基于语义的文档特征提取研究方法

姜 芳[1,2] 李国和[1,2] 岳 翔[3]
(中国石油大学(北京)地球物理与信息工程学院 北京 102249)[1]
(中国石油大学(北京)油气数据挖掘北京市重点实验室 北京 102249)[2]
(中海油研究总院信息数据中心 北京 100029)[3]

摘 要 中文文本特征词选取是文本处理的重要方面,对文本分类有重要影响。现有的文本特征提取方法存在生成特征向量维数高、依赖训练集、忽略低频关键词等不足。利用《同义词词林》计算词语之间的语义距离,通过聚类算法筛选类别的主题相关词,最后通过信息增益算法从主题相关词中选取特征词。以宏F值和微F值为评价指标,通过有效性实验和对比实验表明,该方法的文本特征选取效果优于其他经典算法。

关键词 特征词,语义距离,信息增益,文本分类
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.053

Semantic-based Feature Extraction Method for Document

JIANG Fang[1,2] LI Guo-he[1,2] YUE Xiang[3]
(College of Geophysics and Information Engineering, China University of Petroleum, Beijing 102249, China)[1]
(Beijing Key Lab of Data Mining for Petroleum Data, China University of Petroleum, Beijing 102249, China)[2]
(Information & Data Center, CNOOC Research Institute, Beijing 100029, China)[3]

Abstract Feature extraction of Chinese documents is an important part in the document processing, and imposes great influence on the document classification. Pre-existing document feature extraction methods have many shortcomings, such as creating a feature vector of high dimensions, depending on training sets, ignoring low-frequency keywords, and so on. In this paper, the semantic distance between words was calculated based on the synonyms dictionary, and then theme related words of each classification were selected by the density clustering method, and finally the feature words were selected from the theme related words using the information gain algorithm. In order to validate the proposed method, one validation experiment and one comparison experiment were designed and the evaluation indexes including the macro-F value and the micro-F value were calculated. Experiment results show that the proposed document feature extraction method has better performance than other traditional algorithms.

Keywords Feature word, Semantic distance, Information gain, Text classification

1 引言

为了提高中文信息处理的效率,帮助人们全面地掌握自己所需要的信息,特征提取成为研究热点。本文主要介绍基于文档的特征提取,文档特征可以代表相应文档,在区分文档类别时能减少处理词语数量、降低向量空间维度、简化计算、提高速度和效率,并且能够清晰、直接地代表文档主题。因为文档特征具有区分文本的能力且数量有限,所以信息的特征提取在中文信息处理中显得尤为重要。

特征提取的常用方法主要分为:基于概率的特征提取方法和基于语义的特征提取方法。基于概率的特征提取方法主要有:文档频率[1-3]、信息增益[4-6]、互信息[7,8]、X^2 统计[9,10]、N-Gram算法[11]等。基于语义的特征提取方法有:基于本体论的特征提取方法[12,13]和基于知网概念的特征提取方法[14,15]。现对文档特征提取方法进行总结,如图1所示,基于概率的文档特征提取的步骤是先对文档进行分词,过滤掉停用词,并用相应的传统算法对词语权重进行计算,得到权重的排序后,从高到低选取相应的特征作为代表该文档的特征。而基于语义的文档特征提取方法是在过滤掉停用词后对词语构建网络结构,通过词语网络结构得到特征权值计算公式,从而得到最终文档特征。

基于概率的特征提取方法简单、易于实现,不依赖具体领域和语言,但都有其相应的缺点:文档频率方法虽计算量小、速度快,但可能会错误地过滤某个低频特征词,造成判断失误,影响精度;互信息方法虽直观且易于理解,但没有考虑特征词的发生频率,在互信息评估时会倾向选择罕见的特征词;

到稿日期:2015-01-20 返修日期:2015-05-18 本文受国家高新技术研究发展计划(2009AA062802),国家自然科学基金(60473125),中国石油(CNPC)石油科技中青年创新基金(05E7013),国家重大专项子课题(G5800-08-ZS-WX)资助。
姜 芳(1984-),女,博士生,主要研究领域为智能信息处理等,E-mail:jiangfangzhang@163.com(通信作者);李国和(1965-),男,博士,教授,博士生导师,主要研究领域为人工智能、知识发现;岳 翔(1988-),男,硕士,主要研究领域为知识发现。

信息增益方法计算繁琐；X^2 统计量方法处理低频特征不太可靠；N-Gram 算法会产生错误的汉语词语切分。

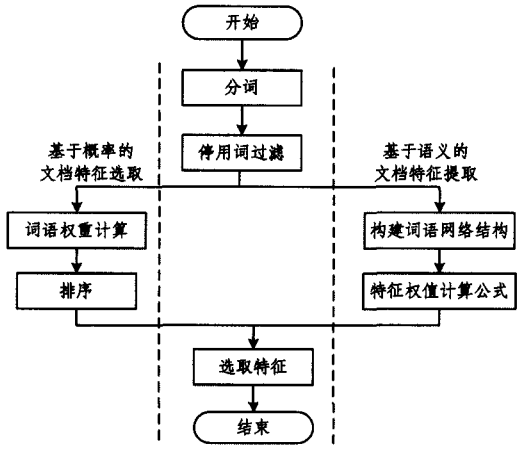

图 1　现有文档特征提取方法

将基于概率的特征提取方法的缺点总结如下：

①特征提取计算量大、效率低。

②特征提取后，生成的特征向量维数高。

③特征项假设独立，没有对上下文信息进行关联。

④高低频率词对特征词都具有影响，但未能在现有方法中体现。可能选中出现频率较高但对分类贡献较小的特征；可能约简掉出现的频率低但对分类贡献大（和文档主题密切相关）的特征词。

⑤缺乏对文档语法语义的分析。

⑥需要用一个很庞大的训练集才能获得几乎所有的对分类起关键作用的特征。

⑦需要消耗大量的时间和空间资源。

在基于语义的文档特征提取方法中，基于本体论的文档特征提取方法利用构建词语网络结构，得到特征权值计算公式，从而得到文档特征，该方法充分考虑特征所在文档位置和特征之间的关系。基于知网概念的文档特征提取方法利用知网概念对文档进行部分语义分析，利用语义分析合并同义词，并对词语进行聚类，最后得到文档特征。基于语义的特征提取方法降低特征向量维度，与基于概率方法不同，其不仅去除了无用信息或噪声数据，同时还解决了同义词和多义词的问题，很好地表示了词之间的内在联系。但基于语义的文档特征提取方法仍然存在如下问题：

①基于知网概念的文档特征提取方法利用同义词和近义词匹配构建网络。由于汉语文献作者使用语言的多样性，表达同一主题的关键词大多不是同义词或近义词，使同主题的词语大部分未能得到语义关联，导致语义在关键词提取中不能发挥应有作用；并且利用词频与区域特征进行关键词选择，容易忽略出现频率低但具有关键意义的特征。

②在构建词语网络时，基于本体论的文档特征提取方法如有分词错误，将会对文档特征提取产生影响。

基于以上方法的不足，提出一种基于语义距离的文档特征提取方法。首先利用基于语义距离的方法提取主题相关词，然后利用信息增益算法从主题相关词中提取出文档特征。该方法首先利用语义距离准确缩小特征范围，之后再利用信息增益算法高效地提取文档特征。

2　基于语义的文档特征提取方法

首先对文档进行分词，通过哈工大《同义词词林》扩展版对词语计算语义距离，然后根据词语语义距离对词语进行密度聚类，确定主题相关词的集合，最后利用信息增益算法得到代表文档的特征集合。

定义 1　集合 $D(N)=\{D_i \mid 1\leqslant i\leqslant n\}$ 为同一类别 R 的文档集，其中 D_i 为该类别 R 中的一篇文档，n 为该文档集中文档的数目。对集合 $D(N)$ 进行分词，得到集合 $D(W)=\{D_i(W)\}=\{w_1,w_2,\cdots,w_n\}$，其中 $D_i(W)$ 为对文档 D_i 进行分词后的分词结果集，w_i 为一个词语，n 为对集合 $D(N)$ 分词后的词语个数。

2.1　词语语义距离计算

图 2 是哈工大《同义词词林》扩展版的词语树形结构。由图 2 可以看出，《词林》对词进行了详细的分类，每个词语都有相应的编码 $Code_i=X_{i1}X_{i2}X_{i3}X_{i4}X_{i5}F_i$，分别是“大类”、“中类”、“小类”、“词群”、“原子群”，标志位 F_i 的标记有 3 种，分别是“=”、“#”、“@”，“=”代表“相等”、“同义”；末尾的“#”代表“不等”、“同类”，属于相关词语；末尾的“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。由图可知，每个词语都有对应的编码，如小农\老农的编码为：Ae07A02，佃农\佃户的编码为 Ae07A03。通过词语的对应编码，可以计算出词语间的语义距离[16,17]。

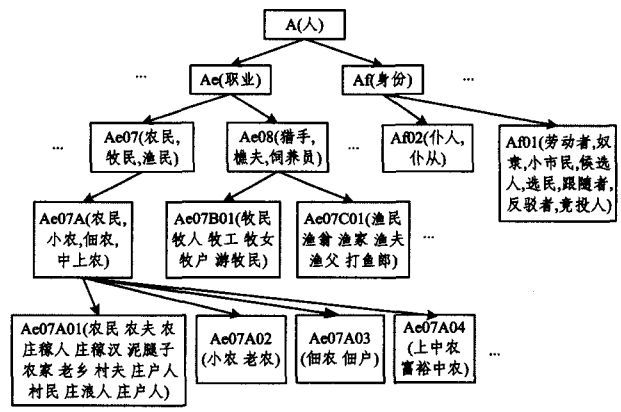

图 2　哈工大《同义词词林》扩展版词语树形结构

定义 2　假设词语 w_1 在词典中共有 m 个编码，分别为：$Code_{11},Code_{12},\cdots,Code_{1m}$，词语 w_2 共有 n 个编码，分别为：$Code_{21},Code_{22},\cdots,Code_{2n}$，则词语 w_1 和 w_2 的语义距离 $Dis(w_1,w_2)$ 定义为：

$$Dis(w_1,w_2)=\min_{i=1,2,\cdots,m;j=1,2,\cdots,n} Dis(code_{1i},code_{2j})$$

定义 3　假设两个编码 $Code_1$，$Code_2$ 从第 $i(1\leqslant i\leqslant 5)$ 层上编码不同，由编码的特性可知：i 越高时，语义距离越大。所以本文给每个层次分配不同的权重。定义权重组 $weights=[W_1,W_2,W_3,W_4,W_5,W_f]$，其中，$W_1>W_2>W_3>W_4>W_5>W_f$，本文中 $weights$ 定义为 [1.0, 0.5, 0.25, 0.125, 0.06, 0.03]。

定义 4　编码 $Code_1$，$Code_2$ 的距离 $Dis(Code_1,Code_2)$ 定义为：

$$\begin{cases} init\_dis, & F_1=\text{“@”}\ \text{或}\ F_2=\text{“@”} \\ 0, & Code_1=Code_2\ \text{且}\ F_1=F_2=\text{“=”} \\ weights[5]\times init\_dis, & Code_1=Code_2\ \text{且}\ F_1=F_2=\text{“\#”} \\ weights[i-1]\times init\_dis, & Code_1\ \text{和}\ Code_2\ \text{从}\ i\ \text{层开始编码不相同} \end{cases}$$

其中，$init\_dis$ 为自定义距离初始值，本文中 $init\_dis$ 取为 10。

2.2 基于密度的聚类

得到词语语义距离后，对所有词语进行聚类[18,19]。

因密度聚类算法不需要设置聚类的数目，并且能发现任意形状的聚类，而且密度聚类对噪声不敏感，可以把噪声归成一个单独的类，因此文档选用密度聚类算法。算法步骤如下：

输入：ε为半径，MinPts为给定点在ε领域内成为核心对象的最小领域点数，D为集合

输出：目标类簇集合

方法：repeat 判断输入点是否为核心对象

核心对象｛

找出核心对象的ε领域中的所有直接密度可达点

｝

非核心对象｛

如果p是一个边界对象，p被标记为噪声

｝

until 所有输入点都判断完毕

repeat 针对所有核心对象的ε领域所有直接密度可达点找到最大密度相连对象集合，合并密度可达对象。

until 所有核心对象的ε领域都遍历完毕

针对本文的应用，半径ε取经验值6，$MinPts$ 取集合 D 中词语个数的10%。

2.3 词语数据约简

完成聚类后，选择聚类结果中包含词语数据最多的 N 个子集作为提取特征的主题相关类。

定义5 $M(D)=\{M_1,M_2,\cdots,M_k\}$ 为对 $D(W)$ 密度聚类后得到的主题相关类集合，其中 $M_k=\{w_i,w_j,\cdots,w_k\}$ 为一个主题相关类，w_i 为主题相关类 M_k 的主题相关词语。

定义6 $U(W)=\{w_l,w_m,\cdots,w_n\}$，其中 w_n 为《同义词词林》未登录词。

2.4 特征词提取

得到主题相关类集合 $M(D)$ 后，通过信息增益方法，对词语进行评价，选取若干分类效果最好的词语作为特征词。

信息增益(Information Gain，IG)是一种有效的特征选择方法。在信息增益中，重要性的衡量标准就是看特征能够为分类系统带来多少信息，带来的信息越多，该特征越重要。因此选取信息增益来对特征词进行提取。

$$InfGain(F)=P(W)\sum_i P(Cl_i|W)\log\frac{P(Cl_i|W)}{P(Cl_i)}+$$

$$P(\bar{W})\sum_i P(Cl_i|\bar{W})\log\frac{P(Cl_i|\bar{W})}{P(Cl_i)}$$

定义7 利用信息增益方法对主题相关类集合 $M(D)=\{M_1,M_2,\cdots,M_k\}$ 和未登录词集合 $U(W)=\{w_l,w_m,\cdots,w_n\}$ 进行评价后，按照评价结果信息，从高到低选择 n 个特征词语，得到一个有序的序列 $D(W_t)=\{w_1,w_2,\cdots,w_n\}$，其中 w_i 为利用信息增益方法评价后筛选出的特征词语，w_i 的排序越靠前说明其所携带的分类信息越多，分类效果越好。

3 系统整体结构与流程

本文在特征提取过程中融入语义特征，提出综合语义信息和统计信息的文档特征提取方法。算法逻辑结构如图3所示，主要由5个模块组成：文本预处理模块、判断《同义词词林》未登录词模块、词语语义距离计算模块、词语聚类模块和词语数据约简模块。

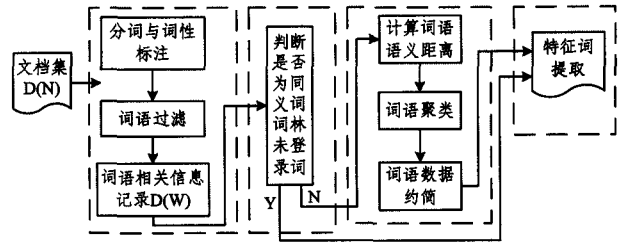

图3 基于语义的文档特征提取方法

算法首先对一类文档集 $D(N)$ 进行分词；然后对分词结果进行停用词过滤，得到分词结果集 $D(W)$；之后判断结果集 $D(W)$ 中的词语是否为《同义词词林》的未登录词，如是未登录词则加入未登录词集合 $U(W)$，如不是未登录词则根据哈工大《同义词词林》扩展版计算词语的语义距离，对词语集进行密度聚类，得到主题相关类集合 $M(D)$，再对主题相关类集合 $M(D)$ 和未登录词集合 $U(W)$ 中的词语利用信息增益进行筛选，得到特征词集合 $D(W_t)$。

算法处理步骤如下：

输入：文档集 D(N)

输出：文档集 D(N)的特征

1. 对文档集 D(N)进行分词和词性标注，获得候选词语列表 Candidate Words。
2. 去除 Candidate Words 中的停用词后，获得分词结果集 D(W)。
3. 判断分词结果集 D(W)中的词语是否为《同义词词林》未登录词，如为《同义词词林》未登录词，则将其加入未登录词集合 U(W)。
4. 根据2.1节中的算法计算词语语义距离。
5. 根据2.2节中的算法对分词结果集 D(W)中登录词进行密度聚类，确定主题相关类集合 M(D)。
6. 根据2.4中的算法对主题相关类集合 M(D)和未登录词集合U(W)中的特征进行筛选，得到代表这类文档的特征集合 $D(W_t)$。

4 实验验证

4.1 实验基础

实验数据采用复旦大学计算机学院提供的文档集，其类别数 $|C|=20$，文档数 $|D|=19637$。采用 ICTCLAS 分词系统进行分词，得到特征词数 $|T|$ 约13万。采用 TFIDF 对所有文档进行加权[20]：

$$\omega(d,t)=TF(t,\{d\})\times\log(\frac{|D|}{|D(t)|})$$

上式表示文档 d 的特征词 t 的权值。

对文档集 D 中的所有文档进行统一加权后，采用5-交叉验证实验，即所有文档随机均分成5组，1组为测试集，其他4组为训练集，共进行5次实验，最后将分类评价指标的平均值作为特征词选取的依据。

分类器选用 KNN[21]。K 值为15，以待判定文档 d 最邻近的 K 个文档类别。文档 d 与 d' 的相似度计算如下：

$$Sim(d,d')=\frac{\sum_{\forall t\in T}\omega(d,t)\times\omega(d',t)}{\sqrt{\sum_{\forall t\in T}(\omega(d,t))^2\sum_{\forall t\in T}(\omega(d',t))^2}}$$

文档 d 与类别 c_i 的相似度计算：

$$f(d,c_i)=\sum_{d'\in K\_set}Sim(d,d')\times g(d',c_i)$$

其中，K_set 为 d 最邻近的 K 个文档。当 d' 属于 c_i 类时，$g(d',c_i)=1$；当 d' 不属于 c_i 类时，$g(d',c_i)=0$。

文档 d 的所属类别 c 的判定式为：

$$Class(d,c)=\arg\max_{i=1}^{|C|} f(d,c_i)$$

4.2 效果评价标准

文档分类的评价标准有精确率、召回率、F 值。设 $x(c)$ 为测试文档的测试结果与真实类别均为 c 类的文档数，$y(c)$ 为测试文档的测试结果为 c 类的文档数，$z(c)$ 为测试文档类别为 c 类的文档数，则文档分类的评价标准定义如下。

精确率：$pre(c)=\dfrac{x(c)}{y(c)}$

召回率：$rec(c)=\dfrac{x(c)}{z(c)}$

F 值：$F(c)=\dfrac{2\times pre(c)\times rec(c)}{pre(c)+rec(c)}$

宏精确：$macro\_pre=\dfrac{\sum_{\forall c\in C} pre(c)}{|C|}$

宏召回率：$macro\_rec=\dfrac{\sum_{\forall c\in C} rec(c)}{|C|}$

宏 F 值：$macro\_F=\dfrac{\sum_{\forall c\in C} F(c)}{|C|}$

微精确率：$micro\_pre=\dfrac{\sum_{\forall c\in C} x(c)}{\sum_{\forall c\in C} y(c)}$

微召回率：$micro\_rec=\dfrac{\sum_{\forall c\in C} x(c)}{\sum_{\forall c\in C} z(c)}$

微 F 值：$micro\_F=\dfrac{2\times micro\_pre\times micro\_rec}{micro\_pre+micro\_rec}$

可以看出，宏 F 值和微 F 值综合了召回率和正确率，因此采用宏 F 值和微 F 值对特征词选取进行评价。

4.3 特征词分类能力有效性实验

根据基于语义距离的特征词提取方法(SFE)对每一特征词的分类能力进行评估，并根据评估值从大到小对所有特征词进行排序。为了证明此特征选取方法的有效性，从有序的特征集中分别“从前到后”(即正向选取)、“从后到前”(即反向选取)和“随机”(即随机选取)选取特征词 n 个，构成 3 个 n 维特征向量，分别进行文档分类效果实验。特征向量维数 n 的范围为 100～4000。每隔 100 个特征词做一次 5-交叉实验，实验结果如图 4 和图 5 所示。

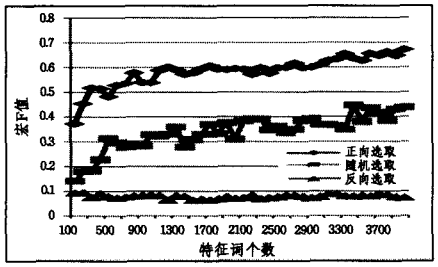

图 4 宏 F 值反映特征集分类能力

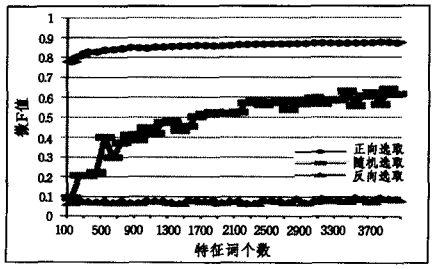

图 5 微 F 值反映特征集分类能力

从实验结果得出：①正向选取的特征词集分类效果好于反向选取特征词集的分类效果，而随机选取特征词集的分类效果介于正向选取和反向选取的分类效果之间；②随着特征词数的增加，正向选取特征词的文档分类效果逐渐变好，而反向选取特征词的分类效果基本不变，说明反向选取的特征词分类能力特别弱；③特征词数目大于 4000 以后，正向选取特征集的分类效果基本保持不变，随机选取特征集和反向选取特征集的分类效果在缓慢逐渐增大，但最大值也难于接近正向选取特征集的分类效果。其他特征词分类能力的评价实验结果与宏 F 值和微 F 值测试结果具有相似的变化趋势。以上说明有序特征集中靠前的特征词分类能力比较强，基于语义距离的特征选取方法 SFE 能够对特征词分类能力进行有效评估，成为特征词选取的依据。

4.4 文本特征选取方法对比实验

分别采用基于语义距离的特征选取方法 SFE、基于知网概念的特征提取方法 BHN、基于本体论的特征提取方法 BOL、文档频 DF、信息增益 IG、互信息熵 MI、统计量 χ^2 (CHI)、文档证据权 WET、期望交叉熵 ECE 和 DIS 进行文档分类效果对比实验，实验结果如图 6 和图 7 所示。

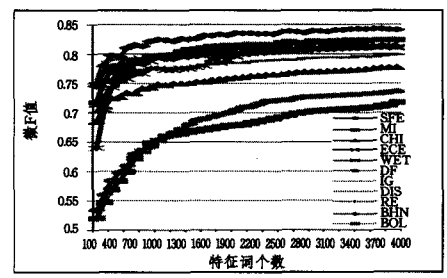

图 6 不同特征词选取方法比较(微 F 值)

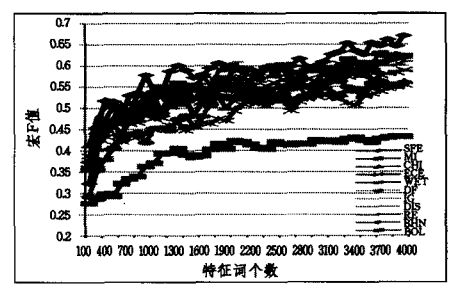

图 7 不同特征词选取方法比较(宏 F 值)

由图 6 可知，所有特征选取方法的微 F 值随着特征词的增多都能达到一个比较稳定的效果，但是 SFE 方法优于其他方法，最快达到稳定值和最大值。由图 7 可知，上述所有方法的宏 F 值随着特征词的增多都呈现明显的上升趋势，但是 SFE 方法上升趋势更明显，且达到的极值最大。当特征词多于 4000 个以后，微 F 值和宏 F 值就没有明显增大趋势，文档分类效果基本达到极值。其他特征词分类效果评价的变化趋势与微 F 值和宏 F 值实验结果变化趋势相似。

结束语 本文介绍了现有的基于概率的特征提取方法和基于语义的特征提取方法，由于其存在效率低、特征向量维度高、计算复杂、忽略低频关键词等问题，本文提出基于语义距离的文本特征提取方法，该方法首先利用《同义词词林》计算某一类别中所有词语的语义距离，然后使用密度聚类算法筛选出该类别的主题相关词，利用信息增益算法对主题相关词的分类能力进行排序，最终选取前若干个主题相关词作为该类别的特征词。算法有效性实验表明，上述方法选取的特征

词能够有效区分不同类别的文档；对比实验表明，上述方法所选取的特征词的分类能力优于其他现有方法，该方法是一种高效的文本特征选取方法。

参 考 文 献

[1] Can Do-gan, Shrikanth S N. On the computation of document frequency statistics from spoken corpora using factor automata [C]// INTERSPEECH 2013-14th Annual Conference of the International Speech Communication Association. 2013:6-10

[2] Yang Kai-feng, Zhang Yi-kun, Li Yan. Feature selection method based on document frequency[J]. Computer Engineering, 2010 (10):33-35,38(in Chinese)
杨凯峰，张毅坤，李燕. 基于文档频率的特征选择方法[J]. 计算机工程，2010(10):33-35,38

[3] Zhang Hai-long, Wang Lian-zhi. Automatic text categorization feature selection methods research[J]. Computer Engineering and Design, 2006(2):3838-3841(in Chinese)
张海龙，王莲芝. 自动文本分类特征选择方法研究[J]. 计算机工程与设计，2006(2):3838-3841

[4] Ren Yong-gong, Yang Rong-jie, Yin Ming-fei, et al. Information-gain-based text feature selection method[J]. Computer Science, 2012,39(11):127-130(in Chinese)
任永功，杨荣杰，尹明飞，等. 基于信息增益的文本特征选择方法[J]. 计算机科学，2012,39(11):127-130

[5] Guo Ya-wei, Liu Xiao-xia. Study on information gain-based feature selection in Chinese text categorization[J]. Computer Engineering and Applications, 2012(27):119-122,127(in Chinese)
郭亚维，刘晓霞. 文本分类中信息增益特征选择方法的研究[J]. 计算机工程与应用，2012(27):119-122,127

[6] Vatsavai R R, Cheriyadat A, Gleason S. Supervised Semantic Classification for Nuclear Proliferation Monitoring[C]// 2010 IEEE 39th Applied Imagery Pattern Recognition Workshop (AIPR). 2010:1-10

[7] Tang Liang, Duan Jian-guo, Xu Hong-bo, et al. Mutual information maximization based feature selection algorithm in text classification[J]. Computer Engineering and Design, 2008(13):130-133(in Chinese)
唐亮，段建国，许洪波，等. 基于互信息最大化的特征选择算法及应用[J]. 计算机工程与设计，2008(13):130-133

[8] Zhou Hai-fang, Du Yun-fei, Yang Xue-jun, et al. Study and Implement of Parallel Region-based Registration Algorithm Based on Mutual Information for Remote-sensing Images[J]. Journal of Image and Graphics, 2010(1):174-180(in Chinese)
周海芳，杜云飞，杨学军，等. 基于互信息的遥感图像区域配准并行算法的研究与实现[J]. 中国图象图形学报，2010(1):174-180

[9] Xiong Zhong-yang, Zhang Peng-zhao, Zhang Yu-fang. Improved approach to CHI in feature extraction[J]. Journal of Computer Applications, 2008(2):513-514,518(in Chinese)
熊忠阳，张鹏招，张玉芳. 基于 χ～2 统计的文本分类特征选择方法的研究[J]. 计算机应用，2008(2):513-514,518

[10] Mao Xiao-li, He Zhong-shi, Xing Xin-lai, et al. Entity relation extraction based on feature selection[J]. Application Research of Computers, 2012(2):530-532(in Chinese)
毛小丽，何中市，邢欣来，等. 基于特征选择的实体关系抽取[J]. 计算机应用研究，2012(2):530-532

[11] Liu Feng-chen, Liu Qing-wen, Hu Yue, et al. Space and time optimized algorithm of n-Gram/2L index structure[J]. Computer Engineering and Applications, 2008(5):180-183(in Chinese)
刘凤晨，刘庆文，胡玥，等. n-Gram/2L 索引结构的存储与时间优化算法[J]. 计算机工程与应用，2008(5):180-183

[12] Xu Hong-tao. The research of Web image semantic analysis and Automatic tagging [D]. Shanghai: Fudan University, 2009 (in Chinese)
许红涛. Web 图像语义分析与自动标注研究[D]. 上海：复旦大学，2009

[13] Liu Duan-yang, Wang Liang-fang. Extraction Algorithm Based on Semantic Expansion Integrated with Lexical Chain[J]. Computer Science, 2013,40(12):264-269,291(in Chinese)
刘端阳，王良芳. 结合语义扩展度和词汇链的关键词提取算法[J]. 计算机科学，2013,40(12):264-269,291

[14] Liu Jie. The research of food safety incidents cross-media information semantic analysis and classification [D]. Beijing: Beijing University of Posts and Telecommunications, 2013(in Chinese)
刘杰. 食品安全突发事件跨媒体信息的语义分析与分类研究[D]. 北京：北京邮电大学，2013

[15] Yan Le-lin. The video semantic analysis and retrieval technology based on visual and auditory information research[D]. Beijing: Beijing University of Posts and Telecommunications, 2012 (in Chinese)
闫乐林. 基于视听信息的视频语义分析与检索技术研究[D]. 北京：北京邮电大学，2012

[16] Wu Xu-dong. Subjective and objective combination of semantic similarity algorithm and its application[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013(in Chinese)
吴旭东. 主客观结合的语义相似度算法及其应用研究[D]. 南京：南京邮电大学，2013

[17] Zhai Yan-dong. The research of essay semantic web mining algorithm Based on WordNet [D]. Changchun: Jilin University, 2012 (in Chinese)
翟延冬. 基于 WordNet 的短文本语义网挖掘算法研究[D]. 长春：吉林大学，2012

[18] Wu Fang-fang, Zhao Yin-liang, Jiang Ze-fei. Novel support vector machine classifier based on density clustering[J]. Journal of Xi'an Jiaotong University, 2005, 39 (12): 1319-1322, 1348 (in Chinese)
武方方，赵银亮，蒋泽飞. 基于密度聚类的支持向量机分类算法[J]. 西安交通大学学报，2005,39(12):1319-1322,1348

[19] Li Xia, Jiang Sheng-yi, Zhang Qian-sheng, et al. A Dynamic Density-Based Clustering Algorithm Appropriate to Large-Scale Text Processing[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013,49(1):133-139(in Chinese)
李霞，蒋盛益，张倩生，等. 适用于大规模文本处理的动态密度聚类算法[J]. 北京大学学报(自然科学版)，2013,49(1):133-139

[20] Li Xue-ming, Li Hai-rui, Xue Liang, et al. TFIDF algorithm based on information gain and information entropy[J]. Computer Engineering, 2012,38(8):37-40(in Chinese)
李学明，李海瑞，薛亮，等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程，2012,38(8):37-40

[21] Liu Song-hua, Zhang Jun-ying, Xu Jin, et al. Kernel-kNN: A New kNN Algorithm Based on Informational Energy Metric[J]. Acta Automatica Sinica, 2010,36(12):1681-1688(in Chinese)
刘松华，张军英，许进，等. Kernel-kNN：基于信息能度量的核 k-最近邻算法[J]. 自动化学报，2010,36(12):1681-1688