 文法关系可确定的语料权值是文法关系不可确定的语料权值的 8 倍.对每句相关语料所得出的匹配概率求加权平均值,作为当 W_i 不是目标属性名时,W_j 也不是目标属性名的条件概率.

$$P(A_j \mid A_i)=\frac{\sum\limits_{W_i\in R_k,W_j\in R_k}\alpha_k P^k(A_j \mid A_i)}{\sum\limits_{W_i\in R_k,W_j\in R_k}\alpha_k},$$

其中,α_k 表示搜索结果中第 k 句α_k 的语料权值.

如果某对词的搜索结果中没有同时包含两个词句子,则认为两词不是目标属性名的事件相互独立.此时,直接令 $P(A_j|A_i)$为随机匹配时的条件概率,即 $P(A_j|A_i)=\dfrac{n-2}{n-1}$.

经过上述步骤后,本文得到了一个相对条件概率矩阵 A,其中各元素 $a_{ij}=P(A_i|A_j),i\neq j$.然后,基于相对条件概率场使用 Page Rank 算法迭代出最终匹配概率.

例如,针对属性值“橡胶”的候选属性名 W_1=“材料”,W_2=“种类”,W_3=“价格”,W_4=“绝缘”,W_5=“密度”,W_6=“规格”和 W_7=“科技”构建的相对条件概率矩阵见表 2.

Table 2 Relative conditional probability matrix for candidate attribute names of value “rubber”

表 2 “橡胶”的候选属性名的相对条件概率矩阵

	W_1 材料	W_2 种类	W_3 价格	W_4 绝缘	W_5 密度	W_6 规格	W_7 科技
W_1 材料	–	0.833	0.589	0.833	0.571	0.344	0.213
W_2 种类	0.833	–	0.833	0.833	0.833	0.833	0.945
W_3 价格	0.882	0.833	–	0.556	0.951	0.884	0.868
W_4 绝缘	0.833	0.833	0.889	–	0.833	0.833	0.833
W_5 密度	0.885	0.833	0.482	0.833	–	0.748	0.833
W_6 规格	0.931	0.833	0.432	0.833	0.900	–	0.850
W_7 科技	0.787	0.278	0.742	0.833	0.833	0.850	–

属性值“橡胶”各候选属性名的初始匹配概率为

$$\begin{cases}P_0(W_1)=0.481\\ P_0(W_2)=0.510\\ P_0(W_3)=0.626\\ P_0(W_4)=0.441.\\ P_0(W_5)=0.434\\ P_0(W_6)=0.415\\ P_0(W_7)=0.566\end{cases}$$

使用 Page Rank 通过相对条件概率场迭代后,属性值“橡胶”各候选属性名的最终匹配概率为

$$\begin{cases}P(W_1)=0.584\\ P(W_2)=0.456\\ P(W_3)=0.493\\ P(W_4)=0.446.\\ P(W_5)=0.485\\ P(W_6)=0.478\\ P(W_7)=0.532\end{cases}$$

可见,如果只通过初始匹配概率得到选举位序,则正确的属性名“材料”将排在第 4 位;而基于相对条件概率场使用 Page Rank 算法迭代出最终匹配概率后,正确的属性名“材料”排在了第 1 位,匹配效果显著提升.

4 实 验

本文针对服饰(衬衫与鞋)、奶粉、电子产品(手机与电脑)和球拍(乒乓球拍与羽毛球拍)这 4 类商品类型,使用百度搜索引擎搜索相应的商品描述,取得来自 TMALL 台湾、JD 等电子商务平台及百度贴吧等交流平台的商品描述语料共 15 000 余句,包含 4 类目标商品在各电商平台上所有主流的 184 个非量化属性值(其中包括 16 个主流的商品品牌).然后,同样使用百度搜索引擎搜索到的语料生成候选属性名及分析文法并进行结构化实验.实验考察以下两个方面:(1) 属性值与候选属性名自动抽取与生成效果;(2) 属性值-属性名的匹配效果.

对于属性值与候选属性名自动抽取与生成效果,本文使用属性值的查全率、查准率以及属性名的查全率这 3 个指标来考察.对于候选属性名的自动生成,本文将基于搜索引擎搜索属性值,并在包含属性值的语句及上下文中抽取一般名词作为候选属性名的生成方法,与只在描述句中抽取一般名词作为候选属性名的生成方法作对比,来验证基于搜索引擎的方法具有较高的查全率.

对于无监督生成的含有大量干扰词的属性名候选集,将本文基于文法过滤并使用相对条件概率场的无监督匹配方法与现有的结构化方法中基于依存关联[2,19]或词权重[23]的无监督匹配方法进行实验对比.其中,基于依存关联是监督或半监督商品属性结构化中常用的匹配方法,它分析属性值与候选属性名的文法依存关系,并根据关联规则置信度进行匹配.基于词权重的方法是统计学上寻找相关语料的关键词的常用方法;本文使用 NLPIR_GetKeyWords 计算属性值相关语料中各候选词的权重,并将权重作为匹配依据.为了对比在属性名候选集质量较高时各方法的效果,本文人工去掉了属性名候选集中不属于属性名类别的词,并再次进行 3 种方法的比较.此外,为了验证相对条件概率场对匹配准确度的影响,实验还对比了单纯基于属性值、商品类型与属性名文法关系的匹配方法的效果.实验使用 3 个常用的指标对以上方法进行效果评价:Rank-1 准确率,Rank 前三的准确率及平均 MRR 值.

4.1 属性值与候选属性名自动抽取效果

本文基于小概率事件原理判断文法的属性值自动抽取方法,属性值查全率为 85.7%,查准率为 81.1%.

对于候选属性名的自动生成,只在描述句中抽取一般名词作为候选属性名的方法的属性名查全率为 61.4%;基于搜索引擎搜索属性值,并在包含属性值的语句中抽取一般名词作为候选属性名的方法的属性名查全率可达 85.3%.这是由于在描述商品属性时,相应的属性名显式出现的概率理论上相当于单次伯努利试验成功的概率;而在使用搜索引擎搜索包含属性值的相关语料中,出现相应属性名的概率理论上相当于所取语料数次伯努利试验中至少成功 1 次的概率.因此,后者的概率显著高于前者.

4.2 属性值-属性名匹配方法的效果对比

表 3~表 6 分别给出了服饰(衬衫与鞋)、奶粉、电子产品(手机与电脑)和球拍(乒乓球拍与羽毛球拍)这 4 种商品的属性值与通过搜索引擎自动生成的候选属性名进行匹配的结果.带*的词指分词器未能正确识别的词,候选属性名词数∞指候选词集不包含目标属性名,这两种情况都需由人工将目标属性名加入词库或候选集后进行配对.粗体数字表示对于相应属性值的属性名匹配,该方法得出的正确属性名的位序是各匹配方法中的最优位序(注:由于基于依存关联与基于关键词权重的匹配方法都不涉及专有名词类非量化属性值[2,19],因此表 7 中以上方法的匹配效果不包括对非量化属性值中 16 个商品品牌的匹配.此外,由于商品品牌大多依靠第 3.3 节中相对文法分析的第(2)条才能有效确定,因此表 8 中单纯基于初始匹配概率进行匹配的方法在对商品品牌匹配失败时不计入有效匹配总数.).

Table 3 Value-Attribute matching result of clothes

表 3 服饰类属性值-属性名配对实验结果

属性值	属性名	#候选属性名	正确属性名 Rank		
			基于依存关联	基于关键词权重	基于相对条件概率场
女款	款式	14	13	11	**1**
纯棉	面料	18	1	3	**1**
圆领	领型*	8	4	8	**3**
长袖	袖长*	9	9	9	**5**
灯笼袖	袖型	14	6	11	**5**
青年	年龄段	16	12	10	**3**
免烫*	工艺	10	3	6	**1**
橡胶	材料	7	7	6	**1**
白色	颜色	11	2	5	**1**
麂皮	面料	11	3	1	**2**
化纤	面料	15	1	1	**1**
XXL	尺码	9	1	1	**1**

Table 4 Value-Attribute matching result of milk powder

表 4 奶粉属性值-属性名配对实验结果

属性值	属性名	#候选属性名	正确属性名 Rank		
			基于依存关联	基于关键词权重	基于相对条件概率场
婴幼儿	年龄段	41	17	23	**3**
全脂*	脂肪含量	24	3	4	**3**
3 200g	净含量	∞	1	1	**1**
胆碱	添加剂	59	46	15	**1**

Table 5 Value-Attribute matching result of electronics

表 5 电子产品类属性值-属性名配对实验结果

属性值	属性名	#候选属性名	正确属性名 Rank		
			基于依存关联	基于关键词权重	基于相对条件概率场
22nm	工艺	13	2	4	**1**
FX9590	CPU/处理器	15	1	4	**6**
Pascal	架构	8	4	6	**1**
四核	核心数	22	20	22	**1**
高通骁龙 801	CPU/处理器	9	1	1	**4**
GPS	功能	17	12	10	**10**
蓝牙*	功能	19	5	3	**3**
3G/WCDMA	网络制式*	20	17	3	**1**
4G	内存	10	1	1	**1**
16G	内存	8	1	1	**1**
4.95 英寸	屏幕尺寸	7	1	1	**1**
3 000 mAh	电池容量	6	1	1	**1**
1920×1080	分辨率	7	1	1	**1**
445PPI	像素密度*	8	5	6	**1**
7 200 转	硬盘转速	7	3	2	**1**
2.8GHz	主频	6	1	1	**1**
Sony Xperia	品牌	∞	>50	>50	**1**

Table 6 Value-Attribute matching result of rackets

表 6 球拍类属性值-属性名配对实验结果

属性值	属性名	#候选属性名	正确属性名 Rank		
			基于依存关联	基于关键词权重	基于相对条件概率场
碳素	材料	12	3	3	**1**
正胶	胶皮	18	2	3	**1**
4U	重量	12	3	2	**4**
26 磅	磅数	14	12	2	**1**

Table 7 Value-Attribute matching result comparison (non-quantization)

表 7 属性值-属性名匹配效果对比(非量化属性值)

	基于依存关联	基于关键词权重	基于相对条件概率场
Rank-1 准确率(%)	19.05	9.52	**52.17**
Rank 前三准确率(%)	52.38	47.62	**79.41**
平均 MRR	0.360	0.297	**0.679**

通过实验发现,由于量化属性值在不同的商品类型或描述中较易产生分歧,一般会在描述句中直接指明相应的属性名,因此基于依存关联和词权重的方法准确度非常高.对于非量化属性值,在监督或半监督的结构化方法中,由于所生成的属性名候选集质量很高,因此基于依存关联和词权重的方法效果较好;但是在属性名候选集的质量受无监督方法制约或是语料库质量欠佳时,基于依存关联和词权重的方法的匹配方法的效果会严重衰退.

但是,以下情况会使基于文法关系的相对条件概率场的效果受到影响,甚至低于单纯考虑属性值-属性名及商品类型-属性名间文法关系进行匹配的效果.(1) 属性值或属性名没有规范的称法.如,“全脂”和“4U”实际上是根据“脂肪含量”和“重量”所划分出的种类,而并不是“脂肪含量”和“重量”的量化属性本身;“GPS”是“全球定位系统”的英文缩写,因此“GPS 定位”“GPS 卫星定位”“GPS 通信”“GPS 系统”和“GPS 定位系统”等都是错误的短语,但是却频繁出现于搜索语料中.(2) 搜索引擎返回的语料发生断句错误,分词器未能识别到属性名短语而错将其组成词拆开或因语料句式的极端不规范而发生分词错误和词性误判.这些问题可以由人工指定易错词以排出候选集来解决.理论上,只要属性值或属性名具有规范的称法,并且分词器能够正确地分词与判断词性,就不需要对候选集进行人工的预处理.因此,本文的方法与在任何情况下都需要一定量人工标记并进行机器学习的半监督、有监督的结构化方法有着本质的不同.表 7 总结了对于非量化类属性值,这 3 种方法的效果对比.

(1) 相对条件概率场对匹配效果的影响

文法关系是本文进行属性值-属性名匹配的核心.然而,如果只是考察属性值、商品类型与属性名之间的文法关系以及属性值与属性名之间的依存关联支持度与置信度,则没有真正解决匹配效果受语言习惯、句意逻辑以及语料库质量制约的问题.这是由于判断相应属性名所使用的属性值、商品类型文法并非是充要的,甚至对属性名类别的判断都不是充要的(即置信度不为 1),并且文法特征依然受到语言习惯、句意逻辑和语料库质量的影响.表 8 给出了针对非量化属性值,基于相对条件概率场进行匹配与直接根据初始匹配概率 $P_0(W_i)$进行匹配的实验效果对比.

Table 8 Matching result with relative conditional probability field compared to using initialize probability $P_0(W_i)$ (non-quantization)

表 8 基于相对条件概率场配对与直接根据初始匹配概率 $P_0(W_i)$配对实验效果对比(非量化属性值)

	基于初始匹配概率 $P_0(W_i)$	基于相对条件概率场
Rank-1 准确率(%)	44.12	**52.17**
Rank 前三准确率(%)	76.47	**79.41**
平均 MRR	0.610	**0.679**

实验结果表明,基于相对条件概率场的匹配方法能够改善单纯基于属性值、商品类型与属性名文法关系的方法的准确率.

(2) 各方法对高质量属性名候选集的匹配效果对比

为了对比在属性名候选集质量较高时各方法的效果,本文人工去掉了属性名候选集中不属于属性名类别的词,并再次进行 3 种方法的比较,结果见表 9.

Table 9 Value-Attribute matching result comparison for high quality attribute candidates (non-quantization)

表 9 使用高质量属性名候选集的属性值-属性名匹配效果对比(非量化属性值)

	基于依存关联	基于关键词权重	基于相对条件概率场
Rank-1 准确率(%)	56.52	56.52	**82.61**
Rank 前三准确率(%)	91.30	86.96	**95.65**
平均 MRR	0.716	0.714	**0.902**

4.3 实验结论

实验结果表明,对于非量化属性值,当属性名候选词集质量较差时,使用本文提出的基于相对条件概率场的属性值-属性名匹配方法,与基于依存关联的方法相比,Rank-1 的准确率提高 30%以上,平均 MRR 提高 0.3 以上.当属性名候选词集质量较好时,由于依然存在语言习惯、句意逻辑及语料库质量等其他因素制约着属性值-属性名的依存关联,因此基于相对条件概率场的属性值-属性名匹配方法仍然较优,Rank-1 的准确率提高了 20%以上.

致谢 在此,我们向对本文的工作给予支持和建议的同行和马雪超硕士表示感谢.

References:

[1] Huang JM, Wang HX, Jia Y, Fuxman A. Link-Based hidden attribute discovery for objects on Web. In: Proc. of the 14th Int'l Conf. on Extending Database Technology. ACM, 2011. 473−484. [doi: 10.1145/1951365.1951421]

[2] Ghani R, Probst K, Liu Y, Krema M, Fano A. Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter, 2006,8(1):41−48. [doi: 10.1145/1147234.1147241]

[3] Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Proc. of the 3rd IEEE Int'l Conf. on Data Mining (ICDM 2003). IEEE, 2003. 427−434. [doi: 10.1109/ICDM.2003.1250949]

[4] Tokunaga K, Kazama J, Torisawa K. Automatic discovery of attribute words from Web documents. In: Proc. of the Natural Language Processing (IJCNLP 2005). Berlin, Heidelberg: Springer-Verlag, 2005. 106−118. [doi: 10.1007/11562214_10]

[5] Hu MQ, Liu B. Mining opinion features in customer reviews. In: Proc. of the 19th National Conf. on Artificial Intelligence (AAAI 2004). AAAI, 2004. 755−760.

[6] Popescu AM, Nguyen B, Etzioni O. OPINE: Extracting product features and opinions from reviews. In: Proc. of the HLT/EMNLP on Interactive Demonstrations. Association for Computational Linguistics, 2005. 32−33.

[7] Zheng Y, Ye L, Wu GF, Li X. Extracting product features from Chinese customer reviews. In: Proc. of the 3rd Int'l Conf. on Intelligent System and Knowledge Engineering (ISKE 2008). IEEE, 2008. 285−290. [doi: 10.1109/ISKE.2008.4730942]

[8] Liu T, Liu BQ, Xu ZM, Wang XL. Automatic domain-specific term extraction and its application in text classification. Acta Electronica Sinica, 2007,35(2):328−332.

[9] Ren X, El-Kishky A, Wang C, Tao FB, Voss CR, Ji H, Han JW. Clus type: Effective entity recognition and typing by relation phrase-based clustering. In: Proc. of the KDD. 2015.

[10] Huang HZ, Cao YB, Huang XJ, Ji H, Lin CY. Collective tweet wikification based on semi-supervised graph regularization. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). ACL, 2014. 380−390. [doi: 10.3115/v1/P14-1036]

[11] Lin T, Mausam, Etzioni O. No noun phrase left behind: Detecting and typing unlinkable entities. In: Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 893−903.

[12] Nakashole N, Tylenda T, Weikum G. Fine-Grained semantic typing of emerging entities. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). ACL, 2013.

[13] Huang RH, Riloff E. Inducing domain-specific semantic class taggers from almost nothing. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. ACL, 2010. 275−285.

[14] Han JW, Wang C, El-Kishky A. Bringing structure to text: Mining phrases, entity, topics, and hierarchies. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. [doi: 10.1145/2623330.2630804]

[15] Han JW, Wang C. Mining latent entity structures from massive unstructured and interconnected data. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. 2014. 1409−1410. [doi: 10.1145/2588555.2588890]

[16] Guarino N. Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge bases. Data & Knowledge Engineering, 1992,8(3):249−261. [doi: 10.1016/0169-023X(92)90025-7]

[17] Su Q, Xu XY, Guo HL, Guo ZL, Wu X, Zhang XX, Swen B, Su Z. Hidden sentiment association in Chinese Web opinion mining. In: Proc. of the 17th Int'l Conf. on World Wide Web. ACM, 2008. 959−968. [doi: 10.1145/1367497.1367627]

[18] Qiu G, Zheng M, Zhang H, Zhu JK, Bu JJ, Chen C, Hang H. Implicit product feature extraction through regularized topic modeling. Journal of Zhejiang University (Engineering Science), 2011,45(2):288−294 (in Chinese with English abstract).

[19] Hao BY, Xia YQ, Zheng F. OPINAX: An effective product attribute mining system. In: Proc. of the 4th National Conf. on Information Retrieval and Content Security (NCIRCS 2008), Vol.1. NCIRCS, 2008. 281−290 (in Chinese with English abstract).

[20] Gupta N, Kumar P, Gupta R. CS 224N final project: Automated extraction of product attributes from reviews. 2009. http://www-nlp.stanford.edu

[21] Wong TL, LamW, Wong TS. An unsupervised framework for extracting and normalizing product attributes from multiple Web sites. In: Proc. of the Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. 2008. 35−42. [doi: 10.1145/1390334.1390343]

[22] Yi J, Niblack W. Sentiment mining in WebFountain. In: Proc. of the 21st Int'l Conf. on Data Engineering (ICDE 2005). IEEE, 2005. 1073−1083. [doi: 10.1109/ICDE.2005.132]

[23] Zhang HP. NLPIR/ICTCLAS Chinese lexical analysis system. 2002 (in Chinese). http://ictclas.nlpir.org/docs

[24] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1-7):107−117. [doi: 10.1016/S0169-7552(98)00110-X]

[25] Bharucha-Reid AT. Elements of the Theory of Markov Processes and Their Applications. New York: McGraw-Hill, 1960.

附中文参考文献:

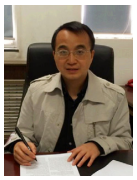
[18] 仇光,郑淼,张晖,朱建科,卜佳俊,陈纯,杭航.基于正则化主题建模的隐式产品属性抽取.浙江大学学报(工学版),2011,45(2):288−294.

[19] 郝博一,夏云庆,郑方.OPINAX:一个有效的产品属性挖掘系统.见:第4届全国信息检索与内容安全学术会议论文集(上卷).2008. 281−290.

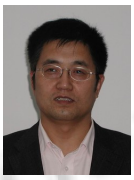
[23] 张华平.NLPIR/ICTCLAS 汉语分词系统.2002. http://ictclas.nlpir.org/docs

侯博议(1990－),男,陕西西安人,博士生,主要研究领域为数据质量,流形.

杨婧颖(1990－),女,硕士,主要研究领域为数据质量.



陈群(1976－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为大数据管理,物联网信息管理.



李战怀(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.