基于分词的关联规则预测系统研究

王志超 孙建斌 秦瑞丽
(航天长征化学工程股份有限公司 北京 101111)

摘 要 定向文本预测,往往存在预测不准、数据量大、针对性不高等问题。提出基于分词的关联规则预测方法,以煤化工行业为例进行发展方向的预测。对预测主题近期新闻标题进行层叠隐马尔可夫模型的初步分词,对得到的词集进行虚词修剪及喻词实化完成语义统一形成参与关联规则的事务集;通过提出的基于分词的关联规则算法 Apriori_Split 对事务集进行计算,最终得到预测结果。实验表明,该预测方法简单有效,可以极大提高预测准确性。

关键词 分词 关联规则 预测模型 文本分词 马尔可夫模型

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.12.027

ASSOCIATION RULE PREDICTION SYSTEM BASED ON WORD SEGMENTATION

Wang Zhichao Sun Jianbin Qin Ruili
(*Changzheng Engineering Co., Ltd., Beijing 101111, China*)

Abstract There are many problems in directional text prediction, such as inaccurate prediction, large amount of data and low pertinence. In this paper, an association rule prediction method based on segmentation was proposed to predict the development direction of coal chemical industry. We used Cascade Hidden Markov model to preliminarily segment recent news headlines of predicted topic, and functional word pruning and figurative materialization on obtained word set were used to complete semantic unification to form transaction sets participating in association rules. The transaction set was computed by the proposed segmentation-based association rule algorithm Apriori_Split, and the predicted result was obtained. The experiment shows that the method is simple and effective. It can greatly improve the accuracy of prediction.

Keywords Word segmentation Association rule Prediction model Text segmentation Markov model

0 引 言

由于媒体的丰富和网络的快速传播,新闻已经成为大数据的一个重要组成部分。新闻包括传统报刊杂志新闻,如《人民日报》《化工报》等;也包括各门户网站和新闻网站的实时新闻,如新浪、百度等。在我国,最重要的新闻报道往往聚焦于政府政策动向的跟踪和及时传播,时效性影响重大。然而,重大政策的形成往往有一个过程,包括初期的调研论证、理论研究、政策试探、舆论反应、试点安排、政策修订、正式颁布实施等阶段,每一个阶段都有大量新闻的跟踪和报道。对该类新闻进行深度挖掘和学习,利用关联规则研究其频繁项集,可以根据产生的频繁项集得到一段时间内关注任务的关联因子的变化,由此可以提前预测相关政策和动向的变化。

利用新闻进行预测,是国内外学者对舆情监控的研究热点之一。唐晓波等[1]提出在互联网新闻文本信息挖掘中,融合新闻热度和读者态度建立高频情感词典,在新闻文本预测分析中对预测结果利用情感频度加权排序,可以获取更好的准确性。然而该方法在新闻推荐等领域可用,在缺乏“情感频度”或不宜收集“情感频度”方面效果不明显。庞有明等[2]在研究信用债估值时引入新闻舆情语料,并重点关注舆情的情绪变化,然而该方法对于实际的应用效果不太明显。Patel 等[3]在股票市场走向预测分析中,引入新闻舆情

收稿日期:2018-01-29。**王志超**,工程师,主研领域:信息化建设,数据挖掘,信息应用。**孙建斌**,高工。**秦瑞丽**,工程师。

监控,并利用分类、还原和统计技术进行研究,用于指导投资。Xu 等[4] 利用极端机器学习和灰度 Verhulst 模型理论在热点新闻点击率预测上应用有一定效果。然而,对于新闻的理解,分词技术的应用是基础。张洪刚等[5] 在分词方法中利用双向长短时记忆模型,但该方法较为复杂。李雪莲等[6] 提出基于门循环单元神经网络的中文分词法,试图解决长短时记忆模型的复杂性。

本文提出利用基于隐层马尔可夫模型的中文分词方法[7,9] 并以报纸及刊物新闻作为数据源,通过对新闻数据预处理及关联规则挖掘[10-11],进行行业政策和发展方向预测研究,并以"煤化工"行业为例。首先,选取新闻标题作为本文预测任务的原始数据集,对新闻标题进行分词;其次,将分词所得的每组新闻标题词集进行语义统一,作为关联规则研究的项集;最后,利用 Apriori 算法对事务集进行关联规则挖掘,得到以"煤化工"等为关键字满足最小支持度和最小置信度的关联规则,并以此作为煤化工政策和发展方向的预测依据。实验证明,基于新闻分词的关联规则挖掘对政策和方向预测具有很好的作用。

1　新闻标题分词技术研究

分词是指将完整的一句话根据其语义分剪成一个词语项集,该词语项集作为参与关联规则挖掘的基本单元[5-6]。语义分词分两步:

(1) 基本分词　对新闻标题做初始分词,如 2014 年 8 月 22 日《中国化工报》行业时评刊文标题《传统煤化工要有"世界级"理想》,进行初步分词后其结果为:

"传统 /j 煤化工 /n 要 /v 有 /v " /w 世界级 /b " /w 理想 /n"。

其中,各词后面所标注"/"为词性,根据英文文法词性标注。

(2) 词语修剪及语义统一　将基本分词所得词集中无实意的虚词及一字动词等剪掉,只保留部分实词,包括动词、名词、量词、代词等,无意义词通过词性标注即可识别,如标注为"/w"即为标点符号,而一字动词则通过词性和词长识别。词语实化即对初始分词集合进行语义统一,如代词(词性为"/p")变为实词、比喻中的喻意词(词性为"/m")变为本意词等,该例中词语修剪后的词集不存在代词等,所以保留修剪后词集结果不变。该步结果为:

"传统 /j 煤化工 /n 世界级 /b 理想 /n"。

语义统一是将代词实化、喻词本意化,同时也是建立关联规则类的一个关键步骤。

1.1　基于 ICTCLS 的基本分词

ICTCLS 是中科院计算所研发的汉语分词系统,采用了层叠隐马尔可夫模型以完整统一理论框架进行分词。本文以此为基础做新闻标题的初始分词。

定义 1　(分词句集)　设 $S=\{s_1,s_2,\cdots,s_i,\cdots,s_n\}$,其中,$s_i$ 为第 i 个原始句串,$1<i<n$。本文中,数据输入即为定义 1 的分词句集 S。

对分词句集 S 中第 k 个原始句串 s_k 进行基于层叠隐马尔可夫模型的 ICTCLS 分词,$s_k\in S$。首先对 s_k 进行原子切分,即将原始句串标记开始结束,并将各字单独为词;其次对原子切分序列进行 N-最短路径粗切分,找到相邻单字组词后序列联合概率最优 N 结果;对最优 N 结果进行人名地名识别;最后对识别后的 N 结果优化并标注类及词性,输入分词结果,表示为 M_k,$M_k=\{m_{k1},m_{k2},\cdots,m_{kj}\}$,其中 $m_{kh}(h=1,2,\cdots,j)$ 为原始句串 s_k 分词结果集中的第 h 个词语。M_k 作为中间项集进行分词修剪及实化,而分词句集产生的每一个 M_k 组成了预事务集。

定义 2　(预事务集)　设 $M=\{M_1,M_2,\cdots,M_n\}$,其中 $M_k(k=1,2,\cdots,n)$ 为分词句集第 k 个句子的分词结果集,$M_k=\{m_{k1},m_{k2},\cdots,m_{kj}\}$,其中 $m_{kh}(h=1,2,\cdots,j)$ 为分词结果集中的第 h 个词语。M 作为 ICTCLS 分词结果的预事务集,进行下一步的词语修剪及语义统一。

1.2　分词修剪及语义统一

ICTCLS 算法分词后,得到预事务集 M 作为本节进行分词修剪和语义统一的对象。由于原始新闻标题分词后存在无实意词,如虚词"的、地",单字动词"有、做"等。分词修剪及语义统一的目的即为生成适宜关联规则算法处理的数据集,将无实意词去掉后的数据集大小少于处理前,使得算法处理的干扰减小且计算速度更快,而语义统一即是将预事务集标准化,得到更为准确的事务集,也使得关联规则计算更加精确。

定义 3　(项集)　定义 2 所得 M_k 经分词修剪和语义统一后的词语集合即为项集,用 I_k 表示。$I_k=\{i_{k1},i_{k2},\cdots,i_{kj}\}$,其中 $i_{kh}(h=1,2,\cdots,j)$ 为二步分词所得词语,是参与关联规则的元数据。

定义 4　(事务集)　设 $D=\{I_1,I_2,\cdots,I_n\}$,其中 $I_k(k=1,2,\cdots,n)$ 为项集,则 D 为参与关联规则挖掘的事务集。

具有修剪及语义统一分词算法 ICTCLS_TRIM 算法描述如下:

```
BEGIN
INPUT S                                    //S 为分词句集
```

```
FOR k = 1 to n
  M_k = ICTCLS(S_k)
            //对每一项句集应用 ICTCLS 做初始分词
  I_k = Reduce&Unify(M_k)
            //对预处理项集进行分词修剪及语义统一
ENDFOR
Split(D, Array(Keywords))
            //关键字修剪,将非目标项集过滤掉
OUTPUT D                    //D 为任务相关事务集
END
```

由于本文针对特定行业特定方向的新闻分词关联规则挖掘,所以在上述算法中,利用 Split(D, Array(Keywords))将非含关键字和关键义的项集修剪掉,使得关联规则挖掘数据集更加精确,事务集 D 作为关联规则挖掘的数据录入。

2 分词关联规则 Apriori_关联规则算法

本文采用改进的 Apriori 算法对形成的事务集进行关联规则分析,Apriori 算法是由 Rakesh Agrawal 和 Ramakrishnan Srikant 两位博士在 1994 年提出的关联规则挖掘算法[12]。该算法主要用于对频繁项集的递归挖掘,在所有满足最小支持度的频集中,发现满足最小可信度的强关联规则。

定义 5 (支持度) 即某项集 X 在事务集 D 中出现的概率,用 $Supp(X)$ 表示,如下所示:

$$Supp(X) = Occor(X)/Count(D)$$

最小支持度即为满足最小 $Supp(X)$ 的项集,当给定最小支持度 $Supp(CONST)$ 时,如果存在 $Supp(X) > Supp(CONST)$,则称 X 为频繁项集。

定义 6 (置信度) 即在频繁项集 X 出现的条件下,频繁项集 Y 也出现的条件概率,表示为 $Conf(X \rightarrow Y) = Supp(X \cup Y)/Supp(X)$。

新闻分词关联规则算法,基于分词的关联规则算法 Apriori_Split 描述如下:

```
BEGIN
INPUT S                     //S 为原始新闻标题
D = ICTCLS_TRIM(S)
            //通过具有语义修剪的 ICTCLS 进行分词
L_1 = Large_Supp(D,Supp_THRESHOLD)
            //选取满足最小支持度的 1_项集
FOR k = 2 to n
  C_k = apriori-gen(L_k-1)
  FOR d_i ∈ D do
    C_i = subset(C_k,d_i);      //事务 d_i 中包含的候选集
  for C_i ∈ C_t do
    C_i. count ++
  ENDFOR
  L_k = {C_i ∈ C_k | C_i. count3minsup}
ENDFOR
```

算法 Apriori_Split 中,首先利用具有语义修剪的新闻分词算法 ICTCLS_TRIM 将原始新闻集变成适于关联规则的事务集。通过计算支持度和置信度产生频繁 1-项集 L_1,对各 1-项集进行关联规则的计算。在第 k 次循环中,过程先产生候选 k-项集的集合 C_k,C_k 中的每一个项集是对两个只有一个项不同的属于 L_{k-1} 的频集做一个 $(k-2)$-连接来产生的。C_k 中的项集是用来产生频集的候选集,最后得到频集 L_k,而 L_k 也必然存在 $L_k \in C_k$。算法经过两次循环,其算法复杂度为 $O(n^2)$。

3 实 验

本文基于新闻分词的关联规则,实验数据集选取 2014 年 7 月 31 日至 2014 年 9 月 3 日期间,包括《中国化工报》、《山西日报》、《中国煤炭报》、《山西经济日报》、《昌吉日报》、《中国国土资源报》等在内的众多报刊中标题、摘要、正文中存在“煤化工”关键字的前 100 项新闻标题为本次实验数据集。

本实验中,为提高计算速度,为多关键字进行编号并处理,如本文实验数据集:

{煤,煤化工,煤科,煤层气,粉煤,…},即编号为{1,2,3,4,…},则试验中 ID 为 1 的关键词即为“煤”,而如果某一新闻分词组中出现“煤”的次数为 2 则数据标记即为{1 2},该表示方式{*ID Times*},*ID* 为数据集编号,*Times* 即为出现次数,如此将实验数据集进行处理。

发展方向类 = {甲醇,煤制气,煤气化,煤油气,热变换,煤电,聚丙烯,…}

关联规则的类的设定属于半监督,该处基于专家知识形成,即分类越科学,规则生成越准确。由此,根据本文所定规则及数据集进行试验,前五个规则结果如表 1 所示。

表 1 预测实验结果表前五项结果

序号	规则	置信度
1	(煤,清洁高效)	0.81
2	(煤化工,煤气化)	0.75
3	(煤化工,煤制油)	0.66
4	(煤,烯烃)	0.61
5	(粉煤,航天炉)	0.56

结果分析:本文试验中以置信率大于 0.5 进行结果的筛选,并展示了前五项试验结果。其中置信率最高的为(煤,清洁高效),这也反映了当前环保的趋势,(煤化工,煤气化、煤制油)反映了煤化工产业的工艺方向,而(媒,烯烃)则反映的是当前中国煤化工的产品结果,(粉煤,航天炉)则是粉煤应用较多的技术标准。实验表明,本文方法给出的预测方向同当前的方向是匹配的,极大地提高了预测的准确性。

4 结　语

本文提出基于分词的关联规则预测方法,首先对待预测方向近期新闻标题进行层叠隐马尔可夫模型的初步分词,对得到的词集进行虚词修剪及喻词实化等语义统一,该步骤得到的事务集通过分词修剪和语义实化后更加精确,降低了无义词的干扰。最后通过本文提出的基于分词的关联规则算法 Apriori_Split 对事务集进行计算,得到预测方向的规则,并以此为依据形成对未来发展的预测。该方法由于对参与关联规则的事务集的精确处理,有效提升了关联规则预测的准确性。

参 考 文 献

[1] 唐晓波, 叶晨孟. 一种融合新闻热度和读者态度的情感分析方法[J]. 图书馆学研究, 2017(10):81-90.

[2] 庞有明, 蒋洪迅. 基于新闻舆情的信用债估值修正模型及其应用[J]. 山西大学学报(自然科学版), 2017, 40(1):1-13.

[3] Patel H R, Parikh S M, Darji D N. Prediction model for stock market using news based different Classification, Regression and Statistical Techniques: (PMSMN)[C]//International Conference on ICT in Business Industry & Government. IEEE, 2017:1-5.

[4] Xu J, Feng J, Sun X, et al. Hot News Click Rate Prediction Based on Extreme Learning Machine and Grey Verhulst Model[C]//Proceedings of ELM-2016. Springer,2016: 89-97.

[5] 张洪刚, 李焕. 基于双向长短时记忆模型的中文分词方法[J]. 华南理工大学学报 (自然科学版), 2017, 45(3):61-67.

[6] 李雪莲, 段鸿, 许牧,等. 基于门循环单元神经网络的中文分词法[J]. 厦门大学学报(自然版), 2017(2):237-243.

[7] 王坤, 刘鹤飞, 蒋成飞. 隐马尔可夫结构方程模型及其贝叶斯估计[J]. 数理统计与管理, 2018(2):272-279.

[8] 张和平,陈齐海.基于灰色马尔可夫模型的网络舆情预测研究[J].情报科学,2018,36(1):75-79.

[9] Zhang H P, Liu Q, Cheng X Q, et al. Chinese lexical analysis using hierarchical hidden Markov model[C]//Sighan Workshop on Chinese Language Processing. Association for Computational Linguistics, 2003:63-70.

[10] 刘军煜, 贾修一. 一种利用关联规则挖掘的多标记分类算法[J]. 软件学报, 2017, 28(11):2865-2878.

[11] 张海涛, 汪佩佩, 张波波,等. 快照查询匿名集关联规则的概率化挖掘方法[J]. 南京邮电大学学报(自然科学版), 2017, 37(5):68-73.

[12] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases[C]//International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 1994:487-499.

(上接第 94 页)

[2] Zografos K G, Androutsopoulos K N. A heuristic algorithm for solving hazardous materials distribution problems[J]. European Journal of Operational Research, 2004, 152(2): 507-519.

[3] Current J, Ratick S. A model to assess risk, equity and efficiency in facility location and transportation of hazardous materials[J]. Location Science, 1995, 3(3):187-201.

[4] Cappanera P, Gallo G, Maffioli F. Discrete facility location and routing of obnoxious activities[J]. Discrete Applied Mathematics, 2003, 133(1):3-28.

[5] 帅斌, 钢铁. 基于时变网络下危险品危害减灾系统 LRP 研究[J]. 交通运输工程与信息学报, 2011, 9(3): 1-4.

[6] 乔联宝, 朱华桂. 危险品配送中心选址及路线选择问题研究[J]. 物流科技, 2013, 36(4):37-42.

[7] Xie Y, Lu W, Wang W, et al. A multimodal location and routing model for hazardous materials transportation. [J]. Journal of Hazardous Materials, 2012, 227/228:135-141.

[8] Assadipour G, Ke G Y, Verma M. Planning and managing intermodal transportation of hazardous materials with capacity selection and congestion[J]. Transportation Research Part E, 2015, 76:45-57.

[9] 开研霞, 王海燕. 危险品运输网络中运输方式和路径优化研究[J]. 中国安全生产技术, 2009, 5(1): 37-41.

[10] 辛春林, 冯倩茹, 张建文. 危险品配送选址-多式联运路径优化[J]. 中国安全科学学报, 2016, 26(9): 73-78.

[11] 付晓凤, 杨丽萍. 危险品多式联运方案优化的探讨[J]. 危险品运输, 2016, 34(3): 54-57.

[12] Leonelli P, Bonvicini S, Spadoni G. Hazardous materials transportation: a risk-analysis-based routing methodology [J]. Journal of Hazardous Materials, 2000, 71(1):283-300.

[13] Gross D, Shortle J F, Thompson J M, et al. Fundamentals of queueing theory[M]. New York: Wiley-Interscience, 2013.