

基于重要事件的文本分类方法研究

仲兆满, 李存华

(淮海工学院 计算机工程学院, 江苏 连云港 222005)

摘 要: 在分析了文本中重要事件识别和文本分类方法的基础之上, 提出了一种基于重要事件的文本分类方法. 重点研究了该方法涉及到的两个关键技术: 以重要事件表示文本和获取文本类别的模板. 在中文事件语料 CEC 上, 使用本文介绍的文本分类方法得到的平均准确率达到 80%, 而使用传统的以词为特征的文本分类方法得到的平均准确率为 72%.

关键词: 文本分类; 文本表示; 重要事件; SVM

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2012)03-0006-04

A Method of Text Classification Based on Key Events

ZHONG Zhao-man, LI Cun-hua

(Department of Computer Science, Huaihai Institute of Technology, Lianyungang 222005, China)

Abstract: After analyzing the methods of identifying important events from texts and text classification, we present a method of text classification based on key events. Two key technologies are mainly introduced: representing text using key events and getting text classification templates. We apply two methods to classify texts in Chinese Event Corpus (CEC). The average accuracy achieves 80% using the method proposed, and the average accuracy is 72% by the traditional method taking words as text features.

Key words: text classification; text representation; key event; SVM

1 引言

事件与概念相比, 包含事件触发词、时间、地点、人物等要素, 被认为是更大的语义单元. 从事件的角度研究文本信息的处理已经引起一些学者的关注, 相关的领域主要包括事件表示、事件抽取、事件关系识别、基于事件的文本相似度计算、基于事件的自动文摘、面向事件的信息检索以及事件本体等^[1-6].

本文从重要事件的角度研究文本的分类方法, 在提取文本事件的基础上, 识别文本中的重要事件, 然后以重要事件为文本的特征, 在计算事件要素的权重后向量化表示文本, 最后通过计算文本的重要事件的向量与文本类别模板的相似度, 决定某篇文本是否属于某个文本类别.

2 相关工作

2.1 重要事件的识别

从文本中识别重要事件是文本信息处理的一项基础工作. 基于文本中的重要事件, 可以进行问题的自动回答、选取重要事件生成文本摘要、评价构建的领域本体是否合理等.

从文本中识别重要事件的研究主要出现在自动文摘的相关文献中. 文献[7-8]认为如果事件包含了相同的要素, 就认为事件之间是有关系的, 在构建事件关系图的基础上计算事件的重要度. 周文^[9]提出了基于事件和形式概念分析(FCA)的自动文摘方法, 借鉴 FCA 用类似于描述概念之间关系的方法描述事件之间的关系. 文献[10]在分析了上述方法存在问题的基础上, 给出了一种利用事件影响关系识

别文本集合中重要事件的方法(记作 HARank),该方法利用了事件之间特有的时间变迁关系,综合考虑了事件节点的出度和入度迭代计算事件的重要度。

2.2 文本分类

文本分类的任务是基于内容将自然语言文本自动分配给预定义的类别,文本分类既是一种文本挖掘任务,也是对文本进行深层次挖掘的预处理步骤^[11]。文本分类属于有指导的机器学习,是根据给定文本的内容,将其判别为事先确定的若干个文本类别中的某一类或某几类的过程。

进行文本分类时首先要考虑的是文本的表示模型。常用的文本表示模型是向量空间模型 VSM,它简便有效,且在实践中取得了很好的效果。机器学习是获取文本集合特征模板的主要手段。由于向量空间模型和机器学习算法在文本分类领域中进行了紧密的结合和成功的运用,使得基于向量空间模型的文本表示方法迅速成为文本分类研究领域中文本表示的主流方法。

近年来提出了一些适合于不同应用的分类算法,如基于向量的距离法、贝叶斯分类算法、支持向量机分类算法、决策树分类算法和神经网络分类算法等,还出现了基于粗糙集理论的利用分明矩阵的文本分类算法和一些结合多种方法的混合分类方法^[12-14]。

3 基于重要事件的文本分类方法

3.1 方法的流程

基于重要事件的文本分类方法的流程如图 1 所示。

图 1 所示的文本分类流程包括提取文本中的事件、计算文本中的重要事件、机器学习、类别模板获取等部件,其中的核心是:(1)以重要事件表示文本;(2)获取文本类别的模板。下面分别介绍这两个核心技术的实现方法。

3.2 以重要事件表示文本

目前,自动提取文本中事件要素的效果还是非常不理想的。赵妍妍等^[15]将不同类别的事件的要素作为一个类别,在 ACE 的中文语料上得到的 F-measure 为 64.64%。Tan^[16]等使用多层次模式识别事件要素,在 ACE 的中文语料上得到的 F-measure 为 48.86%。

本文中的事件语料采用的是上海大学语义智能实验室制作的中文事件语料库(CEC)。CEC 与国际

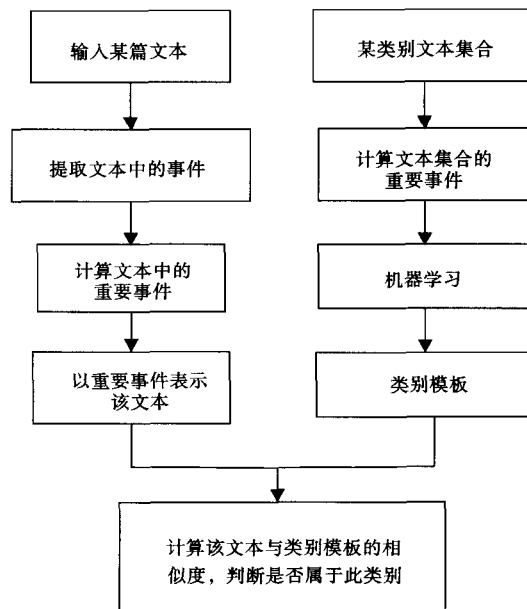


图1 基于重要事件的文本分类方法的流程

上有影响的语料库 ACE 和 TimeBank 相比,CEC 语料规模偏小一些,但对文本中的事件要素和事件关系的标注最为全面,ACE 仅标注了语料中的 8 大类 33 小类事件,TimeBank 重点围绕动词标注事件,但很多情况下动词并不一定代表了事件。有关该语料涉及到的类别、文本篇数等内容请参考本文的第 4 部分。

从 CEC 语料库中按照事件要素的 XML 标签提取出事件后,单篇文本中重要事件的识别方法采用文献^[17]介绍的在构建事件关系图的基础上使用 PageRank 算法计算事件的重要度。在识别出文本中的重要事件后,再提取重要事件包含的要素作为文本的特征,文本最终的表示采用 VSM 模型。事件各个要素的权值取依据文献^[17]中计算所得事件的重要度。如果同一要素因出现在不同的事件中而有不同的权重,其权值取最大的权重。

3.3 获取文本类别的模板

每个类别都有相应的事件语料,按照 CEC 约定的 XML 标签格式提取所有的事件,文本集合中重要事件的计算采用文献^[10]介绍的利用事件影响关系识别文本集合重要事件的方法。文本集合中重要事件要素的权值计算类似于 3.2 中介绍的单篇文本中重要事件要素的权值的取值方法。

基于数据的机器学习是智能技术中的重要方面,现有的机器学习方法共同的理论基础之一是统计学。支持向量机 SVM 是根据统计学理论提出的一种基于结构风险最小化原理,其特点是具有出色

的学习性能,只需要较少的样本就可以迅速训练出具有相对较高性能指标的分类器. SVM 正在成为继神经网络研究之后新的研究热点,并将有力地推动机器学习理论和技术的发展.

本文中文本类别模板的获取采用 SVM 机器学习的方法. 基于 SVM 的文本分类模型如图 2 所示.

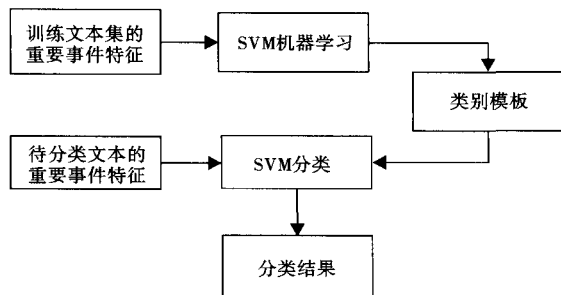


图 2 基于 SVM 的文本分类模型

图 2 所示的文本分类模型中主要包括 SVM 学习和 SVM 分类两个部分. 本文中使用的 SVM 分类器工具包是 LibSVM, LibSVM 是台湾大学林智仁等开发设计的一个简单、易用和快速有效的 SVM 模式识别与回归的软件包. 另外, Matlab 中也有成熟的 SVM 工具箱可用, 其中运行效率较高的是 SVM OSU 系列的工具箱.

4 实验及分析

4.1 实验语料

中文事件语料库(CEC)围绕突发事件从互联网上收集文本语料. 针对文本语料中所有提及的事件、事件要素、事件关系进行标注, 而不仅仅局限于标注某几类事件. 而且, 该语料的制作方法是建立在中文句法分析和语义分析基础之上的, 符合中文的特点. 评测实验表明, 采用该方法标注得到的语料可以取得较高的一致性. CEC 语料基本情况如表 1 所示.

表 1 语料基本情况介绍

类型	篇数	事件数	事件要素数
地震	45	682	1026
交通事故	54	798	1201
恐怖袭击	30	456	741
食物中毒	45	700	1127
火灾	31	496	783
总数	200	3133	4878

在 XML 标签的基础上对事件原始语料库进行标注, 形成了事件、事件要素和事件关系标注分明的中文事件语料库 CEC. CEC 与国际上有影响的语料

库 ACE(中文 633 篇/英文 599 篇)和 TimeBank(英文 300 篇)相比, CEC 语料规模虽然偏小, 但对文本中的事件、事件要素和事件关系的标注最为全面, ACE 仅标注了语料中的 8 大类 33 小类事件, TimeBank 重点围绕动词标注事件. 标注的事件对文本内容具有较高的覆盖率, 对已标注的 200 篇突发事件领域的新闻语料的统计分析表明, 标注的事件覆盖了文本中 89.1% 的相对比较重要的句子.

4.2 实验结果

针对每个类别, 取一半的文本作为训练语料, 通过 SVM 学习得到分类模板, 另一半文本作为正例集测试语料, 从其他类别随机抽取和正例测试集同样多的文本作为它的反例集测试语料. 为了考查本文所提方法的有效性, 用两种方法进行了实验对比: (1) 以词为文本的特征, 词的权重取 $TF * IDF$ 值, 该方法记作 M1; (2) 本文介绍的基于重要事件的文本分类方法, 记作 M2. 通过实验确定的文本集合中重要事件选取的比例是 50%, 单篇文本重要事件选取的比例是 60%. 实验结果如表 2、3 所示.

表 2 方法 M1 的文本分类结果

语料类别	训练文档数	测试文档数	正例集正 确分类数	反例集正 确分类数	准确率
地震	23	22	18	15	75%
交通事故	27	27	23	20	80%
恐怖袭击	15	15	10	8	60%
食物中毒	23	22	16	17	75%
火灾	16	15	10	11	70%
平均准确率					72%

表 3 方法 M2 的文本分类结果

语料类别	训练文档数	测试文档数	正例集正 确分类数	反例集正 确分类数	准确率
地震	23	22	20	16	82%
交通事故	27	27	26	19	83%
恐怖袭击	15	15	12	9	70%
食物中毒	23	22	20	19	89%
火灾	16	15	12	11	77%
平均准确率					80%

实验结果采用的评估指标是准确率, 准确率是所有测试的文本中与人工分类结果吻合的文本所占的比率.

从表 2 和表 3 可见, 基于重要事件的 SVM 文本分类方法 M2, 较之传统的以词为特征的 SVM 的

文本分类方法 M1, 准确率有明显提高. 方法 M2 的平均准确率为 80%, 方法 M1 平均准确率为 72%, 平均准确率提高了 8%. 另外, 无论对方法 M1, 还是方法 M2, “恐怖袭击”和“火灾”这两个类别的准确率都偏低, 主要原因是这两个类别标注的事件语料较少, 用于 SVM 分类器训练的语料不足.

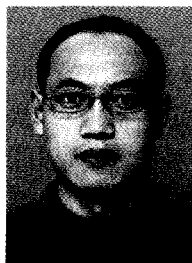
5 结束语

本文以文本中的重要事件为出发点研究文本的分类方法, 用待分类文本的重要事件为其向量特征, 用类别文本集的重要事件为机器学习的输入条件. 对本文所提方法的实现阶段, 采用 VSM 模型表示文本、采用 SVM 算法对 CEC 训练语料进行学习, 实验结果表明了所提方法的有效性. 为了获取更有说服力的结果, 一方面事件语料的规模需要扩展, 另一方面还需要采用多种机器学习的方法进行更进一步的实验比较, 这是下一步研究的重点.

参考文献:

- [1] Chang J. Event structure and argument linking in chinese [J]. Language and Linguistics, 2003, 4(2):317-351.
- [2] 郑家恒, 王兴义, 李飞. 信息抽取模式自动生成方法的研究[J]. 中文信息学报, 2004(18):48-54.
- [3] 于江德, 肖新峰, 樊孝忠. 基于隐马尔可夫模型的中文文本事件信息抽取[J]. 微电子学与计算机, 2007, 24(10):92-94.
- [4] Filatova F, Hatzivassiloglou V. Domain-independent detection, extraction, and labeling of atomic events [C] // Proceedings of RANLP. Borovetz, Bulgaria, 2003:145-152.
- [5] Zhou W, Liu Z. Event-based knowledge acquisition for ontology learning [C] // Proceedings of The 6th IEEE International Conference on Cognitive Information (IC-CI'07). Lake Tahoe, California, USA, 2007: 498-501.
- [6] 刘宗田, 黄美丽, 周文, 等. 面向事件的本体模型[J]. 计算机科学, 2009, 36(11):191-195.
- [7] Li W, Xu W, Wu M, et al. Extractive summarization using inter-and intra-event relevance [C] // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL. Australia: Sydney, 2006: 369-376.
- [8] Vanderwende L, Banko M, Menezes A. Event-centric summary generation [C] // Proceedings of the DUC-2004 Workshop. Boston, MA, 2004: 127-132.
- [9] 周文. 若干知识表示模型及其相关方法研究 [D]. 上海: 上海大学, 2007.
- [10] 仲兆满, 刘宗田. 利用事件影响关系识别文本集合中重要事件的方法 [J]. 模式识别与人工智能, 2010, 23(3):307-313.
- [11] 马金娜, 田大钢. 基于支持向量机的中文文本自动分类研究 [J]. 系统工程与电子技术, 2007, 29(3):475-478.
- [12] Yu G, Ostrouchov G, Geirst A, et al. An SVM-based algorithm for identification of photosynthesis-specific genome features [C] // Proceedings of the 2nd IEEE computer society bioinformatics conference. CA, USA, 2003:235-243.
- [13] Migual E R. Hierarchical neural networks or text categorization [J]. Machine Learning, 1997, 8(1):452-463.
- [14] Yiming Y. An evaluation of statistical approaches to text categorization [J]. Journal of Information Retrieval, 1999, 1(1/2):67-88.
- [15] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究 [J]. 中文信息学报, 2008, 22(1):3-8.
- [16] Tan H, Zhao T, Zheng J. Identification of chinese event and their argument roles [C] // Proceedings of the IEEE 8th International Conference on Computer and Information Technology Workshops. Washington, DC, USA, 2008:14-19.
- [17] Zhong Z, Liu Z. Ranking events based on event relation graph for a single document [J]. Information Technology Journal, 2010, 9(1):174-178.

作者简介:



仲兆满 男, (1977—), 博士, 副教授. 研究方向为事件本体、信息检索和人工智能应用.



李存华 男, (1963—), 教授. 研究方向为数据挖掘、模式识别与人工智能.