ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 550－558
DOI:10.16337/j.1004-9037.2017.03.014

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

面向文本分类的有监督显式语义表示

孙　飞　郭嘉丰　兰艳艳　程学旗

(中国科学院计算所网络数据科学与技术重点实验室,北京,100190)

摘　要:文本表示作为文本分类的一个基本问题,一直广受关注。目前文本表示主要有词袋模型、隐式语义表达和基于知识库的显式语义表达3种方式。本文首先分析对比了这3种文本表示方式在文本分类中的效果。实验发现,基于知识库的显式语义表达并没有如预期一样提高文本分类的效果。经分析,其原因在于显式语义表达在扩展文档表达时易引入噪声。针对该问题,本文提出了一种有监督的显式语义表达方法。该方法利用数据集的标注信息识别文档中与分类最相关的核心概念,并扩展核心概念以形成文档显式语义表达。3个标准分类数据集上的结果证实了本文所提文本表示方法的有效性。

关键词:文本分类;文本表达;有监督显式语义表示

中图分类号:TP391　　**文献标志码**:A

Supervised Explicit Semantic Representation for Text Categorization

Sun Fei, Guo Jiafeng, Lan Yanyan, Cheng Xueqi

(Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: As a fundamental problem of text categorization, text representation is widely concerned. Currently, there are three main ways of text representation: bag-of-words model, latent semantic representation and knowledge-based explicit semantic representation. The paper analyzes and compared the effects of these methods applied to text categorization. Experiments show that the knowledge-based explicit semantic representation cannot improve the text categorization performance as expected. To tackle the problem that the knowledge-based explicit semantic representation easily introduces noise in extending text, a supervised explicit semantic representation method is proposed. The dataset label information is used to identify the most relevant concepts in document and the document is represented in explicit semantic based on expanding those key concepts. The results of three datasets confirm the effectiveness of the proposed method.

Key words: text categorization; text representation; supervised explicit semantic representation

引　　言

文本分类作为信息处理关键技术,被广泛应用于信息检索、自然语言处理等领域。而文本表示作为

基金项目:国家重大基础研究发展计划("九七三"计划)(2012CB316303,2014CB340401)资助项目;国家高技术研究发展计划("八六三"计划)(2012AA011003)资助项目;国家自然科学基金重点(61232010)资助项目;国家科技支撑计划子课题(2012BAH46B04)资助项目。

收稿日期:2014-09-13;**修订日期**:2016-09-29

文本分类的一个核心问题,一直受到广泛关注。传统的文本表示主要采用向量空间模型(Vector space model, VSM)[1]。其假设文本中每个单词之间相互独立,俗称词袋模型(Bag of words, BOW)。词袋模型中,每个文档被表示为一个高维空间中的向量,向量的每一维代表了一个单词,其值对应此单词在文档中的权重。数十年来,词袋模型在文本表示方面取得了巨大成功。然而,词袋模型在表示文本时依然有着严重的缺陷。其单词之间相互无关的假设直接导致了它不能处理文本中的同义词。另外,它也不能处理一个单词具有多种含义的情况。隐式语义表达(Latent semantic representation,LSR)的提出部分地解决了词袋模型的问题。早在 1990 年,Deerwester 等提出隐语义索引模型(Latent semantic indexing, LSI)[2]对文档进行 SVD 分解,提升检索效果。随后,Hoffman 等提出概率隐语义分析模型(Probabilistic latent semantic analysis, PLSA) 模型[3],从概率角度建模文档的生成过程。而在 Blei 等[4]提出隐含狄利克雷分布模型(Latent Dirichlet allocation, LDA)后,基于 LDA 的各种话题模型层出不穷,成为信息检索、自然语言处理等领域研究热点。但是,隐式语义分析也有其不足之处,它仍然是一个基于词的表达(在给定话题时,词之间仍然有独立性假设),学习出的话题隐表达语义不清晰,难以建模单词之间的等价、上下位等语义关联,难以引入先验知识,且存在模型计算复杂等问题。鉴于以往以词为基本单元的文本表示存在的缺陷,近年来,研究者们提出了显式语义表达(Explicit semantic representation,ESR)用于文本表示,即利用本体知识库中的概念来显式表达文本的语义[5]。在这类工作中,概念成为表达的基本单元,作为人类知识的基本元素,概念具有清晰语义,概念间有丰富的语义关联。早期,许多工作尝试利用 WordNet 的结构化语义,丰富文本的表达以用于文本分类[6-9]。然而,WordNet 包含词汇量较少,且因单词描述短而难以有效地进行语义消歧。随着 Wikipedia 这样大规模知识库的发展,这两个问题都得以解决。近年来,基于知识库的主流工作大多以 Wikipedia 为基础[5,10-13]。

本文在 3 个标准的文本分类数据集上对比上述方法的分类效果,结果显示,利用大规模知识库对文本进行显式语义表达,与传统的词袋模型和隐式语义表达相比,并没有对分类效果产生预期的提升。分析数据发现,显式语义表达中关键的一步是利用概念间的关系进行扩展表达,然而文档中与语义密切相关的概念只占少数部分[14],假如对识别的概念全盘扩展,将引入大量的噪声数据导致分类性能低下,这一点在知识库规模庞大时(概念庞杂、关系密集)将尤为凸显。针对此问题,本文提出一种面向分类的有监督的文本显式语义表达。其核心思想是充分利用训练数据集中的标注信息来指导显式语义表达中的概念扩展。具体的,本文基于类别信息,定义了概念的区分度值,衡量概念对于分类类别的判别显著性。在此基础上,本文通过一个概率框架定义了文档中概念的核心度,依此识别与文档类别最相关的核心概念(Key concept)并进行扩展表达。最终,在 3 个数据集上的结果均验证了本文提出的文档表示方法对于分类问题的有效性与普遍性。

1　文本表示

1.1　词袋模型

词袋模型将一篇文档分割为一个个单词,并假设这些单词之间相互无关。在此假设下,将一篇文档表示为一个向量,向量的每个元素为一个单词,值为对应元素的权重,则

$$\boldsymbol{\varphi}(d_i)=[w_{i1},\cdots,w_{iM}] \tag{1}$$

式中:w_{ij} 为单词 t_j 在文档 d_i 中的权重,M 为字典的大小。词袋模型中单词的权重计算,一般使用 Tfidf[15],则

$$\mathrm{Tfidf}(t_j,d_i)=\mathrm{tf}(t_j,d_i)\times\log\frac{N}{|\{d_i\in D:t_j\in d_i\}|} \tag{2}$$

式中:$\mathrm{tf}(t_j,d_i)$ 为单词 t_j 在类别 d_i 中出现的次数;$|\{d_i\in D:t_j\in d_i\}|$ 为含有单词 t_j 的文档个数,N 为数据集文档总数。词袋模型是一种经典的文本表示方法,早期文本分类大多基于此模型,如文献

[16-19]。

1.2 隐式语义表达

隐式语义分析的典型方法如 LDA 概率化建模了文档的生成过程。LDA 认为文档是 K 个话题上的多项式分布，而话题是关于单词的多项式分布。一篇文档在生成时，先根据多项分布参数 θ_i 随机选择第 n 个单词的话题 $z_{i,n}$，然后根据多项式分布 $\varphi_{z_{i,n}}$ 选择单词。所以，一篇文档可以表示成一个关于话题的向量，则

$$\boldsymbol{\varphi}(d_i)=[\theta_{i1},\cdots,\theta_{iK}] \tag{3}$$

式中：θ_{ik} 为第 k 个话题在文档 d_i 的权重，K 为文档集设定的话题个数。Blei 的论文首次提出 LDA 作为文本表示模型[4]。其他使用隐式语义表达来进行文本分类的工作见文献[20,21]。

1.3 显式语义表达

显式语义表达以知识库中的概念为基本单元，将文档表示为概念的集合。并利用概念之间的同义、上下位和相关等关系对文档进行增强表达。据此，可以将文档表示为一个概念的向量。显式语义表达的基本过程如下：

(1)基于知识库识别文档 d_i 中所包含的概念。

(2)计算文档中概念的权重为

$$\boldsymbol{\varphi}(d_i)=[w_{i,c_1},\cdots,w_{i,c_M}] \tag{4}$$

式中：w_{i,c_j} 表示概念 c_j 在文档 d_i 的权重。

(3)依据知识库中概念间的关系添加文档中概念的相关概念，扩展文档 d_i 的表达为 $\boldsymbol{D}_E(d_i)$，则

$$\boldsymbol{D}_E(d_i)=[w_{i,c_1},\cdots,w_{i,c_M},w_{i,t_1},\cdots,w_{i,t_N},w_{i,c'_1},\cdots,w_{i,c'_K}] \tag{5}$$

式中：w_{i,c'_j} 为后添加进文档 d_i 的概念 c'_j 的权重。

(4)最终文档表达 $\boldsymbol{\varphi}'(d_i)$ 为 $\boldsymbol{D}_E(d_i)$ 的 l_2 范数归一化表达，则

$$\boldsymbol{\varphi}'(d_i)=\frac{\boldsymbol{D}_E(d_i)}{\boldsymbol{D}_E(d_i)_2} \tag{6}$$

早期显式语义表示的工作主要基于 WordNet[22]，Siolas 等[8]以及 Buenega 等[7]的工作成功地将其用于文本分类。作为目前最大的在线知识库，近年显式语义表示方面的工作，大多基于 Wikipedia。如，Gabrilovich 等[10-11]寻找与文本最相似的 Wikipedia 页面，利用其标题作为新的特征丰富文档的表达，来提高文档分类效果。此外，Wang 等[13]概念化表示文档后，利用 Wikipedia 的网络结构，构建了一个语义核函数进行文本分类。

2 文档表示实验验证

2.1 数据集

2.1.1 测试数据集

本文使用 3 个标准数据集 Reuters-21578，OHSUMED，Movie Reviews 来评价 3 种文本表示方法对于文本分类的效果。每个数据集的描述如下：

(1)Reuters-21578[23]：Reuters-21578 由 1987 年 Reuters 刊载的新闻构成，是文本分类领域使用最广泛的数据集之一。其共有 21 578 篇文档，135 个类别。本文采用 ModApte 划分法来构建训练集和测试集。去除没有标签或正文的文档，实际使用数据为：训练集 7 063 篇文档，测试集 2 742 篇文档，共 117 个类别。

(2)OHSUMED[24]：OHSUMED 数据集由在线医学数据库 MEDLINE 中的参考文献构成。其包含 1987～1991 年间 270 份医学期刊的 348 566 篇论文数据。每篇文档用 500 Mesh 索引项子集标注类别。

本文采用 Joachimes[9] 的设定,将 OHSUMED 1991 年数据,划分为 23 个疾病子类,选择前 10 000 数据作为训练集,后 10 000 数据作为测试集。去除摘要为空的数据,最终训练集大小为 6 286,测试集大小为 7 643。

(3)Movie Reviews[25]:此数据集来源于 IMDB 的电影评论,根据用户对电影标注的星级或分数将其人工标注为正、负两类各 1 000 条数据,广泛应用于情感分类领域。

2.1.2　Wikipedia 数据

本文使用 Wikipedia 作为知识库验证显式语义表示模型的能力。Wikipedia 中,每个页面描述了一个单一的话题,并带有一个简洁规范的标题,类似于传统词典中的词项[12]。本文将其定义为概念,标题为概念的名称,页面内容为此概念的内容。在不引起歧义的情况下,将概念与 Wikipedia 的页面等价混用。Wikipedia 中概念之间的语义关系主要表现为:同义、多义以及一般的相关关系。详细描述如下:

(1)同义:对于一个概念,Wikipedia 只包含一个页面详细描述它。但是,很多概念存在同义表达,或者其名称字面上有许多变形。对于这种情况,Wikipedia 用重定向将这些等价的概念归于一个最常用的概念之下。

(2)多义(歧义):许多概念在字面上是具有多重含义的。对于具有歧义的概念,Wikipedia 会有一个消歧页面,其中列出该名称可能代表的概念。

(3)语义相关:Wikipedia 的文章之间存在着许多超链接。这些超链接某种程度反映了概念之间的语义关联。

本文使用 Wikipedia 2013 年 7 月 8 日的数据。此数据集包含 4 347 801 个页面,5 988 323 个重定向,38 691 个消岐页面,页面之间的超链接共 226 389 440 个。Wikipedia 中也包含了大量无意义的页面以及空页面,如"List of newspapers"这样的网站管理功能性质页面。所以,在使用 Wikipedia 数据之前,本文先进行了一定的清理工作。与文献[13]工作相似,本文只保留那些标题大写、符合规范,或者单个单词标题在正文出现超过 3 次的页面。本文所使用概念集合为 Wikipedia 所有的页面标题,以及指向其的重定向的标题所构成的集合,最终所用概念集合大小为 3 215 886。由于 Wikipedia 的类别体系不够规范,同类别下的概念往往语义关联很弱,在实际应用时容易引入大量噪声信息,所以本文并未利用 Wikipedia 的类别信息。

2.2　算法及评价指标

本节验证 2.1 节所述 3 种文本表达方式对于文本分类的效果。

(1)词袋模型:选择 Tfidf 计算词项权重,并做归一化。

(2)隐式语义表达:选择了 LSI 以及 LDA 模型,LSI 使用 Gensim 实现。LDA 使用 GibbsLDA^{++}[26] 实现,参数 $\alpha=50/K$,$\beta=0.01$,迭代至收敛,否则迭代 2 000 轮。

(3)显式语义表达:本文使用前向最大匹配[27]算法识别文档中的概念,对于匹配成功的单词将其替换为对应的概念,而没有匹配成功的单词依然保留。概念的权重使用 Tfidf 计算策略。使用文本相似度来解决概念的消歧问题。计算每一个候选概念的 Wikipedia 页面内容与文档的 Tfidf 余弦相似度,选择相似度最高的概念作为词项所指概念。对于扩展的概念,其权重为引其进入的概念的 Tfidf 值。当然,这里对于同义、相关概念完全可以选择不同的权重策略。在引入相关概念时,计算被引入概念与引入概念的余弦相似度,本文与文献[13]一样只使用了相似度最高的 10 个概念,以避免引入过多噪声数据。本文根据不同扩展策略,实现了多个显式语义文本表示方法:(1)ESR-noterm:仅使用文本中识别出的概念;(2)ESR:除文本中识别出的概念,还保留没有匹配成功的单词;(3)ESR-s:对文本中概念添加其同义概念;(4)ESR-a:对文本中概念添加其相关概念;(5)ESR-sa:对文本中概念添加其同义和相关概念;(6)Wiki-sk[13]:使用 Wikipedia 语义化表示文本进行分类的经典工作。扩展与文档 Tfidf 余弦相似度最

高的 3～5 个概念，并利用 Wikipedia 的网络结构构建针对数据集中所有概念的语义。对于所有的数据集，所有算法都进行相同的去除停用词预处理。因词形还原会导致很多概念无法被识别(如 Data mining)，因此除显式语义表达，其余算法均采用了相同的词性还原预处理。鉴于 SVM 多年来在文本分类中取得的突出效果，本文使用 SVM 作为分类器。对于 Reuters-21578 和 OHSUMED 数据集，使用 OneVsRest 策略进行多标签分类。本文分类算法均使用 Scikit learn 封装的 Libsvm，并采用了线性核。对于每个算法，在每个数据集的训练集上使用网格搜索确定 SVM 的参数 c 最优值。

2.3 评价标准

本文使用精确度、召回率和 F_1 值来评价各个方法的性能[17]。对于每个文档，其计算公式定义为

$$P=\frac{|l_p \cap l_t|}{|l_p|} \quad R=\frac{|l_p \cap l_t|}{|l_t|}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{7}$$

式中：P 和 R 分别为精度和召回率；l_p 为分类器预测的文档类别；l_t 为文档的真实类别。所有评价指标结果为整个数据集所有文档得分的均值。所有实验均运行 10 次，取 10 次结果均值作为最终结果。Movie Review 数据集因其为二分类数据，故 P，R 和 F_1 值相等，本文只给出 F_1 结果。

2.4 实验结果与分析

从表 1 可以看出，对于文档的隐式表达，随着话题数目的增加，分类效果逐渐递增，但是递增的趋势在逐渐放缓。LSI 随着话题数目的增加，分类效果逐渐接近甚至超过词袋模型，但优势并不明显。这与 SVD 的数学性质相吻合，性能超出词袋模型的部分得益于保留前 K 个话题而去除的噪声。相比之下，同样话题数目下，LDA 的分类效果并不如 LSI，也没能超过词袋模型。对于文档的显式语义表达，如果仅仅使用文档中识别的概念，无论是否保留未匹配成功的词项，在所有数据集上分类的结果，都明显低于词袋模型的结果。原因在于概念通常由多个单词组成，仅用概念表示文档而不作扩展时，相比词袋模型，并没利用特征间的相关性，而使用的特征又更少，因而效果下降。令人意外的是，在使用同义概念与相关概念对文档进行扩展后，分类的结果不但没有变好，反而明显下降。这与期望的结果并不一致。观

表 1 文本表示分类结果

Tab. 1 Results of document classification

算 法	Reuters-21578			OHSUMED			Movie Review
	P/%	R/%	F_1	P/%	R/%	F_1	F_1
BOW	88.74	87.20	87.97	61.45	54.83	57.95	84.90
LSI(k=100)	87.11	87.61	87.36	52.70	44.04	47.98	77.86
LSI(k=500)	90.06	89.03	89.54	61.27	56.31	58.69	83.26
LSI(k=800)	90.21	89.64	89.92	62.59	58.34	60.39	82.27
LDA(k=100)	78.77	78.58	78.67	45.71	36.09	40.33	80.28
LDA(k=500)	84.66	84.02	84.34	56.02	50.23	52.97	80.38
LDA(k=800)	84.78	84.32	84.55	58.00	52.06	54.87	77.80
ESR-noterm	87.14	85.67	86.40	54.82	48.82	51.64	82.73
ESR	87.41	85.39	86.39	60.77	53.48	56.89	85.28
ESR-s	86.75	85.83	86.29	58.12	52.70	55.28	78.93
ESR-a	87.77	85.93	86.84	58.53	52.49	55.34	82.47
ESR-sa	87.69	86.18	86.93	57.72	52.13	54.78	81.75
Wiki-sk	71.75	79.71	75.52	44.16	49.91	46.86	72.47

察数据发现,得益于 Wikipedia 的广泛内容,文档往往可以识别出大量的概念。而这其中许多概念,与文章所要表达的主题并无紧密联系。如 Reuters-21578 中编号 14828 文档中"A survey of 19 provinces and seven cities showed vermin consume between seven and 12 pct of China's grain stocks, the China Daily said."可以识别出"survey","provinces","seven cities","vermin","seven","China","grain","stock","China Daily"概念,但是其中只有"vermin","grain"与文章的主题"grain"相关。这种情况下,扩展"provinces"、"seven cities"、"China"、"China Daily"等概念,会导致文档原本的语义发生稀释偏移,可能会使得扩展后的文档与讲述地理的文档更加相似。此外,作为和本文工作密切相关的 Wiki-sk,由于其代码无法公开,在按照其论文所述方式实现之后其结果要远低于其他方法,也低于其论文中给出的结果。分析其中可能的原因在于本文使用 Wikipedia 的数据规模比文献[13]中所使用数据高出一个量级。因此,文档中能够被识别出的概念要比 Wiki-sk 多很多,直接导致了大量噪声的引入。虽然 Wiki-sk 使用了文档与概念的 Tfidf 余弦相似度,选择最相似的 3～5 个概念进行扩展,但在更大规模的 Wikipedia 上发现,此策略并不奏效,可参照第 4 节实验分析。最关键在于 Wiki-sk 在扩展后,构建了一个关于数据集中所有概念之间的语义核函数。此核函数在一个更大的 Wikipedia 情况下,会导致严重的语义偏移。实验结果表明,使用大规模知识库进行显式语义表达,并没有达到人们期待的性能,分类效果与词袋模型、隐式语义表达相当或略低。分析数据发现,造成结果不佳的原因主要是扩展表达中的语义偏移问题,其主要由两方面造成:(1)扩展的概念可能与文章表达的主题不相关,如上文中的"seven"等概念。(2)扩展的概念虽然与文档的主题有一定关联,但是与分类问题中此文档所属类别无关。许多文档有多个话题。如 Reuters-21578 中 14 828 文档讲的是中国多省害虫消耗粮食以及粮食的浪费问题,但是在此数据集中,其只被标注为"Grain"类别。如果扩展了与浪费相关的概念,虽然与文档的主题也相关,但对于此文档的分类并无帮助。针对显式语义表达存在的这些问题,本文提出有监督的显式语义表示来解决上述问题。

3　有监督显式语义表示

在使用 Wikipedia 等知识库对文档作显式语义表达时,核心问题在于寻找与文档主题相关的概念,本文将其定义为核心概念(Key concept)。对于分类问题,核心概念就是与文档语义类别真正相关的概念。所以提出有监督的显式语义表达(Supervised explicit semantic representation, SESR)方法,其核心思想就是充分利用训练样本中的有监督信息来指导识别核心概念,实现文本的显式语义表达。

3.1　概念的类别区分度

本文首先利用训练数据集的标注信息,定义概念对于分类中每个类别的区分度值(Discriminative score, DS),表征概念对类别的判别显著性。其值越高,代表概念对于当前类别关联程度越高。

(1)类别的表示:本文首先将分类类别表示为类别下所有文档的集合,而文档是由概念组成的,所以类别最终使用此类所有文档包含的所有概念来表示。

(2)计算概念对于类别的区分度,定义为

$$d_s(c_i,l_j)=\mathrm{Tf}(c_i,l_j)\times\log\frac{N}{|\{l\in L:c_i\in l\}|} \tag{8}$$

式中:$\mathrm{Tf}(c_i,l_j)$ 为概念 c_i 在类别 l_j 中出现的次数;$|\{l\in L:c_i\in l\}|$ 含有概念 c_i 的类别数;N 为类别总数。这里使用类似传统 Tfidf 的计算策略,当然也可以使用 χ^2,信息增益等,本文主要为了证明核心概念的有效性与必要性,所以简化处理。

3.2　核心概念的识别

在此基础上,本文提出一个统一的有监督显式语义表达框架,解决文档显式语义表达中扩展概念的选择问题。核心思想是通过对文档中概念核心度的估计来识别文本的核心概念,依此进行扩展。以类

别为桥梁，定义对于文档核心度的概率框架为

$$p(c_i \text{ 是关键概念} \mid d_k) = \sum_{l_j} p(c_i \text{是关键概念} \mid l_j) p(l_j \mid d_k) \tag{9}$$

式中：$p(c_i \text{是关键概念} \mid l_j)$ 为类别 l_j 下，概念 c_i 的核心程度；$p(l_j \mid d_k)$ 为文档的类别隶属度，即给定文档 d_k，其属于类别 l_j 的概率。对于 $p(c_i \text{是关键概念} \mid l_j)$，可以利用概念区分度来估计，其基本思想是对于概念在一个类别下的核心程度，可以通过该概念对类别的判别显著性来表示。具体的，可使用规范化的概念区分度表达类别下概念核心程度，则

$$p(c_i \text{是关键概念} \mid l_j) = \frac{\mathrm{ds}(c_i, l_j)}{\sum_{c_{i'}} \mathrm{ds}(c_{i'}, l_j)} \tag{10}$$

而对于文档的类别隶属度 $p(l_j \mid d_k)$ 的度量，可以采用不同的模型进行估计。在这里，本文利用一个线性分类模型，基于文档词向量来对文档的类别隶属度进行估算。具体的，本文通过最小化目标函数

$$l(w_j) = \sum_{i=1}^{N} L(y_i, f(x_i)) + \alpha R(w_j) \tag{11}$$

来学习线性分类模型 $f_j(x) = \boldsymbol{w}_j^{\mathrm{T}} x + b_j$，其中，$y_i \in \{-1, +1\}$ 表示是否属于类别 l_j，损失函数 $L(\cdot, \cdot)$ 使用修正的 huber 损失函数，$R(\boldsymbol{w}_j)$ 为正则项，本文使用 l_2 范数，α 为正则权重，设为 0.000 1。在学习得到参数 w 和 b 的基础上，估计文档的类别隶属度为

$$p(l_j \mid d_k) = \frac{f_j(d_k) + 1}{2} \tag{12}$$

基于上述估计，通过对各个类别的累加，便可以基于式(1)得到文档中每个概念核心度的概率值，基于该核心度值对概念进行降序排列，并选择前 k 个概念作为文档的核心概念进行扩展表达，得到文档最终的显式语义表示。在最终文本分类器训练时，本文如 1.3 节所述对所有文本进行扩展表示，但只扩展文档中重要度最高的 k 个概念。然后使用扩展后的语义表达训练分类器模型，对扩展表达后的测试数据进行分类。

4 有监督显式语义表示实验及分析

本文通过实验验证 SESR 的分类性能。对于每个文档，参照 Zhang 等的工作[14]，选择重要度最高的前 3 个概念进行扩展。本文实现了 3 种不同的扩展表示：(1)SESR-s：添加同义概念；(2)SESR-a：添加相关概念；(3)SESR-sa：添加同义与相关概念。作为基准的显式语义表达算法，本文使用余弦相似度作为选择文档核心概念的方法，来扩展最相关的 3 个概念：(1)ESR-s3：添加 Tfidf 余弦相似度前 3 的概念的同义概念；(2)ESR-a3：添加 Tfidf 余弦相似度前 3 的概念的相关概念；(3)ESR-sa3：添加 Tfidf 余弦相似度前 3 的概念的同义和相关概念。

实验结果显示，基于余弦相似度选择核心概念的显式语义表达 ESR-s3，ESR-a3，ESR-sa3，与扩展所有概念的显式语义表达 ESR-s，ESR-a，ESR-sa 相比，并无任何优势，且两者都低于不做扩展的方法 ESR。这显示余弦相似度并不能很好地刻画文档的语义信息。从表 2 中可看出，在所有的方法中，有监督的显式语义表达取得了最好的结果 $p_{\text{value}} < 0.01$。对比 SESR-sa 与 ESR，发现利用监督信息对文档的核心概念进行扩展表达相比原始的概念表达，在所有数据集上对分类结果都有显著提高，这说明了扩展的有效性。对比 SESR-sa 与 ESR 的一系列扩展(如 ESR-sa，ESR-sa3)的结果，发现 SESR-sa 明显好于其他两种扩展方式。这说明本文提出的概念类别区分度可以有效地识别文本的核心概念，减少噪声引入，从而提高分类效果。此外，从表 1，2 还可发现 SESR-sa 相比词袋模型在 3 个数据集上 F_1 分别提升了 2.25%，4.16%和 5.05%，比隐式语义表达有明显提升，这主要得益于有监督的显式语义表达更准确地利用了更加丰富的语义信息。从表 2 还可看出，同时扩展同义、相关概念 ESR-sa，在核心概念识别不

正确的情况下,会比只扩展同义或者相关概念 ESR-s,ESR-a 效果更差。因为此时扩展引入的不相关概念更多。但是,如果核心概念识别正确,扩展同义和相关概念 SESR-sa 要比只扩展同义或者相关概念 SESR-s,SESR-a 要好(除 Reuters-21578)。

表 2　显式语义表达对比结果

Tab. 2　Comparison results of explicit semantic representation

算法	Reuters-21578			OHSUMED			Movie Review
	P/%	R/%	F_1	P/%	R/%	F_1	F_1
BOW	88.74	87.20	87.97	61.45	54.83	57.95	84.90
ESR	87.41	85.39	86.39	60.77	53.48	56.89	85.28
ESR-s	86.75	85.83	86.29	58.12	52.70	55.28	78.93
ESR-s3	85.88	84.77	85.32	57.70	52.83	55.16	82.33
ESR-a	87.77	85.93	86.84	58.53	52.49	55.34	82.47
ESR-a3	85.11	83.73	84.42	56.45	51.30	53.75	79.76
ESR-sa	87.69	86.18	86.93	57.72	52.13	54.78	81.75
ESR-sa3	83.94	83.01	83.47	55.54	50.34	52.81	78.42
SESR-s	90.53	88.85	89.68	64.97	57.75	61.15	86.80
SESR-a	90.84	89.89	90.37	65.03	58.83	61.78	89.05
SESR-sa	90.63	89.82	90.22	65.51	59.05	62.11	89.95

5　结束语

本文首先阐述了文本表示的 3 种方式,并在 3 个主流的文本分类数据集上实验这 3 种表示方式的分类效果。结果显示,利用知识库对文本进行显式语义分析,并没有对分类效果产生预期的提升。通过分析数据发现在扩展文本表示时,引入了大量的噪声数据。本文提出 SESR 方法,利用文档的标签信息,识别扩展与文档类别最相关的概念。3 个数据集上的结果均验证了此方法的有效性与普遍性。本文后续工作主要集中在研究扩展概念数量对于扩展表示效果的影响,如何确定最佳的核心概念集合大小。另一个需要改进的地方在于扩展添加进文档的概念的权重计算,本文简单地使用引入其的概念权重。更合理的方式应该根据概念之间的语义关联强弱计算得到相应的权重。

参考文献:

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [J]. Commun ACM, 1975, 18(11):613-620.

[2] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6):391-407.

[3] Hofmann T. Probabilistic latent semantic indexing[C]// Proceedings of SIGIR '99. New York, NY: ACM,1999: 50-57.

[4] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. J Mach Learn Res, 2003,3:993-1022.

[5] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis[C]// Proceedings of IJCAI'07 San Francisco. CA, USA:[s. n.],2007: 1606-1611.

[6] Jing L, Zhou L, Ng M K, et al. Ontology-based distance measure for text clustering[C]// Proceedings of the Text Mining Workshop, SDM'06. Bethesda,MD,USA:[s. n.], 2006:1-8.

[7] Rodríguez M D B, Gómez-hidalgo J M, Díaz-agudo B. Using wordnet to complement training information in text categorization[C]// Proceedings of RANLP'97. Tzigov Chark, Bulgaria:[s. n.], 1997: 150-157.

[8] Siolas G, d'Alché Buc F. Support vector machines based on a semantic kernel for text categorization[C]// Proceedings of IJCNN'00. Washington D C, USA:[s. n.],2000: 205-209.

[9] Ureña-López L, Buenaga M, Gómez J. Integrating linguistic resources in TC through WSD[J]. Computers and the Humanities,2001,35(2):215-230.

[10] Gabrilovich E, Markovitch S. Feature generation for text categorization using world knowledge[C]// Proceedings of IJCAI'

05. San Francisco, CA, USA:[s. n.],2005:1048-1053.
[11] Gabrilovich E, Markovitch S. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge[C]// Proceeding of Twenty-First National Conference on Artificial Intelligence. Boston, M A: AAAI, 2006: 1301-1306.
[12] Milne D, Medelyan O, Witten I H. Mining domain-specific thesauri from wikipedia: A case study[C]// Proceedings of WI' 06. Washington DC, USA:[s. n.], 2006: 442-448.
[13] Wang P, Domeniconi C. Building semantic kernels for text classification using Wikipedia[C]// Proceedings of KDD'08. New York, NY, USA: ACM,2008: 713-721.
[14] Zhang L, Pan Y, Zhang T. Focused named entity recognition using machine learning[C]// Proceedings of SIGIR'04. New York, NY, USA: ACM,2004: 281-288.
[15] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Inf Process Manage, 1988, 24(5):513-523.
[16] Joachims T. Text categorization with support vector machines: Learning with many relevant features [J]. Lecture Notes in Computer Science, 1998,1398: 137-142.
[17] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// Proceedings of ICML'97. San Francisco, CA, USA:[s. n.],1997: 412-420.
[18] Yang Y. An evaluation of statistical approaches to text categorization[R]. CMU-CS-97-127,[S. l.]: Carnegie Mellon University, 1997:69-90.
[19] McCallum A, Kamal N. A comparison of event models for naive Bayes text classification [C]// AAAI-98 workshop on learning for text categorization. [S. L.]:AAAI,1998:41-48.
[20] Yu B, Xu Z B, Li C H. Latent semantic analysis for text categorization using neural network[J]. Knowledge-Based Systems, 2008,21(8), 900-904.
[21] Liu T, Chen Z, Zhang B,at al. Improving text classification using local latent semantic indexing[C]// In ICDM'04. [S. L.]: IEEE, 2004: 162-169.
[22] Miller G. A. Wordnet: A lexical database for English [J]. Commun ACM, 1995, 38(11):39-41.
[23] Reuters Ltd, Carnegie Group, Inc. Reuters-21578 text categorization collection [EB/OL]. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection,1997-09-26.
[24] Hersh W, Buckley C, Leone T J, at al. Ohsumed: An interactive retrieval evaluation and new large test collection for research[C]// Proceedings of SIGIR '94. New York, NY, USA: ACM, 1994:192-201.
[25] Pang B, Lee L, Vaithyanathan S. Thumbs up: Sentiment classification using machine learning techniques[C]// Proceedings of EMNLP '02. Stroudsburg, PA, USA:[s. n.],2002: 79-86.
[26] Xuan-Hieu P, Cam-Tu N. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) [EB/OL]. http://gibbslda.sourceforge.net, 2013-04-15.
[27] Wong P K, Chan C. Chinese word segmentation based on maximum matching and word binding force[C]// Proceedings of COLING'96. Stroudsburg, P A, USA:[s. n.],1996: 200-203.

作者简介:

孙飞(1988-),男,博士研究生,研究方向:文本表示和知识库,E-mail: sunfei@software.ict.ac.cn。

郭嘉丰(1980-),博士,副研究员,研究方向:网络搜索和挖掘、用户数据挖掘以及社交网络。

兰艳艳(1982-),女,博士,副研究员,研究方向:排序学习。

程学旗(1971-),男,研究员,博士生导师,研究方向:网络科学、互联网搜索与挖掘、对等网络、信息安全和分布式系统。