基于词项共现关系图模型的中文观点句识别研究*

王明文，付翠琴，徐凡，洪欢

（江西师范大学，计算机信息工程学院，南昌 330022）

摘要：不同于传统的词项间强独立性假设的词袋模型驱动的观点句识别方法，本文提出了一种新型的基于词项共现关系的图模型方法。该方法通过构建词项共现关系图模型，利用词项与词项之间的共现性和句法关系来描述词项在观点句和非观点句集合中的分布差异，同时采用基于入度的词项权重计算方法来计算词项特征值。上述研究在基准语料上进行实验，实验表明采用基于词项关系图模型方法后，中文观点句识别准确率相比目前基于词袋的方法得到显著提升。

关键词：词项共现；图模型；观点句识别；特征值；有监督学习

中图分类号：TP391　　**文献标识码：**A

A New Chinese Subjective Sentences Recognition Method Based on Word Co-occurrence Relationship Graphic Model

Wang Mingwen, Fu Cuiqin, Xu Fan, Hong Huan

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

Abstract: Different from the traditional term independence assumption-based bag-of-words model, we present a new word co-occurrence relationship-based graphic model. Our model describes the distribution difference among the terms within both subjective and non-subjective sentences sets via the term co-occurence and syntactic information, also integrates a indegree-based term weighting calculation mothod. Evaluation on the benchmark dataset shows the importance of the term co-occurrence graphic model. It also shows that our model significantly outperforms the bag-of-words model currently in the subjective sentence identification field.

Key words: word co-occurrence; graphic model; subjective sentence identification; feature value; supvervised learning

1. 引言

随着 Web2.0 应用的普及，用户生成内容（User-generated content，简称 UGC）与日俱增。通常，这些论坛、贴吧，博客、微博等新型媒介的内容蕴含着大量的用户观点信息，这些观点信息存在巨大的潜在价值。例如，生产商通过 UGC 中的产品评论信息获取用户对产品的情感倾向，并依此作为更新产品的依据。此外，电影投资者通过电影评论来预测电影票房，政府机构根据 UGC 中的事件评论来分析舆情动态等等。

观点句识别(或情感句识别)是情感分析的一个子任务，其旨在从文档中准确抽取出带有情感倾向的观点句子和不带情感倾向的句子，可以被广泛应用于产品调查、市场预测和舆情分析等诸多领域。

主流的观点句识别方法采用有监督的机器学习技术，利用向量空间模型（Vector Space

* **基金项目：**国家自然科学基金资助项目(61272212，61163006，61203313，61365002)

作者简介： 王明文（1964—），男，博士，教授，主要研究方向为计算语言学、信息检索、数据挖掘和机器学习；付翠琴（1990—），女，硕士研究生，主要研究方向为信息检索和中文信息处理；徐凡（1979—），男，博士，讲师，主要研究方向为自然语言处理和云计算；洪欢（1991—），男，硕士研究生，主要研究方向为信息检索和数据挖掘。

Model，简称 VSM）来表示文档，即把每篇文档表示成一个词项向量或特征向量。这种文档特征向量的表示方法基于词项间强独立性假设，并未考虑词项与词项之间的顺序和依赖关系。在英文观点句识别中，采用基于 VSM 的有监督机器学习的分类方法可以取得不错的识别性能。然而，由于中文微博、论坛、贴吧等评论信息都是口语化的文本，表达方式多样，而且评论的长度一般有限，这些缺点导致了手工构建语法库不仅工作量大，而且与日常口语的表达方式仍然存在差异。采用基于 SVM 的有监督机器学习方法进行中文观点句识别并不能取得较好的性能。基于图模型的文本表示方法[1-3]可以很好地捕捉中文文本中词项的依赖和句法关系，该方法在信息检索、文档摘要和词义消歧等已取得较好的效果。鉴于此，本文将一种新型的基于词项共现关系的图模型方法应用于中文观点句识别中。该方法通过构建词项共现关系有向图模型，利用词项与词项之间的共现性和句法关系来描述词项在观点句和非观点句集合中的分布差异，同时采用基于入度的词项权重计算方法来计算词项特征值。本文的方法能够有效地捕捉到中文句子中的语法信息，从而免去了昂贵的手工建立语法库的工作；同时本文结合基于信息检索的复杂特征值计算模型，将词项分布特征及词项间的语法信息融入分类器的训练过程中。上述研究在基准语料上进行实验，实验表明采用基于词项关系图模型方法后，中文观点句识别准确率相比目前基于词袋的方法得到显著提升。

本文后续内容组织如下：第 2 节介绍观点句识别的相关工作，第 3 节重点介绍本文提出的词项共现关系图模型方法和相应的特征值计算方法，并在第 4 节给出了实验设置及详细的结果分析。最后，第 5 节是本文的结论和将来工作部分。

2. 相关工作

Pang 等[4]首次将文本中的一元词、二元词作为特征，并且采用布尔值(二元值)和词频等特征值计算方法，通过训练朴素贝叶斯（Naive Bayes，简称 NB）支持向量机（Support Vector Machine，简称 SVM）、最大熵模型（Maximum Entropy，简称 ME）三种分类器对电影评论进行情感分类，实验结果表明使用一元词作为特征和二元值作为特征值训练的 SVM 分类器的观点句识别效果最好。随后，相关文献分别围绕观点句的特征提取、多分类器的融合和特征值计算三个方面展开研究。

针对特征提取方面，Kushal 等[5]从统计学和语言学的角度来提取文本中的有效特征，并融入 N-grams、文本子串和词的邻近性关系等多种有效的特征。Pang 和 Lee[6]通过建立句子和句子、句子和类别之间的关系图，利用图的最小切割算法来识别文档的主观性部分，从而达到主客观句子分类的目的。Mullen 等[7]利用 WordNet 对标注语料中的特征进行扩展，使用 Turney 提出的语义指向法[8]来提取语法方面的特征，并用这些混合特征训练 SVM 分类器，对句子的情感倾向进行识别。

针对多分类器融合方面，Prabowo 等[9]构建有监督的分类器时融入规则方法，并用于文本的倾向性分类任务。Qiu 等[10]提出了一个自学习的分类模型，该模型分成两个阶段（Phase1 和 Phase2），其中 Phase1 是特征自学习的过程，Phase2 分类过程，并将分类的结果加入 Phase1 以便增强自学习的性能。吕云云等[11]提出基于自举（BootStrapping，简称 BS）的集成分类器的中文观点句识别方法，他们利用 Fisher 线性辨别器提取特征并计算特征值，训练了 NB，SVM 和 ME 三种分类器，并将三类分类器的结果进行集成，同时将集成分类器具有高置信度的分类结果循环用于分类器的训练过程。

针对特征值计算方面，Justin 等[12]提出了一种专门用于情感分析的 Delta Term Frequence-Inverse Document Frequence (D-TFIDF)特征值计算方法，通过分别计算词项在观点句和非观点句中的分布来提高分布差异性大的词项的重要性，减弱分布均衡的词项的影响。在 Justin 的工作基础上，Georgios 等[13]将信息检索中的词项权重计算方法首次应用于英文文本情感分析中，实验证明了基于 D-TFIDF 的 BM25[14]的方法具有最好的情感分类性能。

Deng 等[15]提出了基于词的重要性（importance of a term in a document，简称 TID）和词的情感值（importance of a term for expressing sentiment，简称 TIS）的特征值计算方法，实验表明在分类性能上要优于基于 D-TFIDF 的 BM25 方法。这些特征值计算方法对应的情感分类实验结果均优于 Pang 提出的用二元值方法,说明了采用不同的特征值计算方法会对观点句识别效果产生较大影响。

综合以上观点句识别模型所述，传统词袋模型驱动的方法依赖于词项间的强独立性假设，忽略了词项之间的依赖关系。因此，本文通过构建词项共现关系图模型，着重考虑了词项间的共现性和句法关系，同时研究和扩充了多种基于信息检索的词项特征计算方法。

3. 基于词项共现关系图模型的中文观点句识别方法

本节主要阐述本文提出的新型中文观点句识别方法,主要包括词项共现关系图构建和特征值计算两个方面的内容。

3.1 词项共现关系图构建

已有研究[16 17]表明形容词、动词和名词对句子的观点表达影响更大，而且观点句的表达方式也存在一定的模式，例如：形容词+名词、动词+形容词等结构的句子是观点句的概率更大。然而，微博、论坛、贴吧等评论信息都是口语化的文本，表达方式多样，而且评论的长度一般有限。这些缺点导致了手工构建语法库不仅工作量大，而且与日常口语的表达方式仍然存在差异。

基于此，本文分别构建基于观点句集和非观点句集的词项有向图，自动学习词项在观点句和非观点句中的分布，获得词项之间共现和邻近关系。在有向图中，顶点代表词项，有向边代表词项与词项的共现和邻近关系。若在一个固定窗口中两个词项同时出现，则建立一条词项到词项的有向边，边的方向由词在原文中出现的顺序决定。采用前者指向后者构建的图为前向图（Forward Graph,简称 FG）采用后者指向前者构建的图为后向图（Backward Graph，简称 BG）。本文以句子为单位来构建词项的有向图，即窗口只在一句话中进行滑动，这是因为一句话在内容上相对比较相近，同时表达的语义也更为完整。我们通过固定窗口大小,并设定边的方向，构建出每句话的词项有向图，然后将所有句子的有向图进行合并得到相应的词项共现关系图模型。

为了清晰起见，下面通过观点句例 1 和例 2 来说明 FG 模型的构建过程。

例 1：总体感觉还是不错，看世界杯挺爽。

例 2：效果非常不错，感觉就像真的一样。

首先对每个句子进行预处理，包括分词、删除标点符号和一些无用字符等步骤。得到观点句例 1 的预处理结果为：“总体 感觉 还 是 不错 看 世界杯 挺 爽”，观点句例 2 的预处理结果：“效果 非常 不错 感觉 就 像 真 的 一样”；

然后以句子为单位，按固定窗口进行滑动，窗口内共现的词项之间建立一条边，词项在原文中出现的顺序即为边的方向；

最后将上述句子的有向图进行合并，得到最终的 FG 模型。

图 1 显示了窗口大小为 3 条件下的例句 1 和例句 2 合并后的 FG 模型。

根据图 1 所示的词项共现关系图模型,我们可以计算出每个词项在观点句和非观点句中的分布情况，分布差异大的词项比分布均衡的词项对观点句的判定影响更大。另外，根据词的共现性和邻近性，使用频率高的短语及语法结构也会更加突出。例如：图 1 中，带有明显观点倾向的词项“感觉”和“不错”所连接的边数相对较多,同时与这两个词项有直接连接边的词项也是常用的搭配。

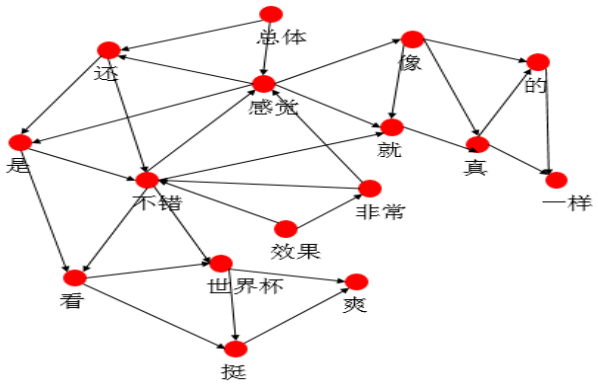

图 1 FG 模型实例

3.2 特征值计算

本节首先介绍已有的基于信息检索的特征值计算方案，然将重点介绍基于上述词项共现图模型文本表示形式下的特征值计算方法。

3.2.1 基于信息检索的特征值计算

在信息检索中，公式(1)是经典的词项权重计算方法，其中 tf_{ij} 表示词项 i 在文档 j 中出现的次数，df_i 表示整个文档集合中出现词项 i 的文档数，N 表示整个文档集合的文档数。公式(1)包含了词频 tf 和逆文档频率 idf 两部分。

$$w(i,j) = tf_{ij} * \mathrm{idf}_i = tf_{ij} * \log \frac{N}{\mathrm{df}_i} \tag{1}$$

结合上述信息检索的词项权重计算方法，本文将特征值的计算分成了词项的权重(term weighting，简称 tw)和词项的分布（inverse document frequence，简称 idf)两部分，根据已有信息检索模型的得分函数，给出了表 1 和表 2 所示的改进后的特征值计算方法。其中 dl_j 表示文档 j 的长度，ave_dl 表示文档集的平均文档长度。

表 1 词项权重计算方案

符号	公式
$tw_0(i,j)$	$\begin{cases}1 & tf_{ij} > 0\\ 0 & other\end{cases}$
$tw_1(i,j)$	$1+\log(tf_{ij})$
$tw_2(i,j)$	$0.5+\frac{0.5 \times tf_{ij}}{\max_k(tf_{kj})}$
$tw_3(i,j)$	$1+\ln(1+\ln(tf_{ij}))$
$tw_4(i,j)$	$\frac{(k+1) \times tf_{ij}}{k \times (1-b+b \times dl_j/ave\_dl)+tf_{ij}}$

表 2 词分布计算方案

符号	公式
$idf_1(i)$	$\begin{cases}1 & df_i > 0\\ 0 & other\end{cases}$
$idf_2(i)$	$\log N+1/df_i$
$idf_3(i)$	$\log\left(N-df_i+0.5/df_i+0.5\right)$

3.2.2 基于入度的词项共现图模型的特征值计算

现有的权重计算方法如 TF-IDF 和 BM25 均以词袋模型的形式来表示文档，其词项权重的计算都是基于词频。然而，本文通过顶点的入度数来确定每个词项的权重。如果指向某个顶点的边数越多，则说明词项的共现次数越多。基于入度的权重计算方法不仅计算简单，而且能更好的捕捉词项与词项之间的关系。已有的工作表明[1-2]，采用基于入度的词项权重可以很好的找到文档的中心（重要词项），相应地，在中文观点句识别时，找到图模型中带有观点倾向的词项（即中心）。

本文分别构建了基于观点句集和非观点句集的词项共现图模型，模型中边的方向定义有两种：FG 和 BG。因此，本文计算词项在两个图模型中的权重值时，分别采用以下两种方法：

1）基于 FG 模型的词项权重称为前向词权重（Forward Term Weighting，简称 FTW）。由两部分组成：基于主观句集 FG 模型上词项的权重 ftw^s 和基于非主观句集 FG 模型上词项的权重 ftw^n。

2）基于 BG 模型的词项权重称为后向词权重（Backward Term Weighting，简称 BTW）。由两部分组成：基于主观句集 BG 模型上词项的权重 btw^s 和基于非主观句集 BG 模型上词项的权重 btw^n。

在上述构建的词项共现图模型中，我们已经分别计算了词项权重 FTW 和 BTW，得到公式(2)和公式(3)所示的基于图模型的权重计算方法。其中，ftw_i 是词项 i 在文档集构建 FG 模型中的权重，btw_i 是词项 i 在文档集构建 BG 模型中的权重。

$$ftw_i = ftw_i^s + ftw_i^n \tag{2}$$

$$btw_i = btw_i^s + btw_i^n \tag{3}$$

结合基于信息检索的权重计算方法，本文针对 FTW 和 BTW 分别给出了八种基于图模型的词项权重计算方案，具体如表 3 所示，词项分布的计算方案采用表 2 一致的计算方法。

表 3 基于图模型的词项权重计算方案

符号	公式	符号	公式
$ftw_1(i)$	$1+\log(ftw_i)$	$btw_1(i)$	$1+\log(btw_i)$
$ftw_2(i)$	$0.5+\frac{0.5\times ftw_i}{\max_k(ftw_k)}$	$btw_2(i)$	$0.5+\frac{0.5\times btw_i}{\max_k(btw_k)}$
$ftw_3(i)$	$1+\ln(1+\ln(ftw_i))$	$btw_3(i)$	$1+\ln(1+\ln(btw_i))$
$ftw_4(i)$	$\frac{(k+1)\times ftw_i}{k\times\left(1-b+b\times\frac{dl_j}{ave\_dl}\right)+ftw_i}$	$btw_4(i)$	$\frac{(k+1)\times btw_i}{k\times\left(1-b+b\times\frac{dl_j}{ave\_dl}\right)+btw_i}$

4. 实验及结果分析

本节将通过实验验证基于词项共现图模型在中文观点句识别任务的有效性，并对实验结

果进行详细的分析。

4.1 实验设置

本文采用第三届中文倾向性分析评测[①]所发布的电子产品评论作为语料集，该语料共包括 2,000 篇电子产品领域的文档，去噪后语料包含观点句 5,662 条，非观点句 9,266 条。为了取得平衡的数据，我们采用了随机裁剪的方法，使得观点句和非观点句在数量上相当。同时，使用中国科学院的分词与词性标注软件[②]对语料进行分词预处理，采用 LIBSVM[③]作为分类器(参数均取默认值)，并采用十折交叉验证方法获取实验结果，词项滑动窗口的取值范围为 2 至 6。

为了验证基于词项共现图模型的中文观点句识别方法的有效性，本文将已有的基于信息检索（IRM）的中文观点句识别方法和吕云云等[11]提出的基于 BootStrapping 的集成分类器（BSM）的中文观点句识别方法作为 Baseline。BSM 是采用 Bootstrapping 的方法扩展训练语料，分别训练贝叶斯、支持向量机和最大熵分类器。然后，通过给三个训练好的分类器赋权获得一个集成分类器。此外，本文还将通过设置两个对比实验来验证，在中文观点句识别当中，基于 FG 模型和基于 BG 模型的性能。

4.2 实验结果及分析

4.2.1 Baseline 实验结果

图 2 显示了 IRM 和 BSM 模型的中文观点句识别性能。在 IRM 模型中，tw 有 5 种计算方法，idf 有 3 种计算方法，共有 15 种计算方案，实验参数取值采用已有文献中求得的经验参数值 k=1.2 和 b=0.75。在 BSM 模型中，语料标注率在 0.05 和 1 之间以 0.05 的步长进行平滑，在标注率分别为 0.4 和 1 时，训练的集成分类器的效果达到最好，分别为 0.7786 和 0.7724。因此，本文只取 $bsm_{0.4}$ 和 bsm_1 的实验结果作对比。

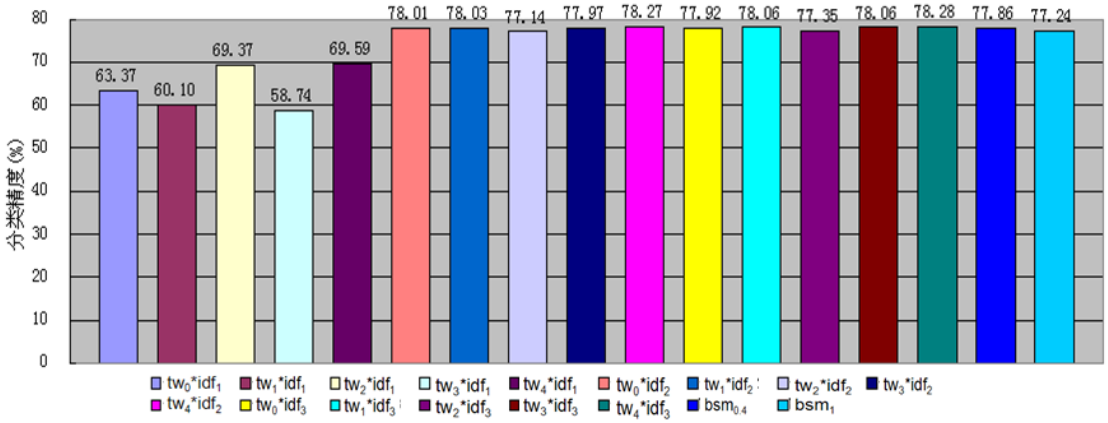

图 2 Baseline 实验结果

图 2 数据表明：

1）在英文观点句识别中，仅使用二元特征值就可以取得很好的效果，分类精度通常可以达到 82.9%，但该方法在中文观点句识别中效果并不好，图 2 中 tw_0*idf_1 的精度只有 63.37%，这一实验结果说明了简单的二元特征值计算方法不适合于中文观点句识别任务；

2）在计算特征值时仅考虑词项权重时总体识别效果都不太好，其中 tw_1*idf_1、tw_2*idf_1、tw_3*idf_1 和 tw_4*idf_1 的分类精度分别为 60.10%、69.37%、58.74%和 69.59%，精度均低于 70%。相比较而言，仅用词项分布来计算特征值的方法分类效果具有显著提升，识别性能均在 77%

① http://www.ir-china.org.cn/
② http://ictclas.org
③ http://www.csie.ntu.edu.tw/~cjlin/libsvm/

以上，其中在 tw_0*idf_2 和 tw_0*idf_3 方案中，精度分别高达到 78.01%和 77.92%，这一实验结果说明词项在语料中的分布情况对于观点句的识别具有重要的作用。同时，BM25 模型（tw_5*idf_3 方法）可以取得更好的识别效果，分类精度达到最高（78.28%）。

3）整体上来说，在中文观点句识别中，使用基于信息检索的特征值计算方法可以显著提升中文观点句识别效果，甚至可以取得比 BSM 模型训练的集成分类器更好的分类性能。

4.2.2 基于 FG 模型的特征值计算

图 3 显示了窗口大小为 3 时的 FG 模型中特征值计算下的中文观点句识别性能。图中实验结果表明：使用基于 FG 模型的特征值计算方法，分类性能都有较大幅度的提升。在不考虑词项分布的 ftw_1*idf_1、ftw_1*idf_2、ftw_1*idf_3 和 ftw_1*idf_4 方案中，分类性能分别达到了 75.98%、66.55、71.68%和 69.98%，与基于信息检索的同等条件下的方案相比，性能都有所提升，这一实验结果充分说明本文构建的 FG 图模型可以捕捉到更多的观点句和非观点句中词项间的依赖及语法关系，并用于分类器的训练。基于 BM25 模型的 ftw_4*idf_2 方案可以达到最好的性能 81.22%，ftw_3*idf_2、ftw_3*idf_3 的分类性能也都在 80%以上。但是，ftw1*idf3 和 ftw2*idf3 方案与同等条件下的基于信息检索的方案相比，性能会有所降低（仅有 73.07%和 75.18%），原因在于一些非观点词的共现频率过高，导致分类器引入更多的噪音。

另外，基于 FG 模型的观点句识别方法也可以获得比基于 BSM 模型更好的分类性能。

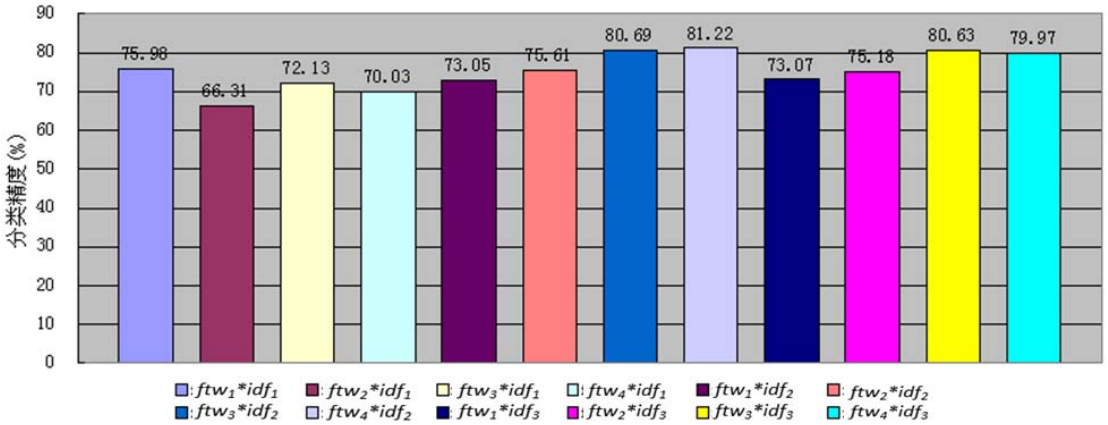

图 3 基于 FG 的特征值计算

4.2.3 基于 BG 模型的特征值计算

图 4 显示了窗口大小为 3 时的 BG 模型中特征值计算下的中文观点句识别性能。

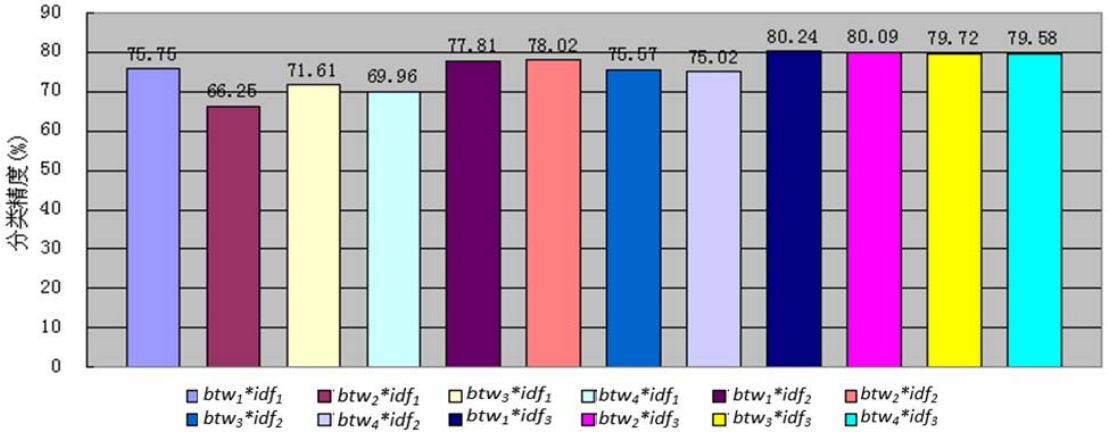

图 4 基于 BG 的特征值计算

实验结果表明，基于 BG 模型的分类性能与基于 FG 的方法的分类性能总体相差不大。分析图 3 和图 4 可得，采用 ftw_3*idf_3、ftw_4*idf_3 和 btw_3*idf_3、btw_4*idf_3 方案计算特征值的性能较好，而且也较稳定。其中，最好的模型分类性能是方案 bw_1*idf_3，高达 80.24%，比基于

FG 的方法最好的性能稍差。ftw_2*idf_1的性能最差，只有 66.31%，分析原因是在计算词项权重时用文档集中最大的权重值进行了平滑，导致部分观点词的影响减弱。

通过上述分析可得，基于 FG 和 BG 模型采用不同的特征值计算方案，得到的观点句识别性能会有相对较大的浮动，原因在于这些特征值计算方案是直接采用信息检索中性能较好的计算模型，未能考虑中文观点句识别中存在的用词习惯、语法及文档长度等属性。相信找到更加适合中文观点句识别这一任务的特征值计算方案可以取得更好的识别性能，这也将作为我们未来的研究工作。

4.2.4　窗口大小的选择

图 5 显示了在 FG 模型中，ftw_4*idf_2特征值计算方法在不同的词项滑动窗口下的分类性能。实验表明词项滑动窗口大小与分类精度并不是简单线性关系。例如：在窗口大小为 2 时，分类精度为 80.3161%；在窗口大小为 3 时，分类精度达到最大 81.2169%。该实验结果可以很好的总结汉语表达的语法习惯，即在中文表达观点时，作者倾向于采用三元成分结构。当窗口逐渐增加，分类精度却有逐渐下降，原因在于随着窗口大小的增大，很多非合法语法搭配结构被当作有用的特征来计算词项的 FTW 和 BTW 值，即加入了很多噪音特征，从而导致分类精度下降。

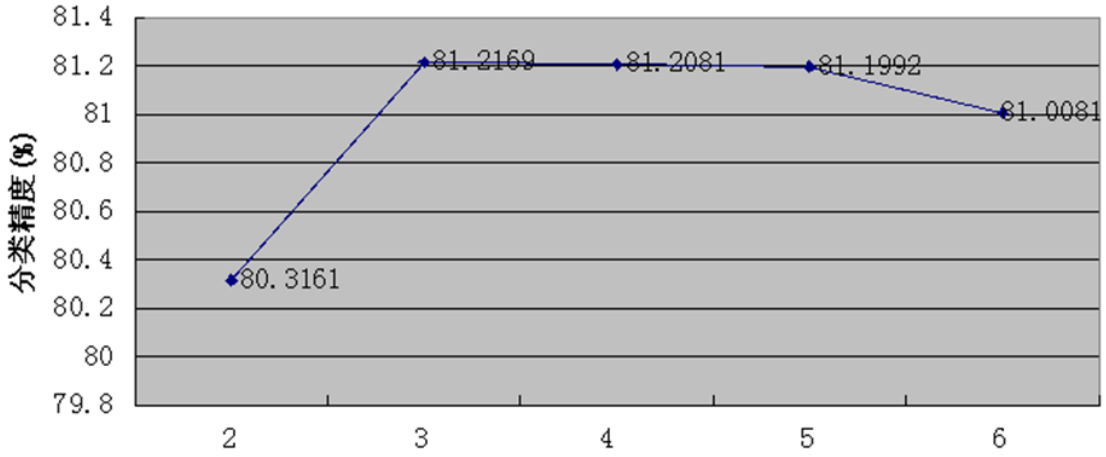

图 5 FG 模型中调整词项滑动窗口大小对性能的影响

5. 总结与展望

针对传统的词项间强独立性假设的词袋模型驱动的中文观点句识别方法的不足，本文提出了一种新型的基于词项共现关系的图模型方法。该方法通过构建词项共现关系有向图模型，利用词项与词项之间的共现性和句法关系来描述词项在观点句和非观点句集合中的分布差异。同时，在构建的图模型上，本文采用基于入度的词项权重计算方法计算每个顶点（词项）的权重，并结合基于信息检索的特征值计算方案计算特征向量的特征值。上述研究在基准语料上进行实验，实验表明采用基于词项关系图模型方法后，中文观点句识别准确率相比目前基于词袋的方法得到显著提升。

将来工作主要包括以下两个方面：(1)在跨领域数据集上验证本文提出的模型性能；(2)将本文提出基于词项共现图模型的特征计算方法与目前已有的特征计算方法进行结合，以体现方法的协同性。

参考文献

[1] Rousseau F, Vazirgiannis M. Graph-of-word and TW-IDF: new approach to ad hoc IR[C]//Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. ACM, 2013: 59-68.

[2] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) . 2004, 4(4): 275.

[3] 洪欢, 王明文, 万剑怡, 等. 基于迭代方法的多层 Markov 网络信息检索模型[J]. 中文信息学报, 2013, 27(5): 122-128.

[4] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the Associate Computational Linguistics 02 Conference on Empirical Methods in Natural Language Processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[5] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th International Conference on World Wide Web. ACM, 2003: 519-528.

[6] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 271.

[7] Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2004, 4: 412-418.

[8] 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. Association for Computational Linguistics, 2002: 417-424.

[9] Prabowo R, Thelwall M. Sentiment analysis: A combined approach[J]. Journal of Informetrics, 2009, 3(2): 143-157.

[10] Qiu L, Zhang W, Hu C, et al. Selc: a self-supervised model for sentiment classification[C]//Proceedings of the 18th ACM Conference on Information and knowledge Management. ACM, 2009: 929-936.

[11] 吕云云, 李旸, 王素格. 基于 BootStrapping 的集成分类器的中文观点句识别方法[J]. 中文信息学报, 2013, 27(5): 84-92.

[12] Martineau J, Finin T. Delta TFIDF: An Improved Feature Space for Sentiment Analysis[C]// Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM). 2009: 258-261.

[13] Paltoglou G, Thelwall M. A study of information retrieval weighting schemes for sentiment analysis[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1386-1395.

[14] Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields[C]//Proceedings of the 13th ACM International Conference on Information and Knowledge Management. ACM, 2004: 42-49.

[15] Deng Z H, Luo K H, Yu H L. A study of supervised term weighting scheme for sentiment analysis[J]. Expert Systems with Applications, 2014, 41(7): 3506-3513.

[16] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2003: 105-112.

[17] Kim S M, Hovy E. Determining the sentiment of opinions[C]//Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, 2004: 1367.

通讯作者信息
姓名：付翠琴
地址：江西省南昌市紫阳大道 99 号江西师范大学瑶湖校区
邮编：330022
电话：15179131398
电子邮箱：fucuiqin@126.com