文章编号: 1001-0920(2009)07-1078-05

基于离散小波变换的数据库水印检测算法

姜传贤, 陈孝威
(贵州大学 计算机科学与技术学院, 贵阳 550025)

摘 要: 为了克服空域水印技术的不足,提出一种基于离散小波变换的关系数据库水印算法. 根据过滤规则,从关系数据库中抽取某一类的数据,组成二维信号,将水印嵌入到二维信号的小波域中. 使用该算法嵌入的数字水印具有很好的隐蔽性,对原始载体的影响很小. 同时,嵌入的数字水印具有较好的鲁棒性,常规的数据库的处理方法对其影响较弱.

关键词: 关系数据库水印; 离散小波变换; 信息检测; 数值型属性
中图分类号: TP391 **文献标识码**: A

New algorithm for relational database watermarking based on DWT

JIANG Chuan-xian, CHEN Xiao-wei
(School of Computer Science and Technical, Guizhou University, Guiyang 550025, China. Correspondent: JIANG Chuan-xian, E-mail: emailfeibai@163.com)

Abstract: In order to enhance the robustness of database watermark, the paper proposes a relational database watermarking algorithm based on discrete wavelet transform (DWT). Some types of data are selected from the relational database according to filtering rules, and two-dimension signal is formed. Then, the watermark is embedded into the wavelet domain of two-dimension signal. Experimental results show that the embedded digital watermarks with the proposed algorithm are invisible and robust enough against the commonly used database processing techniques.

Key words: Relational database watermarking; DWT; Information detecting; Numerical attribute

1 引 言

当前,数据库技术飞速发展,数据库广泛应用,数据库安全问题也越来越受到重视. 由于关系数据库的特殊性[1],此水印技术的研究在国内外还处于起步阶段. 2002 年,Agrawal 等人首次提出关系数据库水印;文献[1,2]提出了对关系数据库中数值型属性值满足条件进行标记的策略;[3]在[1]的基础上将有意义水印信息嵌入数据库中;[4,5]将一幅图像嵌入关系数据库中. 以上这些都是基于最低有效位的空域数据库水印方案[6]的研究. 经过查阅国内外许多文献,较少有基于变换域方法来研究关系数据库的水印.

通过借鉴多媒体水印技术和分析关系数据库本身的特点,提出一种新型基于离散小波变换的关系数据库水印检测算法. 分析了关系数据库离散小波变换的高频系数服从高斯分布的特点,同时使用服从高斯分布的水印模板,为线性相关检测技术使用创造了条件,达到了线性最优检测技术[7,8]和离散小波技术应用到关系数据库领域的目的. 理论和实验分析表明,该方案具有较强的鲁棒性和较好的"不可见性".

2 相关知识

文献[1,2]假设关系数据库中存在一些属性值可以修改,而不影响数据库的正常应用. 例如气象数据、地理数据、经济数据等,需保证这些数据正常使用,且尽可能小的修改. 例如某市年降雨量为 1235.5 mm,通过加入水印降雨量为 1235.9 mm,这种改变不会影响它的使用.

设待嵌入水印的数据库关系用 $R(p, a_1, \cdots, a_v, \cdots)$ 表示,其中 p 为主关键字且不能改变,$a_1, \cdots, a_v$ 为 v 个可嵌入水印的数值型属性列(不包括主关键字),R 由 n 个元组 $r_1, r_2, \cdots, r_i, \cdots, r_n$ 组成,每个元

收稿日期: 2008-09-13; **修回日期**: 2008-11-18.
基金项目: 贵州省科研基金项目(黔科合 20052109).
作者简介: 姜传贤(1978—),男,湖南邵阳人,助教,博士生,从事数字图像处理、模式识别等研究; 陈孝威(1945—),男,贵阳人,教授,博士生导师,从事图像处理与多媒体、计算机视觉等研究.

组 r_i 都有 1 个主关键字 $r.p$ 和 v 个数值型属性值 $r.a_1, \cdots, r.a_i, \cdots, r.a_v$.

定义 1 (过虑规则)　从记录数据库操作的日志文件中,统计分析出某一类元组具有相对较少的更新操作.

3　水印隐藏与检测

本文算法的基本思想是,将水印信息以加性方式嵌入到服从高斯分布的水印载体中,并按照水印载体数据的选择算法,放回到原关系数据库中,实现水印的嵌入. 在检测端,准确重构水印嵌入时所采用的模板序列,并将其连同待检测信号一起输入相关检测器中进行水印检测. 水印隐藏的流程如图 1 所示.

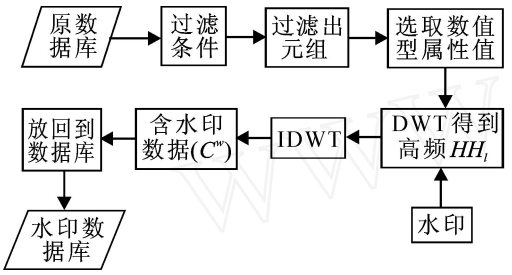


图 1　水印隐藏的流程

水印信息被编码成二进制序列 $B = \{b_1, b_2, \cdots, b_i, \cdots, b_L\}$ $(b_i = 0, 1)$, $|B| = L$, 水印嵌入模型为

$$C^w = C + \alpha F(b) W. \tag{1}$$

其中: C^w 代表嵌入水印信息后的信号; C 代表载体信号; W 代表水印信息模板; C^w, C, W 均为一维矢量; α 是水印嵌入强度因子; L 是长度(可以小于或等于水印载体长度); 为了方便,用 $W1$ 表示 $F(b)W$. 函数 $F(b_i)$ 定义为

$$F(b_i) = \begin{cases} 1, & b_i = 1; \\ -1, & b_i = 0. \end{cases} \tag{2}$$

W 是在水印钥匙 K 的控制下生成的伪随机序列,水印攻击者不知水印钥匙 K,因此无法准确复制被嵌入的水印模板序列,从而无法检测出水印信息. 同时, W 服从标准正态分布,并与载体信号 C 独立. 在载体信号 C 也服从正态分布的理想情况下,线性相关检测在这个水印系统中能够达到最优[7,8]. 下面着重讨论该算法的几个重要过程.

3.1　水印载体数据的选择

为了使用相关性检测器进行水印检测,所选择的数据满足相关假设,并通过以下步骤来获取载体信号. h 表示哈希函数, μ 用来控制某一类元组的大小, $m \times m$ 表示数列元素的个数, v 控制可嵌水印数值型属性的个数.

Step1: 对关系数据库进行排序,产生出伪随机数列 $\{sn_i \mid 0 < i \leq m \times m, i \in N\}$;

Step2: 从原关系数据库中根据过滤规则和嵌入强度因子测出一类元组,从其中取出满足条件 $(h(r.p) \bmod \mu = 0)$ 的所有元组,用序列 sn 标记;

Step3: sn_i 标记的元组中,取出 $r.a_j$ 属性值,组成一个 $m \times m$ 二维信号,用 A (没填满用随机数填满) 表示,其中 $j = h(r.p) \bmod v$;

Step4: 将 A 经过 l 级小波变换,抽取第 l 级高频系数 (HH_l),组成水印载体序列 C,其长度为 N,其中 $N = \frac{m}{l} \times \frac{m}{l}$.

为符合归一化相关检测器的条件,要求选取 l 级小波变换的第 l 级高频系数较好地服从高斯分布. 高频系数尽管在多媒体水印领域中,在其中嵌入的水印鲁棒性不强,但是在关系数据库假设下,能很好地设计出效果较好的数据库水印系统. 因此采用高频系数作为水印载体,选择适当的嵌入强度因子和阈值,系统性能较好.

3.2　嵌入强度因子的选择

由式(1) 可以看出,为降低检测的错误率,应尽可能地提高相关性检测性能. 但是受水印保真度的约束,相关性检测不可能无限地提高. 同时,不同的数值型属性值或者同一数值型属性不同数位,其掩蔽水印的能力也不同. 因此,必须根据关系数据库的具体特征选择相应的强度因子.

在关系数据库中嵌入水印,使得关系数据库发生改变,这些改变是可以接受的,且不影响关系数据库的正常使用. 因此,需要设计一种可以接受的误差规则[2],即

$$(d_i - v_i)^2 < t_i, \quad \forall i = 1, 2, \cdots, n,$$

$$\sum (d_i - v_i)^2 < t_{\max}.$$

其中: $C = \{d_1, \cdots, d_n\} \subset R$ 表示能嵌入水印载体数据, $C^w = \{v_1, \cdots, v_n\}$ 表示嵌入水印后载体数据, $T = \{t_1, \cdots, t_n\}$ 和 $t_{\max}$ 表示可以接受的改变量的阈值,即 T 表示在不影响数据库的正常使用下,某一个数值型属性值可改变的量的范围,而 $t_{\max}$ 表示可以接受的累计误差.

由以上规则可以得出,尽管每一个数值型属性值都有一个值域,但从整体上讲:数值型属性值小,可改变的量小;数值型属性值大,可改变的量相对较大. 因此,在本文算法中,嵌入强度因子被表示为载体信号 d_i 和 t_i 的函数

$$\alpha_i = f(d_i, t_i). \tag{3}$$

3.3　水印检测

在水印检测过程中,虽然不需要原关系数据库,但是数值型属性值统计特性抽取规则和小波高频系

数的抽取规则以及产生水印信息模板序列的密钥 K,却是水印嵌入者和水印检测者必须预先知道的.

水印检测具体算法如下:

Step1: 从待检测的关系数据库中按照水印载体数据的选择算法获取待检测的序列 $C2$;

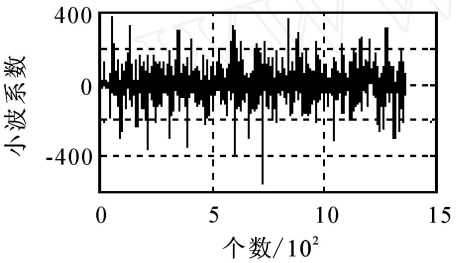
Step2: 将待检测的序列 $C2$ 与水印序列 $W1$ 做相关检测,判定是否有水印存在,即相关系数

$$\tau = \frac{W1^{\mathrm{T}} C2}{\sqrt{W1^{\mathrm{T}} W1}\ \sqrt{C2^{\mathrm{T}} C2}},\ 0 < \tau < 1, \quad (4)$$

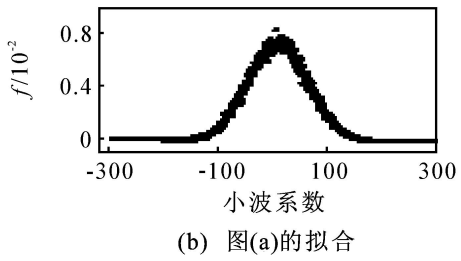
其中 τ_0 为阈值. 当 $\tau > \tau_0$,水印信息存在;否则水印不存在.

4 关系数据库小波系数的分析

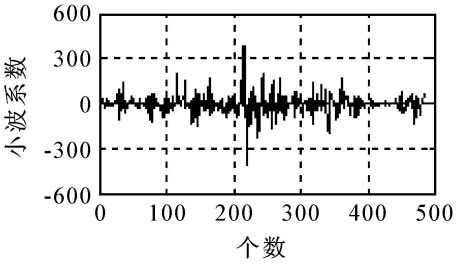
小波高频系数代表图像的边缘和纹理,且当图像的边缘和纹理较丰富时,小波高频系数值就会增大. 因此由数值型属性值组成的载体,可以想象成一幅"噪声图像"的小波高频系数统计分布,且近似于高斯分布. 而从原关系数据库中根据过滤规则过滤一定数目的数值型属性值,组成一个 $m \times m$ 二维向量 A,可以近似看成一幅"噪声图像". 下面给出 A 小波高频系数 G 的统计分布情况,如图 2 所示.



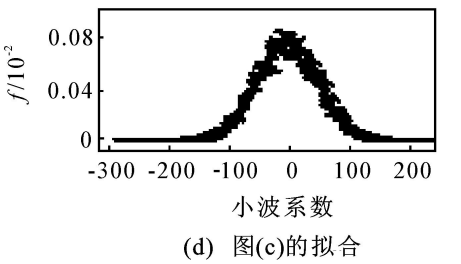
(a) 二级小波高频系数分布



(b) 图(a)的拟合



(c) 三级小波高频系数分布



(d) 图(c)的拟合

图 2 小波高频系数分析

图 2(a) 为二级小波分解的高频系数统计分布,图 2(b) 为三级小波分解的高频系数统计分布. 以上统计分布情况表明,小波高频系数近似服从高斯分布. 为了进一步说明水印载体序列服从高斯分布,通过公式 $f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(t-u)^2}{2\sigma^2}}$ 对图 2(a) 和图 2(c) 进行拟合,分别见图 2 (b) 和图 2(d). 其中 $u = 0$,$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} G_i^2$ 分别对应样本均值和样本方差.

5 算法分析

水印信息匹配产生两种可能结果:"匹配"或者"不匹配",其概率为 1/2. 从概率统计的角度分析,其分布规律为 0-1 分布. 若从 k 个待检测的数值型属性值中检测出了 i 个数值型属性值含有匹配信息,实际上可以看作是一个成功 i 次,失败 $k - i$ 次的 k 重贝努里实验,其概率为

$$b(i,k,1/2) = \binom{k}{i} \left(\frac{1}{2}\right)^{i} \left(1 - \frac{1}{2}\right)^{k-i}. \quad (5)$$

如果信息至少匹配 $k/2$,则有以下概率:

$$B(i,k,1/2) = \sum_{i=k/2}^{k} b(i,k,1/2). \quad (6)$$

显然它的概率值是 1/2. 为判定数据库的版权,设 $\eta = 1/k$,η 的概率分布满足二项分布,η 分布于区间 $[i_0/k, 1]$(即 i 分布于区间 $[i_0, k]$) 的概率为

$$B(i,k,1/2) = \sum_{i=i_0}^{k} b(i,k,1/2) \geq 1 - \delta. \quad (7)$$

选取置信因子 $\delta(0 < \delta < 1)$,δ 满足 $B(i,k,1/2) \leq \delta$ 时,可得出 i_0 的最小值 $i_{\min}$ 和 $\eta_{\min} = i_{\min}/k$. 依据概率统计中的"小概率原理",可以理解当数据库关系在嵌入水印后受到攻击,部分水印信息丢失,但只要 $1 > \eta > \eta_{\min}$,便可判定数据库关系中嵌入了水印,可确定数据库关系的版权. 对置信因子 δ,增大 δ 值,可提高检测率,但可能出现在没有嵌入水印的数据库关系中验证出水印(称误报). 因此,δ 值越大,误报的可能性也就越大;相反,取较小的 δ 值,虽减小了水印系统的误报率,但降低了水印的鲁棒性,应根据具体应用情况选取合适的 δ. 阈值 τ_0 用 $\tau_0 = \mathrm{thr}(1 - \delta)$ 来确定,其中 thr 为映射函数.

常见的对数据库水印的攻击方式有子集选取、子集增加、子集更改、子集打乱等[1,2]. 其中子集选取、子集更改、子集增加的实质都是导致部分属性值的匹配信息丢失,也就是攻击使得 i 值变小. 但只有

当 i 值小于 $i_{\min}$ 时,攻击才算成功.

由以上分析可知,当攻击强度一定时,水印分布在载体中的范围越广,水印受攻击的概率越小,则鲁棒性越好. 本算法采用小波技术,能使"小"水印分布到"大"载体数据库中(即宽的频带上传输带宽很窄的信号),在一定程度上提高了水印不可见性和鲁棒性.

6　仿真实验

为了验证本文算法的有效性,所用的数据库关系由某地天气信息组成,数据库中包含了此天气的区站号、经度、纬度、海拔高度、气压、气温、露点温度、风向、风速、降水等参数的测量信息. 关系数据库(表)中共有 200 000 个元组,每个元组有 39 个属性,选取其中 12 300 条元组和 8 个数值型属性嵌入水印. 实验环境是 Matlab7.0,实验数据是根据水印载体数据的选择算法抽取数值型属性值,组成一个 128 ×128 二维向量 A,$\mathbf{l}$ = 3. 利用私有密钥生成 600 个长度为 N 且服从高斯分布的不同的随机序列 $W_i(i = 1,2,\cdots,600)$,其中所有的序列都是正态分布,W_{300} 作为水印信息模板,水印序列为二进制序列$\{B\}$. 根据以上水印嵌入算法,将水印信息嵌入到关系数据库中. 之所以选 600 个不同的随机序列,而只选取其中一个作为水印信号模板,目的是能直观地反映出线性相关检测的效果. 嵌入水印对关系数据库的影响见表 1,表明水印嵌入对关系数据库的影响较小. 在没有攻击时,水印检测器检测效果见图 3(横坐标表示输入真假水印序列号;纵坐标表示检测出来的相关系数值 τ);在受到攻击时,水印检测器检测效果见图 4.

表 1　水印对数据库的整体影响

属性名	均值	方差	均值改变比例 (1.0e-014)	方差改变比例 (1.0e-007)
A1	115.3067	19703	0.0863	0.1046
A2	115.1391	25091	0.0864	0.1687
A3	114.5901	21014	0.1106	0.2003
A4	117.5825	43005	0.3747	0.3190
A5	126.0792	27839	0.2705	0.3505
A6	114.0572	14262	0.0987	0.1649
A7	119.4280	21442	0.0833	0.1197
A8	136.4670	25229	0.1250	0.1167

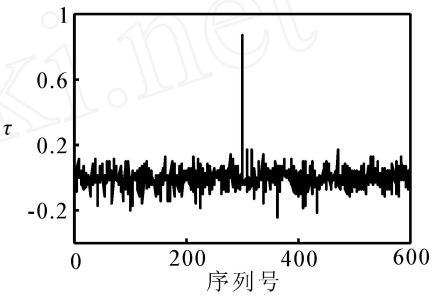
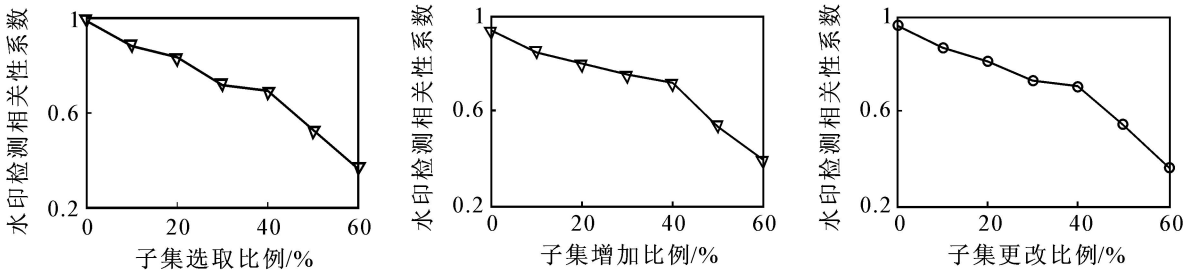


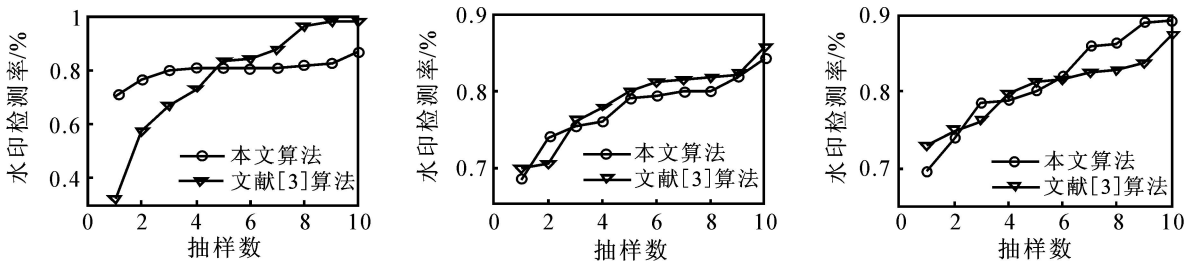
图 3　线性相关检测 τ

最后给出本算法与文献[3]的比较分析:1) 不可见性方面:从文献[3]和本文算法的不可见性实验结果可知,水印嵌入对某些属性修改的较大,即局部修改大. 而本文算法将水印信息嵌入小波域中,使得水印能够均匀地分布到关系数据库中,不可见性



(a) 子集选取水印情况　(b) 子集增加水印情况　(c) 子集更改水印情况

图 4　子集攻击测试



(a) 子集选取20%水印检测情况　(b) 子集更新20%水印检测情况　(c) 子集增加20%水印检测情况

图 5　稳键性能比较

较好(见表 1). 2) 抗攻击能力方面:在遭到攻击时,本文算法水印检测稳定性好. 而文献[3]的算法基于空域技术,水印检测变化较大,不稳定. 实验结果见图 5. 该实验在同一个数据库关系和嵌入等量数位的水印环境中完成,给出了部分实验结果. 当攻击比例增大时,本文算法表现出了较好的鲁棒性.

7 结 论

本文针对空域技术的关系数据库水印方案存在鲁棒性的问题,分析出关系数据库离散小波变换的高频系数服从高斯分布的特点,并根据相关检测技术,提出基于离散小波变换的关系数据库水印检测算法. 对算法的水印不可见性和鲁棒性作了全面的分析,本文算法运算简单方便,不可见性较好且有较强的抵御各种对关系数据库攻击的能力. 利用数字水印技术实现数据库的安全保护,以及在数据库关系的非数值型属性值中嵌入水印是今后要继续开展的研究工作.

参考文献(References)

[1] Rakesh Agrawal, Jerry Kiernan. Watermarking relational databases[C]. Proc of the 28th VLDB Conf. Hong Kong, 2002: 155-166.

[2] Sion R, Atallah M, Sunil Prabhakar. Rights protection for relational data[C]. Proc of the 2003 ACM SIGMOD Int Conf on Management of Data San Diego. California: ACM SIGMOD, 2003: 98-109.

[3] 牛夏牧, 赵亮, 黄文军, 等. 利用数字水印技术实现数据库的版权保护[J]. 电子学报, 2003, 31(12A): 2050-2053.
(Niu X M, Zhao L, Huang W J, et al. Watermarking relational databases for ownership protection [J]. Chinese J of Electronics, 2003, 31(12A): 2050-2053.)

[4] Zhang Z H, Jin X M, Wang J M, et al. Watermarking relational database using image[C]. Proc of the 3rd Int Conf on Machine Learning and Cybernetics. Shanghai, 2004: 1739-1744.

[5] 姜传贤, 孙星明, 易叶青, 等. 基于JADE算法的数据库公开算法的研究[J]. 系统仿真学报, 2006, 18(7): 1781-1784.
(Jiang C X, Sun X M, Yi Y Q, et al. Study of database public watermarking based on JADE algorithm[J]. J of System Simulation, 2006, 18(7): 1781-1784.)

[6] Gupta G, Pieprzyk J. Reversible and semi-blind relational database watermarking[C]. Proc of Int Conf on Signal Processing and Multimedia Applications. Barcelona, 2007: 283-290.

[7] Cox Ingemar J, Kilian Joe, Leighton Tom, et al. Secure spreads spectrum watermarking for multimedia [J]. IEEE Trans on Image Processing, 1997, 6(12): 1673-1687.

[8] Cox Ingemar J, Miller Matthew L, Bloom Jeffrey A. Digital watermarking[M]. SanDiego: Academic Press, 2002.

(上接第 1077 页)

[7] 王坚强. 一种信息不完全确定的多准则语言群决策方法[J]. 控制与决策, 2007, 22(4): 394-399.
(Wang J Q. Group multi-criteria linguistic decision-making method with incomplete certain information[J]. Control and Decision, 2007, 22(4): 394-399.)

[8] 夏勇其, 吴祈宗. 一种混合型多属性决策问题的TOPSIS方法[J]. 系统工程学报, 2004, 19(6): 630-635.
(Xia Y Q, Wu Q Z. A technique of order preference by similarity to ideal solution for hybrid mulitple attribute decision making problems [J]. J of Systems Engineering, 2004, 19(6): 630-635.)

[9] 梁昌勇, 吴坚, 陆文星, 等. 一种新的混合型多属性决策方法及在供应商选择中的应用[J]. 中国管理科学, 2006, 14(6): 71-76.
(Liang C Y, Wu J, Lu W X. A new method on hybrid multiple attribute decision-making problem for choosing the supplier [J]. Chinese J of Management Science, 2006, 14(6): 71-76.)

[10] 巩在武, 刘思峰. 不同偏好形式判断矩阵的二元语义群决策方法[J]. 系统工程学报, 2007, 22(2): 185-189.
(Gong Z W, Liu S F. Group decision making method based on two-tuple linguistic for judgment matrices with different fuzzy preferences [J]. J of Systems Engineering, 2007, 22(2): 185-189.)

[11] 巩在武. 梯形模糊数判断矩阵的二元语义排序方法[J]. 系统工程与电子技术, 2007, 29(9): 1488-1492.
(Gong Z W. On priority of trapezoidal fuzzy number complementary judgment matrix based on two-tuple linguistic[J]. Systems Engineering and Electronics, 2007, 29(9): 1488-1492.)

[12] Herrera F, Martinez L. A 2-tuple fuzzy linguistic represent model for computing with words[J]. IEEE Trans on Fuzzy Systems, 2000, 8(6): 746-752.

[13] Herrera F, Martinez L Sanchez. Managing non-homogeneous information in group decision making[J]. European J of Operational Research, 2005, 166(11): 115-132.