计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(8):10−18 [doi: 10.15888/j.cnki.csa.006456]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向 LDA 主题模型的文本分类研究进展与趋势①

赵　乐[1], 张兴旺[2]

[1](桂林理工大学 信息科学与工程学院, 桂林 541004)
[2](桂林理工大学 图书馆, 桂林 541004)
通讯作者: 赵　乐, E-mail: 1028701995@qq.com

摘　要: 文本分类是自然语言处理领域的一个重要研究方向. 综合分析发现, 文本分类的研究和分析, 有助于对信息进行有效的分类和管理, 并为自然语言处理的应用提供有力的支持. 然而, 已有的研究在理论和方法层面虽然已经取得了一定的成就, 但是文本分类研究涉及内容、领域和技术等多个方面, 各学科研究错综复杂, 因此还有很多缺陷和不足, 需要进一步进行系统和深入的研究. 本文针对文本分类这一研究内容, 探讨了文本分类和 LDA 主题模型的相关理论; 然后, 从技术、方法和应用三个方面分析了面向 LDA 主题模型的文本分类的研究现状, 总结了目前研究中存在的一些问题和研究策略; 最后, 归纳出文本分类未来的一些发展趋势.

关键词: 自然语言处理; 文本分类; LDA; 主题模型

引用格式: 赵乐,张兴旺.面向 LDA 主题模型的文本分类研究进展与趋势.计算机系统应用,2018,27(8):10–18. http://www.c-s-a.org.cn/1003-3254/6456.html

Research Progress and Trend of Text Classification for LDA Topic Model

ZHAO Le[1], ZHANG Xing-Wang[2]

[1](College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China)
[2](Library, Guilin University of Technology, Guilin 541004, China)

Abstract: Text classification is an important research direction in the field of natural language processing. It is found that the research and analysis of text classification can help to classify and manage the information effectively and provide strong support for the application of natural language processing. The existing research has made some achievements at the theoretical and methodological level. Nevertheless, the text classification research involves many aspects such as content, domain, and technology, while the research of each subject is complicated. Therefore, there are many defects and shortcomings, which need further systematic and in-depth research. In this paper, we discuss the related theories of text categorization and Latent Dirichlet Allocation (LDA) topic model for the research of text categorization. Then, we analyze the research status of text classification for LDA topic model from three aspects: technology, method, and application. Some problems and research strategies are presented as well. Finally, future trends of text categorization are summarized.

Key words: natural language processing; text classification; Latent Dirichlet Allocation (LDA); topic model

1　引言

随着互联网的发展和迅速普及, 面对着网络中呈爆炸式增长且杂乱无章的数据, 文本挖掘的工作就显得愈发重要, 人们希望能够从海量的信息文本中准确

① 基金项目: 国家社科基金青年项目 (17CTQ004)
Foundation item: Youth Project of National Social Science Fund (17CTQ004)
收稿时间: 2017-11-27; 修改时间: 2017-12-21; 采用时间: 2018-01-02; csa 在线出版时间: 2018-07-28

的获取想要的信息[1]. 那么, 如何有效的获取有价值的信息, 如何对浩如烟海的文本数据进行自动分类、组织和管理就变得愈发困难[2]. 因此, 面对这些问题和需求, 利用计算机进行智能信息处理便得到了广泛的研究. 文本自动分类技术作为自然语言处理领域的研究热点, 得到了快速发展和广泛应用.

文本自动分类技术作为文本数据挖掘的重要组成部分, 在信息抽取、信息检索、搜索引擎、个性化推荐等多个领域得到发展和应用, 是自然语言处理的热点和关键技术之一[3]. 其中, 文本分类在处理大规模数据时, 如何提高分类速度和准确性, 如何进行特征方法选择实现更好的降维操作, 是当前的重要研究方向. LDA 主题模型具有良好的降维性能, 因此把它作为特征模型, 再结合分类器设计能够达到很好的分类效果.

LDA 主题模型是符合文本生成规律的全概率生成模型, 具有很好的文本表示能力, 提取具有语义信息的主题. 为了解决传统意义上文本分类在语义相似性度量和文档主题分布问题的不足, 应用 LDA 主题模型方法[3]. LDA 主题模型的应用有助于降低特征向量空间维度, 有助于提高文本分类性能. 因此本文主要针对基于 LDA 主题模型的文本分类进行分析.

本文首先介绍了文本分类和 LDA 主题模型的相关理论; 其次, 从技术、方法和应用三个方面分析了面向 LDA 主题模型的文本分类的研究现状; 然后, 分析了目前研究中存在的一些问题和研究策略; 最后, 分析并讨论了文本分类未来的一些发展趋势.

2 研究现状分析

近些年来, 信息资源呈现指数增长, 大数据时代已经来临, 关于文本信息分类处理的研究和应用得到快速发展, 成为自然语言处理领域重要的研究方向.

对于文本分类的研究现状分析可从理论、技术和方法三个角度. 理论分析了当前国内外关于文本自动分类技术和 LDA 主题模型的发展概述; 相关技术对当前在文本分类中应用较为广泛的分类器做了简单介绍, 并指出不足之处; 最后是近几年一些研究者在传统方法的基础上进行改进而提出的方法.

2.1 理论分析

2.1.1 文本分类分析

文本分类 (text categorization), 是在预先划定好的文本类别集合中, 根据文本的主题内容, 把文本划分为不同类别的过程. 因为一个文本可能有一个或多个主题, 所以一个文本也就可能对应一个或多个类别. 一个文本分类系统不仅是一个自然语言处理系统, 也是一个典型的模式识别系统, 因此可以把一个文本分类系统看成是简单的输入输出问题, 系统输入的是文本, 输出是文本对应的类别, 如图 1 所示[4].

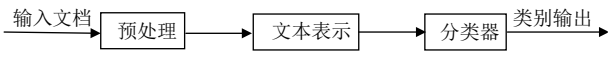

图 1 文本分类系统示意

国外关于文本分类技术的研究起步较早, 发展历程如表 1 所示[5], 当前已得到广泛发展, 应用于信息检索、数据挖掘、模式识别等多种领域. 由于在准确率和稳定性方面具有明显的优势, 基于统计机器学习的文本分类方法日益受到重视.

表 1 文本分类的发展

	第一阶段 (1975 年以前)	第二阶段 (1975–1990)	第三阶段 (1990—)
	早期探索阶段	知识工程阶段	机器学习阶段
	基于词匹配法	基于知识工程	基于统计学习
优点	直观	准确度较高	准确率较高; 稳定性较好
缺点	简单机械, 效果差	依赖于规则; 人力成本高; 完全不具备可推广性	训练集少; 缺乏语义层次的挖掘

在过去的几十年里, 国内外学者提出及改进了一系列经典的机器学习算法, 如朴素贝叶斯 (Naïve Bayes, NB)、支持向量机 (Support Vector Machine, SVM)、K-最近邻法 (K-Nearest Neighbors, KNN) 和神经网络 (Neural Networks, NNet) 等.

这些方法具有很好的可移植性, 将其成功应用于文本分类领域, 取得了良好的效果. 后来提出的 LDA 主题模型, 以及在此基础上改进的半监督和弱监督文本分类算法都取得了较好的分类效果, 文本分类技术也有了很大的进步.

而汉语不同于其他语言, 研究起来比较困难, 所以国内的研究借鉴了国外的一些研究成果, 是在侯汉清[6]

关于自动文本分类技术方面的概述性报告之后才逐渐兴起的. 之后, 一些专家学者开始热衷于文本分类技术的研究, 并提出了一些切实可行, 具有很好分类性能的方法.

2.1.2 LDA 主题模型概述

在 2003 年 Blei 等人在 LSA 和 pLSA 基础上提出了 LDA (Latent Dirichlet Allocation) 主题生成模型[7]. 该模型是全概率生成模型, 内部结构清晰, 即文档-主题-特征词三层结构, 可以利用高效的概率推断算法进行计算, 并且参数空间的规模与训练文本数量无关, 因此可以处理大规模语料. 它的基本思想是:语料库中的每个文本可以看成是若干潜在主题构成的一个概率分布, 每个主题是由若干个特定词汇组成的, 并且以一定的概率出现. 它解决了 LSA 的性能受损和计算复杂性的问题以及 pLSA 模型参数随着文档数量增加出现的过拟合问题, 因此得到了广泛应用.

2.2 相关技术分析

文本分类系统一般包括文本表示、特征选择、权重计算、分类器设计和性能评测等五大功能模块, 而系统中的关键问题就是文本表示和分类器设计.

2.2.1 文本表示

文本是有文字和符号组成的非结构化信息表示方式, 要使计算机能够高效的处理真实文本, 就必须找到一种理想的形式化表示方法, 把非结构化的文本转换为结构化的数学模型. 常用的文本表示模型有布尔逻辑模型、向量空间模型、概率模型等. 目前通常采用应用较多且效果较好的向量空间模型 (Vector Space Model, VSM); 另外, 由 Blei 等人[7]提出的 LDA 主题模型, 因其能够利用隐含主题表示文本, 不仅合理降低了特征词矩阵的维度, 还能保持元数据集的全面性, 不影响分类性能, 也备受人们关注.

1) 向量空间模型 (VSM)

VSM 是由 Salton 等人提出的, 最初用于 SMART 信息检索. VSM 模型将文档用向量$(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$表示, t_k 是特征项, 一个文档可以看成是它含有的所有的特征项的集合, w_k 是特征项的权重, 表示它们在文档中的重要程度. 把特征项看作是 n 维坐标系, 权重就是相应的坐标值, 那么一个文本就表示为 n 维空间的一个向量. 因此就将非结构化的文本信息转化到向量空间来表示.

2) LDA 主题模型

即潜在狄利克雷分布模型, 是一种文档主题生成模型, 也是一种包含词、主题和文档三层结构的三层贝叶斯概率模型. LDA 是一种非监督机器学习技术, 主要是针对离散数据集进行建模, 通过对语料库建模可以用来识别大规模文档集 (document collection) 或语料库 (corpus) 中潜在的主题信息. 它运用词袋 (bag of words) 将每一篇文档视为一个词频向量, 忽略了词与词之间的顺序和文档在语料库中的顺序, 这简化了问题的复杂性, 同时也为模型的改进提供了契机.

2.2.2 分类器设计

分类器实际上就是一个映射函数, 完成从需要映射的文本到预定义的类别集合的映射关系. 常用的分类方法有: 朴素的贝叶斯分类法 (naïve Bayesian classifier)、基于支持向量机 (Support Vector Machines, SVM) 的分类器、K-最近邻法 (K-Nearest Neighbor, KNN)、神经网络法 (Neural Network, NNet)、决策树 (decision tree) 分类法等.

(1) 朴素贝叶斯分类器

朴素贝叶斯分类器是基于贝叶斯定理与特征条件独立假设的分类方法[8], 是利用特征项和类别的联合概率来估计给定文档的类别概率的方法. 它假定词与词之间是独立的, 这在实际情况中很难保证, 因此当假设条件不满足时, 会严重影响分类的准确率和性能. 根据贝叶斯公式, 文档 Doc 属于 C_i 类的概率如公式 (1).

$$P(C_i|Doc)=\frac{P(Doc|C_i)*P(C_i)}{P(Doc)} \tag{1}$$

(2) 基于支持向量机的分类器

支持向量机在解决小样本、非线性及高纬模式识别中有许多优势, 基于支持向量机的分类方法主要用于解决二元模式分类问题, 基本思想是在向量空间中找到一个最优超平面, 即决策平面 (decision surface), 而这个平面能够很好的分割不同类别的数据点, 从而达到分类的目的. 但是在处理具体分类问题时无法选取正确有效的核函数是它的不足之处, 另外, 训练时间与数据集规模大小有关, 训练时计算量通常比较大, 这也会影响分类器的效率.

(3) K-最近邻法

K-最近邻法的基本思想是: 给定测试文档和数据类别, 系统在训练集中查找与目标文本相似度最高的 k 个文本, 并根据这些文本来给其候选类别评分. K-最

近邻算法简单直接, 但计算量大, 时间复杂度较高, 而且训练样本质量对分类器性能有着很大影响.

(4) 基于神经网络的分类器

神经网络 (NNet) 是目前比较成熟的技术之一, 是一种应用类似于大脑神经突触联接的结构进行分布式并行信息处理的数学模型. 其基本思想是: 给每一类文档建立一个神经网络, 输入单词或是特征向量, 通过机器学习获得从输入到分类的非线性映射. 神经网络分类效果比 KNN 和 SVM 较差, 而且时间复杂度比较大, 实际应用很少.

2.2.3 方法

文本分类技术兴起之后, 大批专家学者对此进行了研究, 提出了一些具有良好效果的分类方法. 如 Yang 等人[9]提出的基于聚类的决策树方法, 用于解决在线文本分类问题; Animashree 等人[10]在 LDA 的基础上利用统计中的三元或四元模型, 通过两个奇异值分解来训练文档中的主题模型, 进而实现对文本的分类. Chakraborti 等人[11]通过引入关键词, 提出了基于 LDA 和关键词的弱监督文本分类算法, 也取得了较好的效果.

虽然国内起步较晚, 但也取得了不少成果. 继侯汉清教授之后, 李荣陆等人[12]提出了基于最大熵模型的文本分类算法, 来构建分类器对文本进行分类; 尚文倩等人[13]提出了基于基尼指数的新的文本特征算法. 这些算法的提出在一定程度上提高了分类性能, 推动了文本分类的发展.

LDA 主题模型是一种可以挖掘大型文档数据集合中潜在主题信息和实现文本信息的分类的概率模型, 自从提出以来得到了广泛的应用, 并取得良好效果. 应用至今, 已有不少专家学者对其进一步的改进, 其分类效果得到进一步提升. 因此, 把 LDA 主题模型和其他方法相结合得到了广泛研究, 并取得不错效果.

基于 Labeled-LDA(附加类别标签的 LDA) 模型的文本分类[2], 将类别信息融入传统 LDA 模型, 进而支持文档在全部类别的隐含主题上的协同分配, 解决了传统 LDA 模型用于分类时强制分配隐含主题的缺陷; 基于 mRMR 和 LDA 主题模型的文本分类[14], 预先使用 mRMR 特征选择算法将输入空间映射到低维空间, 去除较大不相关信息和重叠信息, 使得 LDA 能够在更简洁的文本上建模, 从而得到更精确的主题分布; 基于词向量与 LDA 相融合的短文本分类方法[15], 能有效克服短文本的主题聚焦性差及特征稀疏性问题, 提高短文本分类性能. 基于 LDA 的微博生成模型 MRT-LDA[16], 利用微博之间的转发、对话、支持 (赞) 和评论等关系来计算微博之间的相关性, 综合考虑微博之间的相关性和同一用户微博信息间的关系, 来辅助对微博的主题进行挖掘. Fu 等人[17]针对开放类别文本 (文档类别未知) 进行分类, 提出了新的基于多重潜在狄利克雷分配模型的分类系统和方法, 聚类主题并提取关键字帮助分类注释, 最后应用到综合预测类别. Pavlinek 等人[18]提出基于主题模型表示的自训练半监督式文本分类方法, 有助于改进文本分类任务, 这在许多高级专家和智能系统中是必不可少的.

但是随着网络的发展, 文本数量庞大, 内容更为复杂, 因此上述方法中不可避免会存在一些问题. pLSA 模型[19]对文档中主题的混合权重没有做任何假设, 可能会出现过拟合的现象. sLDA 模型[20]为每篇文档关联一个代表着该文档类别标识的变量, 然后用 EM 算法进行最大似然估计, 但是该模型只能处理单一类别标识文档. Labeled-LDA 模型[2]在训练主题模型之前没有去处停用词, 没有考虑到词与其他各类别的关联问题, 并且使用此模型获得的主题分布倾向于高频词, 降低了主题的表达能力等; 使用最大熵模型进行中文文本分类的研究发现, 基于最大熵模型的分类器稳定性比 KNN 方法要差, 使用不同的训练文档测试结果相差较大, 另外实验规模有待扩大; MRT-LDA 模型[16]对于微博中的图片、表情等非文本信息利用不足, 微博信息挖掘有待提高.

2.3 应用现状分析

2.3.1 文本分类

文本分类 (text classification) 是利用计算机系统对文本按照预定义类别进行划分的技术. 文本分类问题的关键技术之一就是文本表示, 目前在文本分类应用中较为流行、分类效果较好的就是 VSM 向量空间模型和 LDA 主题模型. LDA 主题模型是一种无监督的全概率生成模型, 它本身不能直接判断文本类别, 因此它可以把文档表示为一系列潜在主题的概率分布, 然后选择一种合适的分类算法构造分类器.LDA 主题模型实现了对大规模文本数据的降维操作, 能够挖掘文本中潜藏的主题信息、分析语义信息. 传统的 LDA 主题模型在分类过程中可能会存在强制主题分配问题, 因此李文波等人[2]提出了 Labled-LDA 模型, 通过引入类别标签信息, 协同计算新文本在各类别隐含主题的

分配量, 从而克服了传统 LDA 主题模型的强制分配问题. 另外传统 LDA 主题模型没有考虑词顺序问题, 可能会造成词信息的损失, 因此田宝明等人[21]提出了一种基于随机森林的多视角文本分类方法, 利用改进的随机森林方法结合基于词的和基于 LDA 主题的两种文本表示方法, 有效的提高了文本分类性能. 吴建军等人[22]提出的基于互信息的特征项加权朴素贝叶斯算法, 部分消除了特征项独立性和特征项重要性相等假设, 提高了朴素贝叶斯算法的分类效果; 针对短文本信息, 刘泽锦等人[23]提出快速双词主题模型, 用于解决大规模短文本语料库主题模型参数大导致求解慢的问题.

2.3.2 文本聚类

文本聚类 (text clustering) 是依据相同类别的文档相似度较大, 不同类别的文档相似度较小的这一聚类假设提出的非监督的机器学习方法. 文本聚类因为不需要对文本进行训练和分类标注, 所以具有一定的灵活性和自动化处理能力, 应用广泛. 针对热点新闻, 对搜索引擎返回的结果, 对用户感兴趣的文档进行聚类处理, 并且文本聚类还可以用于改善文本分类结果. 对搜索引擎返回的结果进行聚类, 有助于用户快速浏览返回的信息, 找到满足自己需要的信息. 阮光册等人[24]将 LDA 主题模型和 k-means 算法相结合开展了基于主题模型的检索结果应用研究, 利用 LDA 模型实现文本潜在语义的识别, 用于帮助用户快速浏览系统返回的检索结果. 车蕾等人[25]融合新闻命名实体、新闻标题、新闻重要段落、文本语义等多特征影响, 提出基于多特征融合文本聚类的新闻话题发现模型, 并将三种相似度算法最优融合, 改进了用于新闻话题发现的 Single-Pass 算法, 有效提高了算法效率, 并且具有一定的自适应能力. 对于热点话题, 可以先进行聚类分析, 然后利用 LDA 进行建模, 把文档支持率作为话题热度用于区分热点话题和一般话题, 方小飞等人[26]依据这些方法提出了基于 LDA 模型的移动投诉文本热点话题识别等.

2.3.3 情感挖掘

情感挖掘也是文本分类的研究内容, 它是对民众关于社会中一些现象或是问题的态度、观点等的分析, 以此可以了解民众观点, 预测事件走向. 例如销售公司可以利用该技术了解用户对产品的评价、反馈等, 政府部门利用该技术可以分析民众对政府做出的决策或是管理办法的评论, 可以实时的了解大众的态度. 因此, 这需要情感分析作为支撑. 因为人在这过程中并不能完全客观的进行分析, 所以情感分析已经成为情感挖掘的基本技术. 此外, 该技术还涉及文本挖掘、观点挖掘等各方面问题. 对于网络中出现的短文本的情感挖掘, 以微博为代表, 黄发良等人[27,28]提出了基于社交关系的微博主题情感挖掘和基于多特征融合的微博主题情感挖掘, 这两种方法都用 LDA 主题模型进行建模, 更好的挖掘出用户性格情绪特征, 用于分析微博短文本主题情感特征, 把握用户情感动向. 基于在线评论文本, 王伟等人[29]构建较完整的情感词典, 依据情感单元搭配模式, 构建情感单元, 提出了基于 LDA 评论文本情感分类方法, 取得了较好的效果, 但缺乏对更复杂句子语境的讨论.

另外随着网络购物的发展, 用户对商品评价也越来越多, 要从这些评价信息中了解用户对产品的态度, 就要用到情感挖掘, 彭云等人[30]提出了一种基于语义关系约束的主题模型 SRC-LDA, 用于提取商品特征和从用户评价中挖掘出用户情感词, 网络购物平台可以以此来很好的改进自己的商品和服务. 黄章树等人[31]对某通信公司投诉文本进行实验, 提出了改进的卡方统计方法, 并将其运用到特征选择, 通过降低负相关低频词在特征选择算法中的权重, 减小其对模型的影响, 实验表明该方法能更准确的对业务投诉工单进行分类, 进而为通信公司后续改进服务提供数据支持.

2.3.4 个性化推荐

个性化推荐 (personalized recommender) 是根据用户的兴趣爱好或是购买特点, 推荐用户感兴趣的话题信息或是商品. 随着网络中信息和商品的大量增加, 用户在浏览信息或是选择商品时往往需要大量的时间和精力. 为了使用户更便捷的使用社交网络或是购物平台, 个性化推荐系统应运而生. 对于一段文本中可能涉及多个主题, 而 LDA 主题模型主要是挖掘文本中潜在主题, 得到广泛应用. 高明等人[32]基于 LDA 主题模型推断微博的主题分布和用户的兴趣去向, 提出了微博系统上用户感兴趣微博的实时推荐方法; 但是未考虑用户兴趣随时间的变化, 因此陈杰等人[33]提出了一种基于用户动态兴趣和社交网络的微博推荐方法; 对于文献推荐, 杜永萍等人[34]提出了一种基于主题效能的学术文献推荐算法, 利用 LDA 主题模型对候选文献和用户发表的文献进行建模, 挖掘出具有高效能的主题集合, 并根据主题分布计算与用户兴趣间的相似度, 最

后向用户推荐有价值的文献. 王日芬等人[35]通过全局和学科视角的对比来探究基于 LDA 主题模型的科学文献主题识别.

个性化推荐在网络购物平台上应用, 电商可以根据用户的浏览和购买记录推荐一些相关的产品, 省去了用户进行大量浏览的时间; 对于社交平台, 微博、论坛等, 可以向用户推荐一些当前的热点话题, 或是根据用户平时的浏览记录来推荐用户可能感兴趣的话题. 因此崔金栋等人[36]从演化发展角度对 LDA 运行机理进行解析, 分析研究了微博用户信息个性化推荐的主题模型 LDA 演化方向.

2.3.5 网络安全

随着网络的迅速发展和普及, 网络中信息量太过于庞大, 需要对网络中信息进行内容管理、监控和垃圾信息过滤. 这时的文本分类已不再是传统的客观分类了, 这需要分析文本内容的主观因素, 分析作者表达的目的意图, 因此应用到主观倾向性分类. 如何准确的把邮件进行很好的分类, 进而处理掉垃圾邮件是文本分类技术的又一应用热点. 张绍成等人[37]利用 LDA 主题模型对邮件内容进行主题提取, 实现邮件分类, 提出了代价敏感多主题学习的邮件过滤算法, 实现了垃圾邮件过滤. 廖晓锋等人[38]LDA 主题模型和 SVM 支持向量机结合的方法, 在主题向量空间构造一个漏洞分类器, 以国家信息安全漏洞库数据进行测试, 实验表明分类准确度比词汇向量构建的分类器有所提高.

对于网络安全方面, 一般用户的应用主要是过滤垃圾邮件. 对于企业, 公司或是军事领域不仅是要过滤掉垃圾信息, 更重要的是要防止病毒的入侵, 保障机密文件的安全.

3 存在问题和研究策略

通过对文本分类研究现状的分析, 可以发现, 对于文本分类的研究和分析, 有利于对网络中数量庞大的信息进行有效的管理和分类, 方便用户检索和浏览; 有利于分析文本情感倾向, 把握用户情感特征; 有利于分析数据安全特性, 过滤垃圾信息和监管不安全因素. 然而, 已有的研究在理论和方法层面虽然已经取得了一定的成就, 但是目前还存在一些不足, 还需进一步完善和提高.

文本分类存在问题和研究策略分析主要围绕理论体系和方法两个方面进行. 通过对已有的研究进行分析, 总结出文本分类目前存在的一些问题和相应的研究策略.

(1) 理论层面

自然语言处理涉及词法、语法、语义、和语用学等多个层次, 实际上关键问题就是歧义消解和未知语言现象的处理问题. 文本分类的理论研究在国外已经取得重大突破, 趋于完善, 但是我国中文文本分类涉及内容较多, 分类比较困难. 在汉语中, 存在同义词, 一词多义的问题, 而且一个词可有不同词性, 理解词义还需结合上下文语境, 因此给文本分类带来很大困难. 另外, 还存在一些数学模型不够奏效和算法复杂度过高等理论问题. 例如, 文本分类需要处理的数据一般是成千上万的稀疏矩阵, 矩阵维数过于巨大, 因此需要有效的降维操作; 文本的特征词中存在多义词、同义词现象, 还包含大量的噪音, 因此要形成有效的特征矢量; 文本分类在小量数据中应用较好, 但实际应用中数据量是非常巨大的, 因此需要研究大规模文本. 另外在知识资源方面也存在一些问题, 例如, 数据资源匮乏、覆盖率低、知识表示困难等.

近几年来, 中文文本分类研究发展迅速, 一大批专家学者进行了分析研究, 并且提出了很多切实可行的改善理论和方法. 基于统计机器学习的文本分类方法在准确率和稳定性方面具有明显优势, 日益受到重用. 目前文本表示、特征选择和分类方法众多, 性能评测指标也愈发成熟. 文本分类的应用也更加广阔, 深入到人们的日常生活, 例如社交网络评价, 舆情分析, 情感挖掘, 个性化推荐等.

(2) 方法层面

常用的文本表示方法词向量空间模型, 存在向量空间维度过高, 词项之间缺乏语义关系等问题. 因此有国外学者提出语义向量空间模型, 尝试利用潜在语义索引技术或本体的概念语义关系挖掘词项之间的语义关系, 构建低维的语义向量空间模型.

通过对面向 LDA 主题模型的文本分类研究进展与趋势的分析, 可以发现, 应用 LDA 主题模型于文本分类, 有利于处理大规模文本, 不仅合理地降低了特征词矩阵的维度, 还能保持原数据集的全面性, 不影响分类器性能, 解决了传统文本分类中相似性度量和主题单一性问题. 然而, 尽管 LDA 主题模型得到进一步改进和完善, 但还尚有一定缺陷和不足. LDA 是非监督学习模型, 不能直接用于文本分类, 因此必须嵌入到合适

的分类算法中. 传统的 LDA 主题模型存在分类过程中将文档强制在单个类别上分配隐含主题的缺陷; 并且由于实际情况中大规模的数据, 可能会出现主题范围过大, 不能对主题单词的潜在语义进行准确定位, 限制了模型的鲁棒性和有效性; 没有考虑词序问题, 是典型的词袋模型等.

另外在分类器设计方面, 朴素的贝叶斯分类法假定词与词之间是独立的, 这在实际情况中很难保证, 因此当假设条件不满足时, 会严重影响分类的准确率和性能. 基于支持向量机的分类器在处理具体分类问题时无法选取正确有效的核函数, 另外, 训练时间与数据集规模大小有关, 训练时计算量通常比较大, 这也会影响分类器的效率. k-最近邻法计算量大, 时间复杂度较高, 而且训练样本质量对分类器性能有着很大影响. 神经网络法分类效果比 kNN 和 SVM 较差, 而且时间复杂度比较大.

针对这些问题, 多种方法的融合、改进可以改善分类效果. 特征选择和特征重构是降维操作的关键技术, 二者融合有助于改善降维效果. 例如把互信息和聚类融合, 通过互信息最大化从原始特征空间中选择次优特征子集, 借助特征空间的聚类来剔除冗余特征, 从而实现特征空间的再次降维. 把多种分类算法相融合, 利用它们的优点, 剔除缺点, 从而可以改善分类性能. 例如 LDA 分别与卡方统计、互信息和信息增益进行结合, 利用改进后的特征提取方法提取特征词, 实验表明结合后的方法比原来的方法分类效果好; 另外随着特征词个数的增多, 每一种方法的分类性能也有提高.

4 发展趋势

根据目前国内外已有的研究成果和存在问题来看, 文本分类已经成为自然语言处理领域的研究热点和重点, 虽然在理论体系和技术层面还不够完善, 但其重要性已经逐步展现出来, 引起了研究者的重视. 基于此, 本文总结归纳出了文本分类未来的一些研究方向, 供读者参考.

(1) 文本分类在对话系统中的应用

人机对话系统有智能聊天、知识问答、任务执行和信息推荐等四个方面的内容. 当前的主要任务就是研究如何能够让对话系统更自然, 具备人一样的情感, 如何能够在场景化任务执行中做到高效的场景切换.

聊天机器人不仅要理解人类语言, 而且还要感知用户情绪变化, 分析用户情感特征, 实现和用户的交流. 通过对大规模聊天语料的标注, 训练和对上下文语境信息的分析, 从而进行分类, 得到对话模型, 计算机可以生成表达不同情绪类别的内容来与人进行对话. 如微软的小冰. 以后聊天机器人不仅要能够通过文字、语音、表情、动作等识别情感情绪信息, 还要进化到道德、精神层面的高级情感, 进行更深层次的自主学习.

对话系统中个性化推荐在很多领域都有广泛的应用, 如电商购物、社交网络、新闻资讯等. 在以后的发展中旨在提高推荐的精准度和更加个性化, 提高用户的满意度.

(2) 文本分类在人工智能知识服务体系中的应用

人工智能知识服务体系就是把分散于个人的知识技能集中起来, 实现知识共享, 把人工智能涉及的技术和领域知识组织起来, 让计算机能够像专家一样, 辅助决策, 成为综合知识集合, 结合人工智能的体系框架、技术方法, 以及涉及到的众多知识学科和应用领域, 将各种显性和隐性知识按照需求进行提炼, 从而解决用户需求的过程. 那么如何获取如此庞大的知识, 并且进行分析整合, 最后反馈给用户呢? 可以使用机器学习, 包括文本分析、自然语言理解、计算机视觉和数据挖掘等技术, 向用户智能推送. 这需要持续累积大量的训练样本和数据, 让机器学习系统不断地学习, 改善和进化.

在信息流的场景中, 人们可以更便捷的获得更多的标注数据和颗粒度更细的标注, 用于帮助自然语言理解和自然语言生成等. 语义化的进一步研究, 使得人工智能能够处理、分析、挖掘和理解信息流里的每一个环节, 可以利用这一技术进行知识的获取、分析和整合, 然后把内容反馈给用户. 以此让人工智能更多元, 更智慧的为人们服务, 例如帮助用户进行内容的创作, 帮助消费, 以及机器阅读等.

(3) 文本分类在文化遗产数字化与数字人文中的应用

对于种类庞杂, 信息总量庞大的文化遗传的采集, 可以把多源数据融合、自动纹理映射和影像建模等技术结合将大规模、高精度文化遗产数字化, 利用文本分类技术对信息进行分类、整理为不同类别, 建立档案库. 然后采用虚拟现实和数字动画技术, 建立虚拟的数字博物馆, 对文化遗产的现象、场景和过程进行复原或再现. 以此做到更好的保证文化遗产数字化档案

质量和客观性.

利用 VR(虚拟现实) 和 AR(增强现实) 技术对文化遗产进行保护, 实现人机交互. 例如, 可以通过 VR 技术进行对非物质文化遗传进行全方位的展现, 可以通过人机交互了解文化遗传的演变与发展等. 利用 AR 技术将现实文化遗产增加一层虚拟维度, 通过复原再现、展示传播等赋予文化遗产鲜活的生命, 具有很高的互动性和参与性.

(4) 文本分类在突发事件监测中的应用

我国每年突发事件频发, 交通事故、火灾等不计其数. 如何对这些突发事件进行监测, 并实施有效的救援, 这是一个难题. 现在网络技术发达, 其实可以把网络信息进行详细分类, 针对网络中出现的信息进行分析、挖掘, 过滤出敏感词汇, 如地震, 失火, 车辆相撞, 追尾等, 分析出可能发生的隐患事件和对已经发生的事件进行追踪, 从而实施有效的预防和救援措施, 保障人们的生命财产安全. 应用于公安系统可以预防犯罪发生和快速破案. 也可应用于军队, 对我国领海、领土、领空进行监测, 一旦发现事故发生或是外部入侵, 可以及时采取有效措施, 保障我国国民和领域安全.

(5) 文本分类在智慧医疗系统中的应用

我国人口众多, 排队看病是一个难题, 病人流量太大, 医院环境嘈杂, 可能会影响病人描述病情和医生进行更有效诊断. 因此, 将文本分类和信息抽取应用于医疗健康系统, 将用户输入的咨询信息进行分类和整理, 提取出用户的病症信息, 然后根据处理后的病症内容进行分类, 诊断出可能的病症名称, 然后推送给不同的科室医生进行在线回复, 还可以根据分析出的病情推荐合理的看病科室. 将文本分类应用于医疗健康后, 病人可以更方便的对自己的病情进行咨询和就诊, 医生也可以根据这些信息对病人病情进行更好、更快捷的诊断. 这不仅对病人、医生, 还是医院都提供了有利的条件, 因此可以在这方面进行更深一步的研究.

5 总结

文本分类是自然语言处理的热点研究内容之一. 文本分类的研究和分析, 有助于对网络中数量庞大的信息进行有效的管理和分类, 方便用户检索和浏览; 有助于分析文本情感倾向, 把握用户情感特征, 对于商家可以据此提高产品质量, 提升服务水平; 有助于分析数据安全特性, 过滤垃圾信息和监管不安全因素, 政府、高校、公司等可以据此来提高部门数据安全, 防止不利或是有害信息传播, 并为自然语言处理的应用提供有力的支持. 然而, 已有的研究在理论和方法层面虽然已经取得了一定的成就, 但是文本分类研究涉及内容、领域和技术等多个方面, 各学科研究错综复杂, 因此还有很多缺陷和不足, 需要进一步进行系统和深入的研究.

本文针对文本分类这一研究内容, 探讨了文本分类和 LDA 主题模型的相关理论; 然后, 从技术、方法和应用三个方面分析了面向 LDA 主题模型的文本分类的研究现状; 总结了目前研究中存在的一些问题和研究策略; 最后, 展望了文本分类未来的一些发展趋势.

文本分类的最终目的还是为自然语言处理服务, 因此, 可以将文本分类的研究成果应用到信息检索、信息抽取、舆情分析和个性化推荐、网络安全等研究中, 以期取得更好性能.

参考文献

1 张勇. 基于词性与 LDA 主题模型的文本分类技术研究. 合肥: 安徽大学, 2016.

2 李文波, 孙乐, 张大鲲. 基于 Labeled-LDA 模型的文本分类新算法. 计算机学报, 2008, 31(4): 620–627.

3 李锋刚, 梁钰, GAO XZ, et al. 基于 LDA-wSVM 模型的文本分类研究. 计算机应用研究, 2015, 32(1): 21–25.

4 宗成庆. 统计自然语言处理. 2 版. 北京: 清华大学出版社, 2013.

5 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述. 图书情报工作, 2013, 57(14): 134–139. [doi: 10.7536/j.issn.0252-3116.2013.14.022]

6 侯汉清. 分类法的发展趋势简论. 情报科学, 1981, 2(1): 58–63, 30.

7 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3(1): 993–1022.

8 李航. 统计学习方法. 北京: 清华大学出版社, 2012.

9 Yang YM, Pierce T, Carbonell J. A study of retrospective and on-line event detection. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia. 1998. 28–36.

10 Anandkumar A, Foster DP, Hsu D, *et al*. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. arXiv: 1204.6703, 2012.

11 Hingmire S, Chakraborti S. Sprinkling topics for weakly

supervised text classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, MD, USA. 2014. 55–60.
12 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类. 计算机研究与发展, 2005, 42(1): 94–101.
13 尚文倩, 黄厚宽, 刘玉玲, 等. 文本分类中基于基尼指数的特征选择算法研究. 计算机研究与发展, 2006, 43(10): 1688–1694.
14 史庆伟, 从世源. 基于 mRMR 和 LDA 主题模型的文本分类研究. 计算机工程与应用, 2016, 52(5): 127–133.
15 张群, 王红军, 王伦文. 词向量与 LDA 相融合的短文本分类方法. 现代图书情报技术, 2016, (12): 27–35. [doi: 10.11925/infotech.1003-3513.2016.12.04]
16 庞雄文, 万本帅, 王盼. 基于 MRT-LDA 模型的微博文本分类. 计算机科学, 2017, 44(8): 236–241, 259. [doi: 10.11896/j.issn.1002-137X.2017.08.040]
17 Fu RJ, Qin B, Liu T. Open-categorical text classification based on multi-LDA models. Soft Computing, 2015, 19(1): 29–38. [doi: 10.1007/s00500-014-1374-x]
18 Pavlinek M, Podgorelec V. Text classification method based on self-training and LDA topic models. Expert Systems with Applications, 2017, 80: 83–93. [doi: 10.1016/j.eswa.2017.03.020]
19 Hofmann T. Probabilistic latent semantic analysis. Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden. 1999. 289–296.
20 Blei DM, McAuliffe JD. Supervised topic models. Advances in Neural Information Processing Systems, 2007, 3: 327–332.
21 田宝明, 戴新宇, 陈家骏. 一种基于随机森林的多视角文本分类方法. 中文信息学报, 2009, 23(4): 48–54.
22 武建军, 李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法. 计算机系统应用, 2017, 26(7): 178–182.
23 刘泽锦, 王洁. 同主题词短文本分类算法中 BTM 的应用与改进. 计算机系统应用, 2017, 26(11): 213–219.
24 阮光册, 夏磊. 基于主题模型的检索结果聚类应用研究. 情报杂志, 2017, 36(3): 179–184.
25 车蕾, 杨小平. 多特征融合文本聚类的新闻话题发现模型. 国防科技大学学报, 2017, 39(3): 85–90. [doi: 10.11887/j.cn.201703014]
26 方小飞, 黄孝喜, 王荣波, 等. 基于 LDA 模型的移动投诉文本热点话题识别. 数据分析与知识发现, 2017, 1(2): 19–27. [doi: 10.11925/infotech.2096-3467.2017.02.03]
27 黄发良, 于戈, 张继连, 等. 基于社交关系的微博主题情感挖掘. 软件学报, 2017, 28(3): 694–707. [doi: 10.13328/j.cnki.jos.005157]
28 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘. 计算机学报, 2017, 40(4): 872–888.
29 王伟, 周咏梅, 阳爱民, 等. 一种基于 LDA 主题模型的评论文本情感分类方法. 数据采集与处理, 2017, 32(3): 629–635.
30 彭云, 万常选, 江腾蛟, 等. 基于语义约束 LDA 的商品特征和情感词提取. 软件学报, 2017, 28(3): 676–693. [doi: 10.13328/j.cnki.jos.005154]
31 黄章树, 叶志龙. 基于改进的 CHI 统计方法在文本分类中的应用. 计算机系统应用, 2016, 25(11): 136–140. [doi: 10.15888/j.cnki.csa.005393]
32 高明, 金澈清, 钱卫宁, 等. 面向微博系统的实时个性化推荐. 计算机学报, 2014, 37(4): 963–975.
33 陈杰, 刘学军, 李斌, 等. 一种基于用户动态兴趣和社交网络的微博推荐方法. 电子学报, 2017, 45(4): 898–905.
34 杜永萍, 杜晓燕, 姚长青. 基于主题效能的学术文献推荐算法. 北京工业大学学报, 2015, 41(2): 215–222.
35 王曰芬, 傅柱, 陈必坤. 基于 LDA 主题模型的科学文献主题识别: 全局和学科两个视角的对比分析. 情报理论与实践, 2016, 39(7): 121–126, 101.
36 崔金栋, 杜文强, 关杨, 等. 微博用户信息个性化推荐主题模型 LDA 演化分析研究. 情报科学, 2017, 35(8): 3–10.
37 张绍成, 刘威, 程子傲, 等. 代价敏感多主题学习的邮件过滤算法. 华中科技大学学报 (自然科学版), 2016, 44(S1): 176–180.
38 廖晓锋, 王永吉, 范修斌, 等. 基于 LDA 主题模型的安全漏洞分类. 清华大学学报 (自然科学版), 2012, 52(10): 1351–1355.