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:

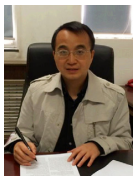
[18] 仇光,郑淼,张晖,朱建科,卜佳俊,陈纯,杭航.基于正则化主题建模的隐式产品属性抽取.浙江大学学报(工学版),2011,45(2):288−294.

[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.

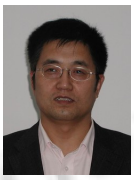
[23] 张华平.NLPIR/ICTCLAS 汉语分词系统.2002. http://ictclas.nlpir.org/docs

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.