一种嵌入分布信息的 Web文档相似性度量

孙春红, 杨 明

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] Web文档间的相似性度量是 Web文本分类的关键, 有效的相似性度量策略可改进 Web文本分类的精度. 经典的向量空间模型 (VSM)仅考虑网页中单词的出现频率, 未有效利用单词的分布信息, 因而影响了网页的分类精度. 论文计算了网页中单词分布位置的均值和方差, 并将之引入到网页的相似性计算中, 提出了一种直接嵌入分布信息的新的网页相似性度量方法. 该方法因合理利用单词的出现频率及其分布信息, 可有效改进和拓展经典的网页相似性度量策略. 实验结果表明, 该网页相似性度量方法是有效可行的.

[关键词] Web网页的相似性度量, VSM, 分布信息, Web网页分类

[中图分类号] TP391. 1 [文献标识码] A [文章编号] 1672-1292(2008) 03-0066-05

A Novel Similarity Measurement for Web Pages by Incorporating Distribution Information

Sun Chunhong Yang Ming

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract The similarity measurement for Web pages is a key issue for Web pages categorization. Effective similarity measurement strategies can efficiently improve the accuracy of Web pages classification. Traditional Vector Space Model (VSM) only uses the frequency of each selected word in the pages, does not make efficient use of the distribution information such as the average position and bias of the word, hence the method has a great impact on the accuracy of the pages classification. Therefore, in this paper, the means and variances of the words in the document, which are applied into the similarity measurement method, are computed, and a novel method for the similarity measurement of Web pages, that is directly embedded by the distribution information, is present. This approach can effectively improve and extend the classically similarity measurement strategies for Web pages, which properly incorporates the distribution information into the similarity measurement of Web pages. Experimental results show that the method of this paper is efficient and flexible.

Key words similarity measurement of Web pages; VSM; distribution information; Web page categorization

为使用户快速浏览一个大规模 Web文档数据集或从大量的 Web文档中发现感兴趣的网页, 研究者运用机器学习的方法, 对 Web文档进行有效的分类和聚类[1,2], 其中经典的 Web文档分类方法有 Naive Bayes, 决策树、神经网络等[3]. 同时, 研究者还提出了一些有效的基于 SVM 和 AdaBoost的网页分类算法[4,5], 这些算法一定程度上改进了网页的分类效果. 然而, 这些算法主要是从分类器设计的角度出发, 通过模型优化等策略来改进 Web网页的分类精度, 而采用的 Web文档表示为经典的 VSM 模型, 忽略了 Web文档的其它有用信息.

VSM 采用一个向量表示一个文档[6], 即用单词在文档中出现的频率来表示向量特征空间的一维或分量. 由于 VSM 仅考虑单个单词在文档中出现的频率, 未考虑其在文档中的分布情况, 因而在某些情况下使得基于 VSM 的网页相似性度量的分类精度下降, 如 here you are 和 you are here 两个句子, 由于单词的摆放位置不同, 使得句子的含义大相径庭. 此外, 一个单词出现在标题中和出现在文档正文中甚至出现在

收稿日期: 2008-01-25.
基金项目: 国家自然科学基金 (40771163)资助项目.
通讯联系人: 杨 明, 教授, 博士, 研究方向: 数据挖掘、机器学习和粗糙集理论及应用研究. E-mail: myang@ njnu. edu. cn

段首、段尾对整篇文档来说其重要性是不同的, 即单词在文档中的出现位置, 将对文本的分类效果有很大的影响. 为有效利用 Web 文档相应单词的位置等分布特征信息, 文献[7]引入了 分布特征 的概念, 并提出相应的 Web 分类器设计策略, 但该策略未将位置等分布信息直接用于 Web 网页的相似性度量.

为此, 本文在采用位置等分布信息建立扩展向量空间模型(Extended Vector Space Model, EVSM)后, 提出一种新的基于 EVSM 的 Web 文档相似性度量方法, 该方法有效嵌入位置的分布特征信息到文档的相似性度量中, 因而可有效改进和拓展经典的网页相似性度量策略, 提高 Web 网页的分类精度. 此外, EVSM 也为 Web 文档的知识表示提供了一个新的途径. 实验结果表明, 本文提出的网页相似性度量方法是有效可行的.

1 相关工作

为有效利用单词出现的先后次序对 Web 网页分类造成的影响, Lewis David D 将文档看作是单词的序列[8], 即假设一个文档是关于(或不关于)某个主题的条件下, 文档出现的概率就变成文档中关于(或不关于)该主题的每个单词的概率的乘积. 然而, 由此形成的 文档轮廓 通常不易于清楚的分割. 为此, Sauban 和 Phahringer[9] 在此基础上提出了一个新的文档表示方法, 该方法首先计算每个单词的区分值; 继而根据每个单词的输入顺序, 将一篇文档表示为一条曲线, 称之为 文档轮廓 ; 进而根据一个固定的距离将这个轮廓转化为数值特征. 可见, 该方法没有从单词中抽取出新的特征, 也没有提出单词的位置信息, 但考虑到了一篇文档中单词的输入序列, 因而一定程度上可改进经典 VSM 的文档表示, 增强文档相似性度量的有效性.

为更有效地利用 Web 文档提供的位置等分布信息, 文献[7]引入了单词的 分布特征 的概念, 以此描述了一个单词在文档中的分布. 所谓 分布特征 是指单词在文档中首次出现的位置及其密集程度, 有关 分布特征 的描述如下:

设文档 D 中包含 n 个句子, 单词 t 的分布序列是 array= $[c_0, c_1, \ldots, c_{n-1}]$, c_i 为单词 t 在句子 $i+1$ 中出现的次数; 单词 t 的紧缩度(ComPact)和首次出现位置(FirstApp)分别定义为:

$$\mathrm{count}(t, d) = \sum_{i=0}^{n} c_i, \tag{1}$$

$$\mathrm{centroid}(t, d) = \sum_{i=0}^{n-1} \frac{c_i \cdot i}{\mathrm{count}(t, d)}, \tag{2}$$

$$\mathrm{ComPact}(t, d) = \frac{\sum_{i=0}^{n-1} c_i \cdot |i - \mathrm{centroid}(t, d)|}{\mathrm{count}(t, d)}, \tag{3}$$

$$\mathrm{FirstApp}(t, d) = \min_{i=0,\ldots,n-1} (c_i > 0 \mid i). \tag{4}$$

标准的 TFIDF 公式表示如下:

$$\mathrm{TFIDF}(t, d) = \mathrm{Importace}(t, d) \cdot \mathrm{IDF}(t).$$

当使用不同的特征时, $\mathrm{Importance}(t, d)$ 相应使用不同的值. 当特征指某单词出现的频率时, 使用 TF; 当特征指某单词的密集程度时, 使用 CP; 当特征指某单词的首次出现位置时, 使用 FA. TF、CP、FA 分别如下计算:

$$\mathrm{TF}(t, d) = \frac{\mathrm{count}(t, d)}{\mathrm{size}(d)}, \tag{5}$$

$$\mathrm{CP}(t, d) = \frac{\mathrm{ComPact}(t, d) + 1}{\mathrm{len}(d)}, \tag{6}$$

$$\mathrm{FA}(t, d) = 1 - \frac{\mathrm{FirstApp}(t, d)}{\mathrm{len}(d)}. \tag{7}$$

式中, $\mathrm{size}(d)$ 和 $\mathrm{len}(d)$ 是分别表示文档 D 中单词的总个数及文档 D 中句子的总个数. 可见, 分布特征 有效考虑了单词的位置信息. 实验结果证明, 分布特征 可有效改进文档分类的效果. 然而, 文献[7]提出的方法未能直接将单词的位置信息运用于 Web 文档的相似性度量. 为此, 本文在引入 Web 网页的扩展向量空间模型表示(EVSM)后, 提出一种新的网页相似性度量(a Novel Similarity Measurement for Web

pages, NSM), 以使文档中单词的位置信息更直接地影响文档的相似度度量.

2　嵌入分布信息的相似性度量

2 1　*VSM* 及其经典的相似性度量

经典的 W eb文档表示方法采用向量空间模型, 即对某个文档 D_i, 其向量 $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \cdots, d_{in})$, 其中 $d_{ik}(k = 1, 2, \cdots, n)$ 为相应单词在 D_i 中出现的频率信息.

对给定的两个文档 D_i 和 D_j, 相应的向量分别为 $\boldsymbol{d}_i$ 和 $\boldsymbol{d}_j$, D_i 和 D_j 的相似度就可以借助于向量之间的某种距离来表示, 其中经典的度量方法采用向量之间夹角的余弦函数来计算:

$$\mathrm{SM}(\boldsymbol{d}_i, \boldsymbol{d}_j) = \cos(\boldsymbol{d}_i, \boldsymbol{d}_j) = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{|\boldsymbol{d}_i| \cdot |\boldsymbol{d}_j|}, \tag{8}$$

可以看出, 根据式(8), here you are 和 you are here 这两个句子是完全相似的, 显然这不符合实际应用的需要. 造成度量失准的主要原因是: 该度量策略未能有效地利用各单词的分布特征信息. 同理, 对于两个文档, 若其各单词出现的次数相同, 则依据式(8), 它们是完全相似的, 这不符合人们的直觉. 因此, 如何利用式(5)、(6)、(7)提取到的文档分布特征信息进行有效的网页相似性度量, 是本文的重要研究内容.

2 2　*EVSM* 及 *NSM*

为克服经典的 VSM 表示 W eb文档造成的不足, 本文采用文档中各单词出现的频率、平均位置及相应的方差对应的 3组向量来表示一个文档. 为方便计, 本文将文档的这种表示简称为扩展向量空间模型(Extended Vector Space Model, EVSM).

对给定的文档 D_i, 其 3组向量分别为 $\boldsymbol{d}_i = \{v_{i1}, \cdots, v_{in}\}$, $\boldsymbol{\mu}_i = \{\mu_{i1}, \cdots, \mu_{in}\}$, $\boldsymbol{\sigma}_i = \{\sigma_{i1}, \cdots, \sigma_{in}\}$. 其中, v_{ij} 为文档中各单词在 D_i 中出现的频率, μ_{ij} 和 σ_{ij} 分别是相应单词的平均位置和方差组成的向量, n 是向量的维数. 有关单词的平均位置和方差的计算策略见下面的公式.

设文档 D 中有 n 个句子, 单词 t 在每个句子中出现的次数为 c_i, 则单词出现的平均位置 $\mu(t, d)$ 和方差 $\sigma(t, d)$ 分别为:

$$\mu(t, d) = \frac{\sum_{i=0}^{n-1} c_i \cdot i}{\mathrm{count}(t, d)}, \tag{9}$$

$$\sigma(t, d) = \sum_{i=0}^{n-1} \frac{c_i \cdot (i - \mu(t, d))^2}{\mathrm{count}(t, d)}. \tag{10}$$

为有效地将网页中单词的位置和方差这些分布特征信息嵌入到网页相似性度量中, 本文提出一种基于 EVSM 的文档相似性度量策略. 设文档 D_i 和 D_j 为给定的 2个 W eb网页, 其中 D_i 的 3组向量分别为 $\boldsymbol{d}_i = \{v_{i1}, \cdots, v_{in}\}$, $\boldsymbol{\mu}_i = \{\mu_{i1}, \cdots, \mu_{in}\}$, $\boldsymbol{\sigma}_i = \{\sigma_{i1}, \cdots, \sigma_{in}\}$, n 是向量的维数; D_j 的 3组向量分别为 $\boldsymbol{d}_j = \{v_{j1}, \cdots, v_{jn}\}$, $\boldsymbol{\mu}_j = \{\mu_{j1}, \cdots, \mu_{jn}\}$, $\boldsymbol{\sigma}_j = \{\sigma_{j1}, \cdots, \sigma_{jn}\}$, n 是向量的维数; 则新的基于 EVSM 的网页相似性度量策略函数表示如下:

$$\mathrm{NSM}(\boldsymbol{d}_i, \boldsymbol{d}_j) = \frac{\sum_{k=1}^{n} w_{ijk} \cdot v_{ik} \cdot v_{jk}}{\sqrt{\sum_{k=1}^{n} v_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} v_{jk}^2}}, \tag{11}$$

其中,

$$w_{ijk} = e^{-\frac{(\mu_{ik} - \mu_{jk})^2}{b_1}} \cdot e^{-\frac{(\sigma_{ik} - \sigma_{jk})^2}{b_2}} \quad (b_1, b_2 \neq 0), \tag{12}$$

这里, b_1, b_2 是一个可以控制 $\mu(t, d)$ 和 $\sigma(t, d)$ 对相似度函数影响程度的可调参数, 该参数的值可通过计算得到, 但为降低计算的复杂度, 本文设 $b_1 = b_2$. 由式(11)和(12)可得性质 1.

性质 1　对任意给定的两个 W eb文档 D_i 和 D_j, 有 $0 \leqslant \mathrm{NSM}(\boldsymbol{d}_i, \boldsymbol{d}_j) \leqslant 1$ 成立.

证明　由式(12)可知 $0 < w_{ijk} \leqslant 1$, 因而有 $0 \leqslant \mathrm{NSM}(\boldsymbol{d}_i, \boldsymbol{d}_j) \leqslant \cos(\boldsymbol{d}_i, \boldsymbol{d}_j)$, 而 $\cos(\boldsymbol{d}_i, \boldsymbol{d}_j) \leqslant 1$, 故有 $0 \leqslant \mathrm{NSM}(\boldsymbol{d}_i, \boldsymbol{d}_j) \leqslant 1$ 成立. 证毕.

由式 (11) 可以看出, 文档 D_i 和 D_j 越相似, 则 w_{ij} 越大, NSM ($\boldsymbol{d}_i$, $\boldsymbol{d}_j$) 值越大; 反之亦然. 特别地, 当文档 D_i, D_j 完全一致时, w_{ij} = 1 可见, 新的网页相似性度量是基于 VSM 的相似性度量的推广和扩展, 也是基于 VSM 相似性度量的有效改进.

以上分析可知, 本文提出的新的相似度计算方法, 能更直接地体现单词在文档中的位置信息及其对文档相似性的影响, 从而更直接地影响到网页分类的效果.

3 试验及结果分析

为进一步验证本文提出的网页相似性度量的有效性, 本文采用 WebKB、Newsgroups和 Reuters 3个数据集, 通过 K-means和 KNN (K-NearestNeighbors) 算法进行算法性能的测试.

3.1 数据集

WebKB数据集[10]包含着从 4个学校获得的 8 282个 Web网页, 本文采用了其中的 4 000个网页. Newsgroups数据集包含从新闻组收集到的 19 997篇文章, Schapire和 Singer删除了其中的副本[7], 并把多标签的文章用 Xrefs 标记出来, 最后还有 19 465篇. 本文使用了其中的 5类共 50 000个文档. Reuters数据集包含 21 578个取自路透社新闻专线的文章, 本文抽取了 3 680篇. 3个数据集分别随机取其中的 50% 作为训练集, 另外 50% 作为测试集.

3.2 评价方法

为了更准确度量 NSM 方法的有效性, 试验中采用了 F1-measure方法.

$$\text{Recall}=\frac{\text{符合正确分类且被分类器分到此类别的文档个数}}{\text{该类别的文档个数}} \tag{13}$$

$$\text{Precision}=\frac{\text{符合正确类别且被分类器正确分到此类别的文档数}}{\text{系统所判别的属于该类别的文档个数}} \tag{14}$$

$$\text{F1-measure}=\frac{2\ \text{Recall}\ \text{Precision}}{(\text{Recall}+\text{Precision})} \tag{15}$$

对于多类别的问题, 一般采用平均的方法: 微平均 (micro-average)和宏平均 (macro-average). 微平均统一计算全部类别的召回率、准确率和 F1值, 即从整体上平均. 宏平均计算每一类的召回率、准确率和 F1值后取算术平均值. 宏平均值更多地受到稀有类别 (包含文档较少或出现概率较小的类别)的影响.

3.3 实验结果

为参数调整方便, 取式 (12) 中的系数 $b_1 = b_2 = b$. 算法在 WebKB、Newsgroups和 Reuters 3个数据集上的错误率如图 1、图 2所示. 从图中可以看出, b的取值不同, 对算法性能有很大影响. 对于 b_1 和 b_2 分别取不同值时对算法性能的影响将是下一步研究的一个方向. DK-means, DKNN分别代表采用周志华教授有关分布特征的向量模型的算法. IK-means, IKNN 分别表示本文提出的改进算法.

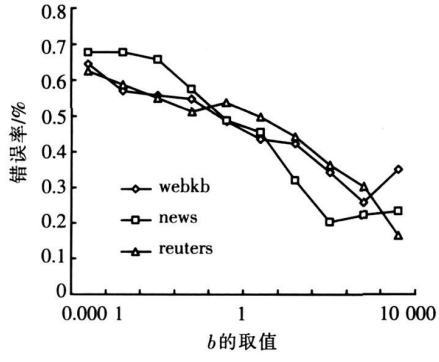

图 1 b 值的变化对 IK-means 算法精度的影响

Fig. 1 Value of b in the IK-means algorithm

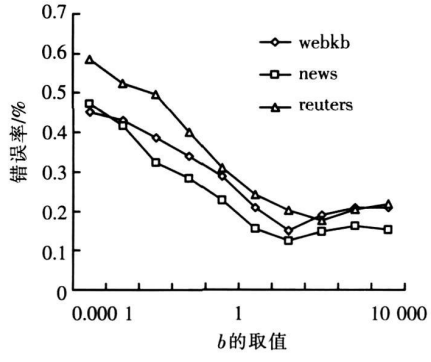

图 2 b 值的变化对 IKNN 算法精度的影响

Fig. 2 Value of b in the IKNN algorithm

图 3、图 4列出了 K-means, KNN 算法及其改进算法在 3个数据集上的分类效果.

表 1、表 2列出了本章中各算法在 3个数据集的精度.

由图 3、图 4及表 1、表 2可见, 在 Newsgroups和 Reuters数据集上, 采用本文新相似性度量的算法性能

优于基于 VSM 的相似函数的算法性能. WebKB数据集上,算法效果不明显甚至有所降低,原因是 WebKB 数据集中的文档是普通的网页,根据通常的写作习惯,单词越早被提及,与文档主题的相关性越大[7], FA 能有效地体现这一特征. 但本文提出的文档相似性度量仅考虑单词的均值和方差信息,而未考虑单词出现的先后次序对文档相似性的影响. 因此,如何融合单词出现的先后次序到文档相似性度量中是未来研究的内容之一.

表 1　不同数据集上的 K-means及其改进算法

Table 1　The performance of K-means in the different dataset

	K-means		DK-means		IK-means	
	mic-fl	mae-fl	mie-fl	mac-fl	mic-fl	mae-fl
WebKB	0.781 75	0.782 926	0.863 992	0.863 329	0.736 952	0.742 381
Newsgroups	0.736 327	0.735 484	0.781 362	0.783 136	0.804 225	0.805 103
Reuters	0.791 248	0.792 329	0.804 161	0.802 779	0.828 232	0.821 94

表 2　不同数据集上的 KNN及其改进算法

Table 2　The performance of KNN in the different dataset

	KNN		DKNN		IKNN	
	mic-fl	mae-fl	mie-fl	mac-fl	mic-fl	mae-fl
WebKB	0.761 069	0.759 444	0.841 27	0.840 346	0.865 35	0.868 77
Newsgroups	0.801 87	0.802 785	0.821 742	0.823 891	0.882 73	0.880 366
Reuters	0.805 859	0.804 144	0.821 941	0.822 202	0.839 378	0.832 586

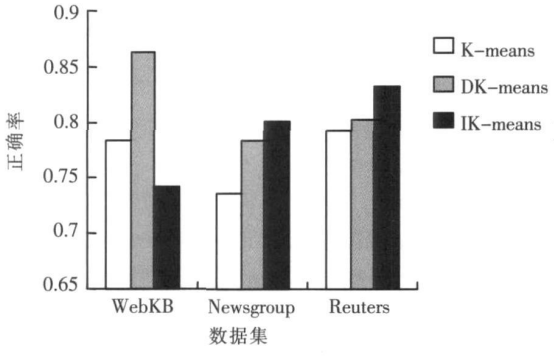

图 3　K-means 及其改进算法性能比较

Fig.3　The performance of K-means and the improved algorithm

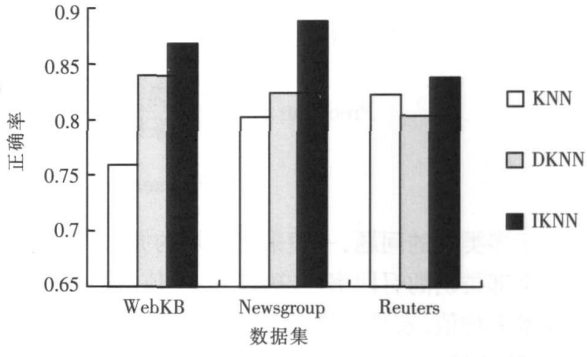

图 4　KNN 及其改进算法性能比较

Fig.4　The performance of KNN and the improved algorithm

4　结语

本文引入一种新的扩展向量空间模型 EVSM 后,提出一种直接嵌入分布信息的新的网页相似性度量方法. 该方法合理利用了单词的出现频率及其分布信息,可有效改进和拓展经典的网页相似性度量,为 Web网页的表示和相似性度量提供一种新的途径. 实验结果表明,本文提出的网页相似性度量方法是有效可行的.

[参考文献] (References)

[1] Cui Zifeng, Xu Baowen, Zhang Weifeng, et al. Web documents clustering with interest links[C] // Service-Oriented System Engineering, IEEE International Workshop, 2005: 111-116.

[2] Zeng Huajun, He Qicai, Chen Zhen, et al. Learning to cluster web search results[C] // Proceedings of SIGIR'04. Sheffield, 2004: 210-217.

[3] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Survey, 2002, 34(1): 1-47.

[4] Joachims T. Text categorization with support vector machines: Learning with many relevant features[C] // Proceedings of ECML-98. Chemnitz, 1998: 137-142.

(下转第 76页)

the 7th European Conference on Machine Learning C. Berlin: Springer, 1994: 171-182.
[15] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems[C] // Slowinski R. Intelligent Decision Support: The Netherlands: Kluwer Academic Press, 1992: 331-362.
[16] Gupta K M, David W A, Philip M. Rough set feature selection algorithms for textual case-based classification[C] // Roth-Berghofer T R. ECCBR 2006, Lecture Notes in Artificial Intelligence, 2006, 4106: 166-181.
[17] Yang Ming, Chen Songcan, Yang Xubing. A novel approach of rough set-based attribute reduction using fuzzy discernibility matrix[C] // Proceedings of 4th International Conference on Fuzzy Systems and Knowledge Discovery. Washington, DC: IEEE Computer Society, 2007: 96-101.
[18] Yang M, Yang P. A novel condensing tree structure for rough set feature selection[J]. Neurocomputing, 2008, 71(4-6): 1092-1100.

[责任编辑: 严海琳]

(上接第 70页)

[5] Schapire R E, Singer Y. Boostexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2/3): 135-168.
[6] Lu Yuchang, Lu Mingyu, Li Fan. Analysis and construction of word weighing function in VSM[J]. Journal of Computer Research & Development, 2002, 39(10): 1205-1210.
[7] Xue Xiaobing, Zhou Zhihua. Distributional features for text categorization[C] // Proceedings of the 17th European Conference on Machine Learning (ECML'06). Berlin: LNAI 4212, 2006: 497-508.
[8] Lewis D D. Naive(Bayes) at forty: The independence assumption in information retrieval[C] // Proceedings of 10th European Conf on Machine Learning. Berlin: Springer, 1998: 4-15.
[9] Sauban M, Pfahringer B. Text categorization using document profiling[C] // Proceedings of PKDD-2003. Berlin: Springer-Verlag, 2003: 411-412.
[10] Craven M, DiPasquo D, Freitag D, et al. Learning to extract symbolic knowledge from the World Wide Web[C] // Proceedings of AAAI-98. Madison: WI, 1998: 509-516.

[责任编辑: 严海琳]