第 43 卷 第 12 期 计 算 机 工 程 2017 年 12 月
Vol. 43 No. 12 Computer Engineering December 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)12-0160-05 文献标志码：A 中图分类号：TP315

基于多特征的微博情感分析研究

刘续乐[a]，何炎祥[a,b]

（武汉大学 a. 计算机学院；b. 软件工程国家重点实验室，武汉 430072）

摘 要：为提高微博情感分类识别的正确率，以网络微博数据作为研究对象，提出一种基于图的情感基准词选择方法。结合知网相似度知识，构建图模型，以图中节点中介性的值为依据，选择出高质量和高覆盖率的情感基准词。根据得到的基准词构建情感分析中所需的情感词典，并给出情感词极性。同时将情感词应用于挖掘短句情感特征，加入到传统支持向量机（SVM）模型中，对微博句子挖掘更多的语义信息从而获取更合理的语义合成函数，捕捉句子情感变化以更好地把握微博整句情感。采用具有特征约束特性的条件随机场（CRF）模型对短句进行分类。实验结果验证了 CRF 模型短句分类的有效性，与多种特征的 SVM 分类方法相比，在不同数据集上具有更好的分类效果。

关键词：微博；情感词；节点中介性；情感分析；机器学习

中文引用格式：刘续乐，何炎祥. 基于多特征的微博情感分析研究[J]. 计算机工程，2017，43(12)：160-164，172.

英文引用格式：LIU Xule，HE Yanxiang. Research on Microblog Sentiment Analysis Based on Multi-feature[J]. Computer Engineering，2017，43(12)：160-164，172.

Research on Microblog Sentiment Analysis Based on Multi-feature

LIU Xule[a]，HE Yanxiang[a,b]

（a. School of Computer；b. State Key Laboratory of Software Engineering，Wuhan University，Wuhan 430072，China）

【Abstract】In order to improve the accuracy of micro-blog emotional classification recognition，regarding the network microblog data as research object，this paper proposes a choice method of emotional basic word based on graph. Combined with similarity knowledge of HowNet，the method builds a graph model to choose high quality and high coverage emotion basic words according to node betweenness centrality in graph. It builds emotional dictionary for sentiment analysis according to selected basic words. The polarity of emotional words is also given. The emotional words are applied to mine short sentence emotional features. Those features will join into traditional Support Vector Machine（SVM） model. More semantic information is mined on micro-blog sentences to obtain a more reasonable semantic composition function. The sentence emotional changes are captured to better grasp the micro-blog emotion of whole sentence. Conditional Random Field（CRF） model that has characteristics of feature constraint is used to classify short sentences. Experimental results verify the effectiveness of CRF model on short sentences. Compared with SVM classification methods with different features，it also has a greater effect on different data sets.

【Key words】microblog；emotional word；node betweenness centrality；sentiment analysis；machine learning

DOI：10.3969/j.issn.1000-3428.2017.12.030

0 概述

在目前主流的网络社交平台中，微博是互联网用户应用最广泛的平台之一，网络平台上的这些海量言论信息可以划分两大类：一类是描述事实的文本，另一类是表达意见的文本。其中表达意见的文本非常重要，比如微博每天产生大量的网络用户，对商品、舆情等的评论信息具有很大的实际价值[1-3]，为商家提供用户反馈进而吸取意见，改进产品以满足更多用户的需求；也可以用来预测社会对政府提出政策的支持与否或者对热点时事的褒贬观点等[4-5]。

情感分析[6-7]以及意见挖掘[8]研究是对大数据的一种分析，旨在挖掘数据中的情感信息。情感分析指判别文本的主客观性，同时对于主观性文本判别其情感倾向，是积极的（褒义）表达还是消极的（贬义）表达。相比传统文本分析，微博情感分析有许多差异：首先微博属于短文本，通常有长度限制；

作者简介：刘续乐（1992—），男，硕士研究生，主研方向为自然语言处理、情感分析；何炎祥，教授、博士生导师。

收稿日期：2016-11-22 修回日期：2017-01-11 E-mail：xuleliu@whu.edu.cn

其次，微博文本中有大量的带有情感的词语，这些词语存在变化多端、文本非结构化，以及用语随意不规范等特点。

针对文本的情感分析存在两个难点，一是情感词典不完备，二是如果直接采用机器学习方法，将传统特征应用到微博中，特征空间太大，但是微博文本短，使得文本在向量空间中表示稀疏，高度不均匀[9]，而且无法捕捉句中情感的变化，这些因素直接影响微博情感分类识别的正确率。本文在传统文本分类技术的基础上，提出针对微博所面临难点的解决思路。

1　相关工作

1.1　情感分析

在情感分析领域，基准词的选择标准之一就是要具有明显的情感倾向。对于给定的一个词，可通过计算一个它和基准词的关系来推断出该词的情感倾向性从而建立情感词典，以此为基础可对句子、段落、篇章等语义单元中的情感极性进行累加，从而判断其情感倾向性。

目前词语的情感倾向判断有许多研究，文献[10]选择那些同时出现并且与文章的内容语义无关的词为基准词来判定词汇的情感倾向，实验结果表明，基准词的挑选对词汇情感倾向性的判别有重大影响。文献[11]同样沿用文献[10]的方法，选择情感明确并且常见的情感词作为基准词。已有的研究基准词多数来自研究者通过经验进行选择，或简单地根据词性、词频等信息[11-12]进行判断，存在着随机性和主观性的缺陷，且难以保证在词典中对语义关系的全面覆盖。本文利用已有的知网研究，提出一种基于情感关系图[13]获取情感基准词的方法。通过采用相同的情感相似度计算方式与已有的方法作对比。

1.2　短文本情感分类

短文本分类是文本分类的一个分支，除了有相同特点之外还面临许多待解决的问题，因为文本的长度短，情感表达变化多端，单纯地把普通文本分类任务算法移植过来并不能取得很好的分类准确率，在近几年的国内评测任务中分析各大机构所用的方法就可以看到效果并不甚理想，其中有使用很多方法来克服短文本带来的特征稀疏问题，比如引入外部资源（借用搜索引擎返回结果）增加文本之间的共享特征，虽然可以在一定程度上解决问题，提高分类效率，但是时间代价太高。对于中文这种情感表达变化快、表达随意的特点，本文利用短文本的特点（短小、情感一致）结合条件随机场（Conditional Random Field，CRF）模型，提取出句子中短文本情感特征，加入到分类器中，试图挖掘句中语义变化，把握整体句子的情感，提高句子分类的正确率。

2　基于图的情感基准词选择

2.1　知网情感词关系图构建

知网（HowNet）[13]是一个从更细致的角度描述词语及词语之间关系的语义关系库，包括概念和义原两部分。对 2 个汉语词语 W 和 W_2，如果 w_1 有 n 个概念 $S_1,S_2,\cdots,S_n$，W_2 有 m 个概念 $S_1,S_2,\cdots,S_m$：，则 W_1 和 W_2 的相似度为各个概念的相似度的最大值为：

$$Sim(W_1,W_2)=\max_{i=1,2,\cdots;n,j=1,2,\cdots,m}(Sim(S_{1i},S_{2j})) \tag{1}$$

而对于概念 S_1、S_2，它们之间的相似度可表示为：

$$Sim(S_1,S_2)=\sum_{1}^{4}\beta_i Sim_i(S_{1i},S_{2j}) \tag{2}$$

其中，$\beta_i(1\leqslant i\leqslant 4)$是可调节的参数，且有

$$\beta_1+\beta_2+\beta_3+\beta_4=1,\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$$

$Sim_1(S_1,S_2)$，$Sim_2(S_1,S_2)$，$Sim_3(S_1,S_2)$，$Sim_4(S_1,S_2)$分别从义原的不同组成来刻画概念之间的相似度。

1）顶点选择

知网情感词关系图中的顶点相当于社交网络中的个体，它们通过边连接起来，边承载着信息流转中介的作用。本文构建的候选褒贬基准词集由属性标注了“良”或“莠”的词语组成。相对而言，这些词有比较明确的褒贬含义，是实现最终方法效果判定的重要基础。候选褒贬基准词集中词汇的情感极性值分别为 1 和 -1，分别用它们作为正负知网情感词关系图的顶点。

2）边的权重设定

知网词汇相似度 $Sim(W_1,W_2)$ 刻画了两个词之间的密切程度。相似度值越大则说明两词间在知网关系图中距离越近，有更大的可能性表达相似的语义情感。

基于此，本文以 $W_{ij}=1-Sim(W_i,W_j)$ 作为图 $G=(V,E)$ 中 2 个节点 W_i、W_j 之间边的权重。两节点联系越紧密，则 $Sim(w_i,w_j)$ 值越大，对应到图中则表示两点间距离 W_{ij} 越小。在求最短路径时，能够尽可能地保证该路径上包含更多的语义相似度更大的节点。

2.2　情感基准词的选择

2.2.1　节点中介性

节点的中介性（Betweeness Centrality，BC）是反映中心性的一个重要指标，由通过该节点的最短路径的数量来刻画。节点的 BC 值[14]反映了该节点在网络中的重要性，值越大说明该节点作为中介在网络连接和消息的传递中起着更突出的作用。

在图 $G=(V,E)$ 中，节点 $v\in V$ 的中介性计算如下：

1）对任一顶点对(s,v)，$s\in V$，$v\in V$，计算两点之间的最短路径；

2)记录以顶点对(s,v)为首尾的最短路径中,包含的其他顶点。

3)重复步骤 1)和步骤 2),记录下任意 2 个顶点之间的最短路径和节点。

定量描述中介性中心性值为:

$$BC(v)=\sum_{s\neq v\neq t\in V}\delta st(v) \tag{3}$$

$$\delta st(v)=\frac{\sigma st(v)}{\sigma st} \tag{4}$$

其中,V 是所有顶点的集合,σst 是顶点 s 和 t 之间最短路径的条数,$\sigma st(v)$是顶点 s 和 t 之间经过点 v 的最短路径条数。

2.2.2 基准词选择依据

以 BC 值的大小作为选择依据,通过计算图中候选基准词的 BC 值,选择值的大小排名靠前的候选情感词作为情感基准词。本文以此构建情感分析的情感词典,通过顶点的选择和边的权重设定来构建知网情感词关系图。

3 微博文本情感分类

3.1 微博文本识别难点

由于微博不同于传统文本,其具有以下特点:1)字符长度有限制,例如新浪微博长度限定在 140 个字符以内;2)内容随意,无结构化,指内容或者情感的表达随意,变化多,没有固定的句式结构。

因此对于微博这种无结构化的语言,难以正确地分析句子的句式和挖掘特征来表达句子的情感变化,从而导致识别微博情绪的正确率较低。

在各种语法实体中,短句具备的语法因素最为齐全,短句既包含词和短语,又具有一定的情感极性,是具有独立性和表述性的最小语法单位。在做情感分析时,可以认为短句表达的意思以及情感具有高度一致性。

本文结合微博的特点,根据中英文标点符号将微博短文本划分成短句,根据短句的情感具有一致性的特点,首先采用 CRF 模型在短句情感分类上进行建模,随后提取出长句中每一个短句的情感特征,加入到特征集合中,再采用支持向量机(Support Vector Machine,SVM)模型对短句之间的情感进行建模,从而挖掘短句与短句之间的语义情感关联关系来达到较高的情感分析准确率。

3.2 基于 CRF 模型的分类器

由于短句的情感具有一致性特点,即短句的某一个特征倾向于某一个情感,其上下文特征也倾向于该情感,这与 CRF 模型存在特征之间的约束特性高度一致,因此本文采用 CRF 模型构造短句情感分类模型。针对短句的情感分类,例如短句“珍爱生命远离烟草”为消极情感,本文用 NEG 和 POS 分别代表积极、消极的标签,其标注结果如下。

原句:珍爱生命远离烟草

CRF 标注后:珍爱/NEG 生命/NEG 远离/NEG 烟草/NEG

识别结果:短句为消极句

CRF 模型通过训练语料的特征学习,以及隐藏变量之间的约束,对单个短句最后的标注结果不会同时出现积极和消极标签,同时 CRF 模型一定程度上解决了文本特征稀疏问题,从而可提高短句的分类准确率。

在短句 CRF 模型中,采用表 1 的特征作为 CRF 特征,其中情感词是情感分析的基础,否定词能反转情感词的情感极性,程度副词可以增强或减弱所修饰情感词的强度。

表 1 CRF 模型特征

特征	特征层级	说明
词语/汉字	短句级	词或者标点
词性	短句级	词语词性
情感词	短句级	情感词
否定词	短句级	否定词
程度副词	短句级	程度副词
转折词	短句级	转折词

CRF 序列标注模型通过特征的选择考虑上下文的语境,同时标注受到上下文环境的约束,能够很好地用来判别短句的情感倾向,后续的实验将证明方法的可行性。

3.3 基于 SVM 的分类模型

本文首先通过构建 CRF 短句模型来识别短句情感,挖掘短句情感特征,然后采用 SVM 模型识别微博整个句子的情感倾向。为了验证短句特征的有效性,选择该领域研究中广泛使用的特征作为实验对比,包括 Unigram、Bigram 联合特征;微博情感特征,包括句子积极消极情感词数,感叹号是否出现,问号是否出现;还有采用基于情感基准词扩充和收集的情感词作为特征;短句特征,除了将构成微博短句的积极、消极情感句个数作为特征外,考虑中文首尾句子极性对整句情感的影响比较大,将句子首尾句的情感极性也作为分类特征,具体的特征选择如表 2 所示。

表 2 SVM 特征选择

编号	特征	特征层级	说明
特征 1	Unigram + Bigram	微博级	特征权重为 tf-idf
特征 2	情感词典	微博级	出现 0;不出现 1
特征 3	积极、消极情感词个数,问号、叹号、转折词是否出现(5 个特征)	微博级	是否出现用 0/1 表示
特征 4	首尾短句情感,积极、消极情感句个数(4 个特征)	微博级	极性用 1/-1 表示

4　实验设置与结果分析

4.1　情感基准词实验

4.1.1　实验设置

本文以知网中包含“良”“莠”的词汇为候选词，利用所提的方法选取 BC 值 top N 对情感基准词。为了使实验效果更具说明性，将其中同时包含“良”“莠”属性的词汇去除，所得词语的褒贬含义更明确，将这类词作为测试集。测试集中褒义词为 3 132 个，贬义词为 3 260 个。

理论上可知，所选情感基准词的质量和数量在一定程度上都会对情感词极性的判定带来影响。为此，在实验 1 中设置不同的情感基准词数量，考察相应情况下对上述测试集中情感词极性判断准确率的变化。

应用聚类的思想进行情感基准词的选择也是较为主流的实验方法[13,15]。在实验 2 中，选择 k-means 聚类算法完成基准词的抽取，聚类中心点数目对应基准词对的数目，并选择那些概念描述中仅有“良”或“莠”的词作为候选情感基准词。设置不同的基准词对数量，应用到测试集上考察相应情况时的判断结果准确率。

4.1.2　实验结果

在本文实验中，以 0 为域值，计算结果大于 0 的认定为褒义词，小于 0 的认定为贬义词。以准确率作为评价标准，$P_{正}$、$P_{负}$、$P_{总}$分别表示对褒义词、贬义词和所有词的准确率。

4.1.3　实验分析

实验 1 分别设置基准词对数为 20、40、60 和 80，以此判断基准词数量对情感极性判断的影响。从表 3可以看出，情感基准词数量的增加会使得情感词极性判断的准确率随之提升。文中方法选择一定规模的基准词即可达到较好的语义覆盖率，对情感词极性判断取得较好的效果，这在一定程度上减轻了大数据环境下的计算量，有一定的使用价值。

表 3　实验 1 评价结果　%

基准词对数量	$P_{正}$	$P_{负}$	$P_{总}$
20	90.68	54.83	76.74
40	93.62	55.62	77.58
60	93.71	55.95	77.79
80	93.27	55.82	77.51

实验 2 选择 k-means 算法对候选基准词进行聚类，选择最终聚类所得各类的中心作为基准词，同样考察基准词数量不同时的结果准确率。由表 4 可以看出，在基于聚类的方法中，基准词对整个情感语义的覆盖率会随其数量的增加而显著提高，并且个别噪音数据造成的负面影响也会得到减弱。

表 4　实验 2 评价结果　%

基准词对数量	$P_{正}$	$P_{负}$	$P_{总}$
20	64.94	52.21	58.60
40	77.93	60.83	69.21
60	80.56	64.69	72.47
80	83.56	69.33	76.29

从表 3、表 4 中可以看出，本文方法无论在整体的准确率还是在对褒贬义词判断的均衡性方面均具有一定的优越性，而且文中方法对基准词数量的敏感性较小。

4.2　微博情感分析

4.2.1　实验数据

实验所有的语料来自近年整理的计算机学会中文情感倾向评测语料，实验中首先需要对语料进行短句划分，具体的划分步骤如下：

1）将微博中空格全部替换为中文的逗号；

2）使用中英文的标点集合“，；。！？ …|，；！？”进行断句；

3）收集所有片段，作为微博的短句集合返回。

实验对微博句子进行情感划分（积极、消极情绪识别），利用短句语料来验证短句 CRF 模型和短句 SVM 模型的效果对比，长句语料用于测试不同方法下微博情感分析的准确率。实验采用交叉验证的方法来验证模型的好坏。模型具体的训练数据和测试数据划分统计为：对短句训练测试数据，正向891 个，负向 1 224 个，总计 2 115 个；对长句训练测试数据，正向 3 717 个，负向 3 680 个，总计 7 397 个；长短句总计9 512 个。

4.2.2　实验结果与分析

实验首先验证 CRF 模型对于短句分类的效果，证明 CRF 短文本分类的有效性。CRF 模型采用表 1 所示特征，SVM 分类器采用 Unigram 和 Bigram 联合特征。本文实验中采用 3 折交叉验证的方法分析模型的效果。

从表 5 的实验结果可以看出，在短句情感分类预测中，CRF 模型相比 SVM 分类模型提高了3.1%，验证了本文提出的短文本特征的一致性以及 CRF 模型的特征约束特性适合短文本分类的思想。

表 5　短句情感分类实验结果

方法	类别	P	R	$F1$	准确率
SVM	正向	0.790 0	0.720 0	0.753 4	0.752 6
	负向	0.840 0	0.680 0	0.751 8	
CRF	正向	0.792 6	0.729 7	0.760 8	0.783 5
	负向	0.790 3	0.822 9	0.806 3	

其次,针对长句数据集采用不同的特征组合用SVM 模型进行分类预测,同时采用 CRF 模型和传统的基于情感词典方法进行对比。传统基于情感词典的分类方法统计句中包含的情感词,然后累加所有情感词的情感总值,如果总情感值大于阈值(本文为0)则为积极句,小于阈值则为消极情感句。

实验对比长句语料下 CRF 模型和不同特征的SVM 分类模型以及传统基于情感词典方法的效果,同时验证短句特征的有效性,对所有的数据采用5 折交叉验证,实验结果如表 6 所示。

表 6 长句情感分类实验结果

方法	类别	*P*	*R*	*F*1	准确率
CRF	正向	0.876 7	0.879 6	0.878 2	0.877 7
	负向	0.878 9	0.875 4	0.877 2	
情感词典	正向	0.791 2	0.699 4	0.742 5	0.738 9
	负向	0.771 5	0.700 4	0.734 3	
特征 1(SVM)	正向	0.884 0	0.898 0	0.890 9	0.890 9
	负向	0.900 0	0.882 0	0.890 9	
特征 2(SVM)	正向	0.852 0	0.768 0	0.807 8	0.815 9
	负向	0.786 0	0.866 6	0.824 0	
特征 3(SVM)	正向	0.748 0	0.796 0	0.771 2	0.761 6
	负向	0.780 0	0.726 0	0.752 0	
特征 4(SVM)	正向	0.830 0	0.810 0	0.819 9	0.820 8
	负向	0.812 0	0.832 0	0.821 8	
特征 4 + 特征 2(SVM)	正向	0.856 0	0.854 0	0.854 9	0.853 9
	负向	0.854 0	0.852 0	0.852 9	
特征 4 + 特征 3(SVM)	正向	0.900 0	0.886 0	0.892 9	0.894 4
	负向	0.890 0	0.902 0	0.895 9	

在实验中,效果最好的为特征 4 和特征 3 组合特征的 SVM 分类模型,准确率分别为 0.894 4。在本文实验中,为了验证了短句特征的有效性,可以看出采用情感词典作为特征的 SVM 模型在加入了短句特征之后分类效果有所提升。同时采用情感词出现个数、否定词出现与否等特征(特征 3)的 SVM 模型加入短句特征之后准确率也有一定的提升,因为短句特征挖掘了长句中情感的变化,一般来说很短的句子表达的情感具有一致的特性,随着句子变长,情感的表达变得不一致,如何抓住句子的核心表达情感是传统的特征无法捕捉的特性。本文加入了长句中短句的情感,供分类器训练,试图捕捉句子情感变化而更好地把握整句的情感,实验充分证明了短句特征的有效性。从实验数据还可以看到,Unigram 和 Bigram 特征效果相对其他特征更好,特征 3 的效果整体是最差的,主要原因在于特征 3 包含的特征数目太少,无法有效地区分情感表达。再者单纯基于情感词典特征(特征 2)也无法取得好的效果,但是在加入短句特征之后分类效果仍不错。

5 结束语

本文提出一种基于图的情感基准词选择方法,并以此构建在后文中所需的情感词典。通过顶点的选择和边的权重设定来构建知网情感词关系图,其中的主要概念有知网相似度和节点的中介性。实验结果显示,本文方法较传统方法所选的情感基准词有更好的语义覆盖率和更好的质量,为后续跨语言情感分析提供保障。

将文本划分为更小的多个短句,采用 CRF 模型预测短句分类,避免了特征选择问题。实验结果证明 CRF 模型在短句分类上比采用 Unigram 和 Bigram 联合特征的 SVM 模型分类效果好。长句分类实验结果表明,SVM 分类模型效果要好于 CRF 模型分类效果,这 2 种模型的分类效果远好于基于统计的情感词典方法。

参考文献

[1] FELDMAN R, GOLDENBERG J, NETZER O. Mine Your Own Business: Market Structure Surveillance Through Text Mining[J]. Marketing Science, 2012, 31(3):521-543.

[2] 韦 航,王永恒. 基于主题的中文微博情感分析[J]. 计算机工程,2015,41(9):238-244.

[3] 陈铁明,缪茹一,王小号. 融合显性和隐性特征的中文微博情感分析[J]. 中文信息学报,2016,30(4):184-192.

[4] 田 野. 基于微博平台的事件趋势分析及预测研究[D]. 武汉:武汉大学,2012.

[5] TONG Wei, CHEN Wei, MENG Xiaofeng. EDM: An Efficient Algorithm for Event Detection in Microblogs[J]. Journal of Frontiers of Computer Science & Technology, 2012,6(12):1076-1086.

[6] 陈 强,何炎祥,刘续乐,等. 基于句法分析的跨语言情感分析[J]. 北京大学学报(自然科学版),2014,50(1):55-60.

[7] 黄 挺,姬东鸿. 基于图模型和多分类器的微博情感倾向性分析[J]. 计算机工程,2015,41(4):171-175.

[8] MENG Xinfan, WEI Furu, LIU Xiaohua, et al. Entity-centric Topic-oriented Opinion Summarization in Twitter[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2012:379-387.

[9] YAN Rui, CAO Xianbin, LI Kai. Dynamic Assembly Classification Algorithm for Short Text[J]. ACTA Electronica Sinica, 2009, 37(5):1019-1024.

[10] TURNEY P D, LITTMAN M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4):315-346.

[11] ZHU Yanlan, MIN Jin, ZHOU Yaqian, et al. Semantic Orientation Computing Based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1):14-20.

(下转第 172 页)

Wiki Table 数据集表头的质量比 Web Table 数据集的质量高。

2)AFD_Model 算法和 PFD_Mode 算法在大表上的表现均好过小表,表明本文提出的近似函数依赖和文献[10]提出的概率函数依赖均可能对大数据量的表格更适用。

3)相比传统的实体列检测方法,APD_Model 算法能实现无表头 Web 表格的多实体列检测,且有不错的表现。

4 结束语

本文提出适用于 Web 表格的近似函数依赖及其评分方法,并将其应用于 Web 表格的实体列发现。与现有的实体列检测方法相比,提出的基于近似函数依赖和规范化的实体列发现算法,可以在更多场景下发现实体列。该方法不仅适用于单实体列的 Web 表格,还可用于多实体列的表格;不仅适用于有表头的 Web 表格,而且适用于没有表头或者利用语义恢复技术也无法恢复出完整表头的 Web 表格。下阶段将研究如何进一步提高函数依赖检测的准确度。

参考文献

[1] CAFARELLA M J, HALEVY A, WANG Z D, et al. WebTables: Exploring the Power of Tables on the Web[C]//Proceedings of the 34th International Conference on Very Large Data Bases. New York, USA: ACM Press, 2008: 538-549.

[2] VENETIS P, HALEVY A, MADHAVAN J, et al. Recovering Semantics of Tables on the Web[C]//Proceedings of the 37th International Conference on Very Large Data Bases. New York, USA: ACM Press, 2011: 1601-1612.

[3] BALAKRISHNAN S, HALEVY A, HARB B, et al. Applying WebTable in Practice[EB/OL]. (2015-12-06). http://webdatacommons.org/webtables/.

[4] LIMAYE G, SUNITA S, CHAKRABARTI S. Annotating and Searching Web Tables Using Entities Types and Relationships[J]. Proceedings of the VLDB Endowment, 2010, 3(3): 1338-1347.

[5] DENG Dong, JIANG Yu, LI Guoliang, et al. Scalable Column Concept Determination for Web Tables Using Large Knowledge Bases[J]. Proceedings of the VLDB Endowment, 2013, 6(13): 1606-1617.

[6] WANG Jingjing, WANG Haixun, WANG Zhongyuan, et al. Understanding Tables on the Web[C]//Proceedings of the 31st International Conference on Conceptual Modeling. New York, USA: ACM Press, 2012: 141-155.

[7] LEE T, WANG Zhongyuan, WANG Haixun, et al. Attribute Extraction and Scoring: A Probabilistic Approach[C]//Proceedings of the 29th International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2013: 194-205.

[8] 任向冉. 网络表格的实体列发现与标识[D]. 北京: 北京交通大学, 2015.

[9] 黎章海, 潘久辉. 基于函数依赖的导出关系候选码计算[J]. 计算机工程, 2016, 42(5): 60-65.

[10] WANG D Z, DONG Luna, SARMA A D, et al. Functional Dependency Generation and Applications in Pay-as-You-Go Data Integration Systems[C]//Proceedings of the 12th International Workshop on the Web and Databases. Berlin, Germany: Springer, 2009: 1654-1655.

[11] WANG J N, LI G L, FENG J H. Fast-Join: An Efficient Method for Fuzzy Token Matching Based String Similarity Join[C]//Proceedings of the 27th International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2011: 458-469.

[12] 萨师煊, 王 珊. 数据库系统概论[M]. 北京: 高等教育出版社, 2002.

[13] LAUTERT L R, SCHEIDT M M, DORNELES C F. Web Table Taxonomy and Formalization[J]. ACM SIGMOD Record, 2013, 42(3): 28-33.

[14] VERGA P, NEELAKANTAN A, MCCALLUM A. Generalizing to Unseen Entities and Entity Pairs with Row-less Universal Schema[EB/OL]. (2016-06-18). https://arxiv.org/abs/1606.05804.

[15] XIA Zhihua, WANG Xinhui, SUN Xingming, et al. A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data[J]. IEEE Transactions on Parallel & Distributed Systems, 2016, 27(2): 340-352.

编辑 顾逸斐

(上接第 164 页)

[12] TURNEY P D. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York, USA: ACM Press, 2002: 417-424.

[13] CHEN Yuefeng, MIAO Duoqian, LI Wen, et al. Semantic Orientation Computing Based on Concepts[J]. CAAI Transactions on Intelligent Systems, 2012, 6(6): 489-494.

[14] BRANDES U. On Variants of Shortest-path Betweenness Centrality and Their Generic Computation[J]. Social Networks, 2008, 30(2): 136-145.

[15] WANG Suge, LI Deyu, WEI Yingjie, et al. A Synonyms Based Word Sentiment Orientation Discriminating[J]. Journal of Chinese Information Processing, 2009, 23(5): 68-74.

编辑 顾逸斐