第16卷 第3期
2006年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 3
Mar. 2006

一种基于TFIDF的网络聊天关键词提取算法

许晓昕,李安贵
(北京科技大学 应用科学学院 数力系,北京 100083)

摘 要:随着Internet的普及,即时通讯软件(IM software)也就是网络聊天软件越来越多地服务于人们的日常生活。利用聊天双方的聊天信息来提供更好的服务成为研究者们的重要课题,而如何提取聊天文本中的关键词又成为此类研究的重点。聊天文本不同于普通的文章,它是一种动态输入的文本,对于这种文本,传统的TFIDF算法存在着缺陷。文中针对传统TFIDF在处理此类文本时的不足之处,利用一个按主题分类的历史缓存来提高TFIDF算法对于这类文本的处理能力。

关键词:TFIDF;文本挖掘;即时通讯软件;关键词提取
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)03-0122-02

A New TFIDF-Based Chat Stream Keyword Extraction Algorithm

XU Xiao-xin, LI An-gui
(Department of Mathematics and Mechanics, School of Applied Science,
Beijing University of Science and Technology, Beijing 100083, China)

Abstract: By the common use of Internet, IM software has more and more affected people's life. How to take advantage of chat text to serve people and how to extract keywords from this text has attracted more and more researchers. Chat text is different from article text. Only using TFIDF algorithm to extract keywords is not well. In this paper, a history cache is introduced to improve the performance of TFIDF in chat text.

Key words: TFIDF; text mining; IM software; keywords extraction

0 引 言

文本挖掘是近年来新兴的,同时也是备受学者们关注的一个领域。它与传统的数据挖掘有本质的区别。由于传统的数据库都有一定的模型,可以根据此模型来具体描述特定数据,故传统的数据挖掘所处理的数据是结构化的。而文本数据是非结构化或半结构化的,为将其转化成特征向量至少需要几万甚至几十万个特征。所以,文本挖掘技术首要解决的问题就是如何合理地、便利地表示出这些文本特征,用尽可能少的特征表达出尽可能多的信息量。

对于文章(一般大于200个词)可以利用TFIDF[1,2](Term Frequency Inverse Document Frequency)算法来进行关键词的识别,其识别效果已经在实际应用中得到过证明。而网络聊天文本是一种简短的、不完整的,却是上下文相关的动态文本。通过实验可以知道,对于网络聊天文本传统的TFIDF算法的性能并不是很好。

文中在TFIDF的基础上提出一种缓存历史的TFIDF算法。这种算法通过缓存按主题划分的文本特征来动态地识别聊天者的聊天主题,以聊天主题为依据,产生出更能代表聊天者意图的关键词。

收稿日期:2005-06-14

作者简介:许晓昕(1982—),男,云南昆明人,硕士研究生,研究方向为模糊数学及计算机软件;李安贵,教授,研究方向为模糊数学。

1 传统的TFIDF算法

传统的TFIDF算法是由Gerald Salton和McGill针对向量空间信息检索范例(Vector space information retrieval paradigm)提出的文档特征表示方法。即一篇文档或一段文本利用特征向量表示,亦即文本被看作是一系列项 t 的集合。对每个项 t,可以加上一个对应的权值,这样文档就由形如〈项 t,权值 W〉的对组成。项$(t_1,t_2,t_3,\cdots,t_n)$代表文档内容的特征项,可以看作一个 n 维的坐标系。权值 $w_1,w_2,w_3,\cdots,w_n$ 表示对应的坐标值,每篇文档 d 都可映射成此空间上的一个特征向量 $V(d)=(t_1,w_1,t_2,w_2,\cdots,t_n,w_n)$。

在此方法中,出现在文档中的文字称为Term(术语),每一个Term对应的权重 w 代表Term在文档识别时的重要程度。Term的权重与Term在文档中出现的频率成正比,而与Term在所有文档中出现的频率成反比,即著名的TFIDF公式:

$$W_j = TF_j \times \mathrm{IDF} \tag{1}$$

其中 W_j 是文档中Term_j的权重;TF_j 表示Term_j在当前文档中出现的频数;DF是在所有文档中Term_j出现的频数,IDF就是DF的倒数,一般来说$\mathrm{IDF}=\log(|D|/\mathrm{DF}_j)$,其

中,D是文档总数。

这种算法,既突出了文档中出现频数较高的词,又消去了在各文档中出现次数都很高的常用词的影响。对于单词数大于200词的静态文本,其效果比较令人满意。

2 缓存历史的TFIDF算法

仅使用TFIDF算法对网络聊天文本关键词进行提取,效果并不是很理想。究其原因,我们认为,由于网络聊天文本是一种动态的文本。其特点是,每一次的输入并不会像一篇文章一样的完整,同时这种文本也是简短文本。因此,对这种文本进行实时的关键词提取时,每一次词出现的频数即TF值并不能很好地反映文本中词的重要程度[3]。

针对这一种情况,利用一个集合 H 来对一段时间之内的聊天信息进行缓存[4],利用缓存的聊天信息来计算新输入文本的TF值。但是这种算法对历史的缓存方式太过于简单,在应用中并没有达到要求。

众所周知,网络聊天都是围绕着一个或者几个特定的主题来进行的。而一个主题可以利用一个词的集合来近似地表示,比如:{ computer, network, software, hardware, website,…} 这样的一个集合可以表示"Information Technology(信息科技)"这样的一个主题。即,一个主题可以由一个文档特征向量来表示。因此,在提取聊天文本关键词的时候,可利用 k 个代表不同主题的文档特征向量 S_i 来缓存历史。而历史可以定义为:

$$H = \{S_1, w_1, S_2, w_2, \cdots, S_n, w_n\}$$

其中 $S_k = (t_{k1}, w_{k1}, t_{k2}, w_{k2}, \cdots, t_{kn}, w_{kn}) \in V(d)$,

$w_i = \max(w_{i1}, w_{i2}, \cdots, w_{in})$。

在对关键词进行分类的时候,同一次输入的文本属于相同的主题。

算法实现如下:

(1)输入文本,利用TFIDF公式选择并生成文本特征向量 V。

(2)利用式(2)计算 $V = (v_1, w_1, v_2, w_2, \cdots, v_n, w_n)$ 与每一个主题 S_i 的相似度 fsim,

$$\text{fsim}_k = \sum_{i=1\cdots n, j=1\cdots n} (v_i * t_j) \qquad (2)$$

其中,$v_i * t_j = \begin{cases} 1, & \text{若 } v_i = t_j \\ 0, & \text{若 } v_i \neq t_i \end{cases}$,$v_i \in V, t_j \in S_k, w_j \in S_k$ 为权值。

(3)若fsim的最大值大于某一个给定的常数,则将 V 加入到fsim最大的主题中,否则认为 V 代表的是一个新的主题,以 V 为特征向量生成一个新的主题 S_{k+1} 加入到 H 中。

(4)对 H 中的词进行衰退处理,若 $t \in H, t \notin V, t$ 的权值 $w = w \times 0.9$,并将 H 中权值低于某一阈值的词删除。

(5)根据权值对 H 中的元素进行排序,并利用算术平均来重新计算 S_k 中 t_i 的权值,即 $w_{ki} = 0.5w_{ki} + 0.5w_i$。

(6)输出 H 中权值最高的几个词作为本次文本输入的关键词。

3 实验结果分析

笔者对TFIDF算法和改进后的算法进行对比实验,在实验中,将从文献[5]中取出100个英文对话共618个句子作为测试样本。

把测试源以对话和句子的形式分开,逐句输入对话,根据提取的效果人为地把结果分成4个等级:A:相关;B:有些相关;C:不相关;D:历史相关。并以按句和按整段对话的方式根据人的经验进行判断。在实验过程中,主观感觉到,改进后的算法提取出来的关键词可以比较好地表达聊天者当前的意图。实验结果如表1所示。

表1　实验结果表

分类	TFIDF		改进后的TFIDF	
	对话	句子	对话	句子
A:相关	80%	49.35%	80%	26%
B:有些相关	18%	49.03%	20%	22.8%
C:无关	2%	1.62%	0%	3.2%
D:历史相关				48%

从结果中可以看出,改进之后的算法被标记为无关的对话为0%,而被标记为相关的句子为74%(相关和历史相关的总和),远远高出改进前算法的49.35%。

4 结束语

网络聊天软件已经逐渐成为人们工作和生活中的一种常用的通讯工具,从网络聊天文本中进行文本挖掘已经愈来愈受到研究者的关注。文中提出了一种利用缓存按主题分类的文档特征向量,来识别出最能够表达聊天都意图的关键词的算法。通过实验表明,这种算法对于处理网络聊天这一种动态文本具有较好的效果。然而,网络聊天语言的特点之一是不完整性,在未来的研究中,将针对这一点来提高算法的性能,并且期望在不完整输入的情况下,根据上下文关系,猜测出聊天者的意图,产生出更能代表聊天者意图的关键词。

参考文献:

[1] 景丽萍,黄厚宽,石洪波.用于文本挖掘的特征选择方法TFIDF及其改进[J].广西师范大学学报(自然科学版),2003,21(1):142-146.

[2] 李　凡,鲁明羽,陆玉昌.关于文本特征抽取新方法的研究[J].清华大学学报(自然科学版),2001(7):98-101.

[3] Brun A, Smaili K, Jean-Paul H. Experiment Analysis in Newspaper Topic Detection[A]. Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)[C]. Curuna, Spain: IEEE Computer Society, 2000.55-64.

(下转第222页)

的,是从自然界固体退火过程中得到启发并从中抽象得来,是解大规模优化问题的一种随机优化算法。该算法最显著的特征是以一定的概率接受使目标函数值增大的移动,即恶化解,所以能从局部最优解的"陷阱"中爬出来而不是简单地终止在局部最优解上,具有全局收敛性[5]。

基于模拟退火算法的思想,设计适合文中模型的算法如下。

步骤1 $X_{00} \in Q$ 是随机产生的初始解的矩阵,其中:

$$Q = \left\{ \begin{vmatrix} x_{11} \cdots x_{1n} \\ \vdots \quad\quad \vdots \\ x_{k1} \cdots x_{kn} \end{vmatrix}, x_{ij} \in \{0,1\} \text{ 且} \sum x_{ij} = 1 \right\}$$

为可能解的集合,x_{ij} 表示第 i 个服务类中第 j 个服务是否被选中的状态。计算相应的目标函数值 W_0;给出控制参数初值 T_0、马尔柯夫链长度 N 以及停止参数 ε(指当控制参数 T 递减到 ε 时算法停止)。

注:在计算目标函数的值时,因为问题的描述中带有约束条件,所以采用惩罚策略来修正目标函数:

$$F(x) = f(x) \pm R_j \sum_{j=1}^{m} \max[0, g_j(x)]^2 \pm R_k \sum_{k=1}^{l} [h_k(x)]^2 \quad (8)$$

式中:$g_j(x)$ 为不等式约束,$h_k(x)$ 为等式约束,R_j,R_k 为加权大数因子;在最大化问题中取"-",在最小化问题中取"+"。因此,本问题中的目标函数修正为:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n} F_{ij} x_{ij} - R_j * \max[0, (\sum_{i=1}^{k} \sum_{j=1}^{n} r_{ij} x_{ij} - R)]^2 \quad (9)$$

步骤2 产生新解并计算新解与当前解的目标函数值之差 Δf。

新解的产生:

在 $1 \sim n$ 之间随机选取 i,j,当前解中若每个服务类中第 i 个服务和第 j 个服务状态相同,则将其中一个和状态不同的服务交换状态;否则,两者交换其状态。即:

$$\begin{cases} x_{ii} = 1 \text{ 或 } x_{ij} = 1, \text{同时置状态为1的服务状态为0} & x_{ii} = x_{ij} \\ x_{ii} = 1 - x_{ij} \text{ 且 } x_{ij} = 1 - x_{ii} & x_{ii} \neq x_{ij} \end{cases}$$

步骤3 由 Metropolis 接受准则判断是否接受新解。

(1) 若 $\Delta f \geqslant 0$,即新解可行且优于当前解,则接受。

(2) 若 $\Delta f < 0$,则按新点接受概率:

$$p(\Delta f, T) = \exp(\frac{\Delta f}{T})$$

取[0,1]区间上均匀分布的随机数 δ,若 $p(\Delta f, T) \geqslant \delta$,则接受新解,否则放弃新解。

步骤4 累计次数 n,若 $n \leqslant N$ 转步骤2,否则转步骤5。

步骤5 判断停止准则是否满足。若不满足则令 $T_{k+1} = \lambda * T_k$,$n = 1$,转步骤3,否则停止算法输出当前解。

式中:$\lambda = 0.8 \sim 0.999$

4 算法分析

从上述算法可以看出,模拟退火算法由于采用了 Metropolis 接受准则,在搜索解的过程中引入了新的随机因素,使算法进程呈现跳跃性而可能跳离局部最优的"陷阱",从而保证了最终解并不依赖于随机选取的初始解,提高了最终解质量的稳定性;对于算法的时间复杂性,可以借助算法的比较次数进行分析。由控制参数 T 的初值 T_0 和衰减函数 $T_{k+1} = \lambda * T_k$,对 T 的每一取值所进行的 N 次迭代及停止准则所规定的控制参数的终值 ε 构成了控制算法有限时执行的冷却进度表,由它的控制作用,使算法的时间复杂性为问题规模的多项式时间[5]。而采用穷尽搜索时,当问题规模很大时需要指数级时间,因此该算法大大提高了运算效率。

5 总 结

文中针对 Web 服务合成中的研究热点问题,即如何从合成服务的 QoS 角度选择满足用户需求的服务,设计了 QoS 全局最优的数学模型,并进行了适合本模型的模拟退火算法的设计。从算法分析的结果可以看出,本算法对求解 Web 服务的选择问题是合适的,是可以在实际中应用的。对于有向无环图型的合成服务,只要计算每一条执行路径的最大效用值,然后再从中选择效用值最大的那条路径作为最优执行路径即可。

参考文献:

[1] Zeng L, Benatallah B, Ngu A, et al. QoS - Aware Middleware for Web Services Composition[J]. IEEE Transactions on Software Engineering, 2004, 30(5): 311 - 327.

[2] Cardoso J, Bussler C, Sheth A. Semantic Web Services and Processes: Semantic Composition and Quality of Service. tutorial at Federated Conferences (CooPIS, DOA, ODBASE), 2002 [EB/OL]. http://lsdis.cs.uga.edu/lib/presentations/SWSP - tutorial - resource.htm, 2002.

[3] Yu T, Lin K J. Service Selection Algorithms for Web Services with End - to - End QoS Contraints[A]. Proceedings of the International Conference on E - Commerce Technology[C]. San Diego, California: IEEE Computer Society Press, 2004. 129 - 136.

[4] 李裕奇. 随机过程[M]. 北京:国防工业出版社, 2003.

[5] 康立山,谢 云,尤矢勇,等. 非数值并行算法——模拟退火算法[M]. 北京:科学出版社, 1994.

(上接第123页)

[4] Bigi B, De Mori R, El - Beze M, et al. Detecting topic shifts using a cache memory[A]. 5th International Conference on Spoken Language Processing[C]. Sydney, Australia: [s. n.], 1998. 2331 - 2334.

[5] 杨 力. 美国口语大观:中英文对照[M]. 合肥:中国科学技术大学出版社, 2001.