一种对聊天文本进行特征选取的方法研究

李盛瑜[1,2] 何 文[1]

(重庆工商大学 重庆 400067)[1] (重庆大学计算机学院 重庆 400044)[2]

摘 要 由于网络聊天文本具有结构松散、简短、上下文相关等特点,对其进行特征选取时使用传统的 TFIDF(Term Frequency Inverse Document Frequency)算法存在较大缺陷。针对这个问题,本文提出了一种通过聊天主题来确定聊天文本的特征选取范围的方法,并通过实验验证了该方法的有效性。

关键词 文本挖掘,聊天文本,TFIDF

A Study on the Method of Feature Selection in Chat Text

LI Sheng-Yu[1,2] HE Wen[1]

(Chongqing Technology and Business University, Chongqing 400067)[1] (College of Computer, Chongqing University, Chongqng 400044)[2]

Abstract Because online chat text is loosely and briefly organized and is context dependent, there are some defects to select features by using traditional TFIDF (Term Frequency Inverse Document Frequency). Aiming at solving the problem, this paper presents a method that the range of feature selection in chat text is obtained by using chat subjects, and the validity of the method is verified through experiment.

Keywords Text mining, Chat text, TFIDF

1 引言

文本挖掘是一个对具有丰富语义的文本进行分析从而理解其所包含的内容和意义的过程,已经成为数据挖掘中一个日益流行而重要的研究领域。文本挖掘所研究的文本数据库,由来自各种数据源的大量文档组成,包括新闻文章、研究论文、聊天文本、书籍、期刊、报告、专利说明书、会议文献、技术档案、政府出版物、数字图书馆、技术标准、产品样本、电子邮件消息、Web 页面等。这些文档可能包含标题、作者、出版日期、长度等结构化数据,也可能包含摘要和内容等非结构化的文本成分[1],而且这些文档的内容是人类所使用的自然语言,计算机很难处理其语义。因此传统的信息检索技术已不适应日益增加的大量文本数据处理的需要,人们提出文本挖掘的方法进行不同的文档比较,以及文档重要性和相关性排列,或找出多文档的模式或趋势等分析[2]。

对于比较长的文本(如 1000 个以上的单词)可以利用 TFIDF(Term Frequency Inverse Document Frequency)[3]算法来进行文本特征选取,其良好的特征选取效果已经在实际应用中得到过证明[4]。但是,网络聊天文本与传统文档中的文本差别较大:网络聊天文本是一种动态的文本,用户每一次的输入并不会像一篇文章一样完整,也不一定遵循严格的语法结构,是一种上下文相关的动态文本,用户可能某一段时间、某几句话表达某一特定主题,主题与主题之间可能完全不相关,并且每句话所形成的文本也是一种比较简短的文本。因此,对这种文本进行文本特征选取时,某关键词在文本中出现的词频 TF 值并不能很好地反映这个关键词在文本中的重要程度。

2 传统的 TFIDF 算法

反比文档频数权重评价算法(Term Frequency Inverse Document Frequency,TFIDF)是 Gerald Salton 和 Mc Gill 针对向量空间信息检索范例提出的广泛应用在对文本集合进行特征选取的方法[5]。它依据某个词的词频(Term Frequency)和其出现过的文本的频率(Document Frequency)来计算该词在整个文本集合中的权重,然后依据权重来进行特征选取。权重越高,说明该词对该文本集合的区分能力越强,否则其区分能力就越弱。

因此,聊天文本集可利用其特征向量来表示,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 w,这样文本就可由形如(t,w)的对组成。项$(t_1,t_2,t_3,\cdots,t_n)$代表文本内容的特征项,可以看着一个 n 维的坐标系。权值$(w_1,w_2,w_3,\cdots,w_n)$表示文本特征项对应的权重,每个文本集 d 都可映射成此空间上的一个特征向量:

$$V(d)=(t_1,w_1,t_2,w_2,\cdots,t_n,w_n)$$

对于词 t 和某一文本 d 来说,t 在 d 中权重的计算公式为:

$$w(d,t)=tf(d,t)\times\log\left(\frac{|D|}{df(t)}\right)$$

其中,$tf(d,t)$是词 t 在文本 d 中出现的数目,$df(t)$是词 t 在文本集合中出现过的文本的数目,$|D|$是整个文本集合中文本的数目。如果一个词在整个文本集合中出现的频度很高,即 $df(t)$趋于$|D|$时,$\log\left(\frac{|D|}{df(t)}\right)$趋于零,从而使得 $w(d,t)$值很小,即该关键词对文本的区分能力比较弱。根据上面的描述,可以根据 $w(d,t)$值从大到小选择用户指定数目的词作为某篇文本的特征,生成文本的特征向量。

李盛瑜 讲师,硕士生,主要研究领域为数据挖掘技术。何 文 讲师。

这种算法一方面突出了文档中的用户需要的关键词，又消除了在各文档中出现次数较高但对文本语义无关的常用词的影响，对于单词数较多的静态文本进行特征选取效果比较好。

基于此，在对网络聊天文本进行特征选取时，如何确定每次进行特征选取的文本范围将直接影响到挖掘效果。进行某次特征选取的文本范围过大，则可能选取的文本为多个聊天主题，进行特征选取时很难准确描述当前用户聊天语义，文本范围过小，又很难通过上下文描述用户聊天片断的准确主题，这直接影响到对文本的分类[6]。因此用传统 TFIDF 对聊天文本进行特征选取存在较大缺陷。

3 改进的 TFIDF 算法

针对这个问题，我们猜想，在对聊天文本进行特征选取时，如果能确定出聊天的主题，然后在该主题范围内对文本进行特征选取，这样，在该主题下，关键词出现的频数 TF 就能够较好地反映该关键词在聊天过程中的重要程度。因此，我们得到图 1 的改进的 TFIDF 算法模型。

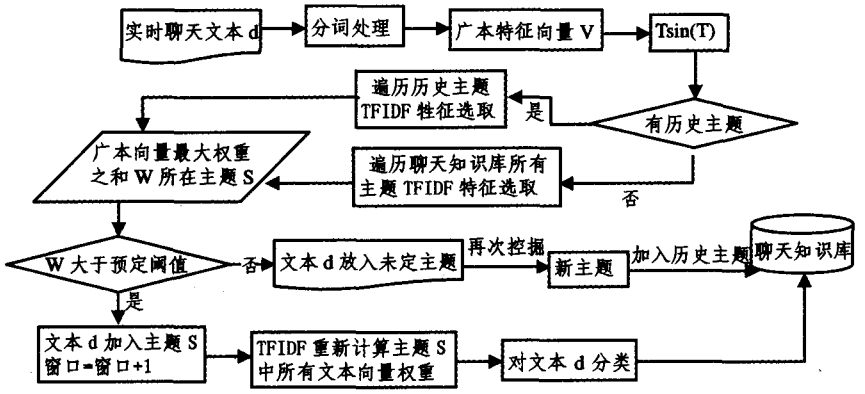

图 1　改进的 TFIDF 算法模型

在该模型中，事先建立一个聊天知识库。假设历史聊天知识库中有 k 个代表不同主题的文档特征向量，即聊天知识库中主题集为：

$$S=\{S_1,w_1,S_2,w_2,\cdots,S_k,W_k\}$$

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

$w_i=\max(w_{i1},w_{i2},\cdots,w_{in})$，为该主题中特征向量值最大的关键词，代表该主题。

对 TFIDF 特征选取算法作如下改进：

1. 输入聊天文本，对该文本进行分词处理后得到其关键词集合 T；

2. 用传统 TFIDF 计算 T 中所有关键词在 S 中的权重，得到其文本特征向量 $V(d)=(t_1,w_1,t_2,w_2,\cdots,t_n,w_n)$；

3. 计算 $V(d)=(t_1,w_1,t_2,w_2,\cdots,t_n,w_n)$ 与每个主题 S_i 的相似度 $TSin(T)$：

$$TSin(T)=\sum_{i=1ton,j=1ton}(v_i*t_j)*w_j$$

其中，$v_i\in V,t_j\in S_k,w_j\in S_k$，若 $v_i=t_j$，则 $v_i*t_j=1$，否则 $v_i*t_j=0$。

4. 若 $TSin(T)$ 的最大值大于某一个给定的常数，则将 $V(d)$ 加入到 $TSin(T)$ 最大的主题 S_x 中，S_x 中的聊天文本样本数增加一个，否则认为 $V(d)$ 代表的是一个新的主题，以 $V(d)$ 为特征向量生成一个新的主题 S_{k+1} 加入到 S 中，该文本成为 S_{k+1} 中的聊天样本。

5. 对加入 $V(d)$ 后的主题 S_x 或 S_{k+1} 中所属文本再次使用 TFIDF 计算其权重，得到新的文本特征向量 $V'(d)=(t'_1,w'_1,t'_2,w'_2,\cdots,t'_n,w'_n)$。

6. 根据权值对 $V'(d)$ 中的权重进行排序，将其中权重低于特定阈值的词去除。最后得到的关键词为本次输入文本的关键词。

7. 用算术平均法修正聊天知识库中该主题关键词权重，即：

$$w_{si}=w_{s1}*0.5+w_{ti}*0.5$$

其中，w_{si} 为聊天知识库中对应主题的某关键词权重，w_{ti} 为该主题下与聊天知识库中对应关键词权重，如果要想让聊天知识库中对应关键词权重变化速度减慢，则可将 w_{si} 比例加大，w_{ti} 比例减小，这可在具体应用中调整。

如果聊天知识库中该主题中无某关键词，则添加该关键词，其权重为在该主题下挖掘出的关键词的权重的 1/2。

8. 聊天结束后，将聊天知识库中的所有关键词的权重都低于预定阈值的新主题删除，并将剩余新主题加入聊天知识库中。

4 实验设计与分析

我们以工商大学网站[7]的某次聊天过程中的聊天文本为测试样本，对 132 句聊天文本用人工分词，并经人工筛选判断得到准确的实际关键词数目为 169 个。使用改进 TFIDF 算法前后进行特征选取得到的关键词用人工方法进行判断为与聊天文本相关情况如表 1。

表 1　TFIDF 改进前后关键词情况

	TFIDF	改进后的 TFIDF
相关	336	350
无关	155	105
历史主题相关	0	38

TFIDF 算法改进前后评估情况如表 2。

表 2　TFIDF 改进前后关的算法评估

	TFIDF	改进后的 TFIDF
查全率	76.02%	91.93%
查准率	68.43%	78.70%

以上测试数据表明，改进后的 TFIDF 算法对文本特征选取的查全率提高了 15.91%，查准率提高了 10.27%，皆优于改进前的 TFIDF 算法。

小结　通过实验表明，这种方法对于处理网络聊天这一

种动态文本具有较好的效果。该方法可用于对聊天过程中的聊天文本进行实时特征选取，也可广泛应用于对短文本进行文本挖掘时的特征选取。但是，随着聊天知识库的增大，如果某新文本为一新主题，遍历聊天知识库查询近似主题将可能非常耗时，并且每次文本特征选取完毕后，对聊天知识库进行修正或改进时同样面临这个问题。不过可将聊天知识库中的所有数据存储在数据库服务器中，通过改进查询算法、使用存储过程等数据库技术来缓解这个问题。

参 考 文 献

1 Han Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. 北京：高等教育出版社，2001. 285～295

2 王继成，潘金贵，张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展，2000(5)：513～520

3 Li Fan, Lu Mingyu, Lu Yuchang. Research about newmethods of text feature extraction[J]. Journal of Tsinghua University (Science and Technology), 2001 (7)：98～101

4 Luo Jie, Chen Li, Xia Delin, et al. Research on Fast Text Classifier Based on New Keywords Extraction Method[J]. Application Research of Computers, 2006(4)：32～34

5 Kantrowitz M, Mohit B, Mittal W. Stemming and its Effects on TFIDF Ranking[J]. In: Proc. of the 23rd annual international ACM SIGIR Conf on Research and development in information retrieval. New York: ACM Press, 2000. 357～359

6 Colliez J, Dufrenois F, Hamad D. Optic Flow Estimation by Support Vector Regression[J]. Engineering Applications of Artificial Intelligence, 2006, 19(7)：761～768

7 http://bbs.ctbu.edu.cn

(上接第 201 页)

分法。内容页一般是用户关心的信息，浏览时间较长。导航页是用户快速找到所需信息而设置的路标，浏览时间较短。通过辅助页面在整个日志中所占的比例的估计，可以使用一个最大可能估计算法来划分辅助页面和内容页面的划分时间。通过对照划分时间，页面就可以划分为内容页或导航页，从而划分为不同的事务。

(2)最大前向引用

有时某些页面含有较多超链接，是用户关心的信息，却被作为内容页，此时可采用 Chen[7] 等人提出的最大向前引用路径(简称为 MFP) 来定义事务。对于每个用户会话，从开始页面为起点，每个最大前向引用路径即为一个事务。这里每一个事务定义为一组页面的访问，从第一次引用，直到在某处向后回溯为止。前向指引定义为一个从未在事务集合中出现的页面，后向指引指已经在前面的事务中出现的一个页面。当一个前向指引出现时，开始一个新的事务。获取最大向前引用序列的算法如下：

剪枝算法描述

Step1：初始化，定义存储最大向前引用序列的字符串变量和循环控制变量。

Step2：扫描数据库，执行连接和剪枝操作操作，输出最大向前引用序列。

Step3：重复执行 Step2，直到没有新的最大向前引用序列为止。

剪枝算法：

输入：大项集阶段转换后的序列数据库

输出：最大序列

Step1：初始化：令 $i=0$，字符串 $S=$ NULL//字符串 T 用来存储当前的向前引用。

Step2：While(s_{i+1}不空且 s_{i+1}比 s_i 更深){连接 s_i 到 S；$i=i+1$；}连接 s_i 到 S；//S 是一个最大向前引用。

把 S 加到数据库中；$i=i+1$；如果 $s_i=$NULL，结束。否则，转到步骤(Step3)。

Step3：while(S_i 比 S_{i+1}大时)//判断是否是向后引用

{$i=i+1$；}$j=i-1$；

while($s[j]$比 $s[i]$大)

{删除 $s[j]$；$j--$；}转到第 Step2。

Step4：End

Answer=Maximal Sequence YS。

2.6 内容和结构数据的预处理

内容和结构数据的预处理是根据具体的应用把 Web 页面中的文本、图像、script 以及 Web 页面间的超链接等数据转化成用于 Web 使用挖掘的格式。例如根据一个 Web 页面的文本内容，描述该页面涉及的概念主题，用于 Web 页面的聚类[8,9]；根据 Web 页面之间的超链接信息构造 Web 站点的拓扑结构图，用于辨识用户。

2.7 数据预处理的结果

经过以上的预处理后，可以得到一个页面集合 $P=\{p_1, p_2, p_k, \cdots, p_n\}$和一个用户事务集合 $T=\{t_1, t_2, \cdots, t_m\}$其中 $t_i \in T$ 是 P 的子集。我们可以把一个事务 t 看成是一个具有 l 长度的序列对 $t=\langle(p_1^t, W(p_1^t)), (p_2^t, W(p_2^t)), \cdots, (p_l^t, W(p_l^t))\rangle$其中 $p_i^t(i=1,2,\cdots,l)$是 P 中的页面，$W(p_i^t)$是页面 p_i^t 在事务中的权重。页面的权重可以根据不同的需要采取不同的策略赋值有两种常用的赋值策略，一种是权值为二进制，表示在一个事务中这个页面是否存在，另一种是使用用户在这个页面上的驻留时间的函数，在协同过滤中，页面的权值是通过用户的评价而计算的。用户的事务可以看成是集合(不考虑用户之间的顺序信息)，也可以看成序列(考虑用户之间的顺序信息)，这需要具体的分析和应用的目标决定。

总结 介绍了 Web 使用挖掘的前期工作——数据预处理的过程，其目的就是尽量使得预处理后的数据比较真实和完整，为后面的数据挖掘打好基础。此外，还可以根据 Web 使用挖掘的目的在此基础上进行改进，为 Web 使用挖掘的预处理提供更好的算法。

参 考 文 献

1 李雄飞，李军. 数据挖掘与知识发现[M]. 北京：高等教育出版社，2003

2 Pyle D. Data Preparation for Data Mining. Morgan Kaufmann Publishers Inc, San Francisco, CA, 1999. 540

3 Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1999, 1(1)：5～32

4 Tan P, Kumar V. Discovery of Web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery, 2002, 6：9～35

5 Jetal S. Web Usage Mining：Discovery and application of usage patterns from Web data[J]. SIGKDD Explorations, 2000, 1(2)：12～23

6 Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns[J]. Journal of Knowledge and Information Systems, 1999, 1 (1)：5～32

7 Chen MS, Park J S, Yu PS. Data Mining for Path Traversal Patterns[A]. In：Proc. of the 16th Int'l Confon Distributed Computing System[C]. Hong Kong, 1996

8 Perkowitz M, Etzioni O. Towards adaptive Web sites：Conceptual framework and case study[J]. Artificial Intelligence, 2000, 118：245～275

9 Etzioni P M. Adaptive Web Sites：Automatically Synthesizing Web Pages[A]. In：Proceedings of Fifteenth National Conference on Artificial Intelligence[C]. Madison, 1997