Text Category Crowdsourcing Solution Filter Research in Reward Mode

Huixing Nie [1,a], Tianshun Wang [2,b,*]
[1,2] School of management, Hefei University of Technology, Hefei, Anhui, China
[a]1225026133@qq.com, [b]wtshfut@163.com

Key words: crowdsourcing, filter, reward mode, solution selection, double-layer filter.

Abstract: A two-level filtering model is proposed to solve the problem of crowdsourcing scheme selection. Firstly, we use word segmentation tool NLPIR to segment crowdsourcing word, then get the text key words and word frequency statistics. Secondly, according to the word frequency method, we extract feature words and establish feature matrix. Thirdly, the vector space model is used to describe the text contents of the crowdsourcing, and the double layer filter is built according to the text relevance theory. Finally, To verify validity of the filtering model, we use it to filter the existing historical crowdsourcing schemes on the crowdsourcing platform, and compare the results of filtering with the results of artificial selection. It shows that this filtering model can realize the semi-automatic selection of crowdsourcing scheme. It can filtrate a large number of ineffective schemes and keep effective schemes. The two-level filtering model has practical application value.

悬赏模式下文本类众包方案过滤研究

聂会星 [1, a]，王天顺 [2, b, *]
[1, 2]合肥工业大学管理学院，合肥，安徽，中国
[a]1225026133@qq.com, [b]wtshfut@163.com

关键词：众包；过滤；悬赏模式；方案选择；双层过滤模型

摘要：针对众包方案选择问题，提出一个双层过滤模型。首先，用分词工具 NLPIR 对众包文本分词，得到词频统计和关键词。其次，根据文档词频法提取特征词，建立特征矩阵。再次，采用向量空间模型描述众包文本内容，根据文本相关度理论建立方案双层过滤器。最后，用该过滤模型对众包平台上现有的历史众包方案过滤筛选，并将过滤结果与人工选择的结果比较，验证过滤模型的可行性。结果表明，该过滤模型可实现众包方案的半自动化选择，能够滤出大量无效方案，保留有效方案，具有实际应用价值。

1 引言

众包指把过去由员工完成的工作以自由自愿的形式外包给大众网络的做法[1]，它是随着互联网的发展而兴起的一种开放式创新模式。众包模式中三个主体，分别是发包方（任务需求者）、网络平台和接包方（方案提供者）。悬赏也叫比稿，是众包模式中发布任务需求征集众包方案的一种模式，接包方根据任务需求提交相应的方案，发包方从中选择一个或几个满意的方案。悬赏众包模式的目的是获得尽可能多的众包方案，其特点是一个任务对应多个方案。众包任务的类型多样，按内容形式将众包分为文本类、图片类、音频类、视频类，最常见的众包任务是文本类，文本类也是其他类的基础。

众包方案的质量具有不确定性。接包方身份匿名，能力和工作态度各不相同，有的没有真心地为任务需求者工作，他们会随机提交众包方案[2]，而有的接包方具有欺骗行为[3]，最终收集到的众包方案的质量低下，属于无效方案，Marsden 认为 90%的众包方案都是无效方案[4]。在悬赏模式下，K 个众包任务需要 M 个接包方提供 N 个方案，这 N 个众包方案中最终只有少量方案符合发包方的需求。如何从大量众包方案中滤出无效方案是众包应用领域亟待解决的问题。

众包质量是众包模式的关键，也是当前国内外众包研究的一个热点问题。学术界对众包质量的研究主要是从众包接包方、众包过程和众包方案三个方面考虑。众包模式的一个主体是接包方，通过对接包方的质量控制可以提高众包的质量[5-7]，给合适的接包方分配合适的任务[8]是提高众包质量的有效方法。一些学者注重众包过程，用过程管控的思想研究众包质量[9]。研究众包质量最直接的方式是研究众包方案的评价与筛选。实际应用中最常见的评价与筛选方式有人工评价与投票原则[10-12]，但人工评价的方式不但效率低，成本高，具有较大的主观性，而投票决定众包方案质量的这种常用方法效果并不是很好[13]，网络投票的方式缺乏有效的监督，存在严重的拉票现象，投票的结果在一定程度上能反映方案的有效性，但不能决定最终的结果。

本文利用信息过滤方法解决众包方案的评价与筛选问题，从众包方案初筛选的角度研究众包质量。提出一个双层过滤模型，通过过滤模型的过滤筛选滤出大量无效方案，实现众包方案的半自动化筛选，辅助人工选择，从而提高众包方案选择的效率，降低人工成本，并有效改善人工评价与投票的不足，获取优质的众包方案。

众包方案过滤属于信息过滤的范畴，信息过滤指根据用户偏好，把用户不关心的信息从信息流中滤出，保留用户感兴趣的部分。信息过滤分为基于内容的过滤、协同过滤和基于规则的过滤[14]。基于内容的过滤是将潜在需求信息构造成用户的信息需求模型，然后与待过滤文本匹配，其中内容相似度的计算来自文本自身[15]。协同过滤应用较广，其关键问题是相似度的计算[16]，基于规则的过滤以预先定义的规则为过滤依据。信息过滤中最基本的是文本过滤，文本过滤主要应用在信息检索、推荐、敏感信息过滤、垃圾邮件过滤等领域。

2 众包方案过滤模型

实现众包方案自动过滤需要对众包任务及众包方案预处理，将内容转换成过滤系统能够处理的格式。依据自然语言处理理论[17,18]，对众包任务及方案文本分词，提取特征，用所提取的特征表示文本内容，然后构建双层过滤器，最后给出对过滤结果评估方法。众包方案过滤模型如图 1 所示。

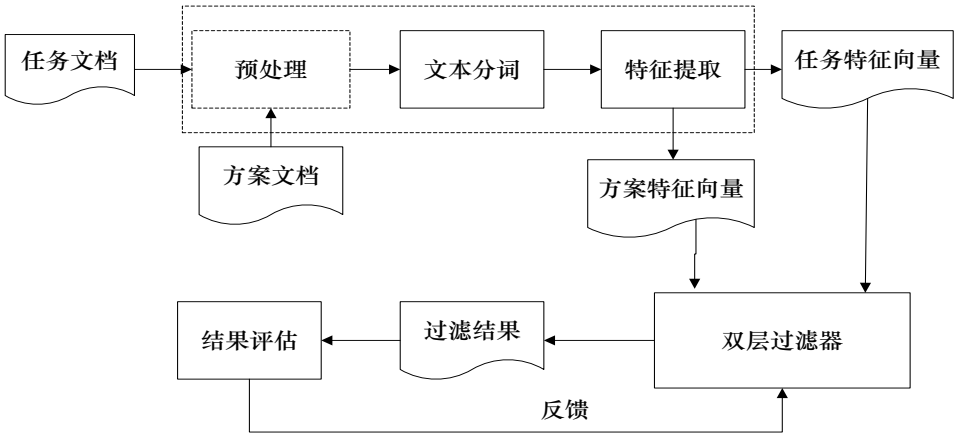

图 1 众包方案过滤模型

2.1 预处理与文本分词

众包任务和方案的原始文档数据格式多种多样，需要进行统一处理成能被过滤模型识别的数据。内容预处理的对象包括众包任务和众包方案，将网页文档和附件中的 PDF 及 WORD 等格式的文档整理并转换为 TXT 文本。

目前有多种自然语言处理工具，汉语词法分析系统 ICTCLAS2013（又叫 NLPIR 汉语分词系统）是比较成熟的一种，已经有十余年的应用历史。NLPIR 的主要功能包括分词标注、词频统计、关键词提取等。本文利用汉语词法分析系统对预处理后的文本内容做分词处理和词性标注，只保留名词、形容词和动词三种词性的词语，将内容文本转换成由词语构成的文本内容表列，并获取词频统计和关键词。

2.2 特征提取

特征是描述内容的属性，由文本分词得到的一系列名词、动词和形容词组成。特征是决定任务和方案相似性的关键，特征提取指从一个已知的特征集中挑选一个子集，使得最终的评价标准最优。

特征提取算法L，数据集S，其中S有 n 个特征$T_1, T_2, T_3, \dots, T_n$，选择 m 个最优特征构成特征子集$T_{opt}$，使得某个评价准则 $F = F(L, S)$达到最优。

特征提取的任务是从特征词集中提取出信息量大、能较好表达文本内容的词汇，用尽量少的特征项表达文本内容，降低空间向量的维数，从而降低计算的复杂度。按特征集评价策略分，特征抽取的方法可以分为两类[19]：一类是滤波器方法(Filter)，另一类叫嵌入式方法(Wrapper)。这两者的区别在于特征抽取的过程是否涉及学习算法，Wrapper 方法利用学习算法实现特征选择，而 Filter 方法不需要用到学习算法。

常用特征提取的主要方法有文档频率、互信息、信息增益、文本证据权和期望交叉熵等。文档频率法选取高词频的词，剔除低频率的词，但往往会忽略某些能很好表达文本内容的低频率的词；信息增益(IG)是基于信息熵的一种评估方法，指某个特征项出现与否而产生的熵差[20]，它是从信息论中信息熵的角度出发，根据各个特征取值情况划分学习样本空间时所获信息增益的多少来选择相应的特征，信息增益法强调某个特征在不同文本中的分布情况；互信息法只考虑与内容正相关的特征项，忽略不相关特征项的影响，最终的性能较差。

NLPIR 汉语分词系统能够方便准确的得到文本的词频统计和关键词，本文的特征抽取采用文档频率法，将高频词选入特征词候选集。另外，某些低频词信息量大，能够较好表达文本内容，这些低频词往往属于NLPIR汉语分词系统提取的文本关键词，这些词也属于特征词。

2.3 文本表示

经过内容预处理及特征提取后文本数据是非结构数据集，依然无法被过滤器识别，文本表示将非结构化数据转换为结构化数据。常用的文本表示方法有概率模型（Probabilistic Model）、向量空间模型(VSM: Vector Space Model) 和布尔模型（Bloolean Model）[21]。

向量空间模型可把非结构化的文档数据内容用向量的形式表示，实现文档的可操作性和可计算性[22]。文本表示采用向量空间模型表述，文本所有特征词构成一个 M 维特征空间，文本 D 的向量表示为：

$$\boldsymbol{D} = [T_1, T_2, \dots, T_i, \dots, T_m] \tag{1}$$

T_i 表示第 i 个特征，m 表示文本集中所含特征数。任务文档 D_1 的特征向量为$D_{1=}(\omega_1, \omega_2, \omega_3, \dots, \omega_m)$，方案文档 D_2 的特征向量为$D_2 = (\mu_1, \mu_2, \mu_3, \dots, \mu_m)$。

在向量空间模型中，特征词值的计算极其重要，它决定相似度计算的结果，最终影响到方案过滤的结果。本文中用 TF 方法[23]确定众包任务特征项值，用 TF-IDF 方法[24]确定众包方案文档特征项的值。

TF 计算法适用于单个文本特征项值得计算，TF（Term Frequency）指文档频度，特征项

T_i在文本中出现的频率越高，说明该特征项描述众包任务文本的能力越强。TF 法的计算公式为：

$$\omega_{TF} = c_t \tag{2}$$

其中c_t指特征 t 在众包任务文本中的词频。

TF-IDF 是向量空间模型经典的计算方法，它由 TF 值与 IDF 值共同决定， 计算公式：

$$\omega_{ij} = f_{TF} * f_{IDF}$$
$$= f_{ij} * \log(\frac{N}{n_i} + \alpha) \tag{3}$$

其中，ω_{ij}表示特征 i 在方文档 j 中的值，f_{ij}表示特征词频，f_{IDF} 为反文档频率，N 为总的众包方案文档数，n_i为出现特征词 i 的文档数，α为经验常数，通常情况下取值 0.01[25]。

将ω_{ij}归一化处理，得到:

$$\omega_{ij} = \frac{f_{ij}*\log(\frac{N}{n_i}+0.01)}{\sqrt{\sum_i^N f_{ij}*[\log(\frac{N}{n_i}+0.01)]^2}} \tag{4}$$

计算特征项值的过程中，IF-IDF 方法通过 TF 值与 IDF 值来考虑特征项对单个方案文本的影响力和其对整个方案文档集中的影响力。

2.4 双层过滤器

众包方案的过滤属于二分类问题，把应该被过滤的方案定义为无效方案，不应该被过滤的方案定义为有效方案。

参考《猪八戒网服务规则》对无效众包方案的判定标准，满足以下任意一点的众包方案属于无效方案：广告信息(X_1)、低俗信息(X_2)、敏感信息(X_3)、没有提交实际作品(X_4)、与需求无任何关联(X_5)、雷同稿件(X_6)、无法满足需求的基本要素(X_7)等。

用数学表达式表示无效方案：

$$X = (X_1, X_2, X_3, \dots, X_7) \tag{6}$$

根据 X 的特点，可把众包方案过滤概括为方案与任务的相关度问题，方案与任务相关度低，则该方案不能满足需求，属于无效方案，理当被滤出。

(6)式中，X_1, X_2, X_3, X_4, X_5可以概括为众包方案的特征与众包任务特征无相同或相似项。

令$Y_1 = X_6$，$Y_2 = X_1, X_2, X_3, X_4, X_5$， $Y_3 = X_7$。

建立双层过滤器，第一层过滤掉Y_1和Y_2类型的无效方案，第二层过滤掉Y_3类型的无效方案，如图 2 所示。

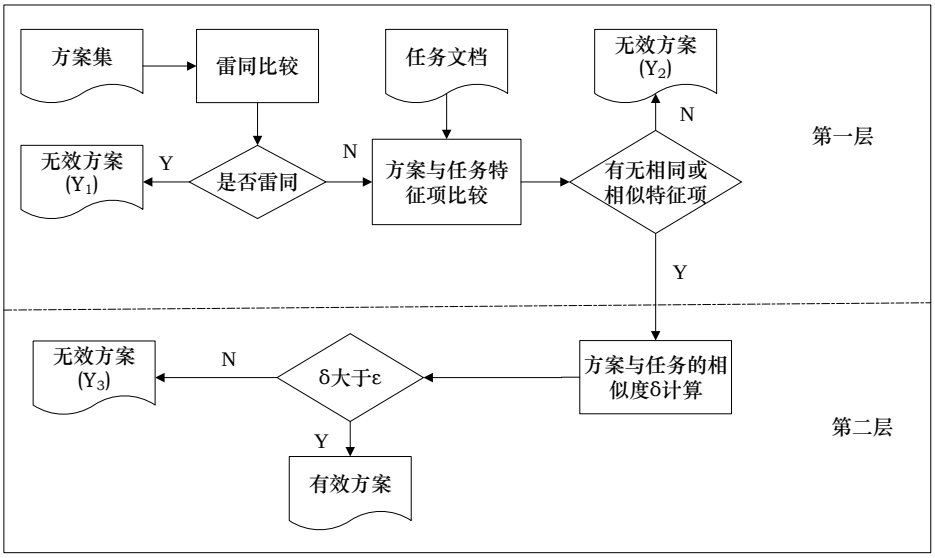

图 2 双层过滤器

采用双层过滤模型可以避免计算部分无效方案特征项的值，从而大大降低过滤器计算的复杂度。在过滤器第一层中，仅通过特征项即可判定是否属于无效方案，不需要计算方案的特征项值。若有两个及两个以上的方案特征集一致，则可判定该方案雷同；若方案文本任意一个特征项都不属于任务文本特征集，则可判定该方案与任务无关，属于无效方案，需被过滤。在过滤器第二层中，利用（2）式计算任务文本特征项的值，利用（4）式计算方案各特征项的值，得到任务和方案的特征向量，再根据向量空间模型相似度计算公式（7）判定方案是否该被过滤。

向量空间模型中，任务与方案直接的的内容相关度$\mathrm{Sim}(\mathrm{D}_1, D_2)$用向量之间夹角的余弦值表示[26]，即：

$$\delta = \mathrm{Sim}(\mathrm{D}_1, D_2) = \cos\theta = \frac{D_1 * D_2}{||D_1|| * ||D_2||} = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{(\sum_{k=1}^{n} {W_{1k}}^2)(\sum_{k=1}^{n} {W_{2k}}^2)}} \tag{7}$$

设定一个过滤阀值为ε，如$\mathrm{Sim}(\mathrm{D}_1, D_2) < \varepsilon$，那么向量所对应的内容符合过滤条件，需要被过滤掉。相似度阈值一旦设立，那些与众包任务文档向量的相似度大于或等于阈值的众包方案就被认为与众包任务相关，应当保留，而其他众包方案被认为与众包任务不相关，属于无效众包方案，应当滤出。

2.5 过滤性能评价

过滤的目的是对收集到的众包方案作初筛选，过滤掉无效方案，保留所需要的方案。因此，主要对经过滤处理后的有效类方案做评价。

定义以下变量：设共有 S 个待过滤众包方案，其中有P_1个有效方案，P_2个无效方案，过滤后，Q_1个有效方案被误判为无效方案，Q_2个无效方案被误判为有效方案，显然，$\mathrm{S} = P_1 + P_2$。

本文采用误判率、召回率（recall）、查准率（precision）和 F 值作为有效方案的性能评价指标。

误判率：

$$e = \frac{Q_1 + Q_2}{S} \tag{8}$$

误判是一个重要的问题，若有效方案被判定为无效方案，会导致错失关键信息，造成损失；若无效方案被误判为有效方案，分类器对方案的过滤失效。显然，误判率越低越好。

召回率：$$r = \frac{P_1 - Q_1}{P_1} \quad (9)$$

召回率反映过滤器发现有效方案的能力，召回率越高，遗漏的有效方案越少。

查准率：$$s = \frac{P_1 - Q_1}{P_1 - Q_1 + Q_2} \quad (10)$$

查准率越高，过滤性能越好。

F 值：$$F = \frac{2rs}{r+s} \quad (11)$$

F 值是查全率与召回率的调和平均值，查准率和召回率是互补的，提高召回率会导致查准率降低，反之亦然。

3 实验验证

为验证上述过滤方法的有效性，利用该双层过滤模型对包平台上现有的众包方案进行过滤筛选，并将过滤结果与人工评选结果比较。

3.1 实验设计

实验数据来自知名众包平台猪八戒网，是网友 kunnwin 在猪八戒网上发布的一个博客征文悬赏任务，该任务最终收集到来自 74 名服务商的 356 个方案，根据任务发布者的人工审核与评判，最终有 120 份方案合格，236 份方案不合格

利用本文的双层过滤器对收集到的 356 份方案做过滤筛选，第一层滤出 73 份无效方案。过滤器第二层中，过滤阈值设置为$\varepsilon = 0.5$时，滤出 104 份无效方案，剩下 179 份方案，过滤结果中有 66 份无效方案被误判为有效方案，7 份有效方案被误判为无效方案。由于剩下的方案较多，需要调整过滤阈值。过滤阈值设置为 $\varepsilon = 0.6$时，过滤出 148 份无效方案，剩下 135 份方案，其中有 29 份无效方案被判定为有效方案，14 份有效方案被误判为无效方案。

3.2 验证结果

双层过滤器第一层操作简单，能够降低计算量，过滤结果误判率较低，过滤器第二层通过调整过滤阈值ε确定过滤量，过滤阈值ε越大，滤出的方案越多。表 1 为过滤阈值ε设为 0.5 和 0.6 时对有效方案的性能评价。

表 1 有效方案的过滤性能评价

	误判率	召回率	查准率	F 值
$\varepsilon = 0.5$	0.205	0.942	0.631	0.756
$\varepsilon = 0.6$	0.121	0.883	0.785	0.831

实验验证结果表明，过滤阈值ε设为 0. 5 时误判率较高，F 值较低，过滤阈值ε设为 0.6 时误判率只有 0.12，F 值为 0.831，过滤阈值ε设为 0.6 较为合理。利用该双层过滤模型能够滤出大量无效方案，同时保留有效方案，过滤结果有效。与众包方案的人工选择比较，该过滤模型具有更高的效率。

4 结束语

本文提出的双层过滤模型可以辅助众包方案选择的人工操作，实现众包方案半自动化初筛选。悬赏模式下的众包方案的种类杂多，该过滤方法适用于方案较多的文本类任务，文本类众包方案越多，过滤效果越明显。另外，常见的众包需求任务的描述由任务标题、关键词、

类别、概述、任务基本要求等组成，但有些众包用户提供的任务初始需求信息量不足，方案过滤的最终结果会产生偏差。因此，当众包用户的需求信息不足时有必要根据众包任务发布者的目的扩展用户的需求信息，提高过滤结果的准确率。

References

[1]Howe,Jeff.The Rise of Crowdsourcing[J].06 Jenkins H Convergence Culture Where Old & New Media Collide,2006,14(14):1-5.

[2]Le J,Edmonds A,Hester V,et al.Ensuring quality in crowdsourced search relevance evaluation[C]//SIGIR Workshop on Crowdsourcing for Search Evaluation.2010:17-2.0

[3]Zhu D,Carterette B.An Analysis of Assessor Behavior in Crowdsourced Preference Judgments[C]//SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation.2010.

[4]Marsden P.Crowdsourcing[J].Hanser Fachbuchverlag,2009,4(3):24-28.

[5] LU Ying-jie,ZHANG Peng-zhu,LIU Jing-fang. Task-oriented Talent Selection in Crowdsourcing [J] Journal of Systems & Management, 2013,22(1):60-66.

[6]Tang Si.Quality Control Algorith and Performance Evalution in Crowdsourcing[D]Hangzhou:ZheJiang University,2015

[7]Venetis P,Garcia-Molina H.Quality control for comparison microtasks[C]// International Workshop on Crowdsourcing and Data Mining.2012:15-21.

[8]Karger D R,Oh S,Shah D. Iterative Learning for Reliable Crowdsourcing Systems[C]//Advances in Neural Information Processing Systems.Neural Information Processing Systems,2011:1953-1961.

[9] ZHANG Zhi-Qiang PANG Ju-Sheng XIE Xiao-Qin.Research on Crowdsourcing Quality Control Strateies and Evalution Alagorithm [J] Chinese Journal of Computers,2013, 36(8):1636-1649.

[10]Franklin M J,Kossmann D,Kraska T,et al.CrowdDB:answering queries with crowdsourcing[C]//Acm Conference on Management of Data. ACM,2011:61-72.

[11]Lee J,Cho H,Park J W,et al.Hybrid entity clustering using crowds and data[J]. Vldb Journal—the International Journal on Very Large Data Bases,2013, 22(5):711-726.

[12] Yue D J,Yu G,Shen D R.Crowdsourcing Quality Evaluation Strategies Based on Voting Consistency[J]. Journal of Northeastern University,2014,35(8):1097-1101.

[13] Lin C H, Mausam, Weld D. Crowdsourcing Control: Moving Beyond Multiple Choice[J]. Eprint Arxiv,2012, 7(9):685-696.

[14]Li Y, Zhou X, Bruza P, et al. A two-stage decision model for information filtering[J]. Decision Support Systems, 2012, 52(3):706-716.

[15] Gao Y, Xu Y, Li Y. Pattern-based Topics for Document Modelling in Information Filtering[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(6):1629-1642.

[16]Wu Z Y,Tang Y,Fang J X, Collaborative Filtering Recommendation Algorithm Based on Obtology Semantic Similarity[J].Computer Science.2015,42(9):204-207.

[17] WANG Yu,SHAO Hong-yu. Current Situation Analysis of Domestic Natural Language

Processing Based on Topic Words Extraction [J].Information Science,2013(3)：151-155.

[18]XU Ge WANG Hou-Feng. The development of the topic model in natural language processing [J]. Chinese Journal of Computers,2011, 34(8):1423-1436.

[19] Lin Zhu. Research on Feature Weighting and Feature Selection Based Dsta Mining Algorithms [D].Shanghai: Shanghai Jiao Tong University,2013.

[20] SHI Hui;JIA Daiping;MIAO Pei. Improved information gain text feature selection algorithm based onword frequency information [J]. Journal of Computer Applications, 2014, 34(11):3279-3282.

[21] LIAO Tao, LIU Zong-tian, WANG Xian-chuan.Research on Event Based Method for Text Represesentation[J].Computer Science, 2012, 39(12):188-191.

[22] HUANG Cheng-Hui,YIN Jian ,HOU Fang. A Text Similarity Measurement Combining Word Semantic Information withTF-IDF Method [J]. Chinese Journal of Computers, 2011, 34(5):856-864.

[23] Wang D,Zhang H,Liu R,et al.t-Test feature selection approach based on term frequency for text categorization[J].Pattern Recognition Letters,2014, 45(11):1-10.

[24]Peng T,Liu L,Zuo W.PU text classification enhanced by term frequency–inverse document frequency-improved weighting[J].Concurrency & Computation Practice & Experience,2014,26(3):728–741.

[25] TAI De-yi,WANG Jun. Improved Feature Weighting Algorithm for Text Categorization [J]. Computer Engineering,2010, 36(9):197-199.

[26]Chen hong-fei. Chinese Text similarity algorithm Research based on Vector Space Model [D].Chengdu: University of Electronic Science and Technology of China,2011.