4.0/data 2010,7(27).

基于图的特征词权重算法及其在文档排序中的应用①

黄 云[1,2], 洪佳明[2], 颜一鸣[1]

[1](吉首大学 软件学院，张家界 427000)
[2](中山大学 信息科学与技术学院，广州 510006)

摘 要：信息检索的核心工作包括文档的分类和排序等操作，如何对文档中的特征词权重进行有效度量是其中的一项关键技术。利用词的共现等关系为每个文档建立文本图，基于邻接词间重要性相互影响的思路，结合文档中特征词的词频特性，迭代计算每个词的权重，进一步结合文本图的密度等全局特性，对信息检索的结果进行排序。实验证实，算法在标准数据集上具有良好的效果。

关键词：文本图; 共现关系；文档排序；特征词权重

Graph-Based Term Weighting for Document Ranking

HUANG Yun[1,2], HONG Jia-Ming[2], YAN Yi-Ming[1]

[1](School of Software, Jishou University, Zhangjiajie 427000, China)
[2](School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract：The core work of information retrieval including document classification and ranking operations, how to effectively compute the term weight of every document is one of a key technology. Use of the word co-occurrence relationship to create a text graph for each document, based on the idea of the importance of interaction between adjacent words, combining the characteristics of the word document word frequency characteristics, we iteratively compute weighting of each word. Further combining the global properties of text graph, such as density, we could rank the results of information retrieval. Experiments confirmed that the algorithm in standard data sets with good results.

Key words：text graph; co-occurrence relation; document ranking; term weight

① 基金项目:湖南省教育厅自然科学基金(06C658)
收稿时间:2011-09-23;收到修改稿时间:2011-11-14

随着网络技术的发展，大量网络信息出现与传播，为信息检索技术提出了巨大挑战。文档分类与排序操作影响信息检索的质量与效率。文档特征词权重的度量是解决文档分类与排序问题的核心技术之一[1]。

TFIDF 算法主要考虑特征词频率及倒排文件频率，若查询词在某文档的词频较高，且对应的倒排文件频率较低，则该文档有较高的排序位置[2]。张瑜等人[3]引入类间偏斜度、类内离散度和权重调整因子，提出基于 WA-DI-SI 的改进的 TFIDF 算法，可对文档进行高效分类。由于 TFIDF 等算法忽略了词间关系，陈翀[4]等人提出共现词汇算法 FDC，通过计算各关键词的共现频度、相对距离和共文档率等，考察共现词对文档排序的影响，算法将文档中的全部关键词作为候选共现词，所需的计算量较大。周进华[5]等人考虑到共现词间的影响，利用词共现图生成多文档摘要。

本文将利用文档中小区域内的词共现关系为每个文档创建文本图，从图的整体结构考虑度量特征词的权重，将其用于信息检索的文档排序操作之中。

1 建立文本图

图 G =(V, E)常用于表示复杂系统的拓扑结构，其中 V 为顶点集，EV×V 表示边集，|V|表示顶点数，|E|为边数。顶点 $v_i \in V$ 表示图中的对象，边$(v_i, v_j) \in E$ 表示对象间关系。有时边(v_i, v_j)对应数值 w_{ij}，称为边的权重，往往用于表示对象之间关系的强度。

对文档的研究中常把特征词、句子等作为对象进行考察，分析他们之间的词法、语义等关系，用于实

现主题发现、词条消重、文本分类等功能[6]。人们经常以段落或句子为单位进行研究；对于外文文献，人们还常对固定单词长度的词序列进行研究，就像有一个固定长度的“滑动窗口”从文档头扫描到文档结尾。

下面将文档中的特征词作为对象，利用特征词之间的共现关系为每个文档建立文本图结构。

定义 1（词的共现关系）：若某文档中的特征词 v_i 与 v_j 同时出现在同一句子中或长度为 N 的“滑动窗口”内，则称词 v_i 与 v_j 之间具有共现关系。文档中词共现的次数 $n(v_i, v_j)$称为其关系权重。

定义 2（文本的共现图，简称文本图）：文本的共现图是一个无向加权图，顶点为文本中的特征词，边表示词间的共现关系，其关系权值为对应的边权重。

2 特征词权重计算及文档排序

2.1 特征词权重计算

文本图确定了特征词之间的关系，基于经验：相邻词间的权重会相互作用，且会沿着边进行传递，传递中作用力会逐渐减弱。在计算特征词的权重时，需同时考虑该词在文档中的词频与文本图中邻接的词对其的影响。由此可知，具有相同词频的两个特征词 v_i 和 v_j，若 v_i 的邻接词权重较大，则 v_i 的权重也相应较大。故顶点 vi 的权重计算公式可表示为：

$$w(v_i)=(1-\lambda)\cdot tf(v_i)+\lambda\sum_{v_j\in V(v_i)}w_{j\to i} \tag{1}$$

在公式（1）中，$V(v_i)$表示 v_i 的邻接点集合，λ 为影响因子。第一项统计词频对权重的影响；第二项计算邻接点的影响，这与邻接点的权重及相互间的关系强度有关，本文使用以下公式定义 $w_{j\to i}$：

$$w_{j\to i}=\frac{n(v_i,v_j)\cdot w(v_j)}{\sum_{v_k\in V(v_j)}n(v_j,v_k)} \tag{2}$$

其中 $V(v_j)$表示顶点 v_j 的邻接点集合。

在求解顶点权重的过程中，由于每个顶点的权重都会动态变化，结合公式（1）和公式（2），可得到顶点权重的迭代公式：

$$w(v_i)^{(l+1)}=(1-\lambda)\cdot w(v_i)^{(l)}+\lambda\sum_{v_j\in V(v_i)}\frac{n(v_i,v_j)\cdot w(v_j)^{(l)}}{\sum_{v_k\in V(v_j)}n(v_j,v_k)} \tag{3}$$

其中 $w(v_i)^{(l)}$表示第 l 次迭代时顶点 v_i 的权重。由于公式（3）不仅统计了词本身出现的数量，还利用了其邻接词的权重进行计算。因此，在文本排序和文本分类中，该权重拥有更多信息容量。

基于以上描述，可以得到计算顶点权重的算法：

1）初始设定每个顶点的权重为其词频 $w(v_i)=tf(v_i)$；

2）利用公式（3）计算顶点权重；

3）当公式（3）的迭代运算次数达到设定值 M，或者对于每个顶点 v，两次迭代运算的权重 w(v)之差均小于某给定阈值 δ，则终止迭代计算。

2.2 特征词权重在文档排序中的应用

在信息检索过程中，文档 doc 与查询 q 之间的相关度可由以下公式进行评估：

$$R(doc,q)\approx\sum_{t\in q}w(t,q)*w(t,doc) \tag{4}$$

w(t, q)表示特征词 t 在查询 q 中的权重。通常在基于特征词的查询中，对于 q 中的特征词 t，w(t, q)=1。因此，文档 doc 与查询 q 之间的相关度简化如下：

$$R(doc,q)\approx\sum_{t\in q}w(t,doc) \tag{5}$$

利用公式（3）计算得到的文档中特征词的权重，结合公式（5），并借鉴 TFIDF 的思想，可以得到：

$$R(doc,q)\approx\sum_{t\in q}\log idf\cdot\log w(doc,v)_{v\leftrightarrow t} \tag{6}$$

其中 idf 为倒排文档频率。$w(doc,v)_{v\leftrightarrow t}$ 表示特征词 t 对应的顶点 v 在文档 doc 中的权重。

通过文档间的对比分析发现，其平均词频等统计特性有较大差异，导致文本图拓扑属性差异较大，进而影响到排序结果。若文本图密度（den(G)=|E| / |V|）较小，即词间的相互影响力总体偏弱，则给定权值的特征词的相对地位则较高。于是，可用以下公式改进公式（6）的文档相关度计算：

$$R(doc,q)\approx\sum_{t\in q}\log idf\cdot\log w(doc,v)_{v\leftrightarrow t}+\frac{\mu}{1+den(\mathbf{G})} \tag{7}$$

公式（7）中 μ 为图密度影响因子，一般取值为所有文本图密度均值的倍数。

3　实验及结果分析

3.1 实验说明及评价标准

本文实验采 LETOR 3.0 数据库中的 TREC2004 做为实验的数据准备，跟踪选择.gov 中的数据作为文档集。该文档集包含已标记的 1 053 110 个网页文档，75 个查询词，每个查询词都有近 1000 个相关网页。

实验采用了 3 种不同的评估指标：MAP，P@n，BPREF。单个主题的平均准确率是每篇相关文档检索准确率的平均值，MAP 是各主题平均准确率的均值。P@n 用于测量查询词的检索结果中排名前 n 的文档与查询词的相关度。BPREF 在相关性判断不完全的情况下，计算已进行相关性判断的文档集合中，在判断到相关文档前需要判断的不相关文档篇数。

在试验中，采用 BM25 方法作为检索基准，其计算文档中特征词权重的概率排序函数如下：

$$w(t,d)=\log idf\cdot\frac{(k_3+1)\cdot qtf}{k_3+qtf}\cdot\frac{(k_1+1)tf}{tf+k_1\cdot(1-b+b\cdot\frac{l}{avl})} \quad (8)$$

其中 idf 为倒排文档频率，qtf 为查询中的特征词频率，tf 为词频，l 和 avl 分别表示文档长度和平均文档长度。此外还有三个可调节参数，，由于一般情况下 qtf 等于 1，故 k3 被消掉。试验中取 k_1=1.2，调整 b 值进行优化选择，b 取值为[0,1]区间，步长 0.05。

3.2 实验方法与实验结果

实验首先测试在定参时，公式（6）和公式（7）的实验结果与 BM25 及文献[7]中基于 R-SVM 及 R-Net 排名模型下方法的比较。其中“滑动窗口”大小 N=4，影响因子 λ =0.5，循环迭代次数 M=100。公式（5）中图密度影响因子 μ =300。结果如表 1~表 3 所示。

表 1　MAP 实验结果

排名方法	BM25	公式 6	公式 7	R-SVM	R-Net
MAP	0.2647	0.2852	0.2871	0.2755	0.2764

表 2　BPREF 实验结果

排名方法	BM25	公式 6	公式 7	R SVM	R-Net
BPREF	0.3402	0.3655	0.3696	0.3564	0.3577

实验接下来测试“滑动窗口”大小 N 的变化对 MAP 和 BPREF 评估指标结果的影响。实验结果如图 1 和图 2 所示。实验结果显示，当“滑动窗口”大小 N 从 2 到 4 时，性能提升迅速，然后缓慢增加。当 N=8 时，效果达到最优，然后，随着 N 的增加性能会缓慢下降，且逐渐趋向稳定。

表 3　P@n 实验结果

排名方法	P@1	P@2	P@3	P@4	P@5
BM25	0.6310	0.6233	0.6187	0.6112	0.6005
公式（6）	0.6423	0.6334	0.6245	0.6198	0.6115
公式（7）	0.6455	0.6346	0.6277	0.6214	0.6120
R-SVM	0.6121	0.5998	0.5911	0.5843	0.5764
R-Net	0.6355	0.6264	0.6201	0.6122	0.6025
排名方法	P@6	P@7	P@8	P@9	P@10
BM25	0.5977	0.5868	0.5811	0.5718	0.5567
公式（6）	0.6056	0.5957	0.5878	0.5792	0.5687
公式（7）	0.6070	0.5968	0.5894	0.5806	0.5696
R-SVM	0.5677	0.5588	0.5497	0.5412	0.5310
R-Net	0.5994	0.5892	0.5833	0.5735	0.5597

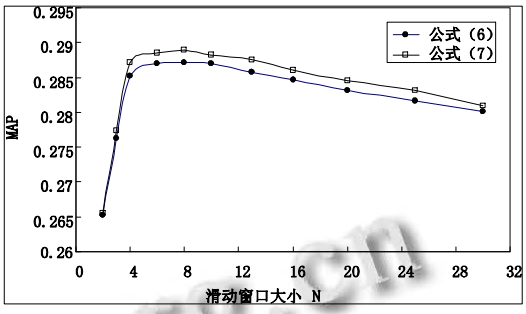

图 1　“滑动窗口”大小对排序性能的影响（MAP）

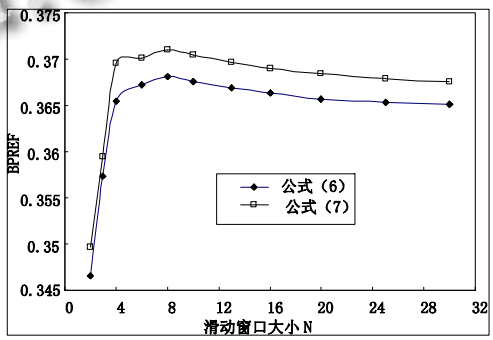

图 2　“滑动窗口”大小对排序性能的影响（BPREF）

4　结语

本文利用文档中特征词的共现关系建立词为顶点的文本图，然后基于词的结构关系计算特征词的权重，并结合文本图的全局特征，用于文档排序过程。实验

结果显示本方法对排序的性能具有一定的改善。接下来，将考虑结合词间的语义关系等改进文本图的结构，

(下转第 194 页)

融合 RFID 的无线传感器网络节能研究①

邬春学，谭石来，刘 磊

(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要：针对节省传感节点能耗和均衡整个网络中各节点能耗的问题，该文提出一种融合射频识别设计与优化路由协议的无线传感器网络节能方法。该方法首先采用 RFID 标签和阅读器分别与无线传感器网络节点以及无线设备融合，然后对该融合策略进行分析与设计，最后结合了 LEACH 算法的思想。实验仿真表明，新方法在延长整个网络生命周期和降低整个网络中的能耗方面明显优于 LEACH 算法。

关键词：射频识别；无线传感器网络；融合；轮换簇头；节能

WSN Energy-Saving Research of Integrating RFID

WU Chun-Xue, TAN Shi-Lai, LIU Lei

(University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: According to the issues of saving energy consumption of sensor nodes and balancing energy consumption of each node in the whole network, this paper proposed a wireless sensor network energy-saving method of integrating RFID design and optimizing the routing protocol. The method firstly respectively integrated RFID Tags and Readers with wireless Sensor Nodes as well as wireless devices, and then made the analysis and design for the strategy of the integration, and finally combined with the idea of the LEACH algorithm. The experiment simulation results demonstrated that the new proposed method was prior to the traditional LEACH algorithm in the terms of prolonging the network life cycle and reducing the energy consumption of the whole network.

Key words: RFID; WSN; integrate; cluster head node; energy-saving

1 引言

传感器网络由传感器节点组成,传感器节点是一个微型的嵌入式系统，它有低功耗，低成本的特点，只能携带有限的电池。近年来，无线传感器网络技术一直受能量的制约，因此，如何高效使用能量来最大化网络生命周期是无线传感器网络的首要设计目标及面临的首要挑战。目前，对无线传感器网络的节能研究主要集中在硬件设计、MAC 协议及网络路由协议等方面。

由文献[1]可知无线传感器网络中无线通信模块在发送状态的能耗最大，在空闲状态和接收状态的能耗接近，略少于发送状态的能耗，在睡眠状态的能量最少。

无线传感器节点发送 kbit 数据到与之相距 d 的另一点所消耗的能量为[2]：

$$E_{Tx}(k,d) = E_{elect} * k + \varepsilon * d^n * k \quad (1)$$

节点接收 kbit 数据消耗的能量为：

$$E_{Rx}(k) = E_{elec} * k \quad (2)$$

其中 Eelec 表示将 1bit 数据进行编码调制等处理的能耗。$\varepsilon * d^n$ 为发送 1bit 数据消耗的放大器能量，由通信距离和误码率决定。ε 为传播损耗系数，n 为传播损耗指数。N 的大小与传输环境有关，通常在 2 到 4

Iobrcbr=imreconstruct(imcomplement(Iobrd),imcomplement(Iobr));

Iobrcbr= imcomplement (Iobrcbr);

⑦ 求取局部极大值

fgm=imregionalmin(Iobrcbr);

se2= strel (ones(5,5));

fgm2= imclose (fgm,se2);

fgm3= imerode (fgm2,se2);

fgm4= bwareaopen (fgm3,20);

gradmag2= imimposemin (gradmag,bgm|fgm4);

LL= watershed (gradmag2);

⑧ 去除边界上不完整的局部图像和小目标，分割后的图像如图 6 所示：

LL1= im2bw (LL);

LL2= imclearborder (LL1,4);

LL2= bwareaopen (LL2,30);

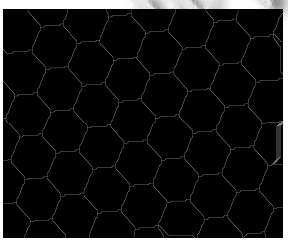

图 5 分水岭界限

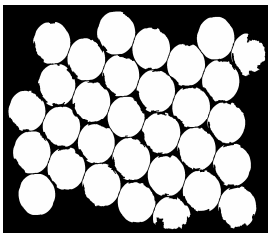

图 6 分割后的图像

3 总结

运用 MATLAB 工具箱对小球图像进行处理，设计了一种改进的分水岭分割算法，能够对粘连的小球进行合理的分割，避免了过分割的问题，最后对分割完成的图像设计程序实现小球形态参数的提取和计算，结果表明，运用这种方法能够比较准确地提取出各小球的面积像素。最终标记了圆心的小球真彩图像如图 7 所示。

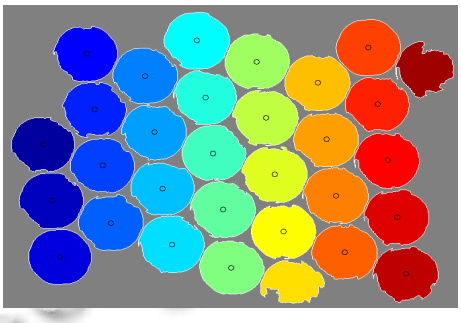

图 7 标记了圆心的小球真彩图像

参考文献

1 李丙春.图像区域边缘形状特征提取.喀什师范学院学报，2007,28(3):62–64.

2 张强,王正林编著.MATLAB 图像处理.北京:电子工业出版社,2009.6.

3 何振宇,张森文等.基于图像处理的 AFM 细胞形态参数的自动获取.电子显微学报,2008,(2).

4 景晓军.图像处理技术及其应用.北京:国防工业出版社,2005.

5 张洪刚.图像处理与识别.北京：北京邮电大学出版社,2006.

6 余松煜,周源华,张瑞.数字图像处理.上海:上海交通大学出版社,2007.

7 Gonzalez RC. Digital Image Processing Using Matlab.北京：电子工业出版社,2004:463–465.

(上接第 218 页)

从而优化特征词权重的计算。

参考文献

1 Agosti M, Crestani F, Pasi G. Lectures on Information Retrieval. Springer Berlin / Heidelberg. 2003.

2 Wang H, Rajman M, Guo Y, et al. NewPR-Combining TFIDF with Pagerank. ICANN 2006, Part II, LNCS 4132, 2006, 932–942.

3 张瑜,张德贤.一种改进的特征权重算法.计算机工程,2011,37(5):210–212.

4 陈翀,彭波,闫宏飞.一种词汇共现算法及共现词对检索系统排序的影响. 清华大学学报(自然科学版),2005,45(S1):1857–1860.

5 周进华,刘贵全.基于衰减词共现图的多文档摘要研究.小型微型计算机系统, 2009,30(1):173–177.

6 Blanco R, Lioma C. Random walk term weighting for information retrieval. Proc. of the 30th SIGIR. Amsterdam, The Netherlands:ACM 2007,829–830.

7 李慧,李存华,王霞.基于特征选择的网页排名算法.计算机工程,2010,36(13):37–39.