一种基于语义标注特征的金融文本分类方法*

罗 明[a], 黄海量[a,b†]
(上海财经大学 a. 信息管理与工程学院; b. 上海市金融信息技术研究重点实验室, 上海 200433)

摘 要: 针对基于词袋的机器学习文本分类方法所存在的高维度、高稀疏性、不能识别同义词、语义信息缺失等问题,和基于规则模式的文本分类所存在的虽然准确率较高但鲁棒性较差的问题,提出了一种采用词汇—语义规则模式从金融新闻文本中提取事件语义标注信息,并将其作为分类特征用于机器学习文本分类中的新方法。实验证明采用该方法相比基于词袋的文本分类方法在采用相同的特征选择算法和分类算法的基础上,F_1 值提高 8.6 %,查准率提高 7.7%,查全率提高 8.8%。本方法融合了知识驱动和数据驱动在文本分类中的优点,同时避免了它们所存在的主要缺点,具有显著的实用性和研究参考价值。

关键词: 文本分类; 金融文本; 语义标注; 词汇—语义模式; 有限状态机
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1001-3695(2018)08-2281-04
doi:10.3969/j.issn.1001-3695.2018.08.010

New approach of financial text classification based on semantic annotation features

Luo Ming[a], Huang Hailiang[a,b†]
(*a. College of Information Management & Engineering, b. Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance & Economic, Shanghai* 200433, *China*)

Abstract: The main problems of traditional machine learning text classification method which based on BOW (bag of words) are high dimension and high sparseness, can not identify synonyms and lack of semantic information etc. Meanwhile, rule based methods have high precision but have weaker robustness. In order to solve these problems, this paper proposed a novel method which based on lexical-semantic patterns to extract event semantic annotations from financial news text, and applied these annotations as features in machine learning method. The experiment shows that this method lifts F_1 value 8.6% than BOW, and the precision is increased by 7.7%, recall is increased by 8.8%, which based on same feature selection algorithm and classification method. This method combines the advantages of the two methods of knowledge driven and data driven in text classification, at the same time avoids the major drawbacks of last two methods, it has a good practical and research reference value.

Key words: text classification; financial text; semantic annotation; lexical-semantic pattern; finite state machine

0 引言

文本分类按其实现原理可分为基于知识驱动(knowledge-driven)和基于数据驱动(data-driven)两大类。基于知识驱动的文本分类主是领域专家依据领域、语法知识,通过人工定义规则的方法来实现文本分类,这种方式虽然具有匹配精度(precision)高、计算量小、不需要大量训练数据、可解释性较强、适合多标签分类等优点,但同时存在人工编写规则模式费时费力、规则的鲁棒性(通用性和灵活性)较差、召回率(recall)提高困难等缺点。基于数据驱动的文本分类是采用机器学习的方法从大量训练样本数据中学习分类模型,由于这种方法具有人工干预少、通用性和灵活性较强且对领域和语言知识要求较少等优点,已成为目前文本分类的主流方法,但这种方法也存在着需要标注大量的训练样本、多标签分类效果不好、计算的时间复杂度和空间复杂度高等缺点。

基于词袋(BOW)表示的向量空间模型[1]是目前机器学习文本分类方法中普遍采用的特征表示方式。词袋法采用文本中经过分词处理后的词条作为特征项,具有原理简单和构造方便等优点,但是也存在着高维度性、高稀疏性、不能识别同义词、语义信息缺失等缺点[2],在事件识别的过程中很容易导致维度灾难、降低分类器效率、过度拟合、易受数据噪声干扰等问题,因此如何提取具有良好分类性能的特征就成为事件识别研究中一个重要的问题。文献[3]先通过 TF-IDF 方法来计算文档中的词权重,然后在 KNN 算法中采用 TF-IDF 值作为计算相似度的参数,显著提高了分类器的执行速度; 文献[4]提出了一种基于主题词的向量空间模型(topic-based vector space model,TVSM),与传统的 VSM 比较,TVSM 模型采用主题词代替一般词条,极大地压缩了向量空间维度,提高了特征向量的语义含量,可以更好地概括出文本的语义特征;文献[5]先采用浅层句法分析来标注语义角色,再根据问题焦点结构和抽取规则获取焦点特征,并在此基础上结合焦点特征中的疑问对象特征所对应的本体类别来确定问题的分类;文献[6]采用知网中的语义相似度来计算特征词间的语义相似度,并赋予不同的权重以提高文本分类的精度;文献[7]采用一种基于语料的叙词表(corpus-based thesaurus)和 WordNet 方法来抽取简单的语义特征,并采用 KNN 算法和 BP 神经网络算法进行分类。综上所述,目前有关特征提取的研究主要集中在探讨如何采用机器学习方法获得特征上[3~5]和研究如何利用 WordNet、知网等通用知识库来分析词条之间的同义、近义和关联关系以获得词条语义特征[6,7]两个方面,对于如何采取语义规则的模式来提取针对某一专业领域的概念语义特征(例如时间、地点、交易金额、收购标的物等),并将其用于文本分类中的研究则较少涉及。

本文从融合知识驱动和数据驱动两种文本分类方式优点的思路出发,面向金融新闻文本领域提出了一种基于语义标注特征的文本分类新方法。这种方法的核心思路是:通过词汇—语义模式方法(lexical-semantic pattern),在不依赖其他外部知

收稿日期: 2017-04-09; **修回日期**: 2017-05-24 **基金项目**: 上海市科技人才计划项目(14XD1421000);上海市科技创新行动计划项目(16511102900);上海财经大学 2014 年研究生创新基金资助项目(CXJJ-2014-438)

作者简介:罗明(1974-),男,重庆人,高级工程师,博士研究生,主要研究方向为自然语言处理、数据挖掘;黄海量(1975-),男(通信作者),教授,博士,主要研究方向为金融领域、电子商务和移动互联网中的大数据分析、用户行为、商务智能和机制设计等(hlhuang@shufe.edu.cn).

识库的前提下,从文本中抽取出一些与专业相关的重要语义标注特征,例如事件发生时间、事件状态、事件标的物、事件动词类型等,然后采用这些特征来构建语义向量空间表示模型,再采用卡方统计[8]、CFS[9]、TF-IDF[10]特征选择算法进行特征选择,最后分别采用朴素贝叶斯(NB)、支持向量机(SVM)和 K 最近邻(KNN)三种分类器进行分类效果验证。

1 处理流程

语义标注特征的提取、特征选择及文本分类处理的总体流程如图 1 所示。

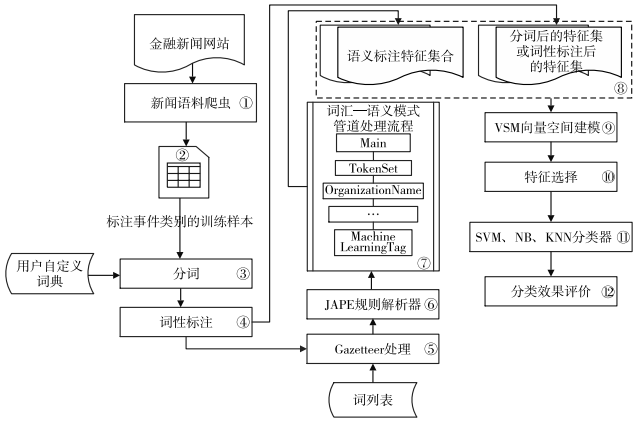

图1 处理流程

流程要点说明:

a)在环节①②处完成原始金融文本语料爬取和训练数据分类标签工作。

b)在环节③④处采用 FNLP 中文自然语言处理软件[11]对文本进行分词和词性标注工作。环节③所加载的用户自定义词典包含一些中英文机构名称和简称,以及金融领域专用名词(例如现金红利、法人)和动词(例如转增、持有)。使用自定义词典的目的是对专业领域文本的分词更加准确。

c)在环节⑤,采用词列表(word list)方法[12,13]来构建概念同义词典,以实现对基础语义概念(lookup 类型)的标注。

采用 word list 的优点是概念词典构建轻便简单,调整扩展方便,不局限于词性类别和词数目(可以由多个词构成的词组来表达某一概念),使用时不依赖于外部通用知识库(WordNet 或 HowNet 等)。它比较适合于面向专业领域、判断逻辑不复杂的情况,构建一个 word list 的流程如下:

(a)在索引文件(list. def)中定义相关词典文件的列表,例如:

event_Verb_Buy. lst:事件动词类型,收购
event_Verb_Cooperation. lst:事件动词类型,合作
…

其中:event_Verb_Buy. lst 代表一个词典文件名;"事件动词类型"代表该文件词汇所对应的主类别(majorType);"收购"代表该文件词汇对应的次级类别(minorType)。

(b)在二级文件中(例如 event_Verb_Buy. lst)中进一步定义主类属于"事件动词类型",次类属于"收购"的所有同义词组,例如:

收购
购买
竞购
资产 购买
资产 收购
…

同义词组一部分可以从中文同义词典中获取(如 HowNet 或同义词林),但主要由领域专业人员根据经验手工编制来完成。

(c)概念同义词典定义完成后,采用最大后向匹配算法在输入文本的相应词组位置标注上 token 和 lookup 类型的基础语义标注,例如:本文采用 GATE[14]中的 ANNIE 插件将文本中的"收购、购买…"这类词汇标注上类型为 lookup,majorType = "事件动词类型",minorType = "收购"的元语义信息,这些元语义信息将作为基础语义概念供第 3 章词汇—语义标注算法调用处理。

d)环节⑥⑦是采用 JAPE 语言[15]编制的一系列词汇—语义模式规则文件(规则文件的格式见第 2 章样例:OrgRule),并通过管道处理方式逐步提炼出标注类型为 event 的初级和高级语义标注信息,并且在 event 类型基础上生成类型为 MLTag 的语义标注信息,专门用于机器学习。有关词汇—语义标注的模型和标注算法在第 2、3 章着重说明。

e)环节⑧可以将获得的语义标注特征集单独使用,或者与词条组合,形成输入特征集合,并构建基于语义的向量空间模型。

f)在环节⑨~⑫中,采用特征选择和分类算法实现特征过滤和分类效果评价。具体实现方法将在第 4、5 章说明。

2 词汇—语义标注模型

有限状态机[16]是用于描述有限个状态以及这些状态之间的转移和动作的数学模型,例如文献[17]采用有限状态机模型来描述中文地址字符串的标准化处理过程。本文的词汇—语义规则模式的处理过程采用有限状态机模型定义如下:

$$M=(\Sigma,Q,q_0,F,\Delta) \tag{1}$$

其中:

a)Σ 为模型 M 的输入 token 信息的集合,$\Sigma=\{a_1,\cdots,a_n\}$,$a_1,a_2,\cdots,a_n$ 为分词处理后形成的 token 序列。

b)Q 为模型 M 中有限的状态集合,在本文中 Q 指每条规则中的满足匹配语句的状态集合,例如对于词汇—语义规则:

```
Rule: OrgRule
Priority: 100
(
({lookup.minorType == ~"(country|province|city)"}) +①
({token.string != ~"[,,。:::;;、\\d]+",lookup.majorType! =~"(title)"})[1,6] ②
(ORG_KEY_COMPANY) ③
(
{token.string == ~"[((]"}
({token.string != ~"[。]"})[1,15]
{token.string == ~"[))]"}
)? ④
):MyOrg
```

在本条规则中,Q 共有六个状态,即初始状态 q_0 和接受终止状态 q_f、规则中的①~④四条语句判断为真时所对应的状态分别为 $q_1\sim q_4$,因此 $Q=\{q_0,q_1,q_2,q_3,q_4,q_f\}$。

c)q_0 代表模型 M 的初始状态,$q_0=\varnothing$。

d)F 代表模型 M 的最终可接受状态集,$F\subseteq Q$,$F=\{q_f\}$。

e)Δ 表示转移函数集合,δ 为转移函数(transition function),$\Sigma\times Q\xrightarrow{\delta}Q,\delta\in\Delta$,在词汇—语义规则中 Δ 为每条规则中的规则子句集合(如本例中的①~④子句),本例中:$\Delta=\{\delta_1,\delta_2,\delta_3,\delta_4\}$,每条规则子句对应一个转移函数 δ_i,后一状态 q_i 与前一状态 q_{i-1} 的转换关系满足公式:

$$q_i=\delta_i(q_{i-1},a_i) \tag{2}$$

其中:a_i 为当前的输入 token。

f)对一个特定的输入 token 序列,例如:$\Sigma^*=\{$中南建设,6 月,8 日,晚间,公告,公司,拟,出资,10 亿,元,…$\}$,在状态机 M 上的匹配执行结果是一个状态序列:$q_0,q_1,\cdots,q_n$,q_n 表示终止状态,如果 $q_n\in F$ 则表示该 token 序列被状态机接受(即匹配成功),否则被拒绝。

g)为了简化模型表示,M 中不记录拒绝状态和转向拒绝状态的转移函数。

3 词汇—语义标注算法

算法 1 词汇—语义规则标注算法

输入:D 为采用 GATE ANNIE 插件预处理后,已经标注有 token 和 lookup 标注类型的输入文档;P 为满足 JAPE 语法规则的词汇—语义规则文件集合。

输出:MLAnnotateSet 为输出的标注类型为 event 的语义标注集合。

1　for each $phase_i$ in P　//$phase_i$ 为 P 中的某一规则文件

2　　getting all annotations from outAS List of Last $(i-1)$ phase and put them in inAS list　/*将上一个规则文件的处理结果取出放入当前处理序列 inAS 中*/

3　　for each $rule_j$ in $phase_i$. rules //对 $phase_i$ 中的规则进行遍历

4　　　for each D. $Nodes_k$ in D /* D. $Nodes_k$ 为文档 D 中的 token 节点*/

5　　　　put D. $Nodes_k$ in token list L

6　　Initialization finite state machine M_j respect to $rule_j$, let $Q=\{q_0, q_f\}$, $\Delta=\{\delta_1,\cdots,\delta_n\}$, $q_0=\varnothing$

7　　if ($\{L_1,\cdots,L_n\}$ are accepted by M_j) /*当满足规则子句匹配条件时*/

8　a)　feed annotation set in inAS which cover $\{L_1,\cdots,L_n\}$ to RHS for creating new semantic Annotation and put computing results into outAS list　/*将匹配的标注集合送入词汇—语义规则右式(RHS)进行程序逻辑处理,并产生新的语义标注信息*/

　　b)　$L=L-\{L_1,\cdots,L_n\}$;/*继续执行下一段 token 的规则匹配操作*/

9　else

　　search next M_{j+1}　/*查找规则文件中的下一条规则再重新开始匹配操作*/

10　getting all semantic annotation which type is "MLTag" from outAS list, and put them in MLAnnotateSet　/*获得类型为 MLTag 的语义标注集合*/

表 1 是经过语义标注算法处理后,从原始文本数据中获得的语义标注集合的例子。

表 1　标注结果示例

原始文本	分词处理后的 tokens	语义标注集合
中南建设 6 月 8 日晚间公告,公司拟出资 10 亿元设立智能制造分公司,谋划布局高端制造业,开展机器人产业投资	中南建设 6 月 8 日晚间公告,公司拟出资 10 亿元设立智能制造分公司,谋划布局高端制造业,开展机器人产业投资	发布者　时间　事件动词　类型_正式公告　事件状态类型_预案　事件动词类型_投资　交易金额　事件动词类型_投资　投资标的物　机构　事件动词类型_投资

4　文本分类

在完成上述语义标注工作的基础上,文本分类阶段的主要工作有三项,分别是特征选择、分类器学习和分类效果评价,简要说明如下。

4.1　特征选择

特征选择的目的是在获得特征集的基础上,进一步精选出更具有分类能力的特征,从而达到进一步压缩特征空间维度,提高分类器学习效率,降低噪声干扰的目的。本文设计的特征集合有五类,均采用 VSM(向量空间模型)来表示,其中 SEG 表示 segmentation,SEM 表示 semantic,POS 表示 part-of-speech,如表 2 所示。

本文采用的特征选择算法有:

a)卡方统计[8](Chi square statistics,简写为 Chi)

$$\mathrm{Chi}(f,c_i)=\frac{N[P(f,c_i)P(\bar{f},\bar{c_i})-P(f,\bar{c_i})P(\bar{f},c_i)]^2}{P(f)P(\bar{f})P(c_i)P(\bar{c_i})} \tag{3}$$

其中:$\mathrm{Chi}(f,c_i)$ 值越大表示特征项 f 与类别 c_i 越相关;N 表示训练集中的包含所有样本的数量;$P(f,c_i)$ 表示在类别为 c_i 的文档中出现包含特征项 f 文档的概率;$P(\bar{f},\bar{c_i})$ 表示在所有非 c_i 类别的文档中出现不包含 f 的文档概率;$P(f,\bar{c_i})$ 表示在类别为非 c_i 的文档中出现包含特征项 f 文档的概率;$P(\bar{f},c_i)$ 表示在所有 c_i 类别的文档中出现不包含 f 的文档概率;$P(f)$ 表示训练集中所有包含特征项 f 的文档概率;$P(\bar{f})$ 表示训练集中所有不包含特征项 f 的文档概率;$P(c_i)$ 表示训练集中类别为 c_i 的文档概率;$P(\bar{c_i})$ 表示训练集中类别为非 c_i 的文档概率。

表 2　特征集方案

特征集方案	处理方法说明	示例
SEG	训练文本分词并去除停用词后的所有词条特征集合	中南建设 6 月 8 日 晚间 公告 公司 拟 出资 10 亿 元…
POS	训练文本分词并进行词性标注,去除停用词后只保留词性为名词、专有名、实体名、事件名和动词的词条特征集合	公告 公司 出资 设立 智能 制造 分公司 谋划 布局 高端 制造业 产业 投资
SEM	训练文本分词后采用本文词汇—语义规则模式方法生成的语义标注特征集合	发布者　时间　事件动词类型_正式公告　事件状态类型_预案　事件动词类型_投资　交易金额　投资标的物　机构 …
SEM + SEG	SEM 特征集与 SEG 特征集合并后的集合	发布者　时间　事件动词类型_正式公告　事件状态类型_预案…中南建设 6 月 8 日 晚间 公告 公司 拟 出资 …
SEM + POS	SEM 特征集与 POS 特征集合并后的集合	发布者　时间　事件动词类型_正式公告　事件状态类型_预案 …公告 公司 出资 设立 智能 制造…

b)相关性分析特征子集选择[9](correlation-based feature subset selection,CFS)

$$M_S=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \tag{4}$$

其中:M_S 为特征子集 S 的 CFS 值,具有最大 CFS 值的特征子集即为类别 c_i 的最终分类特征;f 为特征项;c_i 为文档类别;k 为 S 中特征项的数目;$\overline{r_{cf}}$ 为特征子集 S 中的所有特征项 f 与类别 c_i 之间的不确定性系数的均值;$\overline{r_{ff}}$ 表示特征子集 S 中所有的特征项两两之间不确定性系数的均值。

c)词频—逆文档频率[10](term frequency-inverse document frequency,TF-IDF)

$$w_{ij}=tf_{ij}\times idf_i=tf_{ij}\times\log_2\left(\frac{N}{df_i}\right) \tag{5}$$

其中:w_{ij} 值越大表示特征项 f_i 在文档 d_j 中的权重越大,越能作为该文档的分类特征;w_{ij} 表示特征 f_i 在文档 d_j 中的权重值;tf_{ij} 表示特征项 f_i 在文档 d_j 中经过标准化处理后的频率值;idf_i 表示特征项 f_i 的逆文档频率值。

4.2　分类器学习

本文选择机器学习领域最常用的三种文本分类器:NB(naïve Bayes)、SVM(support vector machine)、KNN(K-nearest neighbor)三种分类器算法来进行分类学习。模型的训练、测试评价方法采用 10 折交叉验证方法(10-fold cross validation)。

4.3　分类效果评价

本文采用微平均精确率(micro precision,简写为 P)、微平均查全率(micro recall,简写为 R)、微平均 F_1 作为文本分类效果评价指标来对分类效果进行评价。

5　实验及分析

5.1　实验数据准备

本文从金融垂直门户网站——东方财富网(http://finance.eastmoney.com/yaowen_cgsxw.html)爬取了 2015 年全年的 87 616 条公司新闻报道,并从中随机抽取了 3 000 条不重复的新闻语料。由于文章的核心要点往往会集中在标题或主题句处[18],为了减少分类噪声和干扰,本文选择主题句来进行文本分类识别。本文定义了 15 种类型的金融事件,并请多名专业人士完成了文本事件类别的标签工作,最终得到在 15 类金融事件范围内的训练数据共计 1 434 条,数据分布如表 3 所示。

本文采用最大特征数分别为 1 000、500、100 三种方案进行实验。

5.2　F_1 值的综合比较

在所有的分类结果中,比较最大前 10 个 F_1 值(max F_1)和

最小前 10 个 F_1 值(min F_1),其排序结果如表 4 所示。

表 3 训练数据类别分布

事件分类	总数	事件分类	总数
利润业绩增长事件	279	股票分红转增事件	65
拟募资事件	203	重大合同事件	63
拟收购事件	122	投资通过事件	54
拟投资事件	114	利润业绩亏损事件	52
股东增持事件	104	管理层变更事件	50
违法违规违纪事件	83	募资通过事件	49
股东减持事件	82	收购通过事件	40
利润业绩下滑事件	74		

表 4 F_1 值排序结果

方案	max F_1	方案	min F_1
NB + Chi + 100 + (SEM + SEG)	0.826	KNN + Chi + 1000 + (POS)	0.63
SVM + CFS + 1000 + (SEM + POS)	0.823	KNN + Chi + 500 + (POS)	0.63
SVM + CFS + 500 + (SEM + POS)	0.823	KNN + TFIDF + 1000 + (SEG)	0.637
SVM + CFS + 100 + (SEM + POS)	0.823	KNN + TFIDF + 500 + (SEG)	0.637
NB + Chi + 100 + (SEM + POS)	0.82	KNN + Chi + 1000 + (SEG)	0.639
NB + CFS + 1000 + (SEM + SEG)	0.817	KNN + Chi + 500 + (SEG)	0.639
NB + CFS + 500 + (SEM + SEG)	0.817	KNN + Chi + 100 + (POS)	0.641
NB + CFS + 100 + (SEM + SEG)	0.817	KNN + TFIDF + 100 + (SEG)	0.647
NB + CFS + 1000 + (SEM + POS)	0.809	KNN + CFS + 1000 + (POS)	0.648
NB + CFS + 500 + (SEM + POS)	0.809	KNN + CFS + 500 + (POS)	0.648

注:SVM + CFS + 100 + (SEM + POS) 表示分类器是 SVM,特征选择算法是 CFS,最大特征数 100 个;(SEM + POS) 表示特征集合的构成是语义特征集 + 具有特殊词性的词条特征集

从表 4 可知,F_1 值最大的前 10 个方案都采用了包含本文提取的语义特征集合,而 F_1 值最小的 10 个方案里特征集合都未包含语义特征,由此可以看出本文提取的语义特征对文本分类的效果提升是非常显著的,F_1 最大值与最小值之间相差 19.6 个百分点。

5.3 与基准方法的比较

本文将基准方法定义为:采用 SEG 特征集合(代表通过传统词袋法所获取的特征),采用 TFIDF 作为特征选择方法。将此基准方法与采用相同特征选择方法(TFIDF)、相同分类器方法和相同最大特征数的多个方案进行比较,其结果如图 2 所示。

由图 2 可知在相同横坐标条件下,以 SEG 代表的基准方案比其他任何一种带有 SEM 特征集方案的 F_1 值都低,其中差距最大的是 KNN + TFIDF + 1000 + (SEM)(F_1 值为 0.731)和 KNN + TFIDF + 1000 + (SEG)(F_1 值为 0.637)两个方案,相差幅度达到 14.76%。

本文以表 4 中 F_1 值分类效果排名第一的 NB 作为分类器,在采用相同的 TFIDF 特征选择算法条件下,F_1、R、P 的值的结果如表 5 所示。

表 5 相同 NB + TFIDF 下的分类结果

	SEM + POS	SEM + SEG	SEM	POS	SEG
F_1(1000)	0.773	0.754	0.798	0.748	0.712
F_1(500)	0.773	0.754	0.798	0.748	0.712
F_1(100)	0.78	0.771	0.798	0.748	0.735
R(1000)	0.77	0.75	0.796	0.744	0.708
R(500)	0.77	0.75	0.796	0.744	0.708
R(100)	0.776	0.766	0.796	0.744	0.731
P(1000)	0.789	0.767	0.802	0.765	0.725
P(500)	0.789	0.767	0.802	0.765	0.725
P(100)	0.795	0.787	0.802	0.765	0.749

由表 5 可知在采用相同分类器(NB)、相同特征选择算法(TFIDF),相同最大特征数条件下,含有 SEM 特征集的方案 F_1、P、R 值均高于代表词袋方式的 SEG 特征集方案,其中差距最大的是 NB + TFIDF + 1000 + (SEM) 与 NB + TFIDF + 1000 + (SEG),前一种方案比后一种方案(基准方案)F_1 值提高了 8.6%,P 值提高了 7.7%,R 值提高了 8.8%。

5.4 使用不同特征选择算法的比较

本文采用最大特征数为 100 作为相同条件来测试采用不同特征选择算法对文本分类效果的影响,实验如图 3 ~ 5 所示。

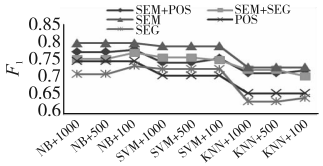

图2 相同TFIDF特征选择方法下的F_1值

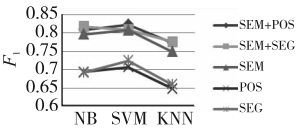

图3 CFS+100的F_1值

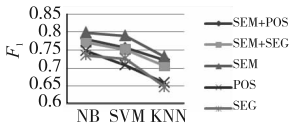

图4 TFIDF+100的F_1值

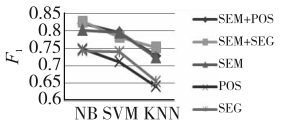

图5 Chi+100的F_1值

由图 3 ~ 5 可知:不同的特征选择算法对文本分类的效果存在显著影响,CFS 和 Chi(卡方)算法分类效果优于基准的 TFIDF 方法;当使用相同 CFS 或 Chi 算法条件下,采用 SEM 与 POS 或 SEG 组成的混合特征分类效果优于只使用 SEM 特征集方案,而使用 TFIDF 算法时,采用 SEM 特征的分类效果最好。产生这种差别的原因在于:CFS 和 Chi 算法在选择特征方面不易受数据噪声干扰,而 TFIDF 算法对数据噪声的抑制能力更弱一些。

通过上述实验可以看出:采用本文方法所提取的语义标注特征,在无论采用哪种分类器或者特征选择算法的条件下,都能够明显提升文本分类的效果,这说明这种基于语义标注的文本分类方法具有较强的鲁棒性。

6 结束语

本文采用词汇—语义规则模式的方法从金融新闻文本中提取了基于事件的语义标注信息,并将其作为一种新的特征用于机器学习的文本分类方法中,实验证明采用这种基于知识驱动和数据驱动的混合方法来进行文本分类,可以显著地提高文本分类的 F_1 值、精确率和召回率指标。采用这种方法一方面能够避免单纯使用规则方式进行文本分类所存在的通用性、灵活性较差的问题,同时又能够避免采用机器学习方法所存在的特征空间高维度、高稀疏性、多标签分类困难等问题,为解决文本分类问题提供了新的思路。本文存在的主要不足是:这种基于语义标注的思路、流程虽然是通用的,但在具体实施中却需要是面向特定专业领域来编制词汇—语义规则。

未来仍有一些问题需要作进一步深入研究,比如:语义标注的先后顺序是否会影响分类效果?语义标注特征结合深度学习能否更加有效地改进文本分类效果?这些都将是笔者继续研究的动力。

参考文献:

[1] Salton G, Yang C S. On the specification of term values in automatic indexing[J]. Journal of Documentation, 1973, 29(4): 11-21.

[2] 张玉芳,万斌侯,熊忠阳. 文本分类中的特征降维方法研究[J]. 计算机应用研究,2012,29(7):2541-2543.

[3] Bruno T, Sasa M, Dzenana D. KNN with TF-IDF based framework for text categorization[C]//Proc of International Symposium on Intelligent Manufacturing and Automation. 2013: 1356-1364.

[4] Becker J, Kuropka D. Topic-based vector space model[C]//Proc of the 6th International Conference on Business Information Systems. 2003: 7-12.

[5] 刘小明,樊孝忠,李芳芳. 一种结合本体和焦点的问题分类方法[J]. 北京理工大学学报:自然科学版,2012,32(5):498-502.

[6] 张国栋,张化祥. 基于语义的文本特征加权分类算法[J]. 计算机应用研究,2012,29(12):4476-4478.

[7] Li Chenghua, Yang Jucheng, Park S C. Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet[J]. Expert Systems with Applications, 2012, 39(1): 765-772.

[8] Chen Y T, Chen Mengchang. Using Chi-square statistics to measure similarities for text categorization[J]. Expert Systems with Applications, 2011, 38(4): 3085-3090.

(下转第 2288 页)

表3 词语相似度对比

词语1	词语2	方法1	方法2	方法3
男人	苹果	0.171	0.313	0.303
男人	经理	0.630	0.530	0.579
男人	工作	0.164	0.148	0.120
男人	鲤鱼	0.208	0.357	0.209
跑	跳	0.444	0.762	0.866
发明	创造	0.615	0.849	0.955
学校	实验室	0.575	0.640	0.591
学校	图书馆	0.575	0.618	0.577
本科	必修课	1.000	0.785	0.501
考	考核	1.000	1.000	1.000

下面对词语相似度计算结果进行分析：

a)从整体上看，方法2与3的结果明显比方法1更符合实际情况，这是因为方法1仅仅考虑了词语中义原距离因素，结果比较粗糙；而方法2和3则在此基础上，不同程度地考虑了义原深度因素。

b)方法3与2比较，大部分数据有所降低，从词语相似度计算方法的角度分析，是因为方法3用概念集合的加权平均值代替了最大值，使得最终结果更加客观。

c)“跑”与“跳”、“发明”与“创造”两组词语在方法3得到的结果高于方法2，更加符合实际。这是因为这两组词语均不含相同的第一基本义原，实质上，方法3提出的加权平均值就等同于最大值，主要是改进的义原相似度算法导致出现这一结果。

d)“学校”与“实验室”、“学校”与“图书馆”两组词语在方法1中的计算结果相同，而方法2与3对这两组词语的相似度有不同程度的区分，均为前者的相似度较大。从义原相似度角度分析原因，是因为前者词语中的义原“学”与“研究”的最小层次深度要大于“教”与“借入”。

e)“本科”与“必修课”词语对在方法1中的相似度为1，这显然不符合客观实际。从词语相似度角度分析，是因为按照方法1得到词语中概念相似度的最大值为1，故词语的相似度也为1。而根据本文提出概念组合的加权平均值取代最大值的方法，得到词语的相似度值分别为0.501，显然后者比较符合客观实际。

f)“考”与“考核”词语对在三种方法中的相似度值均为1，这是由于这两个词语的概念完全相同，从知网的角度分析，实质上是同一个词语间的相似度计算，因此，三种方法的结果均为1。

因此，通过第一基本义原筛选出的词语概念组合，计算结果既不影响计算的精度，而且大大提高了运算速度，尤其对于那些在知网层次结构中解释概念较多和同词性概念较多的词语，这样的速度优势更明显。

4 结束语

为了解决现有词语语义相似度计算方法未考虑义原距离与义原深度的主次关系，通过距离约束最小层次深度因素，并且综合考虑两个义原的层次深度，改进义原相似度计算方法；另外，通过筛选出词语对应的第一基本义原相似度最高的概念组合，再引入动态加权因子，用组合间概念相似度值的加权平均值取代现有方法的最大值，以此提高词语相似度的准确性和客观性。尽管改进后的方法有较好的效果，但由于汉语词汇表达的复杂性、词汇语义概念较强的主观性、具体应用领域的专业性等因素影响，词汇相似度计算仍有很大的研究空间，这也是后续的研究方向。

参考文献：

[1] 葛斌，李芳芳，郭丝路，等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究，2010，27(9)：3329-3333.

[2] Lee L. Similarity based approaches to natural language processing [D]. Cambridge: Harvard University, 1997.

[3] Brown P. Word sense disambiguation using tactical methods[C]// Proc of the 29th Meeting of Association for Computational Linguistics. 1991.

[4] Floreano D, Monidada F. Evolutionary neuro-controller for autonomous mobile robots[J]. Neural Networks, 1998, 11(7/8): 1461-1478.

[5] 王斌. 汉英双语语料库自动对齐研究[D]. 北京：中国科学院计算技术研究所，1999.

[6] 刘群，李素建. 基于知网的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会论文集. 2002：59-76.

[7] 金玉，范学峰. 基于知网的中文DeepWeb模式匹配算法研究[J]. 计算机应用研究，2009，26(10)：3750-3753.

[8] 程传鹏，吴志刚. 一种基于知网的句子相似度计算方法[J]. 计算机工程与科学，2012，34(2)：172-175.

[9] 李峰，李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报，2007，21(3)：99-105.

[10] Lin Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]//Proc of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc.，1998：296-302.

[11] 王小林，王义. 改进的基于知网的词语相似度算法[J]. 计算机应用，2011，31(11)：3075-3077.

[12] 李湘东，曹环，丁丛，等. 利用知网和领域关键词集扩展方法的短文本分类研究[J]. 现代图书情报技术，2015，31(2)：31-37.

[13] 廖志芳，周国恩，李俊锋，等. 中文短文本语法语义相似度算法[J]. 湖南大学学报：自然科学版，2016，43(2)：135-140.

[14] 张沪寅，刘道波，温春艳. 基于知网的词语语义相似度改进算法研究[J]. 计算机工程，2015，41(2)：151-156.

[15] 王小林，王东. 基于知网的词语语义相似度算法[J]. 计算机工程，2014，40(12)：177-181.

[16] 王义，王小林. 基于改进的义原关联度算法的词语相关度计算[J]. 情报学报，2012，31(12)：1271-1275.

[17] 张亮，尹存燕. 基于语义树的中文词语相似度计算与分析[J]. 中文信息学报，2010，24(6)：23-30.

(上接第2284页)

[9] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning[C]//Proc of the 17th International Conference on Machine Learning. [S. l.]: Morgan Kaufmann, 2000: 359-366.

[10] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J]. 计算机学报，2011，34(5)：856-864.

[11] Qiu Xipeng, Zhang Qi, Huang Xuanjing. FudanNLP: a toolkit for Chinese natural language processing[C]//Proc of Meeting of the Association for Computational Linguistics: System Demonstrations. 2013: 49-54.

[12] Gooch P, Roudsari A. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes[J]. Journal of Biomedical Informatics, 2012, 45(5): 901-912.

[13] Hogenboom A, Hogenboom F, Frasincar F, *et al*. Semantics-based information extraction for detecting economic events[J]. Multimedia Tools and Applications, 2013, 64(1): 27-52.

[14] Cunningham H, Maynard D, Bontcheva K, *et al*. GATE: a framework and graphical development environment for robust NLP tools and applications[C]//Proc of the 40th Anniversary Meeting of the Association for Computational Linguistics. Straudsburg, PA: Association for Computational Linguistics, 2002: 168-175.

[15] Cunningham H, Maynard D, Tablan V. JAPE: a Java annotation patterns engine[EB/OL]. http://www.dcs.shef.ac.uk/~hamish.

[16] Gill A. Introduction to the theory of finite-state machines[M]. [S.l.]: McGraw-Hill, 1962.

[17] 罗明，黄海量. 一种基于有限状态机的中文地址标准化方法[J]. 计算机应用研究，2016，33(12)：3691-3695.

[18] Mellouli S, Bouslama F, Akande A. An ontology for representing financial headline news[J]. Web Semantics, 2010, 8(2): 203-208.