融入内部语义关系对文本分类的影响研究

朱建林[1] 杨小平[2] 彭鲸桥[2]

(中国人民大学财政金融学院 北京 100083)[1] (中国人民大学信息学院 北京 100083)[2]

摘 要 为了在不加入外部语义知识的前提下改善向量空间模型的文本分类效果,通过挖掘语料库内部蕴含的词间关系和文本间关系,并以不同的方式融入原始的词文本矩阵,然后选择常用的SVM和KNN算法,在领域性较强的法律语料库和领域性较宽泛的新闻语料库上进行文本分类的对比实验。实验证明,加入词间关系和文本间关系通常能有效改善文本分类的效果,但是对不同的分类方法和领域特征有不同的影响,在实际应用中应该区别对待。

关键词 向量空间模型,文本分类,语义挖掘,特征矩阵

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.015

Research on Effect of Adding Internal Semantic Relationship into Text Categorization

ZHU Jian-lin[1] YANG Xiao-ping[2] PENG Jing-qiao[2]

(The School of Finance, Renmin University of China, Beijing 100083, China)[1]

(School of Information, Renmin University of China, Beijing 100083, China)[2]

Abstract In order to improve the effect of text categorization on the premise of no addition of the external knowledge, this paper presented a feature matrix-based categorization framework. First, the internal knowledge of corpus is mined and added into the original word-text matrix in different ways. Two common algorithms named SVM and KNN are chosen for contrastive experiment of text categorization in highly territorial legal corpus and domain-wide news corpus. Experi-mental results show that it is generally helpful when adding the semantic relationships extracted from corpus into the original matrix, but the adding method should be chosen according to different classification methods and domain chara-cteristics.

Keywords Vector space model, Text categorization, Semantic mining, Feature matrix

1 引言

文本表示模型是信息检索、文本挖掘领域的基础性问题,其中向量空间模型[1]因为形式简单直观、表达能力好,得到了广泛应用。但是,该模型也存在一些缺陷,例如词文本矩阵一般具有高维稀疏性,该模型基于词独立假设,忽略了词间关系等。本文试图在不引入外部知识的前提下,挖掘词间关系和文本间关系,并将其融入原始矩阵,通过对比SVM和KNN的分类效果,研究在具有不同领域特征的语料库上进行文本分类时以不同方式融入语义关系对分类效果的影响。研究内容属于信息检索和文本挖掘等领域的常用基础性方法,因此它具有一定的理论与应用价值。为了改进模型的表达能力,提高文本分类的准确率,研究人员做了很多工作,解决方法可分为两种:引入外部知识和挖掘内部知识。

引入外部知识是指引入语料库之外的背景知识,如通用本体、领域词典等。例如,Hoth等[3]在文本聚类中引入了基于WordNet[4]的背景知识,弥补了词间的语义关系,明显提高了聚类效果;Bloehdorn等[5]使用医疗领域本体提高文本分类和聚类的效果;Gabrilovich等[6]和Huang等[7]利用维基百科将语料库中的词映射为概念,改善了文本相似度计算的效果;Cilibrasi等[8]依据谷歌返回的搜索结果,定义词的"谷歌相似距离",并将其应用于层次聚类、分类和语言翻译等方面。

挖掘内部语义关系法是指利用语料中的词共现现象,计算词间或文本间的关系,将其融入文本表示模型。例如,Deerwester等[9]在20世纪90年代提出了潜在语义模型(Latent Semantic Indexing),通过奇异值分解(Singular-Value Decomposition)挖掘词和文本的内在语义关系;基于文献[9]的工作,Kontostathis和Pottenger[10]把词间关系分为一阶共现、二阶共现和三阶共现,还发现二阶词共现和奇异值分解有很强的相关性,并给出了数学证明;Chen等[11]把文本中不常见的词换为常见的同义词,从而改善了矩阵的高维稀疏性,提高了分类准确率;Figueiredo等[12]把多个特征词组合为区分能力更强的"复合特征",以提高向量的表达能力;Baker[13]、Yang[14]、Forman[15]、Seifert[17]、Lewis[18]、He[23]、Zhang[24]等通过特征选择,压缩特征空间,提高文本的表示能力。

Zelikovitz等[16]把语料库中未出现的领域特征词加入词文本矩阵,得到更高维度的矩阵,再用LSI和SVD降维,并进行分类,实验证明引入外部知识对分类效果的影响显著,尤其当训练集较小时。本文在不引入外部知识的情况下,挖掘语

到稿日期:2015-07-20 返修日期:2015-11-20 本文受国家自然科学基金(71271209),北京市自然科学基金(4132067),教育部人文社会科学青年基金(11YJC630268),河北省自然科学基金项目(A2013410011)资助。

朱建林(1979-),男,博士,讲师,主要研究方向为语义分析、基于知识的推理,E-mail:linjie_zhu@126.com;杨小平(1956-),男,博士,教授,博士生导师,主要研究方向为语义分析、情感挖掘、Web可用性分析等;彭鲸桥(1989-),男,硕士,主要研究方向为语义分析。

料内部蕴含的词间和文本间语义关系，将其融入词文本矩阵，以期提高矩阵的表达能力。

2 融入语义关系的文本表示模型

本文提出的文本表示模型不使用外部知识，拟通过原始矩阵计算出词间关系和文本间关系，并将其融入原始矩阵来改善矩阵的表达能力。首先，将原始的词文本矩阵定义为 X_0，其中行代表某个词的向量表示，列代表某个文本的向量表示，元素表示某词出现在某文本中的 TFIDF 值。

2.1 词相似度矩阵

词的相似度可定义为词向量的夹角余弦值，即如果词相似度矩阵为 Y_0，那么矩阵中的元素 y_{ij} 可由式(1)计算得到：

$$y_{ij}=\cos\langle row\_x_i, row\_x_j\rangle=\frac{row\_x_i * row\_x_j}{|row\_x_i| * |row\_x_j|} \quad (1)$$

其中，y_{ij} 表示 Y_0 中第 i 行第 j 列的值，也表示 X_0 中词 i 和词 j 的相似度，row_x_i，row_x_j 表示矩阵 X_0 中第 i 和 j 行的向量。

2.2 文本相似度矩阵

与式(1)类似，如果将文本相似度矩阵表示为 Z_0，矩阵中的元素 z_{rt} 可由式(2)计算：

$$z_{rt}=\cos\langle column\_x_r, column\_x_t\rangle=\frac{column\_x_r * column\_x_t}{|column\_x_r| * |column\_x_t|} \quad (2)$$

其中，z_{rt} 表示文本 r 与文本 t 的相似度，$column\_x_r$ 和 $column\_x_t$ 表示 X_0 中的第 r 列和第 t 列。

2.3 融入词和文本关系的文本表示模型

目前，我们得到了 3 个基本矩阵：X_0 表示原始词-文本矩阵，Y_0 表示词间关系矩阵，Z_0 表示文本间关系矩阵。为了将语义关系融入原始矩阵，做了如下尝试：

$$X_1=Y_0X_0 \quad (3)$$

$$X_2=X_0Z_0 \quad (4)$$

$$X_3=Y_0X_0Z_0 \quad (5)$$

$$X_4=X_0X_0^TX_0 \quad (6)$$

X_1 等于词-词矩阵乘以词-文本矩阵，相当于把词间语义关系加入原始矩阵。X_2 等于词-文本矩阵乘以文本-文本矩阵，即将文本间关系加入原始矩阵。X_3 等于词-词矩阵乘以词-文本矩阵，再乘以文本-文本矩阵，相当于同时融入了词间关系和文本间关系。X_4 比较特殊，文献[9,19]通过对 LSI 模型的最大似然估计发现，词相似度矩阵可以近似地用 $X_0X_0^T$ 表示，文本相似度矩阵可以用 $X_0^TX_0$ 表示，由矩阵乘法的结合律可知 $X_4=(X_0X_0^T)X_0=X_0(X_0^TX_0)$，因此，矩阵 X_4 也相当于融入了词间关系和文本间关系。其中，4 个矩阵的维度都与 X_0 一致，并都在稀疏性方面较 X_0 有所改善。

为降低文本长度对向量的影响，本文在矩阵相乘中进行了归一化处理。例如，在计算 $X_1=Y_0X_0$ 时，先对 Y_0 的行向量做归一化处理，再对 X_0 的列向量做归一化处理，然后再相乘。对于任意一个向量 v_i，归一化公式定义如下：

$$L_2(v_i)=\frac{v_i}{\sqrt{\sum_{i=1}^{n}v_i^2}} \quad (7)$$

3 实验

3.1 实验设计

本文使用了两个语料库，一个是自建的法律语料库，另一个是中文文本分类语料库 TanCorpV1.0。法律语料库由 4 种罪行共计 1200 篇判决文书组成。TanCorpV1.0 是谭松波等提供的一个开放中文文本分类语料库，本文也选取了 4 类共计 1389 篇新闻。然后，通过分词、去除停用词等预处理过程，计算得到两个语料库的 X_0 到 X_4。

本文选用 SVM 和 KNN 进行分类实验，通过分类效果来衡量这 5 种矩阵的表达能力，并测评特征选择、降维和归一化等方法对矩阵表达能力的影响。

实验中，训练集和测试集按约 3∶1 分配，在法律数据集中，随机选取 900 篇训练和 300 篇测试，在 TanCorpV1.0 中随机选取 1039 篇训练和 350 篇测试。然后，各进行 3 次重复实验，准确率取 3 次实验的平均值。

3.2 实验结果与分析

3.2.1 分类效果对比分析

为了对比 5 种矩阵对分类效果的影响，分别在法律语料库和 TanCorpV1.0 上做了 SVM 与 KNN 的 4 组分类实验，情况如图 1—图 4 所示。

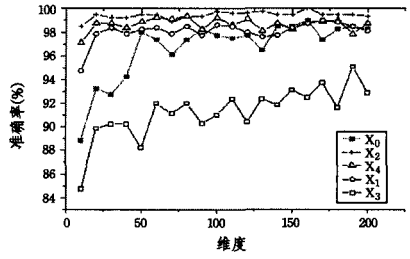

图 1 法律数据集上 SVM 的分类效果(以词频做特征选择，非负矩阵分解降维)

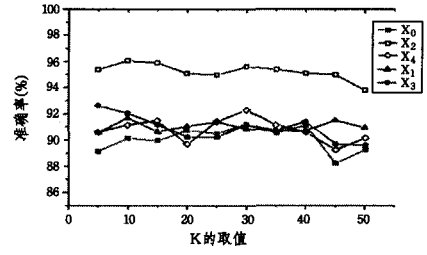

图 2 法律数据集上 KNN 的分类效果(用词频做特征选择)

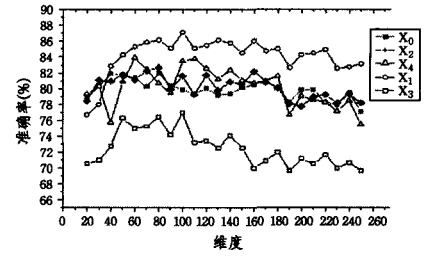

图 3 TanCorpV1.0 上 SVM 的分类效果(用词频做特征选择，奇异值分解降维)

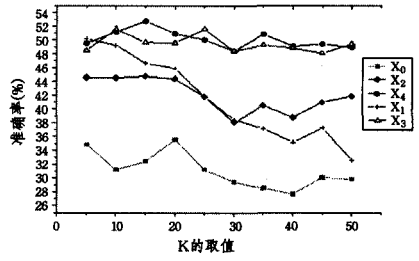

图 4 TanCorpV1.0 上 KNN 的分类效果(用词频做特征选择)

图 1 中，在法律语料库上，用词频做特征选择，非负矩阵

分解降维，SVM 分类的效果：与 X_0 的分类效果相比，X_3 下降明显且一直最差，X_1，X_2，X_4 有明显的提高，X_2 始终最好，尤其当维度较低时差异明显。

图 3 中，在 TanCorpV1.0 上，用词频做特征选择，奇异值分解降维，SVM 分类的效果：与 X_0 的分类效果相比，X_3 有明显的下降且一直最差，X_1，X_2 和 X_4 有不同程度的提高，X_1 效果最好。

图 2 中，在法律语料库上，用 KNN 分类的效果：与 X_0 的分类效果相比，X_1 到 X_4 有不同程度的提高，尤其当维度较小时，提高明显。其中，X_2 的效果最好，且性能稳定。

图 4 中，在 TanCorpV1.0 上，用 KNN 分类的效果：与图 2中法律语料上的实验结果相比，X_1-X_4 对 X_0 分类效果的提高表现得更明显。X_3，X_4 的整体表现更优，X_1 随着 K 的增加下降较快。

对比图 1 与图 3，两组实验都用 SVM 实现分类，用词频做特征选择，选用了不同的降维方法，在不同的语料库上完成。从实验结果分析，SVM 对矩阵中加入语义信息反应敏感，X_1，X_2，X_4 效果更好，X_3 的效果最差。效果变好的原因易于理解，因为加入了更多语义信息。X_3 最差的原因是元素中加入了过多的语义信息，导致降低了元素本身的区分能力。即 $X_3=Y_0X_0Z_0=X_1Z_0$，X_1 的每个元素相当于融入了所有词间关系，再乘以 Z_0 相当于 X_3 的每个元素都融入了所有词和所有文本的关系，同时也就降低了元素的区分能力，而 SVM 是基于最大化类别边界来分类的，所以对元素的区分能力反应敏感。

对比图 2 与图 4，两组实验都用 KNN 算法，在不同语料库上实现分类。从实验结果分析，KNN 算法对矩阵中加入语义信息反应敏感，X_1-X_4 都有更好的表现。图 2 中加入了文本关系的 X_2 效果最好，图 4 中加入了词和文本关系的 X_3 与 X_4 效果更好。X_2 在图 2 中最优，在图 4 中却表现平平，这是因为图 2 中法律数据集的领域性更强，即当语料库的领域性较强时，词之间的共现关系在原始矩阵中已经表现得很强，再融入词间关系时敏感性反而不高，因此文本关系的加入对增加矩阵的语义信息更有价值。图 4 中是新闻语料库，领域更宽泛，词间关系和文本间关系在原始矩阵中表现得都较弱，因此此时同时加入两种关系对分类效果的提高更明显。

对比图 1、图 3 与图 2、图 4，SVM 与 KNN 虽然对矩阵中加入语义信息的反应都很敏感，但是变化略有不同。SVM 是基于最大化类别边界来分类的，它对元素的区分能力反应敏感；即在对元素区分能力影响较小的情况下，加入语义信息能有效地提高矩阵的分类能力；但当向元素中加入过多信息而影响其固有的区分能力时，会降低矩阵的分类能力。KNN 是基于 K 近邻样本来分类的，所以它对样本分布的准确性更敏感，而对元素本身的区分能力要求不高，因此它更希望矩阵中包含更多的语义信息。

对比图 1、图 2 与图 3、图 4，语料库的特点对矩阵表达能力的影响也显而易见。法律语料库因为领域性较强，词共现现象突出，所以原始矩阵中已经包含了较多的词间关系，这时强调文本间关系更能提高矩阵的表达能力。新闻语料库因为领域宽泛，所以原始矩阵中词间关系较弱，因此强调分类边界的 SVM 对词间关系的加入更敏感，而强调样本分布的 KNN 则希望矩阵中包含更多的语义信息。

3.2.2 归一化对分类效果的影响

为了验证归一化对分类效果的影响，选取 TanCorpV1.0 中的 X_1、X_2 和法律数据集中的 X_3、X_4 作归一化对比实验。其中，法律数据集采用 SVM 分类，非负矩阵分解降维。TanCorpV1.0 采用 KNN 分类，结果如图 5 和图 6 所示。

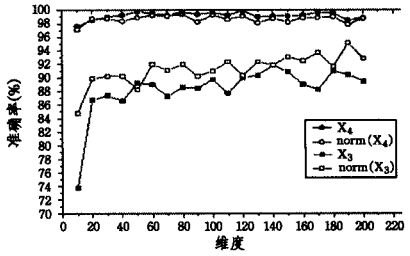

图 5　法律数据集上 SVM 的分类效果(是否归一化)

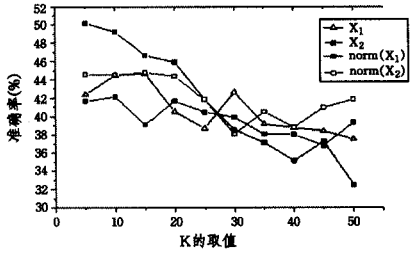

图 6　TanCorpV1.0 上 KNN 的分类效果(是否归一化)

图 5 中，在法律数据集上，使用非负矩阵分解降维，SVM 分类的效果：X_3 归一化后，分类效果改善明显；X_4 归一化前分类准确率已达 99%左右，归一化后准确率略有下降，但整体来看并未影响 X_4 的表达能力，分类效果依然较好。

图 6 中，在 TanCorpV1.0 上，使用奇异值分解降维，KNN 分类的效果：当 K 较小时，归一化对分类效果的改善明显，但随着 K 值的增加，下降速度也更快。

结合图 5 和图 6 可得，归一化能改善 SVM 和 KNN 的分类效果，提高矩阵的表达能力。

3.2.3 降维方法对 SVM 的影响

为降低 SVM 的计算量，需要对矩阵做降维处理，所以为了测评降维方法对 SVM 的影响，本节在法律数据集上分别用奇异值分解(SVD)和主成分分析(PCA)做降维处理，并与图 1 中的非负矩阵分解降维的分类效果进行对比分析。结果如图 7 和图 8 所示。

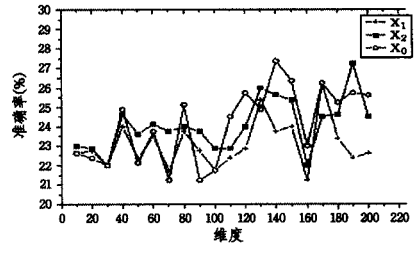

图 7　法律数据集上 SVM 的分类效果(用 SVD 降维)

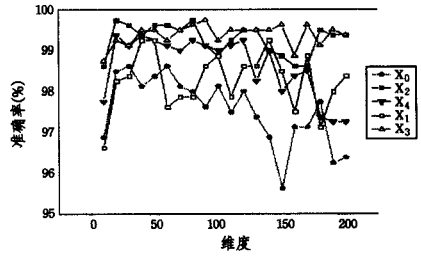

图 8　法律数据集上 SVM 的分类效果(用 PCA 降维)

图 1 中，用非负矩阵分解降维的准确率较高，基本在

90%以上。图7中,用SVD降维的准确率不超过30%。图8中,用PCA降维的准确率最高,一般高于95%。这表明在使用SVM分类时,同一数据集使用不同的降维方法会影响分类效果。其中,使用PCA降维的分类效果优于NMF和SVD,原因是PCA在从高维映射到低维时,样本被广泛散布,使得样本区别更明显,类别边界也更清晰,而SVM是根据类别边界分类的,所以两者合作的分类效果更好。

3.2.4 特征选择对KNN的影响

为了测评特征选择对KNN的影响,在TanCorpV1.0上选择信息增益和开方检验实现特征选择,并与图4中用词频做特征选择的实验结果进行比较,结果如图9和图10所示。

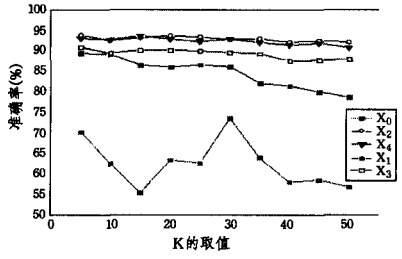

图9 TanCorpV1.0上的KNN分类效果(用信息增益特征选择)

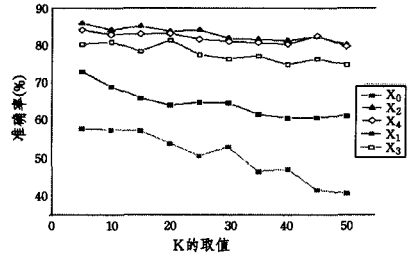

图10 TanCorpV1.0上的KNN分类效果(用开方检验特征选择)

由实验结果可见:1)特征选择方法对KNN分类的准确率有较大影响,词频方式最差,信息增益最好,卡方检验较好。2)虽然分类效果不同,但加入了语义信息的 X_1-X_4 对分类效果的提升作用明显。

结束语 通过上述实验结果与分析,可以得出如下结论:

(1)SVM对加入语义信息反应敏感,分类效果一般会改善,尤其当维度较小时。在对元素区分能力影响较小的情况下,加入语义信息能有效地提高矩阵的分类能力。但因为SVM是基于最大化类别边界来分类的,所以向矩阵元素中加入过多语义信息而影响其的区分能力时,会降低矩阵的分类能力。

(2)KNN对加入语义信息反应敏感,分类效果都会得到改善。KNN是基于K近邻进行分类的,所以向矩阵中加入语义信息会提高样本分布的准确性,改善分类效果。

(3)在专业领域语料库中,因原始矩阵已经包含了较多的词间关系,故单纯强调文本间关系对文本分类的改善更明显。在领域宽泛语料库中,因原始矩阵中词间和文本间的关系都很弱,所以依据边界分类的SVM对词间关系的加入更敏感,而基于K近邻的KNN则希望加入更多的语义信息,即同时加入两种语义信息时分类效果更好。

(4)无论加入语义信息与否,在分类前对矩阵进行归一化操作能有效地改善SVM和KNN的分类效果。

(5)无论加入语义信息与否,降维方法对SVM分类效果有明显影响,且PCA和NMF要明显优于SVD。

(6)无论加入语义信息与否,特征选择对KNN分类效果有明显影响,基于信息增益和卡方检验要明显优于基于词频。

(7)与Zelikovitz等人[16]的工作相比,本文挖掘的是语料库内部语义,没有引入外部知识库,实现起来更简单、方便。

综上,将语料库内部蕴含的词间和文本间语义关系融入原始词文本矩阵一般会改善文本分类的效果,但是会因为分类方法、领域特征、归一化、特征选择或降维方法等不同而不同,在实际应用中应该根据具体情况选择融入语义关系的方法,以期获得更好的分类效果。

参考文献

[1] Salton G, Yang C S. On the specification of term values in automatic indexing[J]. Journal of Documentation, 1973, 29(4): 351-372

[2] Alfred R, Anthony P, Alias S, et al. Enrichment of BOW Representation with Syntactic and Semantic Background Knowledge [M] // Soft Computing Applications and Intelligent Systems. Springer Berlin Heidelberg, 2013: 283-292

[3] Hotho A, Staab S, Stumme G. Ontologies improve text document clustering[C] // Third IEEE International Conference on Data Mining, 2003(ICDM 2003). IEEE, 2003: 541-544

[4] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41

[5] Bloehdorn S, Cimiano P, Hotho A. Learning ontologies to improve text clustering and classification[M] // From Data and Information Analysis to Knowledge Engineering. Springer Berlin Heidelberg, 2006: 334-341

[6] Gabrilovich E, Markovitch S. Wikipedia-based semantic interpretation for natural language processing[J]. Journal of Artificial Intelligence Research, 2009, 34(2): 443-498

[7] Huang A, Milne D, Frank E, et al. Clustering documents using a Wikipedia-based concept representation [M] // Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2009: 628-636

[8] Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

[9] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis[J]. JASIS, 1990, 41(6): 391-407

[10] Kontostathis A, Pottenger W M. A framework for understanding Latent Semantic Indexing (LSI) performance[J]. Information Processing & Management, 2006, 42(1): 56-73

[11] Chen M, Weinberger K Q, Sha F. An alternative text representation to TF-IDF and Bag-of-Words[J]. arXiv preprint arXiv: 1301.6770, 2013

[12] Figueiredo F, Rocha L, Couto T, et al. Word co-occurrence features for text classification[J]. Information Systems, 2011, 36(5): 843-858

[13] Baker L D, McCallum A K. Distributional clustering of words for text classification[C] // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998: 96-103

[14] Yang Y, Pedersen J O. A comparative study on feature selection

in text categorization[C]// Fourteenth International Conference on Mechine Learning. 1997:412-420

[15] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. The Journal of Machine Learning Research,2003,3(2):1289-1305

[16] Zelikovitz S, Hirsh H. Using LSI for text classification in the presence of background text[C]// Proceedings of the Tenth International Conference on Information and Knowledge Management. ACM,2001:113-118

[17] Seifert C,Ulbrich E,Kern R,et al. Text Representation for Efficient Document Annotation[J]. J. UCS,2013,19(3):383-405

[18] Lewis D D. Feature selection and feature extraction for text categorization[C]// Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, 1992:212-217

[19] Ding C H Q. A similarity-based probability model for latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1999:58-65

[20] Lee D D,Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature,1999,401(6755):788-791

[21] Tan S B, Wang Y F. Chinese text categorization corps-TanCorp-V1. 0[OL]. [2014-4-13]. http://www. searchforum. org. cn/tansongbo/corpus. htm

[22] Zhang H P. The Chinese academy of sciences segmentation kit [OL]. [2014-4-13]. http://www. ictclas. org

[23] He L,Wang Z Y,Jia Y,et al. Category candidate search in large scale hierarchical classification[J]. Chinese Journal of Computers,2014,31(1):41-49

[24] Zhang Yu-fang,Wang Yong,Liu Ming,et al. New feature selection approach for text categorization[J]. Computer Engineering and Applications,2013,49(5):132-135

(上接第 70 页)

并且将用户影响力细化到各个领域,根据领域特点进行相应的算法改进。

参 考 文 献

[1] Lin Jia-li, Li Zhen-yu, Wang Dong, et al. Analysis and Comparison of Interaction Patternsin online Social Network and Social Media[C]// Proc of the 21st International Conference on Computer Communications and Networks. Munich, Germany, 2012: 1-7

[2] Wu Xin-dong, Li Yi, Li Lei. Influence Analysis of Online Social Networks[J]. Chinese Journal of Computers, 2014, 37(4): 735-752(in Chinese)
吴信东,李毅,李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014,37(4):735-752

[3] Statistic Report of the 35th China Internet Developing Situation [R]. Beijing: China Internet Network Information Center, 2015 (in Chinese)
第 35 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心,2015

[4] Zhang Qun-yan, Ma Hai-xin, Qian Wei-ning, et al. Duplicate Detection for Identifying Social Spamin Microblogs[C]// Proc of the IEEE International Congress on Big Data. Santa Clara, CA 2013:141-148

[5] Yang Chang-chun, Yu Ke-fei, Ye Shi-ren, et al. New Assessment Method on Influence of Bloggers in Community of Chinese Microblog[J]. Computer Engineering and Applications, 2012, 48 (25):229-233(in Chinese)
杨长春,俞克非,叶施仁,等. 一种新的中文微博社区博主影响力的评估方法[J]. 计算机工程与应用,2012,48(25):229-233

[6] Liang Qiu-shi, Wu Yi-lei, Feng Lei. User Ranking Algorithm for Microblog Search Based on MapReduce[J]. Journal of Computer Applications, 2012, 32(11): 2989-2293(in Chinese)
梁秋实,吴一雷,封磊. 基于 MapReduce 的微博用户搜索排名算法[J]. 计算机应用,2012,32(11):2989-2993

[7] Tang Fei-long, Ye Shi-ren, Xiao Chun. Blogger Influence Ranking Algorithm Based on User Quality in Sina Microblog Community[J]. Computer Engineering and Applications, 2015, 51 (4):128-132(in Chinese)
唐飞龙,叶施仁,肖春. 基于用户质量的微博社区博主影响力排序算法[J]. 计算机工程与应用,2015,51(4):128-132

[8] Meeyoung C, Hamed H, Fabricio B, et al. Measuring User Influence in Twitter: the Million Follower Fallacy[C]// Procof the 4th International AAAI Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2010: 10-17

[9] Brin S, Page L. The Anatomy of a Large Scale Hypertextual Web Search Engine[C]// Proc of the 7th International World Wide Web Conference. Brisbane: ACM Press, 1998: 107-117

[10] Cao Shan-shan, Wang Chong. Improved PageRank Algorithm Based on Links and User Feedback[J]. Computer Science, 2014, 41(12): 179-182(in Chinese)
曹珊珊,王冲. 基于网页链接与用户反馈的 PageRank 算法改进研究[J]. 计算机科学,2014,41(12):179-182

[11] Chen Xiao-fei, Wang Yi-tong, Feng Xiao-jun. An Improvement of PageRank Algorithm Based on Page Quality[J]. Journal of Computer Research and Development, 2009, 46 (Suppl.): 381-387(in Chinese)
陈小飞,王轶彤,冯小军,一种基于网页质量的 PageRank 算法改进[J]. 计算机研究与发展,2009,46(增刊):381-387

[12] Apache Hadoop[OL]. http://hadoop. apache. org

[13] Lammel R. Google's MapReduce Programming Model Revised [J]. Science of Computer Programming, 2007, 68(3): 208-237

[14] Srirama S N, Jakovits P, Vainikko E. Adapting Scientific Computing Problems to Clouds Using MapReduce[J]. Future Generations Computer Systems, 2012, 28(1): 184-192

[15] Chen Gong, Niu Qin-zhou. Research on PageRank Algorithm Based on MapReduce[J]. Microelectronics & Computer, 2012, 29(5): 81-85(in Chinese)
陈宫,牛秦洲. 基于 MapReduce 的 PageRank 算法的研究[J]. 微电子学与计算机,2012,29(5):81-85

[16] Chen Hao, Die Ge. MicroBlog User Ranking Research Based on Hadoop[D]. Shanghai: East China University of Science and Technology, 2014(in Chinese)
陈浩,迭戈. 基于 Hadoop 的微博用户影响力排名算法研究[D]. 上海:华东理工大学,2014