# 基于小波变换的文档相似性检索方法

葛永亮，项颖，耿俊成
(广东工业大学 信号与信息处理专业，广东 广州 510006)

**摘　要：**随着电子文档的出现和发展，电子图书馆在人们的生活中正起着越来越重要的作用，如何从大量文档中获取有用的信息，成为信息检索领域的关键技术。一种改进的文本检索算法使用目前较先进的小波变换算法，有效地结合传统的向量空间法和近似法的优点，并且利用CBW算法来进行加权，有效避免了交叉比较关键词的问题。此算法用以比较文档之间的相似性，实验表明是一套性能较好的文档相似性检索方法。

**关键词：**小波变换；词信号；文本检索；幅值；零相位精准
**中图分类号：**TP391.3　　**文献标识码：**B

# Algorithms for theses similarity retrieval based on wavelet transform

GE Yong Liang，XIANG Ying，GENG Jun Cheng
(Faculty of Information Engineering ，Guangdong University of Technology， Guangzhou 510006，China)

**Abstract：**With the emergence and evolution of electronic documents，electronic library is playing an important role in people's life.How to retrieve useful information from a large amount of information has become a key technology in the information area. In this paper，we present an improved document intrieval algorithm. Using the advantage of vector space methods and proximity search，it is a document similar retrieval technique with a better performance which uses wavelet transform technique and CBW weighting and can be used to compare the similarity of documents.

**Key words：**wave transform；term signal；document retrieval；magnitude；zero pahse precision

随着网络技术的迅猛发展，Web成为人们查询科技论文的主要平台。目前较常用的搜索方式主要有两种：空间向量法和近似法，但是前者仅考虑关键词在文档中出现的次数，其精确度较差，近似方法则只考虑关键词之间的关系，进行交叉比较，运行速度慢。基于离散小波变换，使用CBW[1]加权算法的检索方法则综合二者的优点去其缺点，能有效地提高检索速度和精度。

## 1 背景知识

文本检索是一项交叉学科，从大学科上来看，横跨了计算机、情报、数理统计等学科，从具体研究方向上来看，包含文本检索、自然语言处理、数据挖掘、机器学习等技术。因此 要对文本检索进行全面的综述是一件非常困难的事情。现介绍一些常用的文本检索模型，其中包括基于文字的检索模型和基于结构的检索模型。而基于文本的检索模型又包括:向量空间模型、近似模型、概率模型和统计语言检索模型。而基于结构的文本检索模型又包括:内部结构检索模型[2]、外部结构检索模型。以上检索模型，功效单一，性能较差，仅仅凭借以上单个模型来实现检索不再具有实际使用性。本文结合向量空间模型、近似模型和内部结构检索模型的优点，完善并提出一种新的实用模型——基于小波的检索模型。

本文献检索模型对所采用的一些关键概念信息进行一些阐述，对关键词在文档中进行词信号提取、加权算法、小波变换、 取幅值和零相位精准，最终给出了一个具体的说明。

### 1.1 关键词加权

在查询过程中，由于关键词很可能存在同义词、一词多义等现象，所以检索结果也许并不令人满意，而以前的加权算法也仅仅是寻求文档和关键词之间的关系，不能有效地解决这个问题，影响检索结果精度[3]。所以本文引入一种新的CBW关键词加权算法，它以关键词$q$和NetWord为输入，最终通过融合算法得到所需要的每个关键词的加权值

$CBW_q$，这种加权算法的理论和实践结果证明[4]其能大大提高检索的精度。

WordNet 是普林斯顿大学的一个项目。该项目被称为“英语词汇数据库”，是一个透过将同义词的集合集中到称为同义词集(synonym sets)或同义集(synsets)的群组中来描述及分类单字和概念的系统。因为它的开放性意味着任何工作人员都可以使用它，所以它十分重要。WordNet 有“不受限制的”许可证。它很像 BSD 许可证，因为唯一真正的限制是不可以盗用普林斯顿大学的商标来宣传 WordNet 的任何相关产品。

### 1.2 词信号

首先把一文本分成$B$个组成部分，查询词在文本每一部分出现的次数记作$f,d,B$-1[5]，则整个文本的的词信号组成表示为$f,d,t$=[$f,d$,t,0 $f,d$,t,1 ...$f,d$,t,$B$-1]其中$d$表示文档，$t$表示词信号。后面的实验则取$B$=8，也就是说把文档分成8个部分，分别提取关键词在每一部分出现的次数$f,d$,t,$B$-1，组成了词信号 $f,d$,t。如果两文档中的词信号在相同位置出现则说明两篇文档比较相似，位置相邻则相似性次之。

### 1.3 光谱词信号

对查询词进行小波变换求得光谱词信号[3]，如果只用词信号会降低检索结果的精确度，用光谱信号则是一种更加逼近的方法。

进行小波变换就可以得到查询词的光谱信号。变换后的光谱信号表示为：

$$\overline{\zeta}_{d,t}=[\zeta_{d,t,0}\ \zeta_{d,t,1}\cdots\zeta_{d,t,B-1}]$$

其中$\zeta_{d,t,b}=H_{d,t,b}\exp(\theta_{d,t,b})$，$H_{d,t,b}$是光谱信号的幅值，相位为$\theta_{d,t,b}$光谱信号可以通过其乘积求得。

其中小波变换算法如下：

设词信号为$x$，初始化输出结果$y$=0；初始化计数：$n$=#($x$)；

当$n$不等于1的时候：

(1)对词信号使用2尺度小波函数，然后将$y_{n/2,n-1}=D_2(H(x))$存入小波信号中；

(2)对词信号使用2尺度函数，把它作为新的输入词信号: $x=D_2(G(x))$；

(3)$N=n/2$；应用剩下小波信号的输入值到第零个值$y_{00}=x$。

上面的式中，#()是词信号的数量，$D_2$()是滤波函数，$y_{n/2,n-1}$是词信号$y$的第$n$/2到第$n$个值。小波的高通滤波器系数是[$1/\sqrt{2}$，$-1/\sqrt{2}$]，低通滤波器系数是[$1/\sqrt{2}$，$1/\sqrt{2}$]。

文档的相似性越大，则它们的幅值就会越大，两文档中词信号位置越接近，则表现为相位越相似，所以要分别来处理它们的幅值和相位信息。

## 2 基于小波变换的文档相似性比较模型的建立

综合上述背景知识，可以得到一种文档相似性模型;

(1)找出关键词$q$在文档中的词信号{$f,d$,t,0 $f,d$,t,1 ... $fd$,t,$B$-1}[6]；

(2)加权$wd,t,b=fd,t,b\times CBW_q$($d$是关键词$q$的词信号)；

(3)对$w_{d,t}$进行小波变换得到$\zeta_{d,t}$；

(4)计算信号组成部分的幅值$H_{d,t}=|\zeta_{d,t}|$；

(5)计算信号组成部分的相位$\theta_{d,t}=\dfrac{\zeta_{d,t}}{H_{d,t}}$；

(6)由于文档的幅值越大其相似性越大，而其相位值越相似，则信号位置越接近，所以取其乘积，取一平均值来表达这种相似性，得到下面公式：

$$\phi=\frac{\sum_{i=1,j}^{i=2}H_{d_i,t_j}\times\left|\sum_{i=1,j}^{i=2}\theta_{d_i,t_j}\right|}{4\times\#(t)\times\#(t)}$$

$Hd_i,t_j$是第$i$篇文档第$j$个关键词的幅值。$\theta d_i,t_j$是第$i$篇文档第$j$个关键词的相位精准，#($n$)是查询词的个数，$\sum_{i=1,j}^{i=2}Hd_i,t_j$则用来求取两文档的所有词信号幅值之和，$\sum_{i=1,j}^{i=2}\theta d_i,t_j$用来求取两文档所有光谱信号的相位值，这两个结果分别除以2倍的#($n$)才能得到一个平均幅值和平均相位值；

(7) S=$\|\phi\|$ 对其求范数得到最后分数。

## 3 实验

此例中取$B$=8，用关键词“经济基础，上层建筑”在1、2文档中的词信号来进行分析，如表1所列。

表1 关键词在1、2、3文档中的词信号

| 查询词 | 经济基础 | 上层建筑 |
|---|---|---|
| 文档1 | [0 0 0 1 0 0 0 0] | [0 0 0 0 0 0 1 0] |
| 文档2 | [0 1 0 1 0 0 0 0] | [0 0 0 0 1 0 0 0] |
| 文档3 | [0 1 0 0 0 1 0 0] | [0 0 0 0 0 1 0 0] |

分析文档2分别与文档1、3可以发现，它们的第1个关键词在第4个词信号位置相同且出现次数均为1；而第2个关键词信号出现的位置则相隔一个位置。同样文档2、3的第1个关键词的第2个词信号完全一样，而第2个词信号出现位置则相邻，这说明文档2和文档3有更大的相似性，而实验结果也证明了其正确性。

对表1中的词信号进行小波变换可以得到如表2所列值(此处举例不再使用加权) 。

对1、2文档中的小波信号经济基础和上层建筑分别使用等式3中的求平均幅值和零相位精准部分可以得到：

平均幅值：[$\frac{3}{4\sqrt{2}}$ $\frac{1}{4\sqrt{2}}$ $\frac{1}{4}$ $\frac{1}{2}$ 0 $\frac{1}{2\sqrt{2}}$ $\frac{3}{4\sqrt{2}}$ $\frac{1}{4\sqrt{2}}$]

零相位精准：[1 $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ 0 $\frac{1}{2}$ $\frac{3}{4}$ $\frac{1}{4}$]

文档1与标准文档2相似度的分数 $s2$=0.559，同样可以得到文档3与标准文档2的相似度分数 $s3$=4.518。

(下转第82页)

(上接第78页)

实验与理论都证明文档3与标准文档2更具相似性。

表2 进行小波变换之后的光谱信号

| 查询词 | 经济基础 | 上层建筑 |
|---|---|---|
| 文档1 | $[\frac{1}{\sqrt{2}}\ 0\ \frac{-1}{2}\ \frac{1}{2}\ 0\ \frac{-1}{\sqrt{2}}\ \frac{-1}{\sqrt{2}}\ 0]$ | $[\frac{1}{2\sqrt{2}}\ \frac{-1}{2\sqrt{2}}\ 0\ \frac{-1}{2}\ 0\ 0\ 0\ \frac{1}{\sqrt{2}}]$ |
| 文档2 | $[\frac{1}{\sqrt{2}}\ 0\ \frac{-1}{2}\ \frac{1}{2}\ 0\ \frac{-1}{\sqrt{2}}\ \frac{-1}{\sqrt{2}}\ 0]$ | $[\frac{1}{2\sqrt{2}}\ \frac{-1}{2\sqrt{2}}\ 0\ \frac{1}{2}\ 0\ 0\ \frac{-1}{\sqrt{2}}\ 0]$ |
| 文档3 | $[\frac{1}{\sqrt{2}}\ \frac{1}{\sqrt{2}}\ 0\ 0\ \frac{-1}{\sqrt{2}}\ \frac{-1}{\sqrt{2}}\ 0\ 0]$ | $[\frac{1}{2\sqrt{2}}\ \frac{-1}{2\sqrt{2}}\ 0\ \frac{1}{2}\ 0\ 0\ \frac{-1}{\sqrt{2}}\ 0]$ |

本文给出了一种新的基于小波变换的检索算法，它在检索文档的相似性方面很好地结合了空间向量法和近似法的优点，在运算速度和精度上有一个质的提高，能很好地满足用户的需求。检索过程中，它不用再对论文关键词之间进行交叉比较，大大减少了运算量。而使用CBW加权算法又可以减少关键词多语意和同义词的影响。根据小波变换检索文档的相似性，能把与目标文档最相似的文档返回给用户。

**参考文献**


[1] JOHN Z，BRIJESH V. Concept-based term weighting for web information retrieval [R]. Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05) 0-7695-2358-7/05.

[2] 张文进.文本信息检索中的概率模型[J].情报，2005，(3)：101-110.

[3] 郑 毅.基于概念空间的文本检索系统[J].计算机工程与应用，2002，40(12)：69-71.

[4] 王晓丽，王文杰.基于向量空间模型的文本检索系统[J].微电子学与计算机，2006，23(6)：188-190.

[5] 刘海峰. 基于向量模型的文本检索若干问题研究[J].情报杂志，2006，(10)：57-62.

[6] LAURENCE A.F.P，KOTAGIRI R，MARIMUTHU P.A novel document retrieval method using the discrets wavelet transform[J]. ACM Transaction on Information Systems，2005，23(3):267-298.


(收稿日期:2008-12-13)