一种基于维基百科的文本表示方法

黄浩军[1],王胜清[2]
HUANG Haojun[1], WANG Shengqing[2]

1.北京大学 软件与微电子学院,北京 100871
2.北京大学 现代教育技术中心,北京 100871
1.School of Software & Microelectronics, Peking University, Beijing 100871, China
2.Center for Education Technology, Peking University, Beijing 100871, China

HUANG Haojun, WANG Shengqing. New text represent method based on Wikipedia. Computer Engineering and Applications, 2015, 51(14):127-130.

Abstract: Text representation is the basic task in natural language processing. In general, text representation model can build with sufficient text data. While with insufficient data, it can not complete the task in natural language processing. So, it comes up with a new text represent method to overcome the dilemma. It builds the semantic relationship between words using the link in Wikipedia, and enriches the representation with page rank model diffusing the message to other wiki-items. It verifies that this enrichment can raise the precision, recall and F1-measure of the text classification method.
Key words: Wikipedia; Latent Dirichlet Allocation(LDA); text representation; text classification

摘 要:文本表示是自然语言处理中的基础任务,通常的文本表示模型都是基于训练数据充分的情况下进行。而在训练数据缺乏时,无法完成自然语言处理任务。提出了一种基于维基百科的文本表示方法,引入维基百科词条之间的关系,通过PageRank传播模型,能够一定程度上解决训练数据缺乏时文本表示的问题。通过实验论证了基于维基百科的文本表示能够增强分类方法的准确率、召回率和F1-测度。
关键词:维基百科; 隐含狄利克雷分布; 文本表示; 文本分类
文献标志码:A **中图分类号:**TP391.1 doi:10.3778/j.issn.1002-8331.1406-0071

1 引言

文本表示是将人类可阅读的文字转换成为计算机可以识别的数据结构的过程。文本表示(Text Representation)是文本信息处理中的基础性问题,任何的自然语言处理任务都需要基于文本表示进行,例如文本分类,文本聚类,信息检索,问答系统等。

文本表示的准确度很大程度上决定了自然语言处理任务的结果表现。通常文本表示方法有空间向量模型(Vector Space Model,VSM)和基于超链接模型的网页表示。在可用于训练的数据充足的情况下,两个模型在以上自然语言处理任务中都有很好的表现。但是,在无法提供充足的训练数据情况下,两种表示模型的表现效果退化明显。

本文提出了一种新的文本表示方法,用于解决在训练文本数据缺乏时或者可供训练的文本数据大部分为短文本的情况下,文本表达不充分的问题。在上述情况下,增强文本的表达鲁棒性是解决问题的关键。本文利用维基百科词条链接关系,结合信息传播模型,构建了维基百科词条的文本表示方法,增强了文本表示的鲁棒性。

2 研究现状

2.1 维基百科语义关系研究现状

语义关系是自然语言处理中一个基本问题,如何高效准确判断两个单词之间的语义相似度是自然语言处理中表示学习的核心问题。

在自然语言语料中,维基百科语料是人工编纂,内容质量较高,且容易获得的大规模语料。维基百科词条通过互相之间的链接关系建立联系,其链接关系反映了

基金项目:文化部国家文化科技提升项目(No.201201-02)。
作者简介:黄浩军(1988—),男,硕士研究生,研究领域为自然语言处理;王胜清(1968—),女,博士生,高级工程师,研究领域为数字资源组织与服务。E-mail:littleblack1988@gmail.com
收稿日期:2014-06-06 **修回日期:**2014-12-26 **文章编号:**1002-8331(2015)14-0127-04

词条之间的语义链接关系在自然语言处理各个任务中，基于维基百科词条语义关系的研究都得到应用，例如歧义消解、Wikify、知识图谱构建、实体识别等[1]。

本文通过维基百科词条语义关系构建了文本表示方法。

2.2　文本表示研究现状

在文本表示方法中，常用的方法有向量空间模型（Vector Space Model，VSM）和基于超链接的网页表示模型。同时，随着工业界和学术界对文本表示模型的深入，文本表示模型开始向基于语义的文本表示方法演变。常用的文本表示模型需要在高维空间中进行，且丢失了词与词的相关性和语义信息。为了解决这两个问题，文献[2]提出了基于单词赋权的方法解决词权重问题，并应用于文本分类方法；文献[3-4]提出在给定文本前提下的一种二元表示方法，但是这种方法需要处理一个高位的稀疏矩阵，难以有效地进行文本分类等任务；文献[5]提出了一种基于本体的文本表示方法，能够有效地解决词之间的相互关系，但是无法构造有效的实用本体，文献[6]提出了一种基于LSI的表示方法，文献[7]提出了基于LDA的文本表示方法，能够挖掘文本的潜语义；文献[8]提出了基于*N*-gram的文本表示方法；文献[9]使用UNL（Universal Networking Language）表示一个文本。

3　基于维基百科的文本表示构建方法

基于维基百科的文本表示构建流程分为两个步骤，维基百科词条关系构建和文本表示构建。

维基百科词条关系构建是通过两个方面来进行，首先，根据每个词条内容中的若干个其他词条的导向链接，提取出词条之间的链接关系；然后，通过使用主题模型进行内容分析，得到词条之间的链接权重。

文本表示构建是通过建立的维基百科词条关系来表示文本。文本中的单词需要同维基百科词条建立联系，使得原文本中的词映射到维基百科词条上，这样可以通过之前建立好的词条对文本进行表达。

构建流程如图1所示。

3.1　维基词条关系构建方法

维基百科词条关系构建是通过两个步骤进行的。首先，根据每个词条内容中的若干个其他词条的链接，提取出词条之间的链接关系；然后，通过使用主题模型进行内容分析，得到词条之间的链接权重。

维基百科词条链接权重计算方法可以有两种方法：词频（TF-IDF）和

（1）基于词频计算权重

通过词频计算维基百科词条之间的链接权重，首先需要将每个词条对应的文本用空间向量模型表示；然后通过两个向量之间的 *cosine* 相似度得到词条间的链接权重。

$$Weight(\boldsymbol{wiki}_i, \boldsymbol{wiki}_j) = Cosine([\boldsymbol{vsm}_i \| \boldsymbol{vsm}_j]) \text{ if } i \text{ 指向} j \quad (1)$$

其中，***wiki*** 表示维基百科词条的集合，$\boldsymbol{wiki}_i$ 表示维基词条 i，$\boldsymbol{vsm}_i$ 表示维基词条 i 对应的解释文本的VSM表示。

（2）基于LDA计算权重

每个维基百科词条对应了一篇文档，整个维基百科可以看作是一个语料库。通过LDA进行内容分析，模型输出结果为文档-主题分布、主题-单词分布以及每个单词主题分配。对于每两个有链接关系的词条来说，其链接的权值为：

$$Weight(\boldsymbol{wiki}_i, \boldsymbol{wiki}_j) = D_{\mathrm{KL}}[\boldsymbol{topic}_i \| \boldsymbol{topic}_j] \text{ if } i \text{ 指向} j \quad (2)$$

其中，***wiki*** 表示维基百科词条的集合，$\boldsymbol{wiki}_i$ 代表编号为 i 的维基词条，D_{KL} 表示两个概率分布的KL距离，$\boldsymbol{topic}-\boldsymbol{distribution}_i$ 表示编号为 i 的维基百科词条对应的文档通过LDA分析得到的该文档主题分布。

词频和主题模型构建词条关系网络各有优势，词频方式构建计算复杂度低，不需要引入复杂模型和存储大规模的模型文件，但是词频仅从文本的简单的统计量分析，区分能力有限且文档相似度计算是在高维度空间（空间为词表空间，维度为 $|V|$，其中 $|V|$ 为词表大小），计算量大，同时TF-IDF无法区分词条与词条之间相互指向；而主题模型能从潜语义空间对文档进行建模。在计算相似度时，在潜语义空间中，文档表示向量的维度大大降低，同时其基于KL散度的非对称性能够区分词条

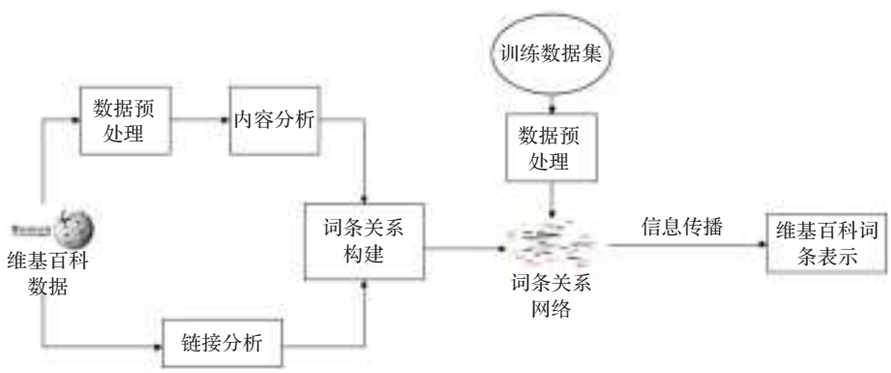

图1　wiki-Representation **构建流程图**

与词条之间的非对称关系，但是主题模型需要存储模型文件，在维基百科数据时需要并行化。

本文采用了基于主题模型的词条权重计算方法，得到词条关系矩阵记为 $\boldsymbol{wiki-relation}$ 。

3.2 文本表示构建

语料库文档需要同维基百科词条网络建立联系，使得原文档映射到维基百科词条上。由于维基百科中每个词条对应一篇解释文档，而词条本身也是一个词，也可能出现在文档中。本文提出一种建立词条和训练文档的映射方法。

映射方法大体思想为：如果文档中的词与该词条一致，则该文档需要和该词条建立强的联系；如果文档中词出现该词条对应的解释文档中，则文档需要和该词条建立一般的联系。对于上述两种情况，本文提出基于有区分赋权思想，将文档与对应的词条关联性量化，如此所有文档均表示为维基百科词条的权值向量。

对于第一种情况，词条的文本和文档中词一致时，文档与对应词条具有很强的关联性。对于文档中的词 t ，其对应词条的所赋权值为 $tf(t,d)\times\gamma$，其中 γ 为经验参数，$tf(t,d)$ 为文档 d 中词 t 的词频。

对于第二种情况，词条对应的解释文档中包含训练文档中词时，训练文档与对应词条具有一般的关联性。对于文档 d 中的词 t，其对应词条的所赋权值 $\frac{tf(t,d)}{N}\times\eta$，其中 N 为维基百科词条解释文档包含的词总数，η 为经验参数。

映射关系和权值计算公式为：对于文档 d 中的词 t 需要与维基百科词条集合和每个词条对应的解释文档进行权值计算：

$$weight_d(\boldsymbol{wiki},t)=\begin{cases}tf(t,d)\times\gamma, \text{if } t\in \boldsymbol{wiki}\\ \dfrac{tf(t,d)}{N}\times\eta, \text{if } t\in \boldsymbol{wiki-doc}\end{cases} \quad (3)$$

其中，$\boldsymbol{wiki}$ 表示维基百科词条集合，$\boldsymbol{wiki-doc}$ 表示维基百科词条对应的解释文档集合。最终，训练文档表示为一个维度为 $|\boldsymbol{wiki}|$ 的向量。

本文挖掘词条之间存在的关联性，并对关联性进行了量化赋权，使得能通过赋权且有链接关系的词条表示文档。在进行文档建模时，由于有部分词条与文档无直接关联或者部分关联词条映射权值不合理。在建立向量时，这部分词条的需要对其权值进行修改和平滑化处理，使得模型能更好地表示文档。

在维基百科词条网络中，每个词条对应一个节点，词条与词条之间的链接关系对应赋权的有向边。

3.3 PageRank

本文通过信息传播方法对词条信息的传播进行建模，可以使词条向量模型在表达文档时更加平滑。

本文采用 PageRank 建模词条之间的信息传播，将前一步得到的文本词条表示。按照词条之间的关系通过 PageRank 模型，得到最终的文本词条表示向量。构建流程举例：假设维基百科词条集合为 $\boldsymbol{wiki}$，那么对于文档中每篇文档 d，按照映射方法，可以表示为一个长度为 $|\boldsymbol{wiki}|$ 的向量 $\boldsymbol{wiki}_d$ 。然后，在 PageRank 模型下，将 $\boldsymbol{wiki}_d$ 看做状态初始值，通过 $\boldsymbol{wiki}$ 词条的链接关系矩阵 $\boldsymbol{wiki-relation}$ ，得到最终的文档表示向量 $\boldsymbol{wiki}_d'$ 。通过 PageRank 进行信息传播后，得到文本表示向量相比之前表示更加平滑，使得某些相关但是初始为 0 的分量上有大于 0 的权重，使得文本的表示更加丰富。

4 实验及结果

本章将以文本分类任务，比较向量空间模型和本文提出的表示方法的效果。本实验中，分类器采用深度信念网络，评价标准为分类方法的精确率、召回率和 F1-测度。

本实验涉及的经验参数包含 LDA 参数（主题数 K ，主题词项分布先验参数 β ，文档主题分布先验参数 α ）、文档和维基百科词条映射关系和权值计算参数 (η,λ) 。

LDA 参数选取参照 A.Chaney 进行的 Wikipedia 分析及实验系统[10]，其中 $K=50$，主题单词分布先验参数 β 设置为默认，$\beta_i=1/V$，即文档主题分布先验参数 α 设置为默认，即 $\alpha_i=1/K$，其中 V 代表词表大小，K 代表主题数。

文档与维基百科词条映射参数 η 和 λ ，用于衡量训练语料文档映射表示的权重。实验采用 Newgroup20 语料，其中将 80%的语料进行训练，20%用于测试和参数选取，测试的 η,λ 组合为 {(2.0，1.0)，(1.5，0.8)，(1.2，0.7)} 三组，通过分类结果的准确率，召回率和 F1-测度进行筛选，结果如表 1 所示。

表 1　η,λ 对文本分类的影响　%

(λ,η)	准确率	召回率	F1-测度
(1.2，0.7)	78.67	80.13	79.39
(1.5，0.8)	79.93	81.54	80.72
(2.0，1.0)	79.36	81.66	80.49

由表 1 数据可知[11]，在 η,λ 选取时，需要考虑训练文档到命中维基百科词条和命中维基百科词条内容的两种情况。对于命中维基百科词条的单词，表示该训练文档与改词条强相关，命中维基百科词条内容的单词，表示该训练文档与维基百科词条存在一定的相关性。如果对第一类命中情况加大相应的权值 (α) ，使得最终表示结果中改词条的权值加大，分类结果会向这个词条对应的类别偏置，极限情况，会对使得分类结果严重的受到词条命中的影响，从而忽略的第二类命中情况。

对比的 DBN 分类模型采用文献[7]中代码，DBN 模型总共四层，其中三层隐藏层，每个层隐藏层包含 1 000 个隐藏单元，每层训练的 epoch 为 100，训练过程采用 min-batch 方式，min-batch 大小设置为 10，非监督训练

学习率设置为0.01,监督训练学习率设置为0.1,使用early-stopping增加模型泛化能力。

本文表示方法中采用词条过滤方法,将不包含数据集中词的词条过滤(词条共1 200),其中内容分析适用Online-LDA,主题数为30,词条权值计算采用KL散度;DBN模型总共四层,其中三层隐藏层,每个层隐藏层包含1 000个隐藏单元,每层训练的epoch为100,训练过程采用min-batch方式,min-batch大小设置为10,非监督训练学习率设置为0.01,监督训练学习率设置为0.1,使用early-stopping增加模型泛化能力。

从图2,图3,图4中,对比本文的表示方法和空间向量模型。在训练数据集充足时(即训练数据和测试数据比率大于等于2:1时),本文表示方法能够保证两者分类效果基本一致;在训练数据缺乏时(即训练数据和测试数据比率小于1:4),本文表示方法文本分类准确率有15%的提升,召回率有8%的提升,F1-度量有9%的提升。综上所述,将文档映射为维基词条进行文档表示转换后,在文本分类任务中,本文的文本表示具有更好的效果。

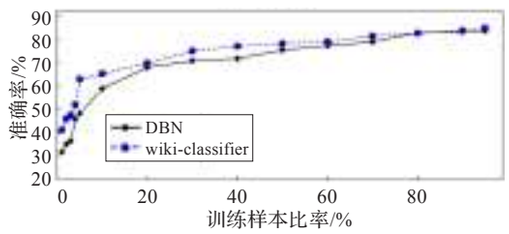

图2 本文方法和空间向量模型分类准确率

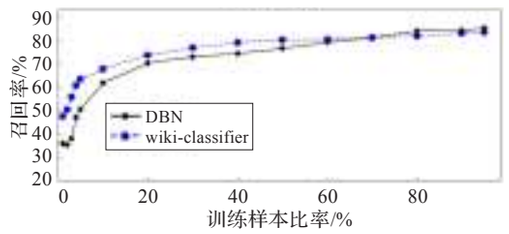

图3 本文方法和空间向量模型分类召回率

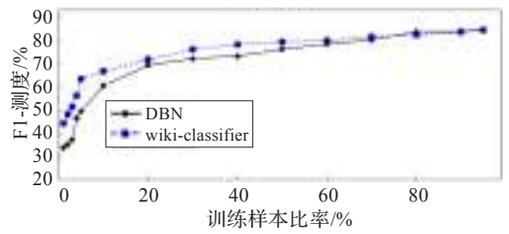

图4 本文方法和空间向量模型分类F1-测度

5 结语

在文本数据的分析挖掘中,文本的表示起到了关键的作用,从传统的文本表示模型(Bag of Words)向语义表示模型(基于WordNet、HowNet等)演变。同时,在通过机器学习建模中,如何将人类知识引入到模型中,不仅仅是让模型学习,而是上模型接受“课程学习(Curriculum Learning)[11-12]”。

参考文献:

[1] Roth D,Ji H,Chang M W,et al.Wikification and beyond:the challenges of entity and concept grounding[C]//Tutorial at the 52nd Annual Meeting of the Association for Computational Linguistics,Baltimore,2014.

[2] Lan M,Tan C L,Su J,et al.Supervised and traditional term weighting methods for automatic text categorization[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(4):721-735.

[3] Altınçay H,Erenel Z.Analytical evaluation of term weighting schemes for text categorization[J].Journal of Pattern Recognition Letters,2010,31(11):1310-1323.

[4] Jain A K,Li Y H.Classification of text documents[J].The Computer Journal,1998,41:537-546.

[5] Hotho A,Maedche A,Staab S.Ontology-based text clustering[C]//Proceedings of International Joint Conference on Artificial Intelligence,2001:30-37.

[6] Wei C P,Yang C C,Lin C M.A latent semantic indexing-based approach to multilingual document clustering[J].Journal of Decision Support System,2009,45(2):606-620.

[7] Blei D,Ng A,Jordan M.Latent Dirichlet allocation[J].Journal of Machine Learning Research,2003,3:993-1022.

[8] Milios E,Zhang Y,He B,et al.Automatic term extraction and document similarity in special text corpora[C]//Proceedings of 6th Conference of the Pacific Association for Computational Linguistics,2003:275-284.

[9] Choudhary B,Bhattacharyya P.Text clustering using universal networking language representation[C]//11th International World Wide Web Conference,2003.

[10] Deep belief networks.[EB\OL].[2014-5-10].http://www.deeplearning.net/tutorial/DBN.html.

[11] 黄浩军.一种基于维基百科的文本分类算法[D].北京:北京大学,2014.

[12] Bengio Y,Louradour J,Collobert R,et al.Curriculum learning[C]//Proceedings of the 26th International Conference on Machine Learning,Montreal,2009.

[13] Huang Hongzhao,Cao Yunbo,Huang Xiaojiang,et al.Collective tweet wikification based on semi-supervised graph regularization[C]//Proc 52nd Annual Meeting of the Association for Computational Linguistics,Baltimore,2014.

[14] Cucerzan S.Large-scale named entity disambiguation based on wikipedia data[C]//EMNLP-CoNLL,Prague,2007.

[15] Milne D,Witten I H.Learning to link with Wikipedia[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management,Napa Valley,2008:509-518.

[16] Chaney A J B,Blei D.Visualizing topic models[C]//ICWSM,2012.