

# Research on Eigenvector Centrality Ranking under Heterogeneous Information Network

Yong Yu<sup>1</sup>, Changgen Chen<sup>1</sup>, Jia Zhou<sup>1\*</sup>, Chuanyu Zang<sup>2</sup>, Shifan Huang<sup>3</sup>

<sup>1</sup>School of Software, Yunnan University, Kunming Yunnan

<sup>2</sup>Central Branch of Kunming, The People's Bank of China, Kunming Yunnan

<sup>3</sup>Central Branch of Dali Prefecture, The People's Bank of China, Dali Yunnan

Email: \*1320835499@qq.com

Received: Sep. 9<sup>th</sup>, 2018; accepted: Sep. 22<sup>nd</sup>, 2018; published: Sep. 29<sup>th</sup>, 2018

---

## Abstract

This paper studies the ranking from the perspective of heterogeneous networks. Traditional ranking research is based on homogeneous networks. However, in real life, the actual networks are heterogeneous. Firstly, after establishing a heterogeneous network based on the real movie dataset, the paper derives the corresponding network model according to the network, and then establishes the meta-path according to the relationship between actors, directors and movies in the movie network. Secondly, the corresponding ranking rules are set in combination with the meta-path, and the eigenvector centrality measurement method is used to conduct the ranking study. Finally, the results of the experiment are given. The innovation of this paper is that the eigenvector centrality method is applied to a heterogeneous network, which can better reflect the characteristics of the real network and compare it with the degree centrality method. It shows that the method has good feasibility.

## Keywords

Heterogeneous Network, Meta-Path, Eigenvector Centrality, Degree Centrality, Ranking

---

# 异构信息网络下的特征向量中心性排名研究

郁 湧<sup>1</sup>, 陈长庚<sup>1</sup>, 周 佳<sup>1\*</sup>, 藏传宇<sup>2</sup>, 黄世反<sup>3</sup>

<sup>1</sup>云南大学软件学院, 云南 昆明

<sup>2</sup>中国人民银行昆明中心支行, 云南 昆明

<sup>3</sup>中国人民银行大理州中心支行, 云南 大理

Email: \*1320835499@qq.com

收稿日期: 2018年9月9日; 录用日期: 2018年9月22日; 发布日期: 2018年9月29日

\*通讯作者。

文章引用: 郁湧, 陈长庚, 周佳, 藏传宇, 黄世反. 异构信息网络下的特征向量中心性排名研究[J]. 计算机科学与应用, 2018, 8(9): 1452-1458. DOI: 10.12677/csa.2018.89159

## 摘要

本文从异构网络角度来进行排名的研究,传统的排名研究是基于同构网络来进行的,然而在现实生活中,实际的网络是异构的。首先,本文根据真实的电影数据集来建立一个异构网络后,根据网络得出相应的网络模式,之后根据电影网络中演员、导演、电影之间的相互关系来建立元路径。其次,结合元路径设定相应的排名规则,结合特征向量中心性度量方法来进行排名研究。最后,给出了实验的结果。本文的创新之处在于,将特征向量中心性方法应用到了一个异构网络中,能更好地反映出真实网络的特征,并与度中心性方法进行对比,展示了本文方法具有良好的可行性。

## 关键词

异构网络, 元路径, 特征向量中心性, 度中心性, 排名

Copyright © 2018 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 绪论

在我们生活的自然界中存在着大量的网络,在学术领域里面,许多的研究内容都能以网络的视角来进行研究,例如可以将一个系统可以用网络表示出来,系统中的所有部件或者构件可以转换成网络中的节点(node),部件之间的关系可以转换成为边(edge),每个节点与节点之间边的重要性可以用度(degree)来进行表示。在我们生活中常见的网络有,技术网络(Internet、交通网、快递网),社会网络,信息网络(万维网、引文网),生物网络(神经网络、生态网)等[1]。当前针对网络的研究大多数是基于同构信息网络(Homogeneous Information Network)来开展的。常超利用 Z-score、Page Rank 以及 HITS 等三个排名算法进行用户影响力排序来寻找社群的权威专家[2],刚家泰在合著网络上提出 New-LeaderRank 算法来进行排名[3]。杜文杰在文献引用网络中,通过在引用网络中加入虚拟节点和对虚拟节点进行分等级,从而提出了基于外部链接的 ELRank 排名算法和扩展的 N-ELRank 排名算法来进行排名研究[4]。本文则是在异构网络的基础之上来进行排名分析。

## 2. 相关定义及描述

### 2.1. 信息网络的形式化描述

网络是庞大且复杂的,对于更好的研究网络,我们引入了图。图是一个重要的数学模型和数据结构,可以来对网络进行合理的描述和建模[5]。

信息网络是指信息数据通过某一种方式连接在一起组成的网络[5],例如我们所熟知的万维网、引文网,包含了部分社会因素的网络也可以看作信息网络,例如邮件网络、微博网络,以及本文所研究的网络。如下定义 1 我们将给出信息网络的形式化描述。

**定义 1 信息网络**[6] (Information Network), 是一个带有对象类型映射函数  $\tau: \mathcal{V} \rightarrow \mathcal{A}$  和链接类型映射函数  $\phi: \mathcal{E} \rightarrow \mathcal{R}$  的有向图  $G = (\mathcal{V}, \mathcal{E})$ , 其中每个对象  $v \in \mathcal{V}$  属于一个特定的对象类型  $\tau(v) \in \mathcal{A}$ , 每个链接  $e \in \mathcal{E}$  属于一个特定的关系  $\phi(e) \in \mathcal{R}$ , 如果两个链接属于同一个关系类型, 那么这两个链接具有相

同类型的开始对象和结束对象。

## 2.2. 异构信息网络

定义 1 已经形式化地描述了信息网络, 在定义 1 的基础之上, 如果对象类型 $|\mathcal{A}| > 1$  或者关系类型 $|\mathcal{R}| > 1$  时, 该信息网络为异构信息网络[6]。为了更好的对异构信息网络进行深入研究定义 2 给出了异构信息网络的元模式即网络模式。

**定义 2 网络模式**[6] (Network Schema), 是带有对象类型映射  $\tau: \mathcal{V} \rightarrow \mathcal{A}$  和链接类型映射  $\phi: \mathcal{E} \rightarrow \mathcal{R}$  的异构网络  $G = (\mathcal{V}, \mathcal{E})$  的元模板, 记为  $T_G = (\mathcal{A}, \mathcal{R})$ , 其中  $G$  是一个定义在对象类型 $\mathcal{A}$ 和关系类型 $\mathcal{R}$ 上的有向图。

## 2.3. 元路径

在异构信息网络中, 定点之间可以由不同的路径来进行链接, 例如本文要研究的网络, 两个演员可以通过导演和电影来进连接, 他们之间的路径可以是“演员 - 电影 - 演员”也可以是“演员 - 导演 - 演员”, 这两条路径的语义分别是: “两个演员共同出演一部电影”, “两个演员同时于同一个导演合作”, 在本文研究的信息网络中, 可以抽取出一条或多条不同路径, 这些路径都被称作元路径, 如定义 3 所示。

**定义 3 元路径**[6] (Meta-path), 是在网络模式  $T_G = (\mathcal{A}, \mathcal{R})$  的图上的一条路径, 其形式为  $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$ , 定义了类型定义了类型  $A_1$  和类型  $A_{l+1}$  间的复合关系  $R = R_1 \circ R_2 \circ \dots \circ R_l$ , 其中“ $\circ$ ”表示关系上的复合运算。

## 3. 网络建模与排名

### 3.1. 电影网络建模

如图 1 所示, 本文在获得数据集之后, 经过数据预处理之后, 对其进行了网络建模。得到电影信息网络。其中红色代表电影, 绿色代表导演, 黑色代表电影。从图中可以看出该异构信息网络包含“电影”、“演员”、“导演”三种节点类型, “拍摄”“参演”两种边类型。

本文是在异构网络下的基于元路来进行研究, 在给出元路径之需要得到电影信息网络的网络模式, 如图 2 所示。

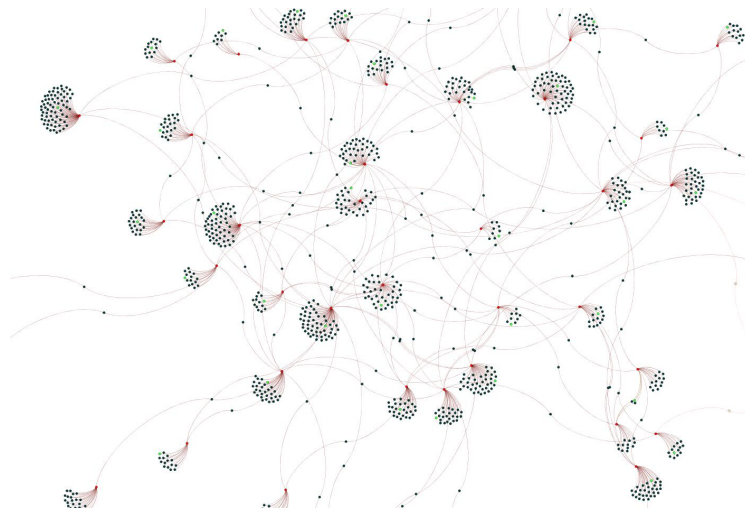


Figure 1. Movie information network  
图 1. 电影信息网络



Figure 2. Network Schema of movie information network  
图 2. 电影信息网络的网络模式

我们用字母  $A$  代表演员,  $F$  代表电影,  $D$  代表导演, 那么, 在本文网络模式下涉及到的元路径如表 1 所示, 不同的元路径代表了不同的语义。

### 3.2. 排名规则

本文的研究目的在于对已经建立的电影信息网络中的电影节点进行排名[7], 设计相关排名规则如下:

**规则 1:** 排名靠前的电影中演员排名比较靠前, 排名靠前的演员参加的电影排名比较靠前;

**规则 2:** 排名靠前的电影会使其导演排名比较靠前, 排名靠前的导演会产生很多排名靠前的电影;

**规则 3:** 排名靠前的电影一般是由排名靠前的导演拍摄, 并吸引很多排名靠前的演员参演。

由规则 1 可知, 电影影响演员、演员影响电影, 此规则符合元路径  $A \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} A$ ,  $F \xrightarrow{\text{参演}^{-1}} A \xrightarrow{\text{参演}} F$ , 由规则 2 可知电影影响导演、导演影响电影, 此规则符合元路径  $F \xrightarrow{\text{拍摄}^{-1}} D \xrightarrow{\text{拍摄}} F$ ,  $D \xrightarrow{\text{拍摄}} F \xrightarrow{\text{拍摄}^{-1}} D$ , 由规则 3 可知电影、导演、演员三者相互影响, 此规则符合元路  $A \xrightarrow{\text{参演}} F \xrightarrow{\text{拍摄}^{-1}} D$ 。

本文是基于元路径的特征向量来对排名进行研究[8], 在本文中我们规定, 演员和电影之间的邻接矩阵为  $M_{AF}(i, j)$ , 电影和演员之间的邻接矩阵为  $M_{FA}(i, j)$ , 电影和导演之间的邻接矩阵为  $M_{FD}(i, j)$ , 导演和电影之间的邻接矩阵为  $M_{DF}(i, j)$ , 其中  $M_{AF}(i, j)$  与  $M_{FA}(i, j)$ ,  $M_{FD}(i, j)$  与  $M_{DF}(i, j)$  互为转置矩阵。基于元路径的关系矩阵如表 2 所示。用  $r_A = \langle r_A^{(1)}, r_A^{(2)}, \dots, r_A^{(m)} \rangle$ ,  $r_F = \langle r_F^{(1)}, r_F^{(2)}, \dots, r_F^{(n)} \rangle$ ,  $r_D = \langle r_D^{(1)}, r_D^{(2)}, \dots, r_D^{(w)} \rangle$  分别表示演员、电影、导演的排名向量, 其中  $r_A^{(i)}$ ,  $r_F^{(j)}$ ,  $r_D^{(k)}$  分别表示第  $i$  个演员、第  $j$  部电影、第  $k$  个导演的排名函数值。

### 3.3. 演员排名函数

根据规则 1 我们可以定义电影与演员之间的排名相互计算的迭代公式为:

$$r_F^{(i)} = \sum_{j=1}^m M_{FA}(i, j)r_A^{(j)}, \text{ 矩阵表示为 } r_F = M_{FA}r_A;$$

$$r_A^{(i)} = \sum_{j=1}^n M_{AF}(i, j)r_F^{(j)}, \text{ 矩阵表示为 } r_A = M_{AF}r_F;$$

将上两式带入合并, 可得:

$$r_A = M_{AF}M_{FA}r_A \tag{1}$$

此式为基于元路径  $A \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} A$  的演员排名函数之间的迭代关系;

$$r_F = M_{FA}M_{AF}r_F \tag{2}$$

此式为基于元路径  $F \xrightarrow{\text{参演}} A \xrightarrow{\text{参演}^{-1}} F$  的演员排名函数之间的迭代关系;

记  $M_{AF}M_{FA}$  为  $M_{AA}$  为元路径  $A \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} A$  的演员之间的合作关系矩阵, 则(1)式可以写为:

$$r_A = M_{AF}M_{FA}r_A = M_{AA}r_A \text{ (*) 即 } r_A^{(i)} = \sum_{j=1}^m M_{AA}(i, j)r_A^{(j)}$$

为了求解演员排名向量  $r_A$ , 我们可以把(\*)式改写成为  $r_A^t = M_{AA}r_A^{t-1}$ , 即每次的  $r_A^t$  值由上次迭代而来,

**Table 1.** Meta-path and semantics

**表 1.** 元路径及其语义

元路径	语义	备注
$A \xrightarrow{\text{参演}} F \xrightarrow{\text{拍摄}^{-1}} D$	某位演员参演了某位导演拍摄的电影	
$A \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} A$	两位演员参演一部电影	
$D \xrightarrow{\text{拍摄}} F \xrightarrow{\text{拍摄}^{-1}} D$	某位导演自身循环的一条路径	本文涉及的电影只有一位导演，在下一章中涉及到导演排名，该条路径无法给出合理的排名。
$F \xrightarrow{\text{参演}^{-1}} A \xrightarrow{\text{参演}} F$	某位演员参演了多部电影	
$F \xrightarrow{\text{拍摄}^{-1}} D \xrightarrow{\text{拍摄}} F$	某位导演拍摄了多部电影	

**Table 2.** Relation matrix

**表 2.** 关系矩阵

元路径	关系矩阵
$A \xrightarrow{\text{参演}} F \xrightarrow{\text{拍摄}^{-1}} D$	$M_{AF}M_{FD}(i, j)$
$A \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} A$	$M_{AF}M_{FA}(i, j)$
$D \xrightarrow{\text{拍摄}} F \xrightarrow{\text{拍摄}^{-1}} D$	$M_{DF}M_{FD}(i, j)$
$F \xrightarrow{\text{参演}^{-1}} A \xrightarrow{\text{参演}} F$	$M_{FA}M_{AF}(i, j)$
$F \xrightarrow{\text{拍摄}^{-1}} D \xrightarrow{\text{拍摄}} F$	$M_{FD}M_{DF}(i, j)$

重复该过程可以得到更好的估计值。重复  $t$  步后可以得到  $r_A^t$  的计算公式为  $r_A^t = (M_{AA})^t r_A^0$ ,  $r_A^0$  是预先设定的一个初始排名向量, 可以将  $r_A^0$  写成矩阵  $M_{AA}$  的特征向量  $v_i$  的线性组合, 即  $r_A^0 = \sum_i c_i v_i$  选择适当常数  $c_i$  可以得到如下式子:

$$r_A^t = (M_{AA})^t \sum_i c_i v_i = \sum_i c_i e_i^t v_i = e_i^t \sum_i c_i \begin{bmatrix} e_i \\ e_i \end{bmatrix} v_i$$

其中,  $e_i$  为矩阵  $M_{AA}$  的特征值,  $e_1$  为最大特征值, 矩阵  $M_{AA}$  为一个元素取值为非负的对称矩阵, 因此要  $e_1 > 0$ 。当式中的  $i \neq 1$  时, 对于所有的  $i$ ,  $\frac{e_i}{e_1} < 1$ 。当  $t$  不断变大时, 除了第一项以外, 其他项都以指数级下降, 当  $t \rightarrow \infty$  时,  $r_A^t \rightarrow c_1 e_1^t v_1$ 。也就是说, 演员排名  $r_A^t$  的极限与矩阵的  $M_{AA}$  的主特征向量成正比。因此, 我们可以等价地认为排名函数  $r_A$  满足:  $M_{AA} r_A = e_1 r_A$ 。

由于矩阵  $M_{AA}$  是一个非负的对称矩阵, 选择适当常数  $c_i$  可以是初始排名向量  $r_A^0 = \sum_i c_i v_i$  中的所有元素都是非负值, 同样  $r_A^t$  的所有值也都是非负的。因此, 演员的排名向量  $r_A$  收敛于矩阵  $M_{AA}$  的主特向量, 可以取矩阵  $M_{AA}$  的主特征向量为演员排名向量的取值, 称为特征向量排名函数。

### 3.4. 导演排名函数

根据规则 2 我们可以定义电影与导演之间的排名相互计算的迭代公式为:

$$r_F^{(i)} = \sum_{j=1}^w M_{FD}(i, j) r_D^{(j)}, \text{ 矩阵表示为 } r_F = M_{FD} r_D;$$

$$r_D^{(i)} = \sum_{j=1}^n M_{DF}(i, j) r_F^{(j)}, \text{ 矩阵表示为 } r_D = M_{DF} r_F;$$

将上两式带入合并, 可得:

$$r_D = M_{DF}M_{FD}r_D \tag{3}$$

此式为基于元路径  $D \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} D$  的导演排名函数之间的迭代关系；

$$r_F = M_{FD}M_{DF}r_F \tag{4}$$

此式为基于元路径  $F \xrightarrow{\text{参演}} D \xrightarrow{\text{参演}^{-1}} F$  的电影排名函数之间的迭代关系；

记  $M_{DF}M_{FD}$  为  $M_{DD}$  为元路径  $D \xrightarrow{\text{参演}} F \xrightarrow{\text{参演}^{-1}} D$  的导演之间的合作关系矩阵，同理，由 3.3 节可知，根据(3)式， $M_{DD}$  的主特征向量可以等价于导演排名向量的取值。但是本文研究的电影只有一位导演，因此  $M_{DD}$  为一个对角矩阵，对角线上的值为导演拍过的电影的数量，将无法合理的给出导演的排名，例如，导演 A 拍摄了多部排名很低的电影，而导演 B 拍摄了一部高排名电影，则 A 的排名比 B 高，这样将不合理。因此，我们将利用 3.5 节计算出来的电影排名来对导演进行重新排名。

在 3.5 节得出电影的排名向量  $r_F$  之后，可以对  $r_D = M_{DF}$  进行平均化处理之后来计算导演排名，改进之后为如下所示：

$$r_D = D_D M_{DF} r_F (**)$$

其中， $D_D$  为对角矩阵，对角线元素为  $\frac{1}{\sum_{j=1}^m M_{DF}(i, j)}$ ，分母为导演各自所拍电影的总数。

### 3.5. 电影排名函数

根据规则 3 演员和导演的排名都会对电影的排名产生影响，因此需要把基于元路径  $F \xrightarrow{\text{参演}} A \xrightarrow{\text{参演}^{-1}} F$  和  $F \xrightarrow{\text{参演}} D \xrightarrow{\text{参演}^{-1}} F$  得到的两个电影排名函数进行迭代加权合并，即把(2)(4)式进行合并，得到电影排名函数如下：

$$r_F = (\alpha M_{FA}M_{AF} + (1-\alpha)M_{FD}M_{DF})r_F (***)$$

其中参数  $\alpha \in [0,1]$  决定演员和导演对电影排名函数的权重， $\alpha$  的赋值可以基于先验知识或者学习训练数据集得到。矩阵  $\alpha M_{FA}M_{AF} + (1-\alpha)M_{FD}M_{DF}$  为基于演员和导演连接的电影之间的带权关系矩阵，记为矩阵  $M_{AD-F}$ ，同理，由 3.3 节可知， $M_{AD-F}$  的主特征向量可以等价于电影排名向量的取值。

## 4. 实验结果与分析

本文研究的数据，经过处理之后共有电影 4545 部、演员 5121 人、导演 2064 人，如表 3、表 4 所示，分别给出了本文排名函数对电影、演员、导演进行排名之后的 Top 5 和度中心性的排名结果。

度中心性在进行排名的时候[9]，各节点的重要性完全取决于各个节点的度，在本文中以演员和导演为例，则可以理解成为演员和导演所拍电影的总数，以电影为例则可以理解成为本部电影的演员和导演的总数。通过实验结果和统计发现，表 4 中的演员和导演的排名取决于他们各自拍摄和参演的电影的总

Table 3. The top 5 of our method

表 3. 本文排名 Top 5

电影	演员	导演
Last Vegas	Robert De Niro	Jerry Zaks
Sleepers	Morgan Freeman	Gary McKendry
Marvin's Room	Brad Pitt	Justin Zackham
The Godfather: Part II	Johnny Depp	Des McAnuff
Brazil	Matt Damon	Jonathan Jakubowicz

**Table 4.** The top 5 of degree centrality method  
**表 4.** 度中心性排名 Top 5

电影	演员	导演
Avatar	Robert De Niro	Steven Spielberg
Pirates of the Caribbean: At World's End	Morgan Freeman	Woody Allen
Spectre	Johnny Depp	Martin Scorsese
The Dark Knight Rises	Bruce Willis	Clint Eastwood
John Carter	Matt Damon	Ridley Scott

数, 这样的排名是不合理的, 例如, 经常跑龙套的演员, 排名就会很高。经常拍烂片的导演的排名就会很高。(由于数据集中限制了每部电影的导演个数为 1, 演员个数为 3, 因此, 分析时不对电影排名进行分析)。本文提出的方法综合考虑了电影、导演、演员之间的排名, 因此, 本文的排名方法具有良好的可行性。

## 5. 结束语

本文研究的内容是基于异构网络来进行完成的, 结合了实际的情况, 利用异构网络充分的表现了现实社会的网络特征, 研究主题是电影, 合理利用了电影、导演、演员三者之间的关系来进行排名。虽然完成了排名, 但是本文依然存在一些不足之处, 在今后的研究过程中, 将会继续深入研究。例如, 可以在网络的排名中加入节点的个性特征, 来使排名更加的准确, 或者将异构网络应用到推荐系统中去[10]。

## 基金项目

国家自然科学基金项目(61462091)。

## 参考文献

- [1] 王桂平, 王衍, 任嘉辰, 等. 图论算法理论实现及应用[M]. 北京: 北京大学出版社, 2011.
- [2] 常超. 基于网络问答社区的专家排名算法分析[J]. 软件, 2015, 36(11): 120-122.
- [3] 刚家泰, 弓浩然, 窦名名. 合著者网络中的网络排名算法[J]. 辽宁工程技术大学学报(自然科学版), 2014(9): 1293-1296.
- [4] 杜文杰. 基于引用网络的学术文献排名算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2013.
- [5] Mark E. J. Newman, 郭世泽, 等. 网络科学引论[M]. 北京: 电子工业出版社, 2014.
- [6] 孙艺洲, 韩家炜. 异构信息网络挖掘原理和方法[M]. 北京: 机械工业出版社, 2017.
- [7] 田鹏伟, 张娴, 胡正银, 等. 异构信息网络融合方法研究综述[J]. 图书情报工作, 2017, 61(7): 137-144.
- [8] 伍转华. 异构信息网络的相似性度量方法[J]. 计算机与现代化, 2016(3): 78-84.
- [9] 郑晓燕. 外部信息驱动的异构信息网络节点影响力分析[J]. 福建电脑, 2016, 32(10): 25-27.
- [10] 丁兴艳. 异构图中的 Top-K 兴趣子图匹配算法研究[D]: [硕士学位论文]. 沈阳: 辽宁大学, 2017.

**知网检索的两种方式：**

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>  
下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 <http://cnki.net/>  
左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：<http://www.hanspub.org/Submission.aspx>

期刊邮箱：[csa@hanspub.org](mailto:csa@hanspub.org)