基于词频统计规律的文本数据预处理方法

池云仙 赵书良 罗 燕 高 琳 赵骏鹏 李 超
(河北师范大学数学与信息科学学院 石家庄 050024)
(河北师范大学河北省计算数学与应用数学重点实验室 石家庄 050024)

摘 要 在大数据时代,文本挖掘面临特征的"高维-稀疏"问题,海量文本词汇与稀少关键特征间的矛盾导致了高时空复杂度和低效率等问题,严重制约了文本挖掘效率,因此在文本挖掘前进行有效的数据预处理至关重要。传统文本挖掘算法在数据预处理阶段只进行分词和去停用词操作。为提高性能,提出基于词频统计规律的文本数据预处理方法。首先,基于齐普夫定律和最大值法推导同频词数表达式;然后,基于同频词数表达式探究各频次词语在文中的分布规律,结果表明词频为1和2的词语与文档的关联度较低,但比重高达2/3;最后,基于词频统计规律进行数据预处理,在预处理阶段去除低频词,减小特征维度。在公共数据集Reuters-21578和20-Newsgroups上进行的实验的结果表明,各频次词语的分布规律是正确的,基于词频统计规律的文本数据预处理方法在分类准确率、精确率、召回率以及F1度量值方面均有提升,运行时间明显降低,文本挖掘效率得到显著提高。

关键词 大数据,文本挖掘,数据预处理,词频统计
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.050

Text Data Preprocessing Based on Term Frequency Statistics Rules

CHI Yun-xian ZHAO Shu-liang LUO Yan GAO Lin ZHAO Jun-peng LI Chao
(College of Mathematics & Information Science, Hebei Normal University, Shijiazhuang 050024, China)
(Hebei Key Laboratory of Computational Mathematics & Applications, Hebei Normal University, Shijiazhuang 050024, China)

Abstract In age of big data, it is a severe problem that feature terms are faced with "high-dimension and sparse" challenge in text mining. Contradiction between enormous scale of terms and scarce of features will cause high-time-space complexity and poor efficiency, and restricts the efficiency of text mining seriously. Thus, it is crucial to preprocess data before mining text. Terms-dividing and stop-words-deleting are operated merely in data preprocessing of traditional text mining algorithms. In order to improve process of data preprocessing, data preprocessing algorithm based on term frequency statistics rules (DPTFSR) was proposed. To begin with, expression about number of terms with identical frequency is deduced based on Zif's Law and rule of maximum area. What's more, regularities of distribution based on terms with identical frequency is explored. It is discovered that proportion of low-frequency terms in documents reach up to 2/3, but there is little relevancy between them. Lastly, data is preprocessed based on terms frequency statistics rules. Low-frequency terms are deleted, and features dimension is decreased greatly. Correctness of term frequency statistics rules and validity of algorithm DPTFSR are verified on data sets from Reuters-21578 and 20-Newgroups. Experimental results show that accuracy, precision, recall and F1 measure are increased, and running time is shortened obviously. Thus, efficiency of text mining is significantly enhanced.

Keywords Big data, Text mining, Data preprocessing, Term frequency statistics

---

到稿日期:2016-07-25 返修日期:2016-09-24 本文受国家自然科学基金项目(71271067),国家社科基金重大项目(13&ZD091),河北省高等学校科学技术研究项目(QN2014196),河北师范大学硕士基金(xj2015003)资助。

池云仙(1987-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:lovebeyond.630@163.com;赵书良(1967-),男,教授,博士生导师,主要研究领域为数据挖掘、智能信息处理,E-mail:zhaoshuliang@sina.com(通信作者);罗 燕(1993-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:luoyan_work@163.com;高 琳(1992-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:gaol520smile@163.com;赵骏鹏(1990-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:zhaojunpeng115@126.com;李 超(1991-),男,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:hbsd_lichao@sina.com。

1　引言

文本挖掘面临特征“高维-稀疏”问题，即词汇规模过大导致较高的时空复杂度，关键词语数目稀少导致稀疏的特征数目，严重降低了文本挖掘效率。多年来，众多学者为提高文本挖掘性能，对文本词汇规律进行了深入的研究。Han 等人[1]提出一种关键词搜索模型，基于语义探求与用户查询意图最相近的兴趣点 POI；Hu 等人[2]提出一种在空间数据上基于用户的查询位置和关键词得到与查询结果相似集合的模糊查询方法；Ren 等人[3]提出一种结合文本语义和文本时效性的多样化结果表示方法；Ding 等人[4]提出一种微博数据挖掘算法，解决了微博文本数据较短和质量较低的问题；Song 等人[5]针对高维数据空间，提出一种基于聚类算法进行快速特征子集选择的方法；ZHAO 等人[6]提出一种无监督特征选择模型，利用图形正则化保留原始数据的局部结构，通过最小化数据重组误差选择保持原始数据相似性和区别信息的最佳特征。

为解决文本挖掘面临的特征“高维-稀疏”问题，选取适当的方法对文本向量空间进行降维对提高文本分类效率至关重要。本文提出基于词频统计规律的文本数据预处理方法(Data Preprocessing based on Term Frequency Statistics Rules，DPTFSR)，研究词频统计规律以找到各频次词语在文档中的分布规律，并利用词频统计规律进行数据预处理。该算法在保证分类精度的前提下，明显降低了时空开销，提高了文本挖掘效率。

本文第 2 节利用齐普夫定律和最大值法推导同频词数 $NTIF_n$ 的表达式；第 3 节依据 $NTIF_n$ 的表达式探究各频次词语在文本中的分布规律；第 4 节基于词频统计规律进行数据预处理；第 5 节对所提方法进行实验分析；最后总结全文。

2　同频词数 $NTIF_n$

定义 1(词频，Term Frequency)　指在文本文档 d 中词语 t_i 出现的次数，记作 TF_n，其中 n 为词频数(即词语 t_i 出现的次数)；$TF_k(t_i)$表示词语 t_i 的词频为 k。

定义 2(同频词数，Number of Terms with Identical Frequency)　指文本文档 d 中具有相同词频的词语总数，记作 $NTIF_n$，其中 n 为词频数。若文本文档 d 中存在一组词语$\{t_1, t_2, \cdots, t_q\}$，满足 $TF_k(t_1)=TF_k(t_2)=\cdots=TF_k(t_q)=TF_k=k$，则 $t_1, t_2, \cdots, t_q$ 为一组词频为 k 的同频词，同频词的总数为 q，即 $NTIF_k=q$。

定义 3(频率，Frequency)　指在文本文档 d 中词语 t_i 出现的频率，即词语 t_i 的出现次数 TF_n 与文本长度 L 的比值，记为 $f_n=TF_n/L$，其中 n 为词频数。

定义 4(词秩，Term Rank)　指对词频 TF_n 的等级排序序号，记作 TR_n。TF_k 与 TR_k 呈一一对应的逆序关系，即当 $TF_k=TF_n$ 时，$TR_k=1$；当 $TF_k=1$ 时，$TR_k=TR_n$。

2.1　利用齐普夫定律推导同频词数表达式

在文本文档中，各频次词语以一定规律分布，词频统计方法利用统计学知识对词汇规律进行描述，齐普夫定律和布次定律是词频统计学方面具有深远影响的两大定律[7-8]。多年来，各领域学者对词频统计规律进行了深入研究，该方法由于简单且实用性强受到众多学者青睐[9-11]。

齐普夫定律(Zipf's Law)由美国学者 Zipf 提出，是文本挖掘领域被广泛应用的文献计量学三大定律之一[12]。

齐普夫定律描述为：给定文本文档 d，L 表示文本 d 的长度(L 足够大)，N_{diff} 表示出现在 d 中的不同词语总数，TF_n 表示 d 中词语的词频(n 为词语在文中的出现次数)，TR_n 表示与 TF_n 相对应的词秩，f_n 表示词语出现的频率，$f_n=TF_n/L$，则：

$$f_n \cdot TR_n = K(K \text{ 为常数}) \tag{1}$$

$$TF_n \leqslant f_n \cdot L < TF_{n+1} \tag{2}$$

其中，$K=1/(\ln N_{diff}+\beta)$(β 为欧拉常数)；K 非定值，其围绕某中心值上下波动[13]。

由齐普夫定律可得以下推论。

推论 1　文本文档 d 的长度为 L，其中出现次数为 n 的词语的词频为 TF_n，出现次数为 $n+1$ 的词语的词频为 TF_{n+1}，则依据齐普夫定律所得的词频为 TF_n 的同频词数 $NTIF_n$ 为：

$$NTIF_n=\frac{K \cdot L}{TF_n \cdot TF_{n+1}} \tag{3}$$

证明：将齐普夫定律式(1)代入式(2)得：

$$TF_n \leqslant \frac{K \cdot L}{TR_n} < TF_{n+1} \tag{4}$$

由此可得，

$$\begin{cases} TRn_{\max} = \dfrac{K \cdot L}{TF_n} \\ TRn_{\min} = \dfrac{K \cdot L}{TF_{n+1}} \end{cases} \tag{5}$$

因此，词频为 TF_n 的同频词数 $NTIF_n$ 满足：

$$NTIF_n = TRn_{\max} - TRn_{\min} \tag{6}$$

将式(5)代入式(6)，即得式(3)所示的同频词数表达式：

$$NTIF_n=\frac{K \cdot L(TF_{n+1}-TF_n)}{TF_n \cdot TF_{n+1}}=\frac{K \cdot L}{TF_n \cdot TF_{n+1}} \tag{7}$$

证毕。

2.2　采用最大值法完善同频词数 $NTIF_n$ 表达式

利用式(3)计算同频词数 $NTIF_n$ 并不能完全适用于词频 TF_n 取任何值的情况，因为其依据是齐普夫定律，但齐普夫定律无法很好地反映词频极低的词语分布规律，当词频 $TF_n=1,2$ 时的波动尤为明显[14]。因此，下文利用最大值法研究 $TF_n=1,2$ 时的同频词计算公式。

推论 2　词秩 TR_n 与词频 TF_n 呈一一对应的逆序关系，采用最大值法确定 TR_n，将词语按词频 TF_n 降序排序，若遇同频词语则顺序随机，词秩取最大值，$N_{\max}$ 为 d 中词语出现的最大频次，那么同频词数 $NTIF_n$ 即为相邻两词秩之间的差值，即：

$$NTIF_n = TR_n - TR_{n+1} \tag{8}$$

依据最大值法得到的词频 $TF_n=1,2$ 时的同频词数 $NTIF_n$ 的表达式为：

$$NTIF_n=\frac{N_{diff}}{TF_n \cdot TF_{n+1}}, n=1,2 \tag{9}$$

其中，N_{diff} 为出现在文档 d 中的不同词语总数。

证明：首先验证齐普夫定律满足最大值法。

由式(1)得：

$$TR_n=\frac{K}{f_n}=\frac{K\cdot L}{TF_n} \tag{10}$$

同理得：

$$TR_{n+1}=\frac{K}{f_{n+1}}=\frac{K\cdot L}{TF_{n+1}} \tag{11}$$

将式(10)、式(11)代入式(8)可得式(3)，显然，基于齐普夫定律推导出的式(3)满足最大值法。

注意：上述验证证实了最大值法也适用于齐普夫定律，但最大值法并非基于齐普夫定律推导而来，因此他不受齐普夫定律的限制。

根据式(8)的最大值法可知，当词频 $TF_n=N_{\max}$ 时，$TR_n=1$；当 $TF_n=1$ 时，$TR_n=N_{diff}$，即 $TR_n\cdot TF_n=N_{diff}(n=1)$。通过对数据集进行统计发现，当 $TF_n=2$ 时，词秩与词频的乘积 $TR_n\cdot TF_n$ 也接近 N_{diff}，即：

$$TR_n\cdot TF_n=N_{diff},n=1,2 \tag{12}$$

将式(12)代入式(8)可得式(9)：

$$NTIF_n=\frac{N_{diff}(TF_{n+1}-TF_n)}{TF_n\cdot TF_{n+1}}=\frac{N_{diff}}{TF_n\cdot TF_{n+1}},n=1,2 \tag{13}$$

证毕。

2.3 同频词数 $NTIF_n$ 的完整表达式

联立式(3)和式(9)得到同频词数 $NTIF_n$ 的完整表达式为：

$$NTIF_n=\begin{cases}\dfrac{K\cdot L}{TF_n\cdot TF_{n+1}}, & n>2\\ \dfrac{N_{diff}}{TF_n\cdot TF_{n+1}}, & n=1,2\end{cases} \tag{14}$$

其中，$K=\dfrac{1}{(\ln N_{diff}+\beta)}$（$\beta$ 为欧拉常数），L 为文本长度，N_{diff} 为不同词语总数，TF_n 为词频，n 为词频 TF_n 的取值。

3 各频次词语在文中的分布规律

3.1 词频分布规律

推论 3 文本文档 d 的长度为 L，不同词语数目为 N_{diff}，其中记出现次数为 n 的词语的词频为 TF_n，依据同频词数 $NTIF_n$ 的完整表达式可得到各频次词语的分布规律：

$$\begin{cases}NTIF_1=\dfrac{1}{2}N_{diff}\\ NTIF_2=\dfrac{1}{6}N_{diff}\\ NTIF_{n>2}=\dfrac{1}{3}N_{diff}\end{cases} \tag{15}$$

证明：对于文本文档 d，词频 TF_n 取遍所有值时，对应的同频词数 $NTIF_n$ 之和等于不同词语总数 N_{diff}，即满足：

$$\sum_{n=1}^{N_{\max}}TF_n=N_{diff} \tag{16}$$

其中，n 表示词频 TF_n 的值，$N_{\max}$ 表示文本中的最大频次(即词频 TF_n 的最大取值)，$1\leqslant n\leqslant N_{\max}$。

当 $n=1,2$ 时，由式(14)得：

$$\frac{NTIF_n}{N_{diff}}=\frac{1}{TF_n\cdot TF_{n+1}},n=1,2 \tag{17}$$

为统计词频 $n=1,2$ 的词语在文中的比重，将 $n>2$ 的词语视为一个整体，从而计算词频大于 2 的词语所占的总比重 $\sum_{n=3}^{N_{\max}}\dfrac{NTIF_n}{N_{diff}}$。

将式(14)代入式(16)得：

$$\sum_{n=1}^{2}NTIF_n+\sum_{n=3}^{N_{\max}}NTIF_n=N_{diff} \tag{18}$$

由式(14)得：

$$\begin{cases}NTIF_1=\dfrac{1}{2}N_{diff}\\ NTIF_2=\dfrac{1}{6}N_{diff}\end{cases} \tag{19}$$

由此得：

$$\sum_{n=1}^{2}NTIF_n=\frac{2}{3}N_{diff} \tag{20}$$

联立式(18)和式(20)得：

$$\sum_{n=3}^{N_{\max}}NTIF_n=N_{diff}-\sum_{n=1}^{2}NTIF_n=\frac{1}{3}N_{diff} \tag{21}$$

联立式(19)和式(21)即得式(15)的各频次词语分布规律，证毕。

各频次词语在文中的分布规律如图 1 所示。

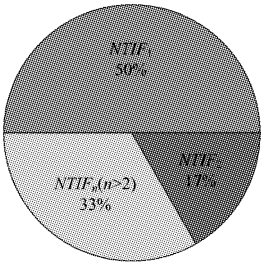


图 1 各频次词语在文中的分布规律

3.2 词频性质

性质 1 特征词语与所在文档的关联度随词语的词频的降低而减小。

证明：TF-IDF 是一种利用词语频率进行信息检索的特征加权及关键词提取方法。基于 TF-IDF 方法计算词语 t_i 的权重为：

$$\begin{aligned}TF\text{-}IDF(t_i)&=TF(t_i,d)\cdot IDF(t_i)\\&=\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF(t_i)}\right)\end{aligned} \tag{22}$$

其中，$TF(t_i,d)=TF_k(t_i,d)/N_{all}$ 为词语 t_i 在文档 d 中出现的频率($TF_k(t_i,d)$ 表示 t_i 在 d 中出现了 k 次，$N_{all}=\sum_{i=1}^{n}TF_k(t_i,d)$ 为 d 中所有词语出现次数的总和)；$IDF(t_i)$ 为反文档频率，其中 N 为文本总数，$DF(t_i)$ 为包含词语 t_i 的文档数。

对于给定的文档 d，所有词语出现次数的总和 N_{all} 为定值。$TF_{k+m}(t_j,d)$ 表示 t_j 在 d 中出现了 $k+m(m\geqslant 0)$ 次，即 $TF_k(t_i,d)=k\leqslant k+m=TF_{k+m}(t_j,d)$。

分以下 3 种情况讨论词语的词频与其 TF-IDF 值间的关系，进而可知词语的词频与文档关联度间的关系。

(1)若 $DF(t_i)=DF_{mid}=DF(t_j)$，即包含 t_i 的文档数与包含 t_j 的文档数目相同，则有：

$$\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right)\leqslant\frac{TF_{k+m}(t_j,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right) \tag{23}$$

即 $TF\text{-}IDF(t_i)\leqslant TF\text{-}IDF(t_j)$。

由此可得：若两个词语所在文档数目相同，则词语对应的 TF-IDF 值随词频数目的减少而减小，即词语的词频越小，权重值越小，与文档的关联度越低。

(2)若 $DF(t_i)<DF_{mid}<DF(t_j)$，则词频较低的 t_i 具有较

低的文档频率 $DF(t_i)$；词频较高的 t_j 具有较高的文档频率 $DF(t_j)$。那么

$$TF\text{-}IDF(t_i)'=\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right)<\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF(t_i)}\right)=TF\text{-}IDF(t_i) \tag{24}$$

$$TF\text{-}IDF(t_j)'=\frac{TF_{k+m}(t_j,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right)>\frac{TF_{k+m}(t_j,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF(t_j)}\right)=TF\text{-}IDF(t_j) \tag{25}$$

由此可得：低频词语 t_i 对应的 TF-IDF 值随文档频率的升高而减小，即低频词语与文档的关联度随所在文档数目增多而减小；高频词语 t_j 对应的 TF-IDF 值随文档频率的降低而增大，即高频词语与文档的关联度随所在文档数目的减少而增大。

(3)若 $DF(t_i)>DF_{mid}>DF(t_j)$，即词频较小的 t_i 具有较高的文档频率 $DF(t_i)$；词频较大的 t_j 具有较低的文档频率 $DF(t_j)$。那么

$$\begin{aligned}&TF\text{-}IDF(t_j)=\frac{TF_{k+m}(t_j,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF(t_j)}\right)\\&>TF\text{-}IDF(t_j)'=\frac{TF_{k+m}(t_j,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right)\\&>TF\text{-}IDF(t_i)'=\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF_{mid}}\right)\\&>TF\text{-}IDF(t_i)=\frac{TF_k(t_i,d)}{N_{all}}\cdot\lg\left(\frac{N}{DF(t_i)}\right)\end{aligned} \tag{26}$$

由此可得：高频词语 t_j 对应的 TF-IDF 值随文档频率的升高而减小，与文档的关联度降低；低频词语 t_i 对应的 TF-IDF 值随文档频率的降低而增大，与文档的关联度提高；但在文档频率相等之前，高频词语对应的 TF-IDF 值始终大于低频词语，即高频词语与文档的关联度高于低频词语。并且，尽管低频词语对应的 TF-IDF 值随所在文档数目的减少而增大，但若低频词语所在文档数目过少，即在整个文档集中均不常出现，则 t_i 很可能为罕见、畸形词，或是与文档关联度较低的词语。

综上所述，特征词语与所在文档的关联度随词语词频的减小而降低。

证毕。

针对低频词与文档关联度间的关系问题，文献[15-16]利用 Microsoft Web N-gram Service 技术获得每个词语的出现概率，进而计算该词语在整个语料库中的分布得分。Q_1，Q_2，Q_3 分别表示 3 个四分位点，即 Q_1 满足 $Pr_{x\in D}(x\leqslant Q_1)=0.25$，$Q_3$ 满足 $Pr_{x\in D}(x\leqslant Q_3)=0.75$，四分位差 $IQR=Q_3-Q_1$ 用来度量中间分布。给定非负参数 a，定义剪除下界：$LF(a)=Q_1-a\cdot IQR$，则得分低于 $LF(a)$ 的为罕见、畸形词，或与文档的关联度较低。将该部分词语进行剪除可提升文本挖掘性能。依据剪除下界对词语进行剪除，词频为 1，2 的低频词基本包含在被剪除词集内。定义非期望剪除率 UPR(即将与文档关联较高的词语进行剪除)与期望剪除率 DPR(即将与文档无关的词语进行剪除)，结果显示所删除词语的 DPR 值较高，UPR 值较低，证明了该剪除下界的正确性。由此可证，低频词与文档关联度较低。

由性质 1 可知，特征词语与所在文档的关联度随词语词频的减小而降低。因此，最小词频(即 $TF_n=1,2$)对应的词语与文档的关联度较低，在进行数据预处理时可进行删除。

通过对同频词语分布规律的研究可知，在相对较长的文本文档中，绝大多数词语的 $TF_n=1,2$(约占全文的 2/3)，这部分低频词在文中的比重较高，但所覆盖的语料库却极低。R. Agrawal 等人也在研究中发现，这些低频词大多与主题的关联度不高，甚至是作为噪声的罕见、畸形词[15-16]，不仅严重制约了文本挖掘的执行效率，甚至会降低挖掘精度。相比之下，词频 $TF_n>2$ 的词语虽仅占全文的 1/3，但覆盖了绝大部分语料库，关键特征大多也包含在这部分词语中。

4　基于词频统计规律的数据预处理方法

数据预处理是影响文本挖掘效率的关键因素之一。大多数已有的文本挖掘算法在数据预处理阶段只进行简单的分词、去停用词操作，但规模庞大的词汇和数目稀少的关键特征词语会带来“高维-稀疏”问题，从而严重降低文本处理效率。通过上述对词频统计规律的研究，确定了文档中各频次词语的分布规律，利用此规律进行数据预处理，去除与文档关联度较低的($TF_n=1,2$)低频词语，可极大提高文本处理性能。

算法 1 给出了基于词频统计规律的文本数据预处理方法。步骤 1 初始化存储词频为 $TF_n=1$，$TF_n=2$，$TF_n>2$ 的字典 *dict*1，*dict*2，*dict*3 及相应的记录不同词频词数的计数器 *count*1，*count*2，*count*3，定义词列表 *TermList* 及计数器 *word_count*；步骤 2－步骤 11 进行分词操作，并记录每个词语的词频；步骤 12－步骤 24 依据不同词频进行归类，并记录各频次词语数目；步骤 25－步骤 37 基于词频统计规律进行数据预处理；步骤 38 返回不同词频词语集合、各集合相应的词语总数和预处理列表。

算法 1　基于词频统计规律的文本数据预处理方法

INPUT：文本文件 Text_file；停用词表 StopWordList

OUTPUT：预处理列表 PreproList，各频次词语对应词集及词语数目 dict1，count1；dict2，count2；dict3，count3

METHOD：

```
1. dict1={},dict2={},dict3={},count1=0,count2=0,count3=0,
   TermList=[],word_count=0/*初始化*/
2. TermList=Text_file.splitWord(空格、回车及其他特征字符)/*文
   本分词*/
3. TermDict=dict.fromkeys(set(TermList),0) /*定义一个字典,其
   中键为词语,键值为对应词频*/
   /* 统计字典中各词语的词频*/
4. FOREACH t in TermDict do
5.      FOREACH i IN len(TermList) DO
6.        IF t=TermList[i] THEN
7.          word_count+=1
8.        END IF
9.      END FOR
10.     TermDict[t]=word_count
11. END FOR
   /*将词语按照词频 TFn=1,TFn=2,TFn>2 进行归类,并记录
   每个集合中的词语数目*/
12. FOREACH t IN TermDict DO
```

```
13.     IF TermDict[t]=1 THEN
14.       dict1[t]=TermDict[t]
15.       count1+=1
16.     ELSE IF TermDict[t]=2 THEN
17.         dict2[t]=TermDict[t]
18.         count2+=1
19.       ELSE
20.         dict3[t]=TermDict[t]
21.         count3+=1
22.       END IF
23.     END IF
24. END FOR
25. PreproList=TermDict /*初始化预处理列表*/
    /*基于词频统计规律进行数据预处理,删除段落中词频小于3的
    词语及停用词*/
26. FOREACH t IN TermDict DO
27.     IF TermDict[t]<3  THEN
28.       TermDict=TermDict-t
29.     END IF
30. END FOR
31. FOREACH i IN PreproList DO
32.     FOREACH t IN PreproList[i] DO
33.       IF low(t) NOT IN TermDict OR IN StopWordList
          THEN/*将单词转换为小写后再进行比较*/
34.         PreproList[i]=PreproList[i]-t
35.       END IF
36.     END FOR
37. END FOR
38. RETURN dict1,count1;dict2,count2;dict3,count3;PreproList
```

图 2 给出了基于词频统计规律的文本挖掘流程图。

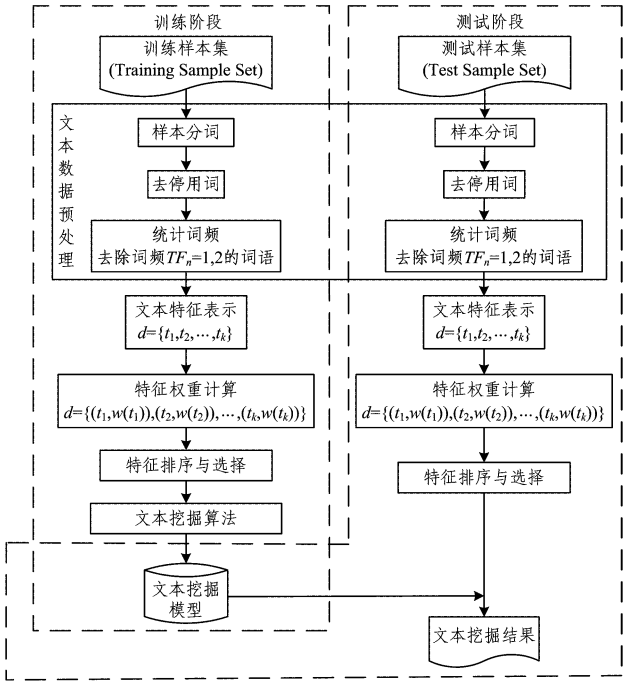


图 2 基于词频统计规律的文本挖掘流程图

通过对词语分布规律的研究发现，TF_n=1,2 的低频词与文档类别的关联度不高，但所占比例较大，这严重制约了文本挖掘性能；虽然 $TF_n>2$ 的词语的比重较小，但与文本类别的关联度较高。因此，在数据预处理阶段，将约占全文 2/3 的 $TF_n\leqslant 2$ 的词语删除，仅利用保留下的 $TF_n>2$ 的词语进行文本挖掘，在保证性能的前提下，大大减小时空复杂度，提升文本挖掘效率。

5 实验

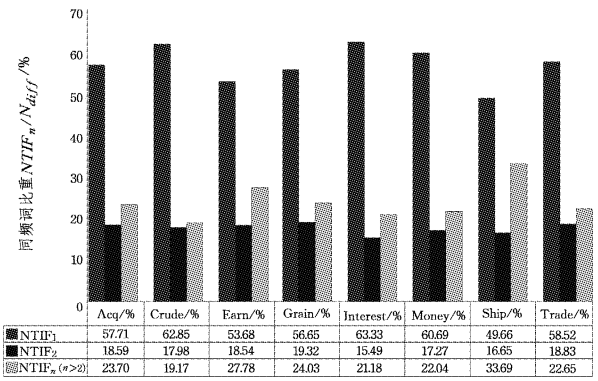
5.1 数据集

数据集选取 Reuters-21578 和 20-Newsgroups。Reuters-21578 包含 135 个主题共 21578 篇文档。实验选取 8 个类别：Acq(1659 篇)，Crude(405 篇)，Earn(2775 篇)，Grain(773 篇，corn，wheat 归入 Grain)，Interest(335 篇)，Money(502 篇)，Ship(200 篇)，Trade(340 篇)。20-Newsgroups 分为 4 大类别：Comp(1162 篇)，Rec(1190 篇)，Sci(1183 篇)和 Talk(975 篇)，每类各有 4 个子类；实验以四大类别一对一形式进行二元分类。训练集与测试集的比例均为 7∶3。

5.2 实验结果

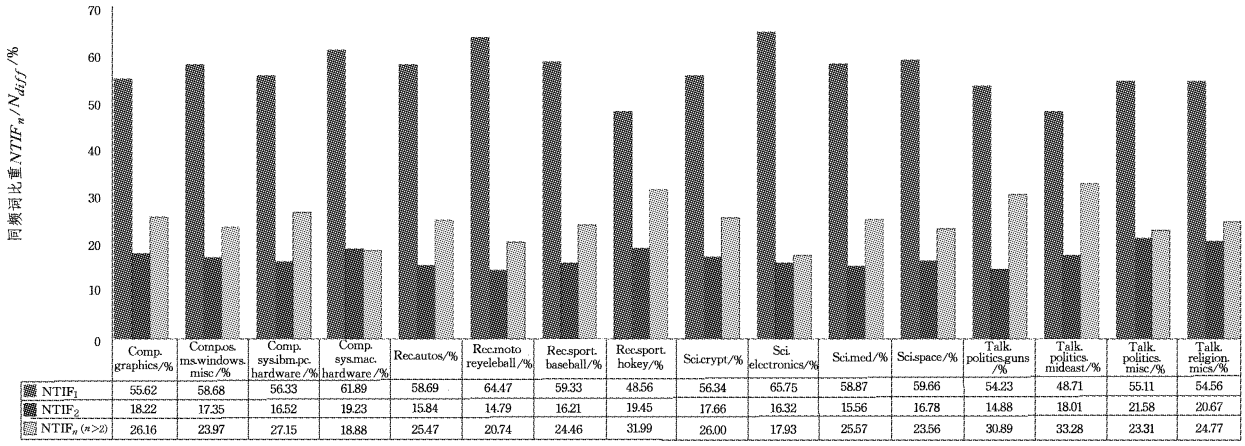
5.2.1 各频次词语在文档中的分布规律

选取数据集 Reuters-21578 和 20-Newsgroups 验证各频次词语在文档中的分布规律。图 3 和图 4 分别给出两个数据集各个子类中词频为 TF_n=1,2 及 $TF_n>2$ 的词语在全文中所占的比重 $NTIF_n/N_{diff}$。图 5 给出两个数据集各子类同频词比重 $NTIF_n/N_{diff}$ 的平均值与模型公式理论值间的对比，横轴为词频 TF_n，纵轴为同频词比重 $NTIF_n/N_{diff}$。由此可知，数据集中的各个子类中 TF_n=1 的词语比重基本维持在 50%以上，平均值分别为 57.89%和 57.30%，接近公式推导的理论值 50%，且分别高出 7.89%和 7.30%。TF_n=2 的词语比重均值分别为 17.83%和 17.44%，接近公式推导的理论值16.67%，且分别高出 1.16%和 0.77%。TF_n=1,2 的总体均值为 75.72%和 74.74%，接近公式推导的理论值 66.67%，且分别高出 9.05%和 8.07%。由此可知，在数据预处理阶段，对 TF_n=1,2 的词语进行去除可明显减少词语数量，降低特征维度。



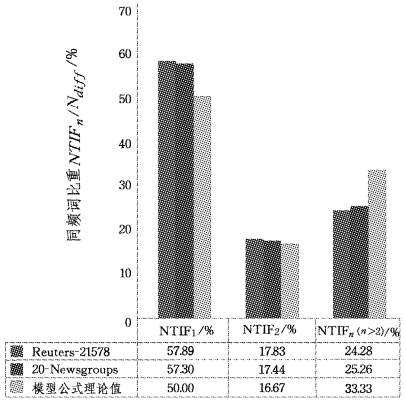
	Acq/%	Crude/%	Earn/%	Grain/%	Interest/%	Money/%	Ship/%	Trade/%
$NTIF_1$	57.71	62.85	53.68	56.65	63.33	60.69	49.66	58.52
$NTIF_2$	18.59	17.98	18.54	19.32	15.49	17.27	16.65	18.83
$NTIF_n$ (n>2)	23.70	19.17	27.78	24.03	21.18	22.04	33.69	22.65

图 3 在数据集 Reuters-21578 上同频词语的分布规律



	Comp. graphics/%	Comp.os. ms.windows. misc/%	Comp. sys.ibm.pc. hardware/%	Comp. sys.mac. hardware/%	Rec.autos/%	Rec.moto reyeleball/%	Rec.sport. baseball/%	Rec.sport. hokey/%	Sci.crypt/%	Sci. electronics/%	Sci.med/%	Sci.space/%	Talk. politics.guns /%	Talk. politics. mideast/%	Talk. politics. misc/%	Talk. religion. mics/%
$NTIF_1$	55.62	58.68	56.33	61.89	58.69	64.47	59.33	48.56	56.34	65.75	58.87	59.66	54.23	48.71	55.11	54.56
$NTIF_2$	18.22	17.35	16.52	19.23	15.84	14.79	16.21	19.45	17.66	16.32	15.56	16.78	14.88	18.01	21.58	20.67
$NTIF_n$ ($n>2$)	26.16	23.97	27.15	18.88	25.47	20.74	24.46	31.99	26.00	17.93	25.57	23.56	30.89	33.28	23.31	24.77

图 4　在数据集 20-Newsgroups 上同频词语的分布规律



	$NTIF_1$/%	$NTIF_2$/%	$NTIF_n$ ($n>2$)/%
Reuters-21578	57.89	17.83	24.28
20-Newsgroups	57.30	17.44	25.26
模型公式理论值	50.00	16.67	33.33

图 5　同频词语比重 STF_n/N_{diff} 的统计值与模型公式理论值的对比

5.2.2　数据预处理方法在文本分类中的应用效果

在数据集 Reuters-21578 和 20-Newsgroups 上，将卡方统计（χ^2）特征选择方法应用于 SVM 分类器，对基于词频统计规律的数据预处理方法在文本分类中的有效性进行验证。

如表 1 和表 2 所列，基于词频统计规律的 SVM 模型的分类准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 度量值在各个子类和宏平均值上均高于传统 SVM 模型，且运行时间明显降低。这是由于在数据预处理阶段将占文档 2/3 的词频 $TF_n \leqslant 2$ 的词语删除以进行特征选择时，仅计算占文档 1/3 的 $TF_n > 2$ 的词语的卡方统计值，大大缩短了运行时间。在数据集 Reuters-21578 中，基于词频统计规律的 SVM 模型的宏平均准确率、宏平均精确率、宏平均召回率和宏平均 F1 度量值分别高出传统 SVM 模型 0.49%，2.39%，2.15%，2.27%。平均运行时间减小 71.61%。在数据集 20-Newsgroups 中基于词频统计规律的 SVM 模型的宏平均准确率、宏平均精确率、宏平均召回率和宏平均 F1 度量值分布高出传统 SVM 模型 2.25%，2.07%，2.49%，2.29%，平均运行时间减小 73.52%。

表 1　在数据集 Reuters-21578 上进行基于词频统计规律的数据预处理前后 SVM 分类器的性能对比

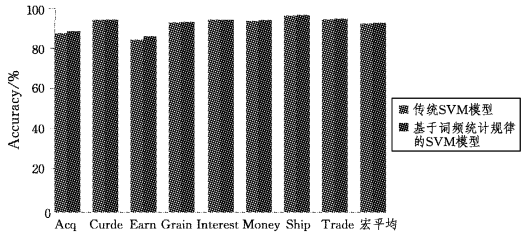
Reuters-21578 Data Sets	传统 SVM 模型					基于词频统计规律的 SVM 模型				
	Accuracy/%	Precision/%	Recall/%	F1/%	运行时间/s	Accuracy/%	Precision/%	Recall/%	F1/%	运行时间/s
Acq	87.89	76.36	70.99	73.58	2279	88.63	78.03	72.53	75.18	611
Crude	94.28	50.61	47.98	49.25	2037	94.55	53.09	49.71	51.34	574
Earn	84.65	84.63	75.02	79.54	2285	86.29	86.99	77.04	81.71	637
Grain	92.94	70.49	61.70	65.80	2196	93.38	72.98	63.22	67.75	605
Interest	94.55	42.75	41.26	41.99	1986	94.28	45.59	43.36	44.45	632
Money	93.92	58.12	54.88	56.45	1959	94.15	59.80	56.74	58.23	553
Ship	96.59	39.74	36.05	37.81	1891	96.72	42.68	40.70	41.67	548
Trade	94.68	45.04	40.41	42.60	2025	94.92	47.73	43.15	45.32	571
宏平均	92.44	58.47	53.66	55.96	2082	92.93	60.86	55.81	58.23	591

表 2　在数据集 20-Newsgroups 上进行基于词频统计规律的数据预处理前后 SVM 分类器的性能对比

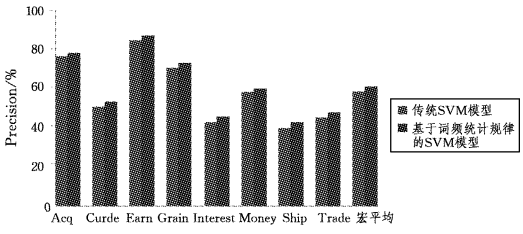
20-Newsgroups Data Sets	传统 SVM 模型					基于词频统计规律的 SVM 模型				
	Accuracy/%	Precision/%	Recall/%	F1/%	运行时间/s	Accuracy/%	Precision/%	Recall/%	F1/%	运行时间/s
Comp vs Rec	87.50	89.12	85.44	87.24	3647	88.78	90.88	86.99	88.89	953
Comp vs Sci	75.31	76.09	73.15	74.59	3418	77.87	78.28	76.59	77.43	892
Comp vs Talk	96.02	96.79	95.87	96.33	4065	96.49	96.81	96.73	96.77	1067
Rec vs Sci	73.58	76.09	69.99	72.91	3734	76.99	78.25	74.12	75.88	981
Rec vs Talk	83.27	86.25	82.78	84.48	3849	86.00	89.01	85.04	86.78	1078
Sci vs Talk	76.14	80.04	75.23	77.56	4276	79.19	83.06	77.94	80.42	1113
宏平均	81.97	84.06	80.41	82.19	3830	84.22	86.13	82.90	84.48	1014

在两个数据集上，利用准确率、精确率、召回率、F1 度量值及运行时间 5 种评价标准对传统 SVM 模型和基于词频统计规律的 SVM 模型的性能进行比较，结果如图 6 和图 7 所示。

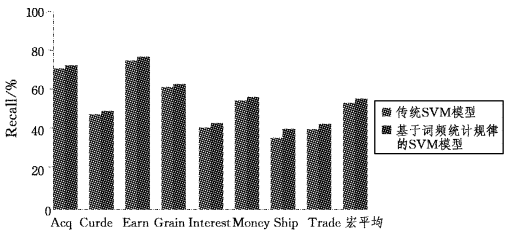
由此可知，基于词频统计规律的文本数据预处理方法，在保证分类精度的前提下极大缩短了运行时间，使分类性能得到了有效提高。



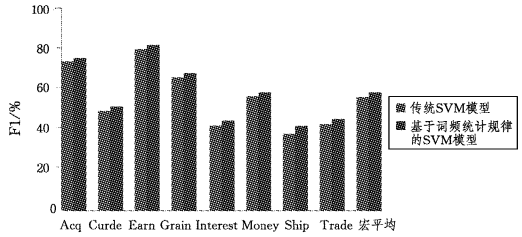
(a)Reuters-21578—Accuracy



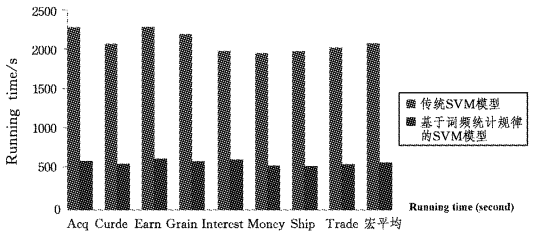
(b)Reuters-21578—Precision



(c)Reuters-21578—Recall

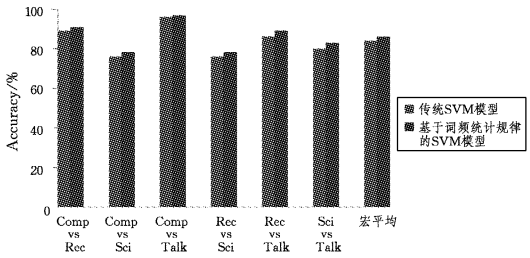


(d)Reuters-21578—F1 度量值

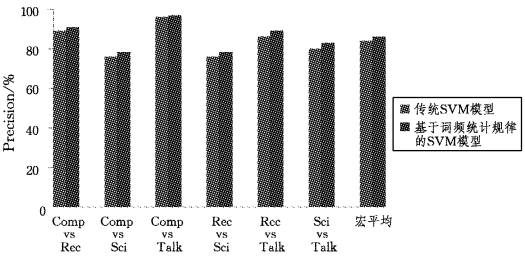


(e)Reuters-21578—Running time

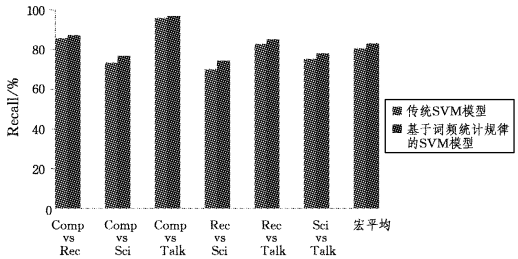
图 6 在数据集 Reuters-21578 上传统 SVM 模型和基于词频统计规律的 SVM 模型利用各评价标准的性能对比



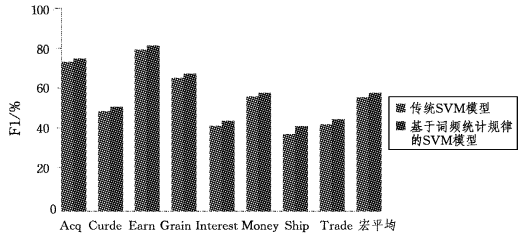
(a)20-Newsgroups—Accuracy



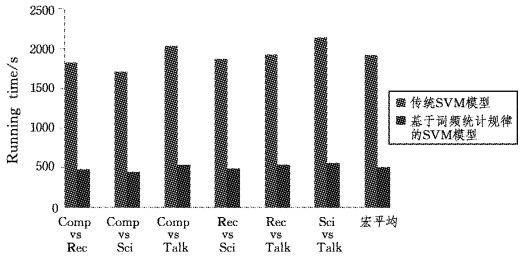
(b)20-Newsgroups—Precision



(c)20-Newsgroups—Recall



(d)20-Newsgroups—F1 度量值



(e)20-Newsgroups—Running time

图 7 在数据集 20-Newsgroups 上传统 SVM 模型和基于词频统计规律的 SVM 模型利用各评价标准的性能对比

(下转第 288 页)

dom field[J]. Journal of Shandong University (Natural Science), 2015,50(11):67-73. (in Chinese)
何炎祥,刘健博,孙松涛,等. 基于层叠条件随机场的微博商品评论情感分类[J]. 山东大学学报(理学版),2015,50(11):67-73.

[12] RAO Y H,XIE H R,LI J,et al. Social emotion classification of short text via topic-level maximum entropy model[J]. Information & Management,2016,53(8):978-986.

[13] ODBAL,WANG Z F. Emotion Analysis Model Using Compositional Semantics[J]. Acta Automatica Sinica, 2015, 41(12): 2125-2137.

[14] WANG X R,ZHANG Q H. Text Emotion Cassification Research Based on Improved Latent Semantic Analysis Algorithm[C]// Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Hangzhou,China,2013:210-213.

[15] ZHANG J M,WANG B,TANG H H,et al. Unsupervised Sentiment Orientation Analysis on Micro-blog Based on Biterm Topic Model[J]. Computer Engineering,2015,41(7):219-223. (in Chinese)
张佳明,王波,唐浩浩,等. 基于 Biterm 主题模型的无监督微博情感倾向性分析[J]. 计算机工程,2015,41(7):219-223.

[16] SU Y,JU S F,WANG Z Q,et al. Semi-supervised Sentiment Classification with Random Feature Subspace Method[J]. Journal of Chinese Information Processing,2012,26(4):85-90. (in Chinese)
苏艳,居胜峰,王中卿,等. 基于随机特征子空间的半监督情感分类方法研究[J]. 中文信息学报,2012,26(4):85-90.

[17] Google 开源深度学习工具 Wordvec[OL]. https://code.google.com/p/word2vec.

[18] 搜狗实验室全网新闻数据(SogouCA)[OL]. http://download.labs.sogou.com/dl/ca.html.

[19] HINTON G E,OSINDERO S,THE Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7):1527-1554.

(上接第 282 页)

结束语 数据预处理方法是影响文本挖掘性能的关键因素。为解决文本挖掘中由文本特征“高维-稀疏”矛盾导致的高时空复杂度与低效率问题,提出基于词频统计规律的数据预处理方法,首先推导出文本同频词表达式,然后探究各频次词语在文中的分布规律,最后以此为基础进行数据预处理。该方法与传统文本挖掘算法在预处理阶段只进行简单的分词和去停用词操作相比,从词频规律入手,进一步探究可改善数据预处理性能的方法。研究发现,词频为 1 或 2 的低频词与文档的关联度较低,但在文中所占比重约为 2/3,在预处理过程中对低频词进行去噪,可在保证文本挖掘精度的前提下,大大减少特征维度,使时空复杂度明显下降,平均运行时间降低了 70%以上,有效提升了文本挖掘性能。该方法所提出的基于词频统计规律进行数据预处理的思想对文本挖掘算法的改进具有重要意义。

参 考 文 献

[1] HAN J,FAN J,et al. Semanti-Enhanced Spatial Keyword Search [J]. Journal of Computer Research and Development,2015,52(9): 1954-1964. (in Chinese)
韩军,范举,等. 一种语义增强的空间关键词搜索方法[J]. 计算机研究与发展,2015,52(9):1954-1964.

[2] HU J,FAN J,LI G L,et al. Top-k Fuzzy Spatial Keyword Search[J]. Chinese Journal of Computers,2012,35(11):2237-2246. (in Chinese)
胡骏,范举,李国良,等. 空间数据上 Top-k 关键词模糊查询算法[J]. 计算机学报,2012,35(11):2237-2246.

[3] REN P J,CHEN Z M,et al. Search Result Diversification Combing Semantic and Temporal Intent[J]. Chinese Journal of Computers, 2015,38(10):2076-2091. (in Chinese)
任鹏杰,陈竹敏,等. 一种综合语义和实效性意图的检索多样化方法[J]. 计算机学报,2015,38(10):2076-2091.

[4] DING Z Y,JIA Y,et al. Survey of Data Mining for microblogs [J]. Journal of Computer Reaserch and Development,2014,51(4):691-706. (in Chinese)
丁兆云,贾焰,等. 微博数据挖掘综述[J]. 计算机研究与发展,2014,51(4):691-706.

[5] SONG Q,NI J,WANG G. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(1):1-14.

[6] ZHAO Z,HE X F,CAI D,et al. Graph Regularized Feature Selection with Data Reconstruction[J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(3):689-700.

[7] ZIPF G K. Human behavior and the principle of least effort:an introduction to human ecology[M]. Addison-Wesley Press,1949:23.

[8] BOOTH A D. A law of occurrences for words of low frequency [J]. Information and Control,1967,10(4):386-393.

[9] EGGHE L. A new short proof of Naranan's theorem, explaining Lotka's law and Zipf's law[J]. Journal of the American Society for Information Science & Technology,2010,61(12):2581-2583.

[10] CHAN P,HIJIKATA Y,NISHIDA S. Computing semantic relatedness using word frequency and layout information of wikipedia[C]// Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM,2013:282-287.

[11] SURYASEN R,RANA M S. Content analysis and application of Zipf's law in computer science literature [C]// 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS). IEEE,2015:223-227.

[12] GEORGE K Z. Human Behavior and the Principle of Least Effort:An Introduction to Human Ecology[M]. New York:Addison-Wesley Press,1949:573-584.

[13] 邱均平. 文献计量学[M]. 科学技术文献出版社,1988:157.

[14] BOOTH A D. A law of occurrences for words of low frequency [J]. Information and Control,1967,10(4):386-393.

[15] AGRAWAL R,GOLLAPUDI S,KENTHAPADI K. Enriching Textbooks Through Data Mining [C]//Proceedings of the First ACM Symposium on Computing for Development. 2010:1-9.

[16] AGRAWAL R,GOLLAPUDI S,KANNAN A,et al. Data mining for improving textbooks[J]. ACM SIGKDD Explorations Newsletter,2012,13(2):7-19.