UNIVERSITY OF TURKU

ABSTRACT

□	Bachelor's thesis
☒	Master's thesis
□	Licentiate's thesis
□	Doctor's thesis

Subject	Information Systems Science	Date	27th June, 2019
Author(s)	Jieqiong Cheng	Student number	609656
		Number of pages	75
Title	Identifying the Main Path Based on Influence Flow in Citation Network		
Supervisor(s)	Prof. Reima Suomi		

With the rapid increase of knowledge as the time goes on, the citation network has evolved into a huge network system. Find the key documents from the huge citation network has become an important issue for analyzing the changes in specific fields. The main path analysis method can better describe the path development process in a target field and effectively avoids the problem that only high-cited nodes will be selected but ignore the fact that the "key documents" may not be connected strongly. The traditional main path weighting algorithms SPX are based on the traversal counts have been applied in various field research nowadays, but these appoaches will also cause some shortcomings. According to the core idea of eigenvector centrality, the concept of influence flow in citation network is summarized at the first and the algorithms to weight the path based on equally-distributed and unequally-distributed influence flow are proposed in this thesis.

Firstly, main paths in desalination field and information security field through traditional traversal counts SPLC algorithm are obtained as a comparative experiment compared to that used equally-distributed influenced flow algorithms. Secondly, this thesis proposes the influence flow algorithms based on PageRank and single traversal from to perspective to divide influence equally. The main paths obtained by the two algorithms are compared with the paths used SPLC to draw the similarities and differences from the aspects of shape, node content, cited times of node and theme evolution. Comparisons show that the proposed two algorithms supplement some key nodes and overcome part of shortcomings to some extent. Thirdly, the thesis proposes another influence flow algorithm based on the coupling strength. It is because different knowledge diffusion effect that citing nodes transmit unequal influence to their cited nodes. Therefore, the influence of the transmission is different according to the closeness between nodes. This algorithm is applied to the existing DNA network and compared the forms and content of the main path with the traditional traversal counts algorithm. Finally, this thesis summarizes the main conclusions of the new proposed algorithms applied in the main path analysis. Also, we discuss the limitations of the research which are not considered and prospects for future research directions on main path analysis.

Key words	Citation network; Main path analysis; Eigenvector centrality; Coupling
Further information	

UNIVERSITY OF TURKU
Turku School of Economics

IDENTIFYING THE MAIN PATH BASED ON INFLUENCE FLOW IN CITATION NETWORK

Master´s Thesis
in Information Systems Science

Author:
Jieqiong Cheng

Supervisor:
Prof. Reima Suomi

27.06.2019
Turku

Table of contents

List of figures

List of tables

1 INTRODUCTION

1.1 Background

Scientific literature is not only the main output form of scientific research activities, but also the main way of recording and disseminating. It bears different scientific research achievements in different periods. However, scientific literature doesn’t exist independently in a certain field, but form a citation network by developing and inheriting the previous literature through citation relationship. The citation network is a sequence diagram of the historical evolution of the discipline from the perspective of time, while it is a subject group from the perspective of space. From the structure of citation network, it can be seen that citation network can reflect the development and evolution of knowledge in specific fields.

Since Garfield (1965) and Price (1965) pointed out the importance of citation networks in scientific research, the analysis of milestones and evolutionary processes in citation networks has become the focus of scholars' research. The amount of knowledge growth sharply as time goes by, and the citation network has evolved into a huge network system nowadays. How to find out the key documents from the huge citation network is an important problem to analyze the changing trend of specific fields. According to the existing evaluation indicators, if the core nodes are selected from a single attribute such as time priority or citation frequency, the relationship between citations will be ignored, and it is difficult to show the evolution process of knowledge. Therefore, when determining whether a node is a critical node, it should examine its " connectivity" throughout the citation network. In order to solve this problem, Hummon and Doreian (1989) proposed the citation network main path analysis method from the perspective of connectivity. The main path based on the connectivity is deemed as the most significant historical path in the target citation network and reflects the main development of the technologies or science. The main path analysis method effectively avoids the high-cited nodes selected only by considering the indegree values and ignores the problem that the "key node" may not be connected strongly with other nodes in citation network, thus better portraying the path development process in a specific field (Han & Jin, 2012).

The results of Goffman (1966), Jahn (1972) and Small (1970) based on the citation network method showed that a major is always defined by a few and extremely important events or people that have emerged in its historical development. The finding is also the core idea of the meaning of main path. The approaches to figure out the main path become more dynamic with the harder challenges emerging in the complexity of technology and the increments of patents. Citation analysis is a wide approach for analyzing these changes for the cited numbers could be used as a vital indicator to reckon the value of the innovation output. Exploring and identifying the knowledge evolution structure within the discipline based on the citation network is an important method of scientometrics. The citation network takes scientific literature as knowledge nodes and constitutes knowledge association in a specific field through citation relationship, which embodies the whole context of knowledge structure and knowledge evolution in the research field. Therefore, the analysis of the main path of the citation network is not only the research needs of the history of science but also the needs of revealing the scientific structure.

From the perspective of the history of science, its main purpose is to reveal the scientific regularities contained in the history of scientific development according to the development of specific disciplines. Through the main path analysis of citation network, we can avoid the non-representativeness of key node selected by single factor and subjectivity. At the same time, citation analysis can identify a series of scientific information (Garfield, 1970), including the positions of key nodes in the development process, the relationships between documents, all of which make the searched nodes more objective. On the other hand, from the demand analysis of revealing the scientific structure, different scholars or institutions have combined visualization tools to reveal the scientific structure through high cited nodes (National Institute of Science and Technology Policy, 2007), co-citation, coupling and other methods, in order to predict the scientific structure and future development trend of specific disciplines. The main path analysis method of the citation network is of great significance for solving key literature or key technologies in specific fields, especially in the situation where the determination of the threshold of highly cited documents puzzles scholars to reveal the scientific structure more objectively.

1.2 Research gap

The paper aims to put forward the algorithm for weighting the path in the citation network. Through analyzing the related documents, we can find that the traditional main path extracted by the methods of traversal counts for each link has the following problems:

1) The key nodes in the specific domain that play an important role may be still ignored because the algorithms of traversal counts can’t search all the vital nodes in the main path (Han & Jin, 2012);

2) The tradition traversal counts methods don’t consider the initial value of each edge in citation network. Because the different importance and knowledge diffusion influence of each node, same traversed counts can’t guarantee the same importance for the edge. Therefore, the neglected initial weight of the edge will cause some deviation for the main path analysis (Wei & Fang, 2016);

3) Main path analysis based on the traversal counts is difficult to realize the fine structure of discipline evolution. The rich and colorful evolution of a domain can be represented by only one linear structure, which may lead to the absence of many evolutionary details. Therefore, it is a basic requirement to reveal the fined evolution structure of a domain whether the main path can be extended to a more detailed evolutionary context on the basis of the skeleton or segregated from a more complex network (Wang, 2013).

1.3 Significance and research questions of the study

1.3.1 Significance

In recent years, with the increasingly close integration of social network and informatics research, more and more scholars use literature citation network and patent citation network to carry out the scientific or technological trajectory and the corresponding research of scientific thought and important technological development. Therefore, using the citation network to trace the main path has important academic research value in the fields of literature or patent information analysis, scientometrics and scientific research.

The purpose to identify main paths mainly focus on finding all related information of the development of science or technologies via the time, especially focus on the related other fields to analyze the evolve directions from the intersectional perspective to further analyze the scientific or technological dependent relationship from those fields. Main path research has been an important approach for scientific and technological management recent decade years. Therefore, identifying the main path has much crucial theoretical and practical significance: 1) Analyzing the whole development of science or technologies in specific domain more comprehensively. According to the main path analysis, researcher could more easily find the vital nodes and the directions the earlier topics or technologies evolved; 2) Identify the main paths is a process of tracing back the history of science, a process of discovering the relationship between scientific or technological inventions and important technological inventions that play an important role in the history of science or technology, and also a process of identifying scientific or technological paradigms; 3) Forecasting the scientific front or technology trend of some special domains which is beneficial for detecting the potential chance. Tracing radical change and incremental development processes through main paths provides essential insights into the evolutionary process and regularities in a specific domain (Martinelli, 2012).

1.3.2 Research questions

In this paper, we are mainly studying on the weighting methods to weight each edge, which is one procedure of the main path analysis to find out the significant development of the target field from the perspective of the influence flow in citation network. In order to identify the main path based on the influence flow, the following four research questions put forward are the main issues and content to explore in this paper.

RQ1: How to measure the influence flow in citation networks?

RQ2: How to divide the influence value from the equal and unequal perspective?

RQ3: How to weight the path based on the influence flow idea in citation network and then identify the main path?

RQ4: What is the difference of the obtained main path results between the traditional traversal counts algorithms and the new proposed algorithms in this paper?

1.4 Main concepts

1.4.1 Citation network

A citation network is a directed acyclic graph (DAG) consisting of a series of nodes and their edges, which means that there is no ring structure between nodes. Citation network is a collection of citation relationships among different kind of literature, which belongs to a complex network. The literature includes journals, scientific literature, patent data, conference papers and other forms. Citation network can reflect the continuity and inheritance of knowledge and can be used to study the history, context and structure of scientific knowledge development, and describe the development and evolution of disciplines or technologies in different fields (Garfield, 1965).

A simple citation network structure is shown in Figure 1-1. The citation network is usually represented by nodes and directed edges. The directed edges from node i to node j indicate that node i is cited by node j. Based on the attributes of the citation network, the characteristics of the citation network can be summarized as follows (Zhuge, 2006):

1) The citation relationship is unidirectional in time and can only be formed by the later literature citing the earlier published literature. If the earlier literature has not been published, the later literature can’t be cited.

2) Once the document is published, it is not modifiable.

3) The two published documents may not be exactly the same.

4) The relationship between the citing node and the cited node embodies the relevance of content and the transmission of knowledge.

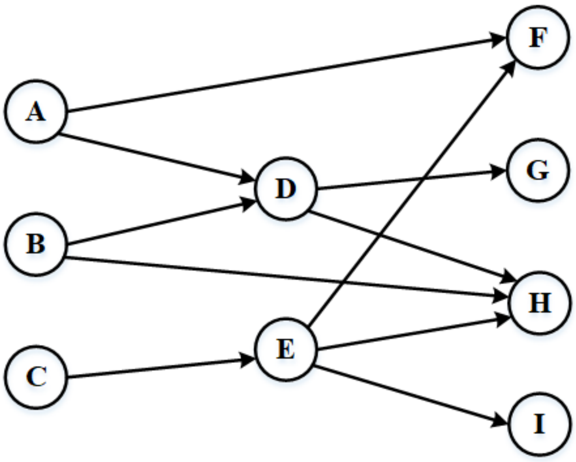

Figure 1-1 Simple citation network structure

1.4.2 Influence flow

The concept of influence flow is extracted through the core essence of the eigenvector centrality（Newman, 2008）used in the directed network that the importance of a node is closely related to the importance of its citing nodes. Inspired by the eigenvector centrality, Renoust et al. (2017) defined the concept of ascending flow assumed that each node will produce some information and this production of information gives credit to their ancestor.

Similarly, we put forward the influence flow as the process that citing nodes transmit part of its influence value to its cited node and the value of the influence flow equals to the weight of the edge obtained in citation network from the equal and unequal perspective in this paper. Which indicates that a document produces some information and this production of information gives the credit to its ancestors.

The idea of PageRank (Page & Brin, 1998) can be also treated as the influence flow process for calculating the influence of pages based on the linked pages. The underlying assumption is that more important websites are likely to receive more links from other important websites. Similarity, we regard the edges as the flow path transmitting the influence value from citing node to cited node in citation network. The influence flow idea draws on the idea that the more important a node is related to “descendent” nodes, the more it affects future development and has a high possibility occurring in the main path. In this paper, the calculation idea of weighting the edge based on the influence flow is that the influence value flowed of the citing node proportional to its own influence value, and inversely proportional to the number of its own references. The influence obtained of each node comes from the sum of the influence transmitted by all their citing nodes.

The value of the influence flow of each edge is proportional to the influence value of the nodes. And the flow procedure can also better show how citations are influenced.

We can suppose that the edge with the high influence flow connects to the influential nodes. Therefore, the main paths we obtained by the proposed algorithms in this paper are the broad meaning of the traditional main path but also aim at reflecting the development process based on the connectivity.

1.4.3 Main path

The original main path in a citation network can be described as the most used path which means the path has the largest overall connectivity among all possible paths of successive edges from the source to the sink nodes based on Hummon and Doreain at first. And this most used path can be usually found by a two-step procedure: the traversal counts for each edge are calculated as the number of different paths between each source and sink nodes that go through this edge and firstly. Second, a search method is used to identify the main path based on the edge traversal counts (Halatechliyski et al., 2014).

Then, the main path extended to some other broad meanings in order to enriching the diversity of main paths for analyzing the development of the target field. Ma and Zhang (2016) redefined the main path which is considered as the path with the largest information and reflect the development for disciplines. Chen, Yang et al. (2015) calculates semantic similarity between nodes, and then weights each link instead of traversal weights in order to get the main path to show the evolving process in patent citation network. In conclusion, the definition of the main path is used differently in various researches but all aim to figure out the most significant evolving and development process of the technologies or themes. (Liu et al., 2014). Nowadays, main path analysis is a mathematical tool for identifying the major path and it is successfully applied to trace many scientific trajectories and technological trajectories in various fields through the patent citations or the bibliographic citations (Verspagen, 2007).

The main path we proposed in this paper roots its contribution in the analysis flows in DAGs (Auber, 2002; Delest et al., 2006). We will select the main path with the most relevant relationship based on the influence flow to better understand how citations are influenced.

1.5 Innovations of the study

The study of finding other approaches to weight the edge is treated as a critical content for main path analysis (Wei & Fang, 2016). There is rising interest in studying how innovation emerges from the blending of accumulated knowledge, and from which path an innovation mostly inherits (Renoust, Claver & Baffier, 2017). Based on the above ideas and existing research gap, the research innovations of this study mainly focus on the new algorithms to weight the path according to the influence flow in citation network and then to analyze the applicability of methods in three different data samples in reality from their corresponding main paths. The main innovations of this paper can be concluded as the following three aspects:

1) Put forward the influence flow idea for main path analysis in citation network. Different from the traditional traversal counts method to weight the links. This paper starts from the essence of eigenvector centrality idea in directed network, then summarizes the cumulative process of node influence thought and mechanism of influence flow in citation network to weight each edge in citation network.

2) Propose the equally-distributed influence flow algorithms to weight the edges in citation network for main path analysis, including the algorithm based on PageRank and the concept of single traversal ideas. In order to gain the most relevant development main path from another perspective, the algorithm weights the edge from the idea of the influence flow but not based on the traditional traversal counts methods. The obtained new main paths compensate the key nodes to some extent that are ignored by the traditional main path traversal counts method.

3) Propose the unequally-distributed influence flow algorithm based on the coupling strength to weight the edges in citation network for main path analysis. The algorithm is from the perspective of distributing the citing nodes' influence based on the closeness relation between nodes aims to avoid the deviation the neglected initial value of edges will bring through the traversal counts methods. According to the coupling strength in citation network, the influence value of the cited node is transmitted in unequal proportion. Compared with the traditional algorithm, it is more reasonable to reveal the refined evolution process of scientific topics in target fields and to strengthen the tightness between nodes on the extracted main path.

1.6 Structure of the study

Content of this paper is divided into six chapters, and the main contents of each chapter are as follows:

Chapter 1 is introduction which mainly introduces the background and significance of the topic and leads to the research problems to be studied in this paper. Then describe some main concepts the paper focus on and elaborates the existing research gap. Chapter 1.6 introduces the innovations and structure of this study.

Chapter 2 mainly introduces the relevant theoretical parts involved in this paper, including the citation network, main path traditional traversal counts algorithms, main path search method and the application of main path analysis.

The main path of SPLC algorithm in the field of desalination and information security is analyzed in chapter 3 which is the basis for the comparison of the main path of the fourth chapter algorithm. The shortcomings of the experimental results based on SPLC algorithm are summarized at the end of the chapter.

Chapter 4 proposed two algorithms of equally-divided algorithms based on influence flow idea to weighting the path in citation network. The two methods are all used in the same data set of desalination and information security, then the similarities and differences between the main paths obtained by the proposed algorithms and the SPLC algorithm are discussed.

Chapter 5 introduces the influence flow algorithm based on the coupling strength and analyzes the main path obtained by this method in the DNA network and the traditional traversal counts algorithm.

The last chapter 6 summarizes the main conclusions of this paper and put forward some suggestions for the future research on main path analysis.

The Research design and methods in the paper is shown in Figure 1-2. The methods adopted in the paper is document analysis, empirical analysis and comparative research. Through analyzing the related document including the citation network, traditional main path traversal counts algorithms and main path search methods, we put forward the influence flow idea and demonstrate how to measure it. Then three algorithms are constructed from the equal and unequal perspective in order to get the main path based on influence flow idea and then compare their results with the traditional traversal counts algorithm.

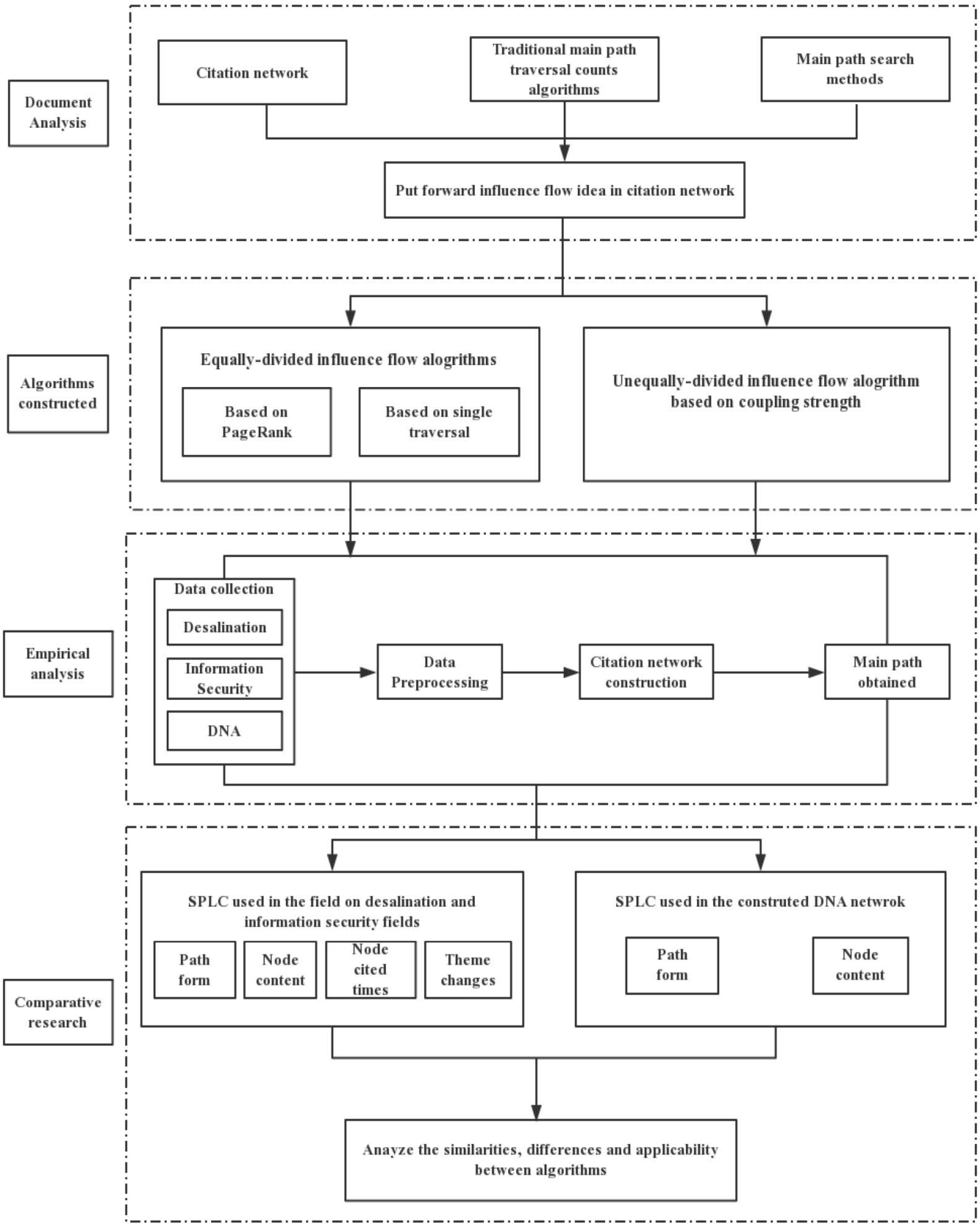
Document Analysis
Citation network
Traditional main path traversal counts algorithms
Main path search methods
Put forward influence flow idea in citation network
Algorithms constructed
Equally-divided influence flow alogrithms
Based on PageRank
Based on single traversal
Unequally-divided influence flow alogrithm based on coupling strength
Empirical analysis
Data collection
Desalination
Information Security
DNA
Data Preprocessing
Citation network construction
Main path obtained
Comparative research
SPLC used in the field on desalination and information security fields
Path form
Node content
Node cited times
Theme changes
SPLC used in the construted DNA netwrok
Path form
Node content
Anayze the similarities, differences and applicability between algorithms

Figure 1-2 Technological roadmap

2 THEORETICAL BACKGROUND

2.1 Citation network

2.1.1 Different types of citation

There are three definitions of citation (Small, 1973): direct citation, co-citation (Marshakova, 1973; Small, 1973), and bibliographic coupling (Kessler, 1963). The citation from the later document links to the previous by one-way edge is called direct citation. Such as the citation between A and E, or A and F as shown in Figure 2-2.

Bibliographic coupling is defined as the edge between two documents citing the same paper(s). Kessler's (1963) theory of coupling is that there must be some connection or similarity in the subject content between the documents with coupling connection, and the strength of the connection or similarity can be measured by the coupling strength which represent the number of the same references between the two papers. The closer the relationship between the two papers with greater coupling strength, the higher the degree of similarity. As shown in Figure 2-1, there is only one common reference for E and F nodes, so the coupling strength of E and F nodes is 1. Once the coupling relationship between the documents is determined, it will not change over time. Documents can be clustered to form a cluster of documents with high similarity according to the coupling relationship of documents and in any document in the formed document cluster, there must be another document with coupling relationship with it. Therefore, some scholars have indicated that on the basis of document coupling, the cluster of documents formed by clustering can reveal the development and evolution process of a certain field or the frontier heat to a certain extent (Jarneving, 2007).

The coupling relationship is a ubiquitous phenomenon, so the subject of Kelsler's document coupling can be expanded such as the journals, authors, subject topics, institutions and so on. For example, if two research institutes cite the literature of another institution, it can be said that the two research institutes have a coupling relationship. The greater the coupling strength, the closer the relationship between the two research institutes is.

Co-citation is defined as the frequency with which two documents are cited together by other documents. The co-citation relationship and the coupling relationship are similar for they are all methods for measuring the similarity between nodes by citation. The stronger the co-citation between two nodes, the closer the relationship between the two nodes (Kessler, 1963). As shown in Figure 2-1, node C and node D are both cited by node G, so a co-citation relationship is formed between node C and node D, and the co-citation strength is 1. A co-citation relationship is a dynamic process because the strength of co-citation changes over time. The co-citation analysis research is mainly divided into three categories in current research. The object of the first type of analysis is literature. If the co-citation strength of the two documents is strong, it will lead to a large number of "new authors" cite them. Therefore, co-citation analysis of literature can be used to describe and explore the active areas of research (Small & Griffith, 1974), which provides an intuitive approach to the frontiers of probing. (Persson, 1994). The second type of analysis is mainly for journals. McCain (1991) used co-citation analysis of academic journals in the field of statistical economics. The third category used literature authors as the unit to analyze, the more co-citation times the authors had, the closer their academic links and the more similar their research fields were. The author's co-citation analysis can reveal the representative theme of the author in the academic field and the position in the discipline (White & Griffith, 1981). Most of the current co-citation subjects are analyzed in terms of literature.

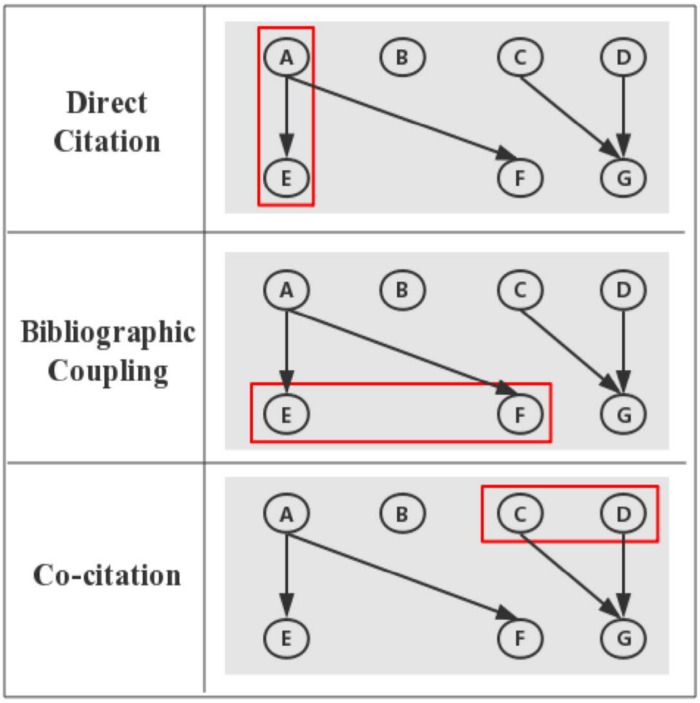

Figure 2-1 Direct citation, co-citation and bibliographic coupling

2.1.2 *Eigenvector centrality of node in citation network*

The eigenvector centrality is an improvement of the degree of centrality because eigenvector centrality considers the fact that the degree of each node is not of equal importance.

Those nodes link to the other nodes with more important nodes are also of higher importance. Therefore, the eigenvector centrality considers the importance of a node is closely related to the importance of other adjacent nodes. Which means if the node is connected to many points with a high degree of centrality, this node also has a high influence (Liu, 2009).

Given the citation network $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, V is the collection of all nodes and E represents the collection of all edges in citation network. $\mathrm{A} = (\mathrm{a}_{\mathrm{i,j}})$ represents the adjacency matrix of the citation network.

$$\begin{cases} a_{i,j} = 1, & i \ is \ connected \ to \ j \\ a_{i,j} = 0, & otherwise \end{cases} \quad (2\text{-}1)$$

And the relative center degree vector x_i of the node i can be expressed as follows:

$$x_i = \frac{1}{\lambda} \sum_{t \epsilon M(i)} x_j = \frac{1}{\lambda} \sum_{t \epsilon G} a_{i,j} x_j \quad (2\text{-}2)$$

Where M(i) represents the collection of all adjacent nodes of node i, and λ is a constant. It can be seen from the formula that the centrality of one node is a function of the center degree of other adjacent nodes. The above formula can be expressed in the form of a matrix as follows:

$$x = \frac{1}{\lambda} Ax \ or \ Ax = \lambda x \quad (2\text{-}3)$$

Thus, the calculation of the centrality of the vector x can be transformed into the calculation of the eigenvalues and eigenvectors of the matrix. In general, there will be multiple solutions of eigenvalues of non-zero vectors and corresponding eigenvectors. But according to Perron-Frobenius theorem (Ninio, 1976), if every element of the adjacent matrix is positive, only the largest eigenvalue λ can satisfy the requirement and the corresponding element i in the eigenvector is the centrality value of node i based on eigenvector centrality. Therefore, a node will obtain a large eigenvector centrality when it is connected to a large number of ordinary nodes or connected with a small number of high eigenvector values.

Google's PageRank is a variant of the Eigenvector centrality measure. PageRank was proposed by Google founder Brin and Page in 1998 and was originally used to measure the importance of the ranking of www pages. The algorithm is based on the assumption that if a web page is linked by another important web page, it reflects that the pointed web page is also an important web page (Page & Brin, 1998). The essence of the PageRank algorithm is to sort all the importance of all pages according to the links between pages. Suppose there are four pages A, B, C and D. The B, C, and D pages all link to page A, then the PR value of page A is the sum of PR value of three pages A, B and C namely $PR(A) = PR(B) + PR(C) + PR(D \ )$. Suppose that page B also links to two other web pages, and page C also links to page D, in that case $PR(A) = \frac{PR(B)}{3} + \frac{PR(C)}{2} + PR(D\ )$.

In the case that there are pages only link to themselves or not linked by any other pages in the World Wide Web, in order to avoid the situation where the PageRank value can't be converged due to these special pages, the PageRank algorithm introduces a damping factor q, users who browse randomly will jump to any random page in the network with a probability of $1 - q$. The values of damping factor are tested in various field, but it is generally assumed that the damping factor $q = 0.85$.

The essence of PageRank is a random walk process, if the WWW are described by the directed graph (Diestel & Reinhard, 2005) $G(V,E)$, V is the set of nodes, E is the set of directed edges in the network: $v_i,\ v_j,\ v_k,\ v_h \ldots\ \in V$, $e_{ij} \in E$. e_{ij} indicates the edge from the node i link to node j, and the PR values of any web page can be expressed as:

$$PageRank(i) = q * \sum\nolimits_{j \in V} \frac{PR(i)}{L(i)} + \frac{1-q}{N} \qquad （2\text{-}4）$$

Where L is the number of the node i link to other pages and N is the total number of nodes in the whole network. PageRank algorithm is still used by Google as one of the symbolic methods to identify important web pages (Massimo, 2011).

2.1.3 Knowledge diffusion in citation network

Knowledge can be divided into two different types: explicit knowledge and tacit knowledge. Explicit knowledge is usually expressed in the form of words, pictures, mathematical formulas and so on. It can be transmitted without the constraints of time and space, such as the citation of documents in citation networks. On the contrary, tacit knowledge is difficult to be expressed by words and voices, so the main carrier of tacit knowledge flow is human, which exists in the mind of the subject and is transmitted through specific situations. Although there are essential differences between explicit knowledge and tacit knowledge, explicit knowledge is often the compilation of tacit knowledge, and explicit knowledge must reflect the flow of tacit knowledge in the process of flow. Therefore, understanding the relationship between these two types of knowledge is crucial to understanding the complete process of knowledge flow. Whether it is a citation network based on patents or papers, the process of knowledge flow brought about by citation relationship belongs to the knowledge flow among explicit knowledge sets, which has become an important field of citation network analysis at present (Chen & Hicks, 2004).

The transfer of knowledge flow is often a process of moving between knowledge subjects through a certain path and medium. Knowledge flow contains three important factors: subject, content and direction (Zhuge, 2006). Knowledge exists in different subjects and subjects can also be understood as the carrier of knowledge. The flow of knowledge cannot exist independently from the subject, and the flow of knowledge is often transmitted continuously through multiple subjects. In the field of scientific research, the carrier of explicit knowledge is mainly journal and literature, while the carrier of tacit knowledge is scientific researchers. Knowledge flow is the process of content transfer, so the value of content determines the value of knowledge flow, and the value of knowledge is often determined by subjective and objective evaluation. The direction of knowledge flow is the direction from the sender of knowledge to the receiver of knowledge. Usually, the receiver of knowledge will also become the sender of knowledge to transfer in the further step.

It can be concluded by the three elements of the knowledge flow that knowledge is suggested to flow from cited nodes to citing nodes which also reflects the process of knowledge flows. Given any arbitrary node, the idea proposed disseminates through the conduits until it hits an end node. In this way, a citation network represents a system of influence generation where the inventive step is provided by the recombination of existing pieces of influence (Martinelli & Nomaler, 2014).

Knowledge flow is a dynamic process in citation network over time. Knowledge flow in different fields will gradually show two trends: the first trend is that the value of knowledge used in the process of knowledge flow in a certain field decreases and shows an aging trend. The specific performance in citation network is that the citation frequency of documents decreases with the time passes by. The second trend is that with the accumulation of knowledge in the process of flow, in-depth research in a certain field gradually gathers to form a small cluster. This trend is called the trend of knowledge gathering, and these small clusters gradually gather to form a larger cluster and covers the whole subject area (Wang & Zhang, 2014).

2.1.4 Research on citation network

At present, citation network is widely used the research of citation network knowledge flow and knowledge sharing, mining technology path, revealing the evolution of scientific structure, optimization of scientific research evaluation indicators and other fields by many scholars (Chen, 2016).

Hummon indicated in the 1990s that the reachability of knowledge flow in citation networks originated from the mutual citation of nodes. Different documents play different roles in the process of knowledge flow transmission. This paper is also one of the papers about knowledge flow field which had a significant impact on the follow-up research earlier. Neoclassical economics assumes that information is free-flowing, cost-free and can be absorbed at a neglected cost (Arrow, 1972). Therefore, in the study of knowledge flow and knowledge sharing in citation networks, David (2005) believes that the innovative ability of a company depends not only on the professional and knowledge development within the company, but also on the knowledge sharing of patent citation networks in the same field. The position of knowledge flow network optimizes knowledge production, thereby driving the company's comprehensive strength.

The main path of technology is the backbone of technology development. It contains the most critical nodes in the process of technology development and the link relationship between these key nodes. The patent literature contains a detailed description of patent innovation, so the main path extracted from the patent citation network. It can better reflect the inheritance and development of technology (Griliches, 1900). Andersen (1998) illustrates how technological evolution has become increasingly interrelated and complex and how typical trajectories of individual technologies explain technological evolution better than conventional aggregate measures through analyzing 1890-1990 US patent data. Verspagen (2007) obtained the main evolutionary paths of different stages in the field of fuel cells in the 1980s through the main path analysis method in the patent citation network, which initiated the research of combining the evolution of the main path with the patent citation network.

Scholars analyze the evolution of the scientific structure of citation networks by studying the regularities of nodes, edges, citations and degrees. Price (1965) pointed out that the number of scientific journals increased exponentially with an annual growth rate of 4.7% and multiplied at an average of every 15 years. He also pointed out that the citation network's node indegree obeys the power law distribution with an index of 2.5 to 3. Lutz Bornmann and Ryudiger Mutz (2015) extracted various subject publications from Web of Science from 1980 to 2012. The research shows that the output of global scientific publications from 1980 to 2012 still satisfies the exponential growth rate, with an average annual growth rate of 3%. The average number of documents doubled every 24 years. Redner (2005) analyzed the citation data of Physical Review journals from 1893 to 2003. The results show that when major discoveries in a field are recognized, there is a sharp increase in citation. There is a positive correlation between the number of citations to a paper and the average age of citations. However, the distribution of citations declines slowly with the increase of publication time.

In the optimization of scientific research evaluation indicators, since Garfield proposed citation network and citation analysis, many scholars have proposed quantitative citation analysis indicators. Commonly used citation analysis indicators have cited numbers, co-citation and coupling, and journal impact factors (Doreian, 1985). The PageRank algorithm (Page & Brin, 1998) is also used to evaluate the quality of citation network nodes.

In addition to the mentioned areas, citation networks are also applied in academic search applications, knowledge map applications and library literature management applications and so on (Hu & Chen, 2004).

2.2 Traditional main path traversal counts methods

2.2.1 Search path link count

In Hummon and Doreian' s paper, on the example of the DNA citation network, which is one form of a directed acyclic graph (DAG). They used the "priority first search" combined with exhaustive search algorithm method to figure out the strongly connected node constructing to the main path. Then proposed three main path methods using traversal counts for each link in the citation network: node pair projection count (NPPC), search path link count (SPLC), search path node pair (SPNP) respectively. Traversal counts measure the times a citation link has been traversed if one exhausts the search from a set of start nodes to another set of end nodes. These three indices all use the weights of each link to identify the important part of a citation network, namely the main path in a citation network. Many new traversal counts draw on Hummon and Doreian have been carried on contributing to the efficiency and effectiveness of searching the main path. In 2003, Batagelj made the further step of the above three indices called search path count (SPC) for the analysis in the large citation networks. Before this application, Hummon and Doreian (1990) also had tried the main path analysis in larger sample. SPC identifies the weight of each link through the times all possible paths traverse from the all sources to all sinks. Then, the SPX is used to represent these above for algorithms. In this section, NPPC will not discussed because it is not suitable for the large network because of its complex computation (Batagelj, 2003). The remaining three traditional traversal counts algorithms will be described.

The core idea of SPLC is to calculate the number of search paths in a citation network from a start point to an end point through a connected path. And the starting point of the search is not necessarily the source of the citation network subgraph, the starting node is the ancestors of the tail node of the edge. Table 2-1 simulates the SPLC algorithm proposed by Hummon and Doreian and calculates the SPLC values of all edges in Figure 1-1.

Table 2-1 SPLC of each edge

Order	Edge	Search Path	SPLC
1	A-D	A-D-G A-D-H	2
2	A-F	A-F	1
3	B-D	B-D-G B-D-H	2
4	B-H	B-H	1
5	C-E	C-E-F C-E-H C-E-I	3
6	D-G	A-D-G B-D-G D-G	3
7	D-H	A-D-H B-D-H D-H	3
8	E-F	C-E-F E-F	2
9	E-H	C-E-H E-H	2
10	E-I	C-E-I E-I	2

2.2.2 Search path node pair

The SPNP algorithm takes into account that the edges connecting more node pair is traversed more times compared to the source and sink. Becasue the source has no references and the sink doesn’t cited by any other documents. Based on this idea, calculate the SPNP value of an edge from the ancestors of the tail node to all its descendants of the head node, the tail node and head node both include themselves of the node pair. As shown in Figure 1-1, the SPNP value of edge D-H is 2, because there are three edges starting from A, B and D which are the ancestor of node D and the descendant is only H. The paths passing through D-H are A-D-H and B-D-H, respectively.

The SPNP calculation results for each edge of Figure 1-1are calculated as following:

A: D (3) F (1)

B: D (3) H (1)

C: E (4)

D: G (2) H (3)

E: F (2) H (2) I (2)

2.2.3 Search path count

The SPC algorithm calculates the SPC value of each edge by traversing all the times from source node s to sink node t in the citation network. We can define citation network as $G(V,E)$, V is the collection of all nodes and E represents the collection of all edges in citation network. Taking the $E(u,v)$ as the count of the number of different paths from source to the sink through the edge (u,v). To compute the $E(u,v)$ easier, two auxiliary quantities are introduced. $e(\mathrm{v})^-$ represents all the paths from node s to the node v and $e(\mathrm{v})^+$ indicates all the paths from node v to node t. $E(u,v)$can be expressed as:

$$E(u,v) = e(u)^- * e(\mathrm{v})^+, (u,v) \in V$$

The SPC values of each side of Figure 1-1 can be obtained as follows:

A: D (2) F (1)

B: D (2) H (1)

C: E (3)

D: G (2) H (2)

E: F (1) H (1) I (1)

In general, the weghting algorithms are improving since the main path analysis was put forward. For example, Choi and Park (2009) suggested an algorithm to identify the patent development path called forward citation node pair (FCNP) by calculating one link's weight through multiplying the number of forward citations of a node pair. Forward citation aids in understanding the economic or technological value compared to the backward citation (Trajtenberg, 2002). Persson (2010) suggested two approaches to measure weight called weighted direct citations (WDC) and normalized weighted direct citation (NWDC) to calculate the link according to the direct citations and co-citations relationship between two nodes. The above three traditional traversal counts are still widely applied nowadays.

2.3 Main path search approaches

2.3.1 Local search

The "priority first search" Hummon and Doreian proposed in 1989 then defined as the local search which is an approach to search the most vital paths in the citation networks through selecting the subsequent follower with the highest SPX weight. Take the Figure 2-2 as an example, the thicker black path shows the main path of Figure 1-1 based on

SPC algorithm. Searching start from the sources A, B and C, because the path C-E has the largest SPC value compare to other two, so this path is extracted as a part of the main path. After reaching the node E, it is calculated that the three paths E-F, E-H, and E-I have the same value of SPC, so three paths are all search by the local search method. There are three main paths in Figure 1-1, C-E-F, C-E-H, C-E-I.

2.3.2 Global search

Contrast to the local search, global search is a new approach to emphasizing the overall knowledge flow which means selecting the path with the largest overall traversal counts among all possible paths in citation networks. This approach was first suggested by Liu and Li in 2012 and similar with the idea of critical path method of James el al. and (1959) for project scheduling and the global search can avoid the local optimization problems caused by the local search which adds more significance in main path analysis. The thicker line in the right figure of Figure 2-2 shows the main path obtained by global search of Figure 1-1 based on the SPC algorithm. There are 7 main paths using the global search, A-D-G, A-D-H, B-D-G, B-D-H, C-E-F, C-E-H and C-E-I, all main paths have the total SPC value of 4 by global search.

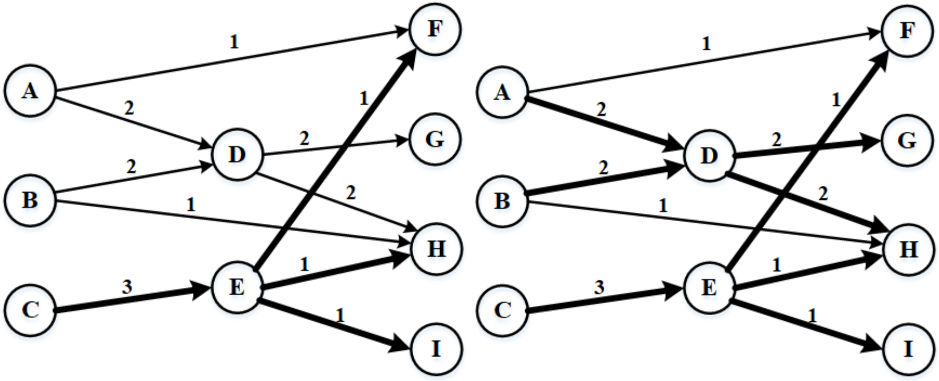

Figure 2-2 Local search and global search based on SPC

2.3.3 Key-route search

There is a common limitation for both local and global search methods that the single path will neglect some significant link that will be not included in the main path which limits the exploration in the development of technology. For purpose of overcoming this potential serious problem, Liu and Lu (2012) suggested the key-route search which can search not only one main path but also contains all the most significant links. Best of all, the numbers of the main path could be set firstly. The main procedures of key-route search can be described as two steps: set the number of edges with the largest traversal count that can only be included in the searched main path and filter out from the citation network firstly. Then we can choose it begin searching from the tail node of the searched critical path until the source of the citation network or choose to search from the source to the tail node. Forward and reverse searches can choose local and global search methods to identify the main path. Xiao and Lu (2014) investigated a social network using these three main path analysis methods and the results show the key route search is better showing more detailed diffusion information than the others. Ho, Lin and Liu (2014), Hung and Liu (2014) also applied the key route search get the similar results.

2.4 The applications of main path analysis

Many researches applied the main path analysis to the specific domain to investigate the development of the specific major using the citation data. In the initial research, main path analysis was applied to academic publications: Hummon and Doreian (1990) applied it to DNA development, Carley et al. (1993) applied it to the social network analysis field. They analyzed scientific influence in the conflict resolution field by using main path analysis. Calero-Medina and Noyons (2008) applied the main path analysis to the field of absorptive capacity and concluded four critical themes from the nodes extracted in the main path. Harris and Luck (2009) analyzed the discovery and delivery in secondhand smoke exposure (SHS) research. Halatchliyshi (2014) applied it to the open learning community and Zhu et al. (2016) used for the online social networks.

There are also various researches which have applied the main path analysis to investigate technological trajectories using citation information: Verspagen (2007) investigated the technological trajectories of fuel cell technology and proposed an historical approach to show the evolution of main path in fuel cell research which could be used to find the life cycle to analyze the trends and understand the main path. Mina et al. (2007) used the main path analysis which was also conducted based on SPX algorithms for getting the emergence, growth and transformation in medical knowledge which is a large-scale empirical analysis. Martinelli (2012) applied a main path analysis to trace the Telecommunications switching industry. Epicoco (2013) examined the long-term evolution of the semiconductor miniaturization trajectory. Ho, Lin and Liu (2014) explored the knowledge diffusion of Membrane electrode assembly technology, and Huenteler (2016) analyzed the long-term pattern of innovation and technological life-cycles in the Wind power and Solar PV fields. Jonathan C. (2014) found the technological barriers and trend in fuel cell field by multiple path analysis.

2.5 Summary

This chapter mainly demonstrate the theoretical research from four aspects: citation network, main path traversal counts algorithms, main path search methods and the application of the main path analysis. From the existing main path analysis research and applications in citation networks, we can see that the main path analysis is based on graph connectivity to discover the key nodes in a specific field which has been used by many scholars to find the evolution trend of the main scientific path or the main technological path. Relevant researchers also have been looking for some new algorithms to weighting the edges and find important nodes.

In order to open up new possibilities in searching more relationships among nodes and greatly increases our capability in exploring the development of a target research domain, this thesis will propose three new algorithms based on the influence flow idea in citation networks for main path analysis. This new weighting approach also provides a new persepctive to overcome the single purpose traditional traversal counts bring.

3 MAIN PATH ANALYSIS BASED ON TRAVERSAL COUNTS METHOD

3.1 Method selection of search path link count

We have already discussed the main idea of the search path link count (SPLC) algorithm for weighting each edge in citation network. SPLC accounts for the number of all possible search paths through the network emanating from an origin node.

SPLC algorithm is selected in this chapter compared to SPNP and SPC to do the control experiments to compare with the proposed new algorithm in this paper because SPLC can avoid some neglection citation relation in the main path caused by SPNP. For example, if there is a citation relation between node 1 and node 2, 2 and there, 1 and 3, the bigger number means the later document. The SPNP will neglect the citation relation between 1 and 3 in the main path, but SPLC will take all the citation relations into account (Lucio-Arias & Leydesdorff, 2008).

SPC is another improved algorithm based on the other two traversal counts methods proposed by H&D. Compared to SPC, though it is more efficient in large citation networks. SPLC preferred more citation chains to enumerate (Liu, Lu & Ho, 2019) which is considered more comprehensive when compare the main paths between different algorithms. Therefore, we select the SPLC algorithm to get the main path in this chapter and compare the results with the proposed new algorithms in the next chapter.

3.2 Data and methods

3.2.1 Data

The experimental data in this chapter are derived from the US-licensed patent dataset in the United States Patent and Trademark Office and the Web of Science core collection of the WOS (Web of Science) database. The purpose of using two different databases is to extract the main path of technology and the main path of science using the main path analysis method.

In the US authorized patent database, the patents in the field of desaltation from January 1976 to June 2018 were selected for analysis, we get 1348 patents by searching the “desalt*” search term from the title and abstract part of the patent. Two reasons are considered to choose desalination as a research area. Firstly, because of the shortage of fresh water resources nowadays, desalination technology as the next promising method of extracting fresh water has been highly valued by relevant scientific researchers (Zhen et al., 2016). Secondly, seawater desalination can be traced back to 1400 B.C. (Zhu, Xue &Xu, 2014), and the rapid development of seawater desalination technology in the 1980s shows that there have been decades of technology accumulation in the field of seawater desalination. Therefore, the amount of patent literature data is large which means that the evolution of the main path of technology will be more obvious for analyzing.

The Web of Science TM core collection is a globally influential database with a total of 10,000 academic journals and more than 100,000 international conferences. We choose literature data to analyze simultaneously in order to verify the applicability of the main path analysis method in the scientific citation network in this paper. A total of 5780 data are retrieved from WoS in the field of information security from January 1900 to February 2019. TS in WoS stands for the meaning of topic, TS= "Information Security" or TS= "Information Safety" or TS= "Information Privacy" or TS= "Information Ethics" are selected as subject search terms. There are also two reasons for us to select the information security as a search object: 1) The earliest literature of “information security” searched from the WoS core collection appeared in 1969, namely there is a 50-year scientific accumulation process in this field until 2019, so it can better analyze the evolution process of key literature in the main path; 2) The era is changing constantly, information plays a very important position in people's real life and the development of all walks of life in 21th century. However, with the emergence of big data, information faces serious security problems. Therefore, analyzing the evolution of the main scientific path in the field of "information security" can help researchers analyze the changes of research topics in the field of information security, and predict potential research fields and derive countermeasures.

The time distributions of the two datasets are shown in Figure 3-1 and Figure 3-2.

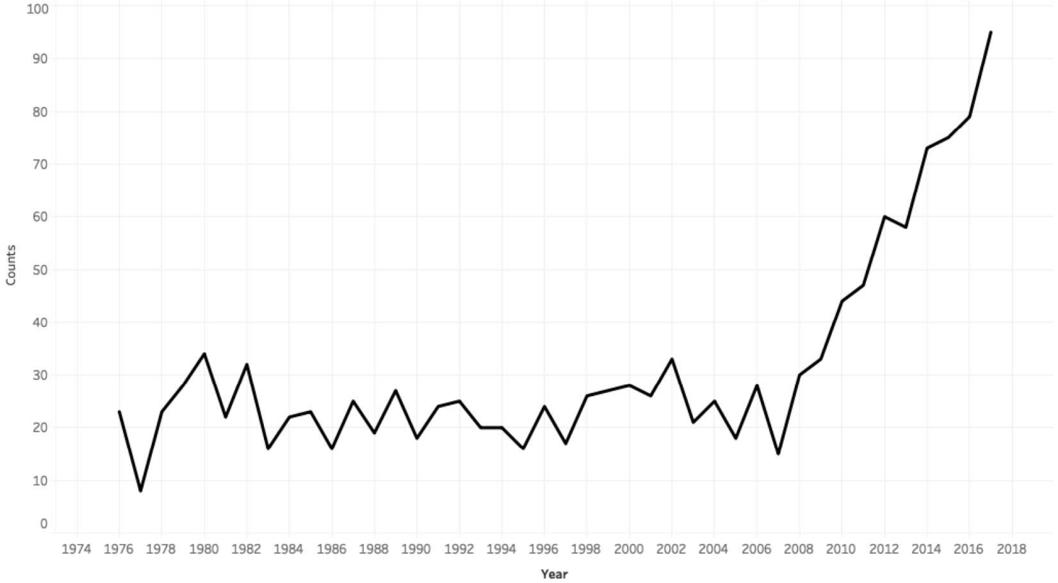

Figure 3-1 The applied patent literature distribution in desalination

From Figure 3-1, it can be seen that the number of patents published in desalination has been in a stable development period until 2007, with the number of patents published fluctuating between 10 and 35. During the 10 years from 2007 to 2017, the technology in the field of seawater desalination developed rapidly, and the number of patents issued reached 95 in 2017.

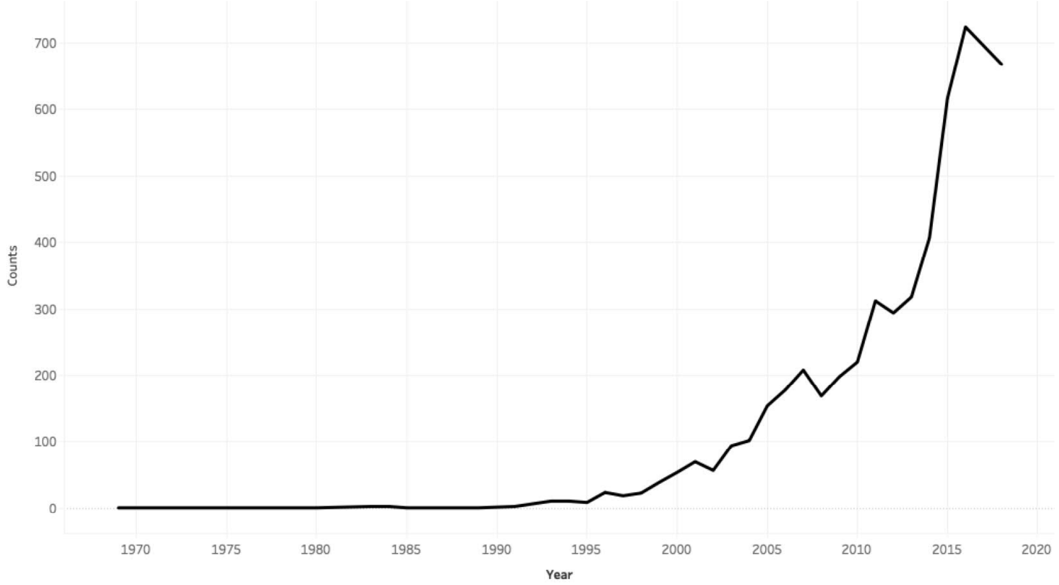

Figure 3-2 The published literature distribution in information security

It can be clearly seen from Figure 3-2 that among all the literature retrieved from WoS core collection, the number of literature related to the field of information security before 1992 was almost one every year and in a state of stagnation. This is largely due to immature development of computer technology before 1992 which caused information security in the field of related research is rare. However, with the development of computer technology and big data, research in the field of information security began to increase rapidly after 1992. In 2004, it broke through 100 related articles. In 2017, it reached the maximum value of 696 in the search year, but from Fig 3-2, it seems that the number of studies in the field of information security has only temporarily volatility, not always falling.

3.2.2 Methods

Specific experimental steps and corresponding tools are shown in Figure 3-3:

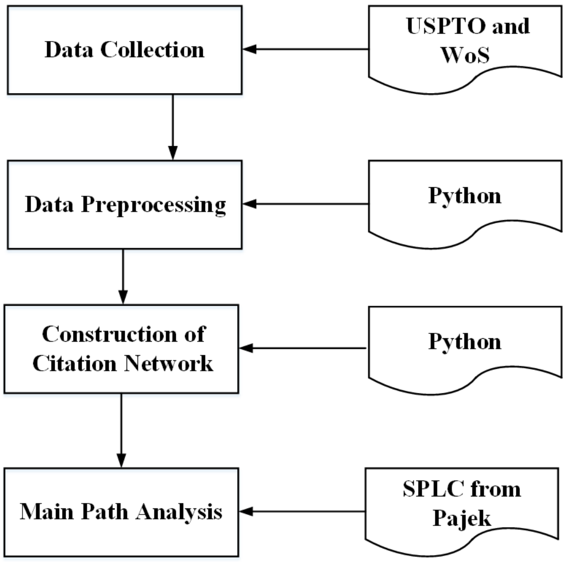

Figure 3-3 Steps and tools in experiment

Step 1: Data preprocessing. The patent data of the processed data includes the patent references, the year of publication, and the country in which the patent is published. The processed literature data includes the serial number corresponding to the literature, the reference of the literature, the author of the literature, and the year in which the literature was published. The preprocessing results of the patents are shown in Table 3-1.

Table 3-1 Patent data preprocessing example

Patent ID	References	Publication Year	Country
4200550	19580600、19730200、19740500…	1980	France
8652303	19480700、19750400、19871000…	2014	Japan
9845252	19690300、19910100、19940800…	2017	America
...	...	....	...

Step 2: Construction of citation network. Construct the citation network as shown in Figure 3-4. The citation network of this chapter is constructed through the collected data and their corresponding references.

Step 3: Obtain the SPLC of each edge. According to the idea of edge traversal of SPLC algorithm, the traversal counts of each edge in the constructed citation network are obtained.

Step 4: Search the main path. Select the global search method to search the main paths in the field of seawater desalination and information security. Since the global search method can avoid the local optimal results, the global search method is selected to search the main path in this chapter.

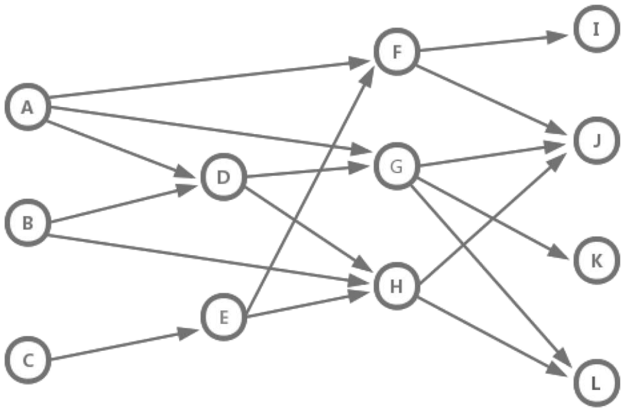

Figure 3-4 Citation network structure

3.3 Result

3.3.1 Technological main path

In Figure 3-5, the ordinate indicates the publication year of patents but the abscissa has no special meaning. The node size in the graph indicates the cited frequency of the node and the thickness of the edge represents the weight of the edges, that is, the value of the traversal count values of edges calculated by SPLC algorithm.

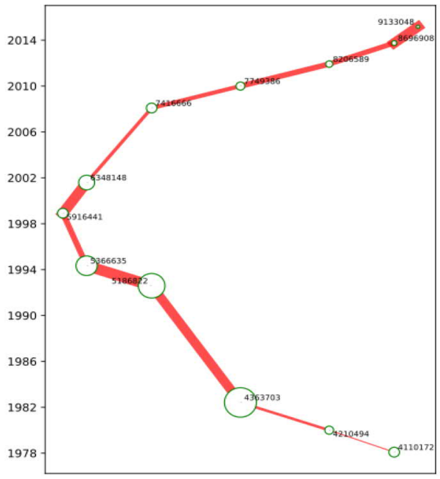

Figure 3-5 The main path of desalination based on SPLC

Along with the time order, it can be clearly seen from the main path in Figure 3-5 that the main path of technological evolution in the patent citation network of seawater desalination starts from patent 4110172, and the sink is 9133048. The time span of the main path is from 4110172 nodes in 1977 to 9133048 nodes in 2015, which shows the technological development in the 38 years of evolution in the field of seawater desalination. A water-containing pond for collecting solar energy for utilization in a process for recovering potable water from non-potable water was applied in 4110172. A solar desalination system comprising a tank having a transparent cover was invented in 1980 (Node 4210494). Patent 4363703 describes a solar energy desalination process utilizing solar radiation directly for the evaporation of salt water in 1982. Tzong applied a desalination apparatus including pressure responsive desalination means, a storage tank and conduit (Node 5186822). Watkins put forward the desalinization apparatus and method for removing salt from sea water (Node 5366635). The main discovery of the patent 5916441 described a desalination system using hydrostatic pressure formed in a vertical mine shaft to function reverse osmosis membranes plumbed so that permeate water from the last membrane is input water to the next. After the Millennium, technology in the field of seawater desalination has developed to removing salt from seawater to produce potable freshwater in 2002（Node 6348148）. In 2002, inventors applied a apparatus comprised of a water intake system, a reverse osmosis system, a concentrate discharge system, a permeate transfer system, a power source, and a control system (7749386).

There are 4 key nodes occurred in the main path, these are patent 7749386, 8206539, 8696908 and 9133048. Patent 7749386 and patent 8206539 were applied by the same inventors. The prior one provides methods and an apparatus for more efficiently and economically producing purified water from sea water or some other salty or brackish water source. The latter one provides methods and apparatus for producing purified water from sea water or some other salty or brackish water source by using brackish concentrate mixed with salty water. Patent 8696908 described methods and apparatus for releasing treated wastewater into the environment by combining the treated wastewater with a concentrated salty water. The latest patent 9133048 has described a seawater desalination method obtaining mixed water by mixing the ultrafiltration membrane-treated water and the membrane bioreactor-treated water.

The incremental innovations are often linked to the further refinement and development of basic breakthroughs that set direction for development for a long time. Since the technological path theory has been put forward for more than 30 years by Dosi (1982), these innovations have been well displayed in the technological paths. Those technologies will be replaced by some new technologies or evolve into other technologies, all of which will create the responding new technological paths from the natural trajectories. The original technologies became an important or key node in the existing trajectories.

Therefore, by extracting the patented nodes in the main path in the field of seawater desalination, we can well analyze the process of alternation of new and old technologies in the field of seawater desalination and the termination nodes which played an important role in the process of technological evolution.

Table 3-2 Nodes on the technological main path used SPLC

Patent ID	Year	Inventors	Title
4110172	1977	Spears, Jr, et al.	Solar energy collecting pond
4210494	1980	Rhodes & William	Solar desalination system
4363703	1982	EIDifrawi, Ahmed A,et al.	Thermal gradient humidification-dehumidification desalination system
5186822	1993	Tzong, Tsair-Jyh, et al.	Wave powered desalination apparatus with turbine-driven pressurization
5366635	1994	Watkins & larry	Desalinization system and process
5916441	1999	Raether & Roger J	Apparatus for desalinating salt water
6348148	2002	Bosley & Kenneth R	Seawater pressure-driven desalinization apparatus with gravity-driven brine return
7416666	2008	Gordon & Andrew W	Mobile desalination plants and systems, and methods for producing desalinated water
7749386	2010	Voutchkov & Nikolay	Desalination system
8206589	2012	Voutchkov & Nikolay	Desalination system and method for integrated treatment of brackish concentrate and seawater
8696908	2014	MacLaggan & Peter	Desalination system and method of wastewater treatment
9133048	2015	Ishihara & Satoru	Seawater desalination method

3.3.2 Scientific main path

The ordinate of Figure 3-6 represents the year in which the literature was published, while the abscissa has no special significance. The size of the node in the figure represents the cited frequency of the node, and the thickness of the edge represents the SPLC value of the edge.

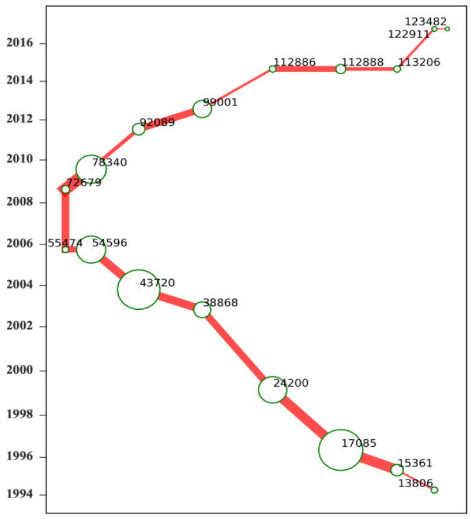

Figure 3-6 The main path of information security field based on SPLC

It can be seen from Fig3-6 that the source of the evolution main path of the information security scientific citation network is 13806, and the sink is 123482. We can draw that there are 4 nodes in the main path before the year of 2000 through analyzing the node contents of the main path in Fig 3-6 and Table 3-3. The earliest node was a review literature (Node 13806), the main content was the integrated social contract theory (ISCT) in business ethics. With the globalization of information system management and data transmission, the relationship among nationality, cultural value, personal information privacy and information privacy monitoring methods was investigated in document 15361 in 1995. The result shows that the hierarchical structure of information privacy seems to be consistent in different countries, and there is no relationship between cultural values and information privacy awareness. The research content of node 17085 mainly built a tool for information privacy to identify and measure individuals' attention to organizational information privacy events. Node 24200 explored the role of procedural fairness in privacy issues that may cause leakage of personal privacy and found that companies can retain more customers by enhancing the fairness of the program. Literature 38868 from the same author Culun of node 24200 proposed a theoretical framework model of equity including a combination of government regulation, industry self-discipline and technological solutions for possible information privacy problems.

With the development of e-commerce, information security on e-commerce websites has become a key concern for users. In 2006, Malhostra constructed a causal model for internet users' information privacy concerns (IUIPC) on e-commerce websites and investigated on-the-spot the reactions and connections of network consumers to various privacy threats on the Internet (Node 43720). Documents 54596, 55474 and 73679 are all from the literature published by Dianev. The authors have constructed corresponding structural equation models (SEM), and studied the factors that users disclose their personal information on e-commerce websites and the factors that cross-cultural differences affect the use of e-commerce websites respectively. The result showed that the construction of these two models are beneficial to the improvement of e-commerce performance to a certain extent. In 2009, Dianev continued to extend the research of cross-cultural differences to users in different countries for comparative research on behaviors between information protection technologies and suggested that cross-cultural differences should be considered in global network information security policies (Node 72679).

Node 78340 focuses on exploring the factors that influence employees' compliance with information policies (ISP) and concluded that employees' compliance with information policies can reduce security risks and enhance corporate credibility. A similar type of study was also carried out in literature 92089 in 2010, indicating that in the field of information security around 2009, enterprises began to put more focus on employees' attitudes towards information security policies as well as the factors and influences causing these attitudes. In 2013, Crossler summarized five categories of future research topics in the field of information security based on the previous hot topics of researchers in the field of information security: insider deviant behavior, unmasking the mystery of the hacker world, improving information security compliance, cross-cultural information security research and data collection and measurement issues. Correspondingly, he summarized the actions that should be carried out and the preventive measures that should be taken by the staff concerned in these five themes. The main path literature 112886 has little similarity with other nodes' content. Node 112886 mainly designed the UI interface based on accountability theory and reduce access policy-violation intentions from the perspective of avoid access-policy violation.

Because the violation of information security has become more and more common in the 21st century, where information is developing rapidly, it is increasingly important and necessary to motivate more people to work on information security protection. Document 112888 proposes three improvement aspects of the existing protection motivation theory (PMT) and the results showed that the proposed model has better applicability than PMT. The research topics of the three latest nodes 113206, 122911 and 123482 in the main path of information security are all about information security protection measures within the organizations. Node 113206 continues the content of PMT which included detailing how organizational commitment is the mechanism through which organizational security threats become personally relevant to insiders and how SETA (security, education, training and awareness) efforts influence many PMT-based components. Node 122911 extended an oft-cited theory in the information security literature PMT to include the relationship of insiders' psychological capital (PsyCap) with the mechanisms of PMT. The last node in the main path is a review article on information security policies within the organizations and built a research framework for 114 influential journal articles in the field of security policy that outlined the structural linkages in the current literature (Node 123482).

Table 3-3 Nodes in the main path used SPLC

No.	Year	Authors	Title
13806	1994	Thomas Donaldson & Thomas W. Dunfee	Toward a unified conception of business ethics: Integrative social contract theory
15361	1995	SJ Mulberg, SJ Burke, et al.	Values, personal information privacy, and regulatory approaches
17085	1996	HJ Smith & SJ Milberg	Information privacy: measuring individuals' concerns about organizational practices
24200	1999	MJ Culnan & PK Armstrong	Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation
38868	2003	MJ Culuna & RJ Bies	Consumer privacy: Balancing economic and justice considerations
43720	2004	NK Malhostra &SS Kim, et al.	Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model
54596	2006	T Dinev & P Hart	An extended privacy calculus model for e-commerce transactions
55474	2006	T Dianev, M Bellotto, et al.	Privacy calculus model in e-commerce-a study of Italy and the United States
72679	2009	T Dianev, J Goo, et al.	User behavior towards protective information technologies: the role of national cultural differences
78340	2010	B Bulgurcu, H Cavusoglu et al.	Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness
92089	2012	Q Hu, T Dinev, et al.	Managing employee compliance with information security policies: the critical role of top management and organizational culture
99001	2013	RE Crossler , AC Johnston, et al.	Future directions for behavioral information security research
112886	2015	A Vance,PB Lowry, et al.	Increasing accountability. Through the user interface design artifacts: A new approach to addressing the problem of access-policy violations
112888	2015	Boss, Scott R, et al.	What do systems years have to fear? Using fear appeals to engender threats and fear that motivate protective security behaviors
113206	2015	C Posey, TL Roberts, et al.	The impact of organization commitment on insiders' motivation to protect organizational information assets
122911	2017	AJ Burns, C Posey, et al.	Examining the relationship of organizational insiders' psychological capital with information security threat and coping appraisals
123482	2017	WA Cram, JG Proudfoot, et al.	Organizational information security policies: a review and research framework

3.4 Summary

In this chapter, the SPLC algorithm in the main path analysis method proposed by Hummon and Doreian (1989) is used to calculate the edge traversal counts, and then the main path is extracted by the path search method of global search. The purpose of using SPLC to search main path is to get the main path result as the control group data and then compare and similarities, differences and applicability of the main paths obtained by different algorithms in Chapter 4.

The traditional SPX main path algorithm has been widely applied to different fields to reveal the technological or scientific main path in the patent citation network or scientific citation network. However, the traditional main path method doesn't pay close attention to the practical significance of the nodes in different types of citation network. It is far from enough to use the method of only traversing the edges directly in patent or scientific literature with substantial significance. Taking the scientific main path we got as an example, the source node and sink node of the main path are two reviews just because the outdegree and indegree of the reviews literature are very high in general, they are unavoidable to be searched in the main path by SPX algorithm. Literature reviews are always treated as a critical method for evaluating the previous literature (Martyn, 2019), also literature reviews are secondary sources and don't report new or original experimental work (Hart, 2018). Therefore, the literature review doesn't do well in helping researchers to discover the evolution of the subject in scientific citation network. The result is consistent with the conclusions some scholars drawed that the main path derived from the traditional traversal counting method is difficult to achieve fine structure revealing in the subject area (Han & Jin, 2012). The search of the main path requires a new perspective to reveal critical nodes that are ignored in the citation network or to show a more elaborate technological or disciplinary development path.

4 EQUALLY-DIVIDED INFLUENCE FLOW ALGORITHMS WEIGHTING THE PATH FOR MAIN PATH ANALYSIS

4.1 Influence flow algorithm based on PageRank weighting path

4.1.1 Problem statement

As we mentioned in the research gap, the calculation of edge based on the traditional SPX algorithm has the following problems: the literature with the higher citation frequencies and more references will get more traversal counts, the higher frequencies of citation do indicate more attention and affirmation by other literature, which is the main indicator to measure the importance of nodes. However, higher references don’t necessarily reflect the importance of the edges passing through this node. Meanwhile, because each citation has a different effect on knowledge diffusion, the importance of edges with the same weight of traversal counts are different in citation network. The traversal counts idea of the traditional SPX algorithm don’t consider the initial weight of the edge will cause some deviation for the main path analysis (Wei & Fang, 2016).

This section proposes an algorithm for calculating each edge weight in citation networks based on PageRank idea—weighting each edge by the different influence value of the citing nodes. Then the global search method is used to get the main path with maximum total influence flow. Finally, we compare the similarities and differences between this method and the traditional main path analysis method through empirical analysis.

4.1.2 Algorithm description

PageRank algorithm can be applied to cyclic and non-cyclic network structure, so it can be applied to citation network. In order to clearly illustrate the applicability of PageRank algorithm in citation network, we take Fig. 3-4 as an example to describe the iteration process. Assuming that a simple complete citation network is composed of five nodes, *A*, *B*, *D*, *G* and *H* in Fig. 3-4, the influence flow is supposed to transmit form the new node to the old node, for example the node *D* flows its influence to node *A* though the edge is from A to D. Firstly, construct the adjacency matrix *H*, $H_{ij} = 0$ indicates node *i* has no references. But in order to avoid the occurrence of non-convergence if there is no link out of a node when the PR value is calculated, node A and B are treated they are link to themselves, *H* is expressed as:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{4-1}$$

Secondly, normalizing the adjacency matrix. Divide each element of the matrix by the sum of each row to get the normalized matrix *A*, as shown below:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix} \tag{4-2}$$

Thirdly, Calculating the transition matrix. Transpose matrix *A* to get transition matrix *T* so that the elements of each column sum up to 1. Matrix *T* is as follows:

$$T = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4-3}$$

Fourth, add the damping factor *q* to the transition matrix and get the new transition matrix *A'*, $A' = q * T + (1 - q) * ee^t/N$.

$$A' = 0.85 * \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + 0.15 * \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix} \tag{4-4}$$

$$\begin{bmatrix} 0.88 & 0.03 & 0.455 & 0.455 & 0.03 \\ 0.03 & 0.88 & 0.455 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.455 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{bmatrix}$$

Lastly, The PR value of each node is calculated iteratively. Assuming that nodes *A*, *B*, *D*, *G*, *H* are randomly assigned to 1, 2, 3, 4, 5, the iteration process for each node is shown in Table 4-1.

Table 4-1 The process of iteration

Iteration	**A**	**B**	**D**	**G**	**H**
0	1	2	3	4	5
1	4.275	5.550	4.275	0.450	0.450
2	6.092	7.176	0.833	0.450	0.450
…	…	…	…	…	…
49	6.633	6.634	0.833	0.450	0.450
50	6.634	6.634	0.833	0.450	0.450

From Table 4-2, we can find the PR value of each node of the citation network composed of five nodes A, B, D, G and H converges to a stable value after fifty iterations.

Citation networks and web networks have similar topological structures, they all satisfy the requirements of Markov property (Gagniuc, 2017). Therefore, citation relationships and links in web are essentially similar (Duan, Zhu & Wang, 2012), reflecting the relationship between the adjacent nodes and the influence relationship: The PageRank algorithm calculates the importance of a web page through a link relationship between web pages, and the citation network flows the influence value through the citation relationship from citing nodes to cited nodes. In a web network, the greater the PR value of the web page, the greater the impact of the pages linked. Similarly, the greater the influence of the citing literature, the greater the influence of the citing nodes in the citation network, at the same time, the more references cited, the less influence each reference literature obtained. We can regard the citation relationship as a process of voting. The process when a node cited other nodes, it can be treated as a process of voting to the cited literature from the citing node. The more important the citing literature is, the more important the votes it will cast. Moreover, if the citing nodes owns more references, the fewer votes it will get for each cited document.

The aim of PageRank algorithm is to sort the importance of pages in web network. However, different from the purpose of sorting nodes by PageRank, the algorithm of this section extracts the main path according to the influence flow value of each edge when considering the connectivity Hummon (1989) proposed. The influence flow through the edge in citation network based on PageRank algorithm can be described as: the greater the influence of the citing node, the greater the influence of the cited node; the higher the number of times the node is cited, the greater the influence of the node; the more the outdegree of the nodes, the smaller the influence of each edge owns. The above relationship is described by the citation network constructed in Figure 3-4: When node *D* cites *A*, *D* transmits a certain proportion of influence to *A*, the influence flows on edge (D, A) is proportional to the influence of node *D*, and inversely proportional to outdegree of node D. That is, the more important node *D* is, the more important nodes *A* and *B* are cited by node D. If it cites too many nodes, each branch will get less influence flow and each cited node will get less influence.

4.1.3 Methods

Step 1: The patent citation network in the field of seawater desalination and the scientific citation network in the field of information security constructed in the chapter 3 are used as the analysis objects.

Step 2: Calculate the influence value *IF* of each node in citation network. According to PageRank algorithm, *IF* values of all nodes in citation network are calculated iteratively, no matter the initial influence value in the patent citation network, the *IF* value will be iteratively calculated and will converge to a stable value. Five different damping factors 0.5, 0.6, 0.7, 0.8 and 0.9 are used in the experiments to verify which one is more suitable. It is found that the nodes in the main path corresponding to different damping factors are completely coincident. Therefore, the damping factors of this algorithm are not discussed much in this chapter, and the general damping factor q = 0.85 is still selected to calculate.

Step 3: Calculate the value $T(ij)$ of all directed edges according to the influence value *IF* of each node obtained in step 2 in two citation networks. Which means the influence flow transmitted from the citing node *i* to the cited node *j* through the link. The influence value of a node flows to its cited nodes is considered to be equally divided into multiple referenced nodes, indicating that the influence value citing node flows to the cited node have no difference. Given $G(V,E)$ to describe the citation network, the influence flow value from node *i* to node *j* and the total influence the node *j* get from all its citing nodes can be expressed as:

$$T(i,j) = q * \frac{IF(i)}{L(i)} + \frac{1-q}{N} \tag{4-5}$$

$$T(j) = \sum\nolimits_{j \in V} T(i,j) = \sum\nolimits_{j \in V} (q * \frac{IF(i)}{L(i)} + \frac{1-q}{N}) \tag{4-6}$$

$L(i)$ represents the number of references of node *i*. The influence flow process is shown in Fig 3-4. The relationship between the twelve nodes in the figure is regarded as the voting process: the process of node *J* cites node *F*, *G*, *H* can be regarded as node *J* votes only one vote to *F*, *G*, *H*. Since there is only one vote *J* owns, *F*, *G*, and *H* can get $\frac{1}{3}$ votes from node *J* respectively. At the same time, node F gets $\frac{1}{2}$ votes cast by another citing node *I*, node *G* gets another 1 vote voted by node *K* and $\frac{1}{2}$ vote of node *L*, node *H* gets $\frac{1}{2}$ votes from *L*. Considering the damping factor:
$T(J,F) = T(J,G) = T(J,H) = q * \frac{1}{3} IF(J) + \frac{1-q}{N}$, $T(F) = q * \left(\frac{1}{3} IF(J) + \frac{1}{2} IF(I\ )\right) + \frac{1-q}{N}$, $T(G) = q * \left(\frac{1}{3} IF(J) + IF(K) + \frac{1}{2} IF(L\ )\right) + \frac{1-q}{N}$, $T(H) = q * \left(\frac{1}{3} IF(J) + \frac{1}{2} IF(L\ )\ \right) + \frac{1-q}{N}$

Through analogy, the $T(ij)$ of each edge can be obtained as shown in Figure 4-1.

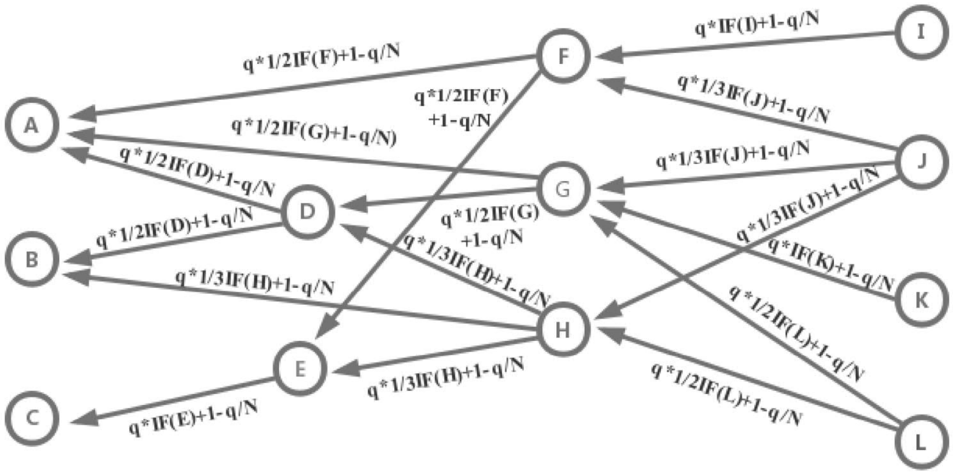

Figure 4-1 The influence flow process in citation network based on PageRank

Step 4: Search for the main path. The main path in two research field are obtained by filtering out the path with the largest edge weight sum through the global search method.

4.1.4 Results analysis

1）Technological main path based on PageRank

The experimental results of the technological main path map in the field of seawater desalination using the influence flow algorithm based on PageRank are shown in Figure 4-2. The ordinate of Figure 4-2 indicates the patent application year and the abscissa in the figure has no special meaning and the node size indicates the cited times. The thickness of the edges represents the weight which means the influence flow value.

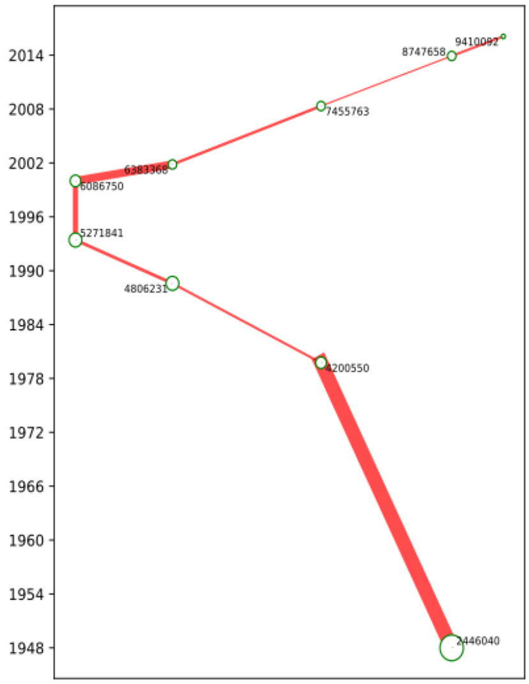

Figure 4-2 The main path of desalination based on the PageRank

Compare the main path results obtained in Figure 4-2 with the main path (Figure 3-5) used SPLC traversal algorithm in Chapter 3. From the path form, it can be seen that the earliest extracted nodes occurred in the main path based on the influence flow algorithm is earlier than then the earliest node through SPLC. The earliest node on the technological main path can be traced back to 1948 which is 40 years earlier when it corresponds to the year 1978 in Figure 3-5. By analyzing the thickest path (4200550, 2446040) and (5366635, 5186822) from Figure 4-2 and Figure 3-5, we can discover that the time span of the former patent pair is larger than the latter. The main content of two patents 4200550 and 2446040 are both about the desalination process from mineral oil, while the other two patents 5366635, 5186822 are related to the equipment such as turbines, which showed that the two experimental algorithms had different bias in the important evolution of the main technology path.

Gross (1927) first proposed that the citation frequency could be used to evaluate the importance of scientific research results, it was widely used to reflect the recognition value of academic papers (Peritz, 1992; Hirsch, 2005) and gradually became one of the most commonly used indicators for evaluating the importance of nodes in scientometrics. Virgo (1977) verified that the size of the citation frequency is positively correlated with the importance of scientific research. Also, the study by Mood et al. (2002) also showed that the higher the cited times is, the greater the value of academic paper is. Therefore, we take the citation frequency as a comparative indicator. Through the comparative analysis of the patent cited times in the main path, it can be concluded that the average cited frequency of the patent searched by the SPLC algorithm is higher than that based on the influence flow algorithm from Table 4-2. But the cited times of the patent is not all lower than the patent extracted by the SPLC, such as the patent 2446040, its cited times is only lower than two patents in the main path through SPLC which indicates that patent 2446040 is omitted to a large extent in the main path based on SPLC. Moreover, it also shows that the nodes on the technological main path based on traversal counts are not the highest in the whole patent network from the point of view of citation frequency. Therefore, the influence flow algorithm based on PageRank add a certain of key nodes to the result obtained by the traditional traversal counts methods.

From the comparative analysis of patent content on two main paths, the patent content of the main path based on the influence flow algorithm based on PageRank is more focused on the chemical process preparation in early stage. The source node 2446040 patent focused on the process preparation of desalination from crude oil, mainly on the process of dissolving inorganic salt from crude oil, and the medium-term desalination technology evolved into a more complex process preparation method. For example, the 4806231 patent applies for the washing technique of desalting at a higher temperature and a larger crude oil ratio and patent 5271841 applied for benzene removal.

The nodes on main path are mainly related to the technology of the new device in later stage, patents 8747658 and 9410092 all involve the use of a separator with a stacked disk centrifuge to separate the emulsified oil and water. The content of nodes in the main path focus on the different aspect in early stage based on SPLC. The analysis leads to the earlier technology that is more biased towards the physical aspect. The desalination techniques involved in the early patents 4110172, 4210494 and 4363803 of the main path are all related to the use of solar or solar radiation for the evaporation of brine technology. In the intermediate stage, the patented technology is biased towards equipment installations, desalination equipment with drive pressure for seawater desalination are involved in patent 5186822 and patent 6348148. Later patents in the main path of technology adopt more professional and advanced equipment and methods, which improve and integrate existing technologies for evolution.

In general, the technological main path obtained by the influence flow algorithm based on PageRank can extract earlier patented technologies, which can be traced back to the root node of the technology in the desalination field. And complements the key nodes missing from the main path obtained by the traversal counts algorithm SPLC to a certain extent. It can also be found that the evolution of the theme of the two algorithms in extracting the main path is different, the content based on the SPLC method is more concentrated which differentiation and evolution effect of technology is less obvious than the main path searched by the influence flow algorithm based on PageRank.

Table 4-2 Key nodes and their cited times in the main path of desalination based on PageRank and SPLC (sorted by cited times)

Patent ID	Cited	PageRank	SPLC
5186822	64		√
4363703	54		√
2446040	30	√	
5366635	25		√
6348148	17		√
4110172	16		√
4210494	15		√
7416666	14		√
4200550	12	√	
5916441	11		√
5271841	9	√	
4806231	9	√	
7749386	7		√
7455763	3	√	
6086750	3	√	
8747658	2	√	
8206589	2		√
6383368	2	√	
8696908	1		√
9133048	0		√
9410092	0	√	

2）Scientific main path based on PageRank

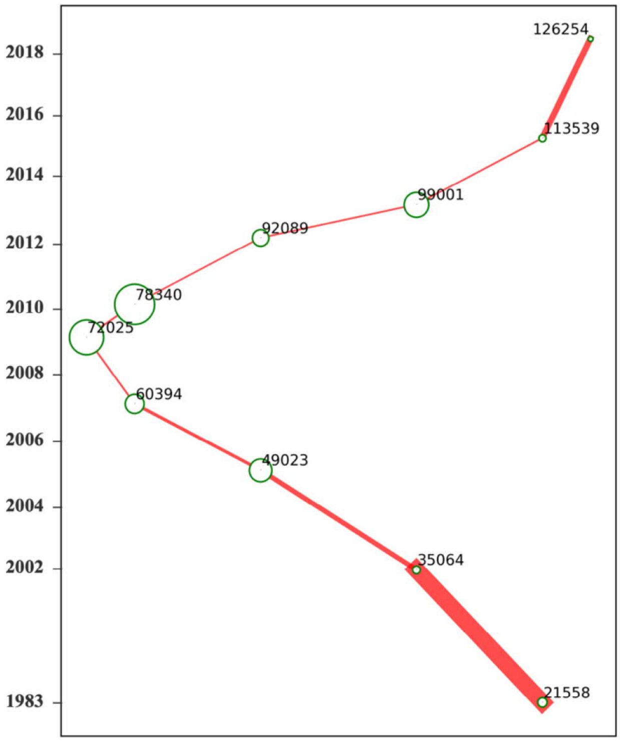

Figure 4-3 Main path of information security field based on PageRank

From the analysis of the main path form in Figure 4-3, it can be concluded that the scientific main path based on influence flow algorithm based on PageRank also can be traced back to earlier node 21558 which publish by DB Parker in 1983 compared with Figure 3-6 in Chapter 3 based on SPLC. Therefore, we can get the earlier research direction and theme in the field of information security through the proposed algorithm.

From the analysis of cited frequency in Table 4-3, although the average cited times of nodes extracted from the main path based on influence flow is lower than the average cited frequency of nodes based on SPLC, it can be seen intuitively that there are two papers in which the cited frequency of key nodes based on SPLC is lower than 10, which are node 122991 and node 123482. While there is only one node 126254 whose cited time are lower than 10 extracted based on the influence flow algorithm in 2018. Since the cited frequency of the recent documents are relatively low compared with other nodes because of the shorter published time, it is difficult to judge whether the latest node can become the key node in the whole field. Therefore, the latest literature doesn't necessarily facilitate the analysis of the accuracy of the main path.

From the analysis of the content of the main path, there are three coincidence points in the method of extracting the main path by the two methods, namely 78340, 92089, 99001. The main content of the nodes 78340, 92089 is to use the survey data to construct the corresponding structural equation model analysis employee's intention to comply with the ISP (Information Security Policy). And the main purpose of node 99001 is to forecast a series of potential crises and thematic directions in the field of information security in the future according to the research on information security at that time. Taking these three nodes as segments, first analyze the node 72025 which earlier than the node 78340 in Figure 4-3. The main content of this node is very similar to the 78340 for analyzing a series of hypotheses such as employee's attitude to information security policy on the basis of building structural equation model. Statistical result shows that the coupling strength of nodes 78340 and 72025 is 14 which means a high tightness. Then we discover that the main content of the previous node 72679 of node 78340 in Figure 3-6 is the cross-cultural difference between Korea and the United States in protecting the behavior of information technology users also based on structural equation model. But the coupling strength of these two documents is only 2 compared to 14. By comparing the coupling strength between the overlapping node 78340 and the previous node in the two main paths, it can be concluded that the relationship between the 72025 literature and the node 78340 in Figure 4-3 is closer, which indicates that the node based on influence flow algorithm has a higher degree of tightness before the stage from different technological development process to overlapping nodes. Comparing and analyzing the contents of the latest overlapping node after 99001,

Node 113539 in Fig 4-3 analyzed the evasive behavior of the students in Australian universities in different scenarios and extended the theoretical model of protection motivation. The main direction of this paper belongs to one of five thematic directions of future information security summarized in 99001 and their coupling strength is 5. Node 112886 which is after node 99001 in Fig 3-6 proposed to use the accountability theory to develop four UI design components to improve the user's accountability perception in the system. This paper aimed at reducing the violation of access policy and maintaining information security from the hardware point of view. The coupling strength between node 112886 and 99001 is 4. Therefore, from the literature content analysis of the front and back nodes of overlapping nodes, the similarity of the two documents before and after the coincidence of the main path extracted based on the influence flow algorithm is higher than that of the SPLC algorithm.

From the overall evolution of the path research theme, the main path extracted based on the influence flow reveals the information security from detecting and predicting the potential internal risk of the company (Node 35064), to the analysis of end-user behavior security (Node 49023 and 60394), developing to the construction of model analysis of employees' attitudes towards information security policies (Node 72025, 78340 and 92089) and final to the prediction of future information security topics and the importance of predictive related topics (Node 99001 and 113539). The theme of the entire main path evolved closely and it is possible to visually analyze the process of inheriting and evolving process between the cited nodes and citing nodes. The main research theme of the main path based on SPLC has evolved from the comprehensive social contract theory

(Node 13806) to the research on personal privacy protection in the information technology era (Node 15361, 17085, 24200, 38868 and 43720). Then extended to the construction of theoretical models to verify the relationship between Internet privacy and e-commerce (Node 54596 and 55474), developing to the latest research on the protection and connection of information security among internal employees in organizations (Node 113206, 122911 and 123482).The evolution process of the research themes based on two methods is different before and after the coincident three nodes, indicating that the core idea of the algorithms of weighting edges are different, which cause the evolution process of the main path is different. But there is no doubt both methods reveal the development process of the information security field with the times from different perspectives. It is worth mentioning that there is no review literature in the main path extracted based influence flow algorithm, which shows the evolution process of the theme more subtly compared to that based on SPLC.

Due to the differences in disciplines, the evolution main path obtained by the same algorithm in different fields will have different biases and different algorithmic will also generate different main paths in the same field in general. There are three common characteristics between the main paths in the field of desalination and the information security based on influence algorithm based on PageRank. The first common characteristic is that the main path extracted can be traced back to the earlier node and the second is that it can supply the important nodes ignored by SPLC. The last common characteristic is that the evolutionary process of the extracted main path is more obvious, especially it can avoid the emergence of the review literature in the main path in the scientific citation network.

Table 4-3 Key nodes and their cited times in the main path of information security based on PageRank and SPLC (sorted by cited times)

Order	Cited	Year	PageRank	SPLC
17085	260	1996		√
43720	250	2004		√
78340	171	2010	√	√
54596	164	2006		√
24200	159	1999		√
72025	143	2009	√	
99001	96	2013	√	√
49023	86	2005	√	
38868	86	2003		√
60394	68	2007	√	
92089	56	2012	√	√
15361	56	1995		√
112888	40	2015		√
72679	28	2009		√
21558	21	1983	√	
112886	19	2015		√
113206	19	2015		√
13806	18	1994		√
55474	17	2006		√
35064	13	1998	√	
113539	10	2015	√	
122911	8	2017		√
123482	2	2017		√
126254	0	2018	√	

4.2 Influence flow algorithm based on single traversal weighting path

4.2.1 Problem statement

Because of that the PageRank algorithm is used in the network with loop at first. Therefore, the random walk process of the PageRank lack of the substantial meaning for those nodes without citation compared to the web network. How to construct a more direct algorithm for directed acyclic networks is the starting idea in this section. Moreover, we can find that it needs 50 times to converge the PR value of each node to a stable value, and then calculate the size of the influence force flow transmission of each side in a small network only composed of 5 nodes shown in chapter 4.1. Therefore, the influence flow algorithm based on PageRank will takes longer calculation time in large citation network which is inefficient. In the process of influence flow transmission, the influence value of each node only comes from its accumulated influence from its citing node, but doesn't include its own creative influence, which will affect the weight value of the edge to a certain extent in citation network.

This section proposes a new algorithm called influence flow algorithm based on single traversal from the perspective of the applicability of directed acyclic network and the purpose to increase the computational efficiency of weighting edge in large-scale network. The algorithm will set the initial influence value of each node and then calculate the weight of all edges in citation network by single traversal.

4.2.2 Algorithm description

In the citation network, each node is regarded as an independent node when it is not connected with any other nodes. In this case, the initial influence value of each node can be regarded as 1, which means that the scientific value created by the literature is equal without citation. But in the actual situation of citation network, there will be links between the documents, and the citation relationship between nodes in the citation network will transmit the influence value from the citing node to the cited node. Therefore, the magnitude of the influence value of each node should be calculated as the sum of its initial influence and all the influence flow value transmitted by all the citing nodes as the final influence of the node in the citation network.

The distribution idea between the influence flow algorithm based on single traversal and PageRank is same. In the process of transmitting influence value of citing nodes, each citing node is regarded as distributing the influence value equally to all the cited nodes. The more important the citing node is, the more important the node it cites to is which is the core idea of eigenvector centrality in citation network.

Consider the links between nodes as the flow path of node influence, and describe the above algorithmic idea using Figure 3-4. The initial influence value of each node is 1, and the four sinks *I*, *J*, *K*, and *L* all distribute their influence value 1 equally to all references in the process of transmitting influence. When *F*, *G* and *H* receive the influence value from the citing documents, they will add it to the initial influence value as their final influence value then continuing pass the final influence to their cited nodes until all the sources obtained the influence and all edges are assigned the corresponding weight of the influence flow value.

The influence flow algorithm based on PageRank needs to introduce damping factor to avoid traps with PR values of 0 for all nodes after multiple iterations of independent nodes which have no node to transmit influence. This problem is effectively avoided by assigning the original value of each node to 1 in the single traversal algorithm in this section. This algorithm only needs a single traversal process to make all nodes in the citation network get the final influence value and then calculate the influence flow value of each edge. The complexity of the single traversal is $O(E)$, *E* is the number of edges in the citation network.

4.2.3 Methods

Step 1: The patent citation network in the field of desalination and the scientific citation network in the field of information security constructed in the chapter 3 are also used as the analysis objects in this section.

Step 2: The initial influence of each node in the citation network constructed is assigned to 1 and set each of the citing nodes equally transmits its own influence value to the cited nodes.

Step 3: Calculate the final influence value *IF2* of each node in the constructed citation network.

Step 4: Calculate the value $T2(ij)$ of all directed edges according to the influence value *IF2* of each node obtained in step 3 in two citation networks. The sum of the influence flows passed from the citing node *i* to the cited node *j* and the influence values *j* obtained can be expressed as:

$$T2(i,j) = \frac{IF2(i)}{L2(i)} \tag{4-7}$$

$$T2(j) = \sum\nolimits_{j\in V} T2(i,j) = \sum\nolimits_{j\in V} \frac{IF2(i)}{L2(i)} \tag{4-8}$$

$L2(i)$ represents the number of references of node *i*. If the influence flow transmission process of the 12 nodes of the citation network in Figure 3-4 is regarded as the voting process: each node initially has one vote, but the process only start from the sinks *I*, *J*, *K*, and *L* and vote to their previous layer, namely their cited nodes. It can be calculated that node *F* not only gets one vote of *I*, but also one third of the vote from node *J*. Then $IF2(F) = 1 + T2(IF) + T2(JF) = 1 + IF2(I) + \frac{1}{3}IF2(J)$, and $\mathrm{T2}(FA) = \frac{1}{3}IF2(F)$. Similarly, the influence flow value $T2(ij)$ of each edge is shown in Figure 4-4 and Table 4-4:

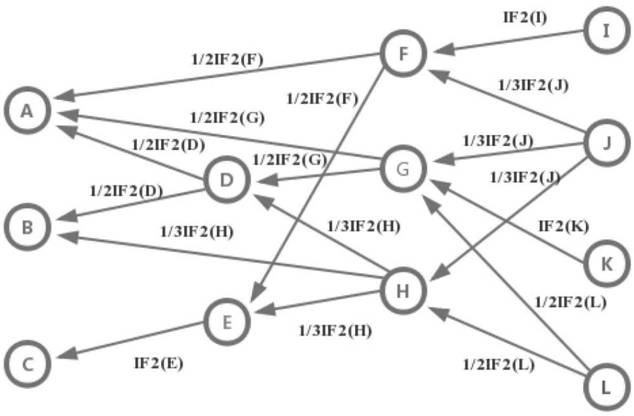

Figure 4-4 The influence flow in the citation network based on single traversal

Step 5: Search the main path using the global search. Compare the main path results between the two algorithms in this chapter.

Table 4-4 The IF2 and accepted influence flow value of each node in Figure 4-4

Node	**A**	**B**	**C**	**D**	**E**	**F**
Original IF2 value	1	1	1	1	1	1
Accepted influence flow value	4.1	2.13	2.78	2.03	1.78	1.33
Sum	5.1	3.13	3.798	3.03	2.78	2.33
Node	G	H	I	J	K	L
Original IF2 value	1	1	1	1	1	1
Accepted influence flow value	1.83	0.83	0	0	0	0
Sum	2.83	1.83	1	1	1	1

4.2.4 Results analysis

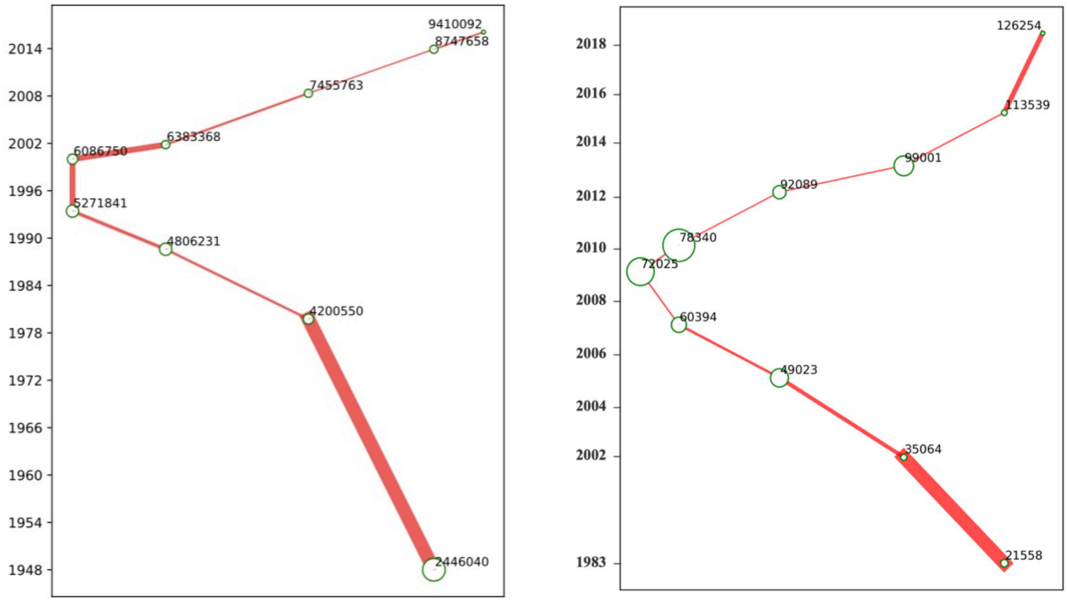

Figure 4-5 The desalination technological main path (left) and the information security scientific main path (right)

Comparing the left and right main paths in Figure 4-5 with those in Figure 4-2 and Figure 4-3 based on PageRank algorithm in Chapter 4.1. We can find that the main paths obtained by the influence algorithm based on the single traversal in the fields of seawater desalination and information security are the same as the extracted nodes in the main path based on PageRank algorithm. From the point of view of the content and evolution of the main path, the main path obtained by the single traversal algorithm has the all the content advantages the PageRank algorithm owns compared to SPLC. From the point of applicability, the single traversal algorithm is only used in the directed acyclic network, so it is suitable for the calculation of the main path edge weight in the citation network. From the computational complexity analysis, the single traversal algorithm greatly improves the computational efficiency compared to the algorithm based on PageRank.

4.3 Summary

The influence flowing processes of the two algorithms proposed in this chapter are both based on the idea of equal distribution for the transmission of influence from the citing nodes to the cited nodes. Comparing the two proposed algorithms, it can be clearly found that the main path obtained by the single traversal algorithm and the main path based on the PageRank algorithm do not have any difference in the rendering results. The difference of the algorithm is that compared with the algorithm based on PageRank, the single traversal algorithm can only be applied in the network with non-cyclic network structure, such as citation network that the node generally represents literature or patent. Therefore, it can’t be applied to the network with loop in it which means this network may exist the occurrence of self-citation. Moreover, the computational complexity of a single traversal is much less than that of the influence flow algorithm based on PageRank.

Compared with the classic SPLC algorithm in Chapter 3, the two equally-divided influence flow algorithms provide a new perspective to weight the edge in citation network not based on the traversal count but ensure the overall connectivity of the main path and discover the different development of the technology or scientific themes. The main path results of two proposed algorithms have the following characteristics compared with SPLC:

1) Although the main path extracted by traditional SPx algorithm can extract the main path including the key nodes, a single main path will ignore a certain of important nodes that have not been identified. The algorithm in this chapter starts from the angle of equally distributing the influence of edge, which provides a new way to analyze the evolution process of the main path and in the search main path, there are also a number of nodes with higher cited times than the SPX derived which compensate for the key nodes ignored by the traversal counts algorithm to a certain extent.

2) The two proposed algorithms in this chapter can be traced back to the earlier nodes. The evolution of the main path has a longer time span, which can help researchers to analyze the early key technologies or research achievements in specific fields.

3) The scientific citation path searched based on the equally-divided influence flow can avoid the literature review appearing in the main path. Compared with the content of the main path node obtained by SPLC algorithm, the content of the main path node is more objective and reveals the evolution process of scientific topics is more detailed.

However, there are still some shortcomings of these two proposed equally-divided influence flow algorithms. According to the nodes’ sizes in the main paths in sections 4.1 and 4.2, it can be seen that the number of high cited nodes identified by two proposed algorithms is less than that of SPLC.

5 UNEQUALLY-DIVIDED INFLUENCE FLOW ALGORITHMS BASED ON COUPLNG STRENGTH WEIGHTING PATH FOR MAIN PATH ANALYSIS

5.1 Problem statement

In the process of flowing influence from the citing nodes to the cited node refereed in chapter 4, it is concluded that the more important the citing node is, the more important the cited node is. Although this idea reflects the tightness between the nodes, knowledge inherited between different references and the citing node is not the same. Also, as we mentioned in the research gap, the neglected initial weight of the edge will cause some deviation for the main path analysis.

Due to the different close relationship between various cited documents and citing nodes, the value of the influence flow of all the citing nodes passed to the cited node can't be considered equal. Therefore, from the perspective of the different tightness between the citing nodes and the cited nodes, the influence flow values transmitted from the citing nodes are considered unequally allocated to the references based on the coupling strength between them, then calculate the influence value and corresponding influence flow value of each edge in citation network.

5.2 Algorithm descriptive

We have already verified the applicability and computational simplicity of the single traversal algorithm in the citation network in chapter 4.2. Therefore, we still adopt the core idea of the single traversal algorithm in this chapter, the initial influence value of each node is set to 1 firstly, and then calculate the cumulative influence value after the process of flowing and get the influence flow value of each edge.

Because the relationship between different citations is different, the influence obtained by different cited nodes from the same citing node will be different in the process of transmitting the influence. We can assign the influence value according to the coupling strength between the two nodes in the citation network. Higher the coupling strength indicates more same reference node pair own and closer relationship between them. Suppose that node A cites two node B and C, B has the higher coupling strength with A than that of B, we can assume that A will transmit more influence value to B because the cited node B may support more helpful theoretical knowledge which means more knowledge diffusion from B to A compare to C to A. Therefore, two nodes with higher coupling strength, the greater the influence value transmitted by the citing node.

The citing node transmits its influence value to its references according to their tightness relationship in proportion. The influence transmitted by citing nodes is directly proportional to their own influence, inversely proportional to their numbers of references, and directly proportional to the coupling strength. We can describe the flowing process of the algorithm using Figure 3-4: select 3 node pairs $(\boldsymbol{H},\boldsymbol{B})$、$(\boldsymbol{H},\boldsymbol{D})$ and $(\boldsymbol{H},\boldsymbol{E})$, if the importance of node *H* is huge, the importance of cited nodes *B*, *D* and *E* are also big, but if the node *H* have more than three cited nodes, each branch will get fewer influence than the references are three. The influence of transmission is proportional to the coupling strength between nodes. In Figure 3-4, the coupling strength of node pair $(\boldsymbol{H},\boldsymbol{D})$ is 1, and the other two node pairs are both 0. Therefore, citing node *H* will transmit more influence through the link connected to *D*, $(\boldsymbol{H},\boldsymbol{D})$ accounts largest proportion of the influence flow. There is a certain number of node pairs in citation network where there is no coupling relationship between the citing node and the cited node just like node pairs $(\boldsymbol{H},\boldsymbol{B})$ and $(\boldsymbol{H},\boldsymbol{E})$. However, the cited node with 0 coupling strength with the citing node will not be passed any influence value if the allocation of the influence value just based on the absolute coupling strength ratio which is contrary to the fact in the real citation network.

In order to solve the above problem, the algorithm will not adopt the original coupling strength as the basis when transmitting the influence value. First, the original coupling strength value between each pair of nodes is added to 1, and then the influence flow is transferred according to the ratio of coupling strength.

5.3 Data and methods

5.3.1 Data

In order to analyze the difference and the applicability between the main path identified by the influence flow algorithm based on coupling strength and the classical SPX algorithm in citation network, we choose a classical DNA citation network composed of 40 landmark nodes (Appendix 1), which was constructed by Garfiled, Sher and Torpie (1964) and summarized by Garfield (1979), the 40 nodes are all documents that have great influence on DNA technology in 1820-1964. Although the main path based on the traditional traversal counts algorithm can reflect the overall connectivity of the network, the special citation network composed of all such important influential nodes lacks the reflection of the close relationship between the citing nodes and the cited nodes which is also one of the development paths that relevant researchers need to analyze.

5.3.2 Methods

Step 1: Set the initial influence 1 of each node in DNA citation network.

Step 2: Calculate the influence value *CIF* of each node in DNA citation network. Each citing node proportionally transfers its own influence value according to the coupling strength between the node pairs, until the influence value reaches all the sources in the citation network.

Step 3: Calculate the influence flow value $\boldsymbol{CT(i,j)}$ of each edge based on the *CIF* of each node. $\boldsymbol{CT(i,j)}$ indicated the how much influence flow from node *i* to node *j*. The $\boldsymbol{CT(i,j)}$ and the total influence node *j* obtained by all its citing nodes can be expressed as:

$$CT(i,j) = \frac{CS_{(i,j)} + 1}{\sum_{i,j\in V}(CS_{(i,j)} + 1)} * CIF(i) \tag{5-1}$$

$$CT(j) = \sum\nolimits_{i,j\in V} T(i,j) = \sum\nolimits_{j\in V}\left(\frac{CS_{(i,j)} + 1}{\sum_{j\in V}(CS_{(i,j)} + 1)} * CIF(i)\right) \tag{5-2}$$

$\boldsymbol{CS_{(i,j)}}$ represents the coupling strength between node *i* and node *j*. Take the Figure 3-4 as an example, the transfer influence process and the value of the *CIF* were shown in the Figure 5-1 and Table 5-1.

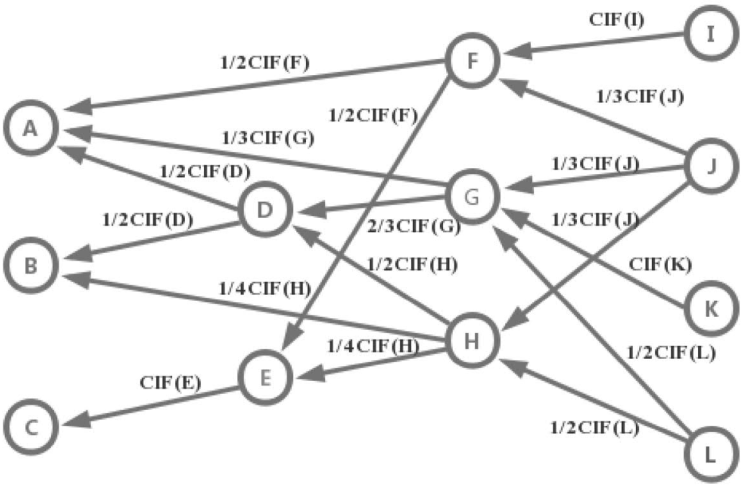

Figure 5-1 The influence flow process in citation network based on the coupling strength

Step 4: Search the main path. We adopt the local search method to identify the main path based on the coupling strength algorithm because the control group we take also used the local search method.

Table 5-1 The CIF and accepted influence flow value of each node in figure 15

Node	A	B	C	D	E	F
Original CIF value	1	1	1	1	1	1
Accepted influence flow value	4	2.36	2.62	2.80	1.62	1.33
Sum	5	3.36	3.62	3.80	2.62	2.33
Node	G	H	I	J	K	L
Original CIF value	1	1	1	1	1	1
Accepted influence flow value	1.83	0.83	0	0	0	0
Sum	2.83	1.83	1	1	1	1

5.4 Results analysis

Hummon and Doreain weight each edge in the DNA citation network based on their proposed traversal counting algorithm, and then use the local search method to identify the main path in the network. The main idea of local search is to use priority first search to identify the path with the largest weight in the next search paths staring from the source node until the sink, all the searched paths form the main path in the DNA network. Searching from from the source node 3 with the largest subgraph, Figure 5-2 clearly shows that the main path based on traversal counts is 3→5→12→20→21→22→27→ 32→36→40. The source of the main path is node 3, and the publication time is 1869 which is mainly about isolation of nucleic acid. The technology of DNA field developed to study of purine and pyrimidine content of nucleic acid (Node 5) in 1886. Levene discovered certain nucleic acids containing DNA (Node 12) in 1929. And the next node along the path is 20, Avert et al. discovered DNA carried genetic information that can change a strain into another in 1944. Nodes 21 and 22 are both from Chargaff's research on purines and pyrimidine present in unequal quantities within nucleic acids and on different nucleotides in the chain are in random order in 1947 and 1950 respectively. In the 1950s, Watson and Crick constructed model of spatial molecular configuration of DNA (Node 27), plus Ochoa isolated a bacterial enzyme producing polynucleotide strands of RNA in 1955（Node 32）. Hurwitz manufactured messenger RNA in test tube (from DNA, nucleotides, enzymes) in 1963. The latest node in the main path is 40 which is about the leverage of the verification of triplet code.

The main path extracted based on the coupling strength (Figure 5-3) is also based on the local search method used in the Hummon and Doreian paper. The main path searched from the source node 3 is 3→5→12→15→29→32→33→35→39. It can be found that there exist 4 coincident nodes with the main path in Figure 5-2, the first three nodes of the two main paths are both 3, 5 and 12, and there is also a coincident node 32 occurred in the middle of the main path. Then we analyze the discoveries of these remaining nodes which are not the coincident with the other main path used the influence flow based on the coupling strength: Levene proposed formulae assigning linkages between the nucleotides in 1935(Node 15). Then Todd confirmed Levene's formulae through chemical synthsis (Node 29) in 1955, which means that the output of Levene had an extremely influence for next 20 years research in DNA field. Kornberg produced synthetic polynucleotides of RNA from an enzyme (Node 33). Node 35 is a study of the existence of second RNA by Jacob and Monod in 1960. The sink node 39 of the main path was discovered by Mirsky and Albrey in 1962 about messenger RNA isolated from mammalian cells.

The extraction of these forty nodes in the DNA network is firstly the 65 specific research productions listed in the book by Asimov (1963), and then 40 milestones reorganized by Garfield (1964) from 65 paper. Therefore, these 40 nodes are 40 milestones in the DNA network from 1820 to 1962 and each node has far-reaching influence in stimulating the development of DNA. Any connected path from node 3 to tail node in the subgraph of maximum connection strength has great influence on subsequent research in the DNA field in the DNA theory network composed of these 40 critical nodes. For a citation network with special properties for this type of node, searching only one main path lacks substantial meaning for exploring development in a particular domain. Although the algorithm based on traversal counts has verified the applicability in many fields from the perspective of overall connectivity, and the corresponding results of the main paths show that it can better analyze the development of themes or technologies in the specific field. However, it is equally important for relevant researchers to analyze the evolutionary relationship between the important nodes based on the influence and the strong tightness of the nodes to discover the development of the theme or technology in such networks where each node has a strong influence.

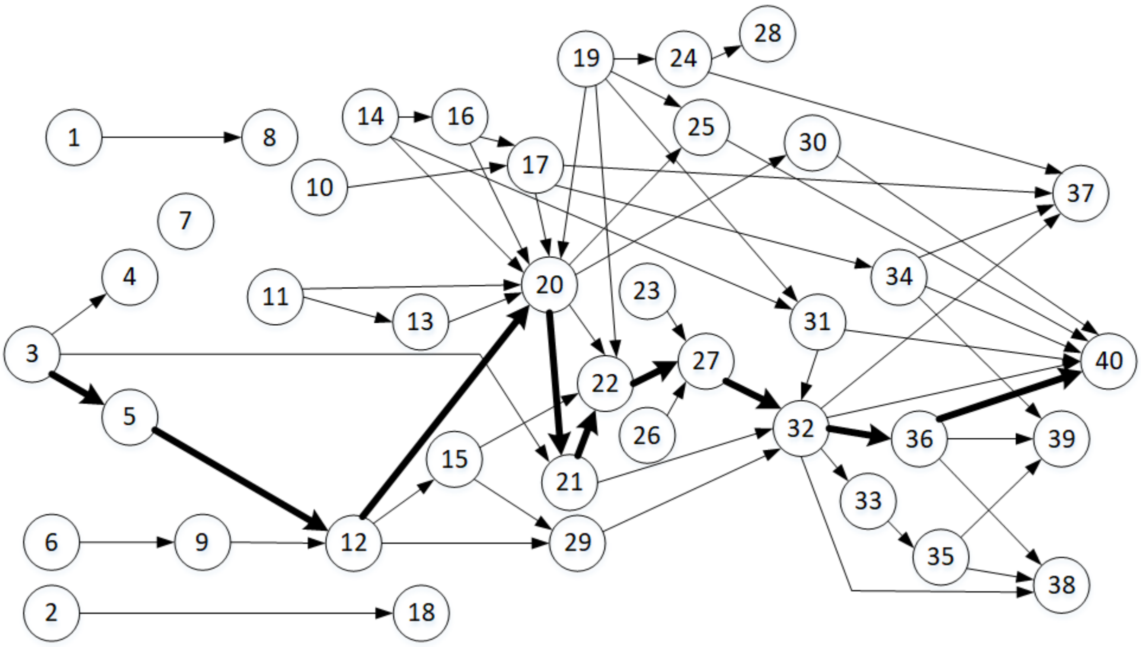

Figure 5-2 DNA main path based on the SPLC

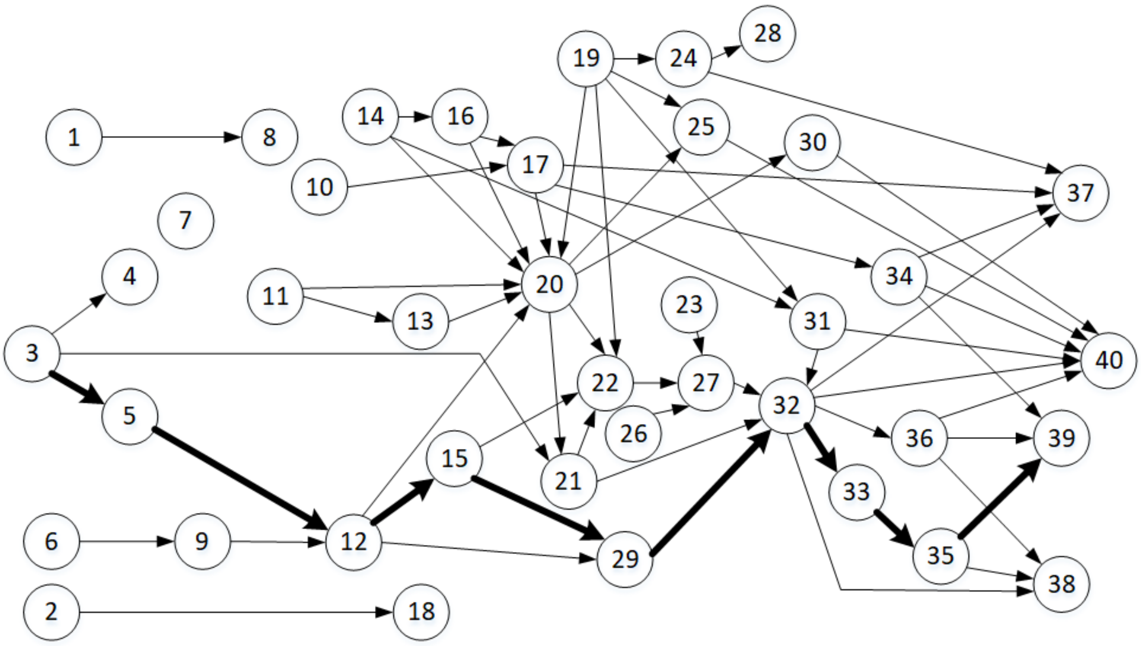

Figure 5-3 DNA main path based on the coupling strength

5.5 Summary

According to the different attributes of the citation network, there will be multiple research purposes which have a certain degree of difference. The traditional traversal counts used for the main path analysis often neglects the practical significance of the node in different citation networks because of its simplicity. Through the core idea of the coupling, the method of identifying the main path based on the coupling strength is more suitable for the citation network that needs to reveal a more refined evolution structure. This new approach can help researchers study closer and deeper relationship between nodes especially in some special networks with the high importance of all nodes. From the comparison of the results in Figure 5-2 and Figure 5-3, it can be found that the main path weighting based on the coupling strength and the path searched by the traversal counts are different in the given DNA citation network, the new main path guanartees both the influence flow and the largest coupling strength.

6 CONCLUSIONS

6.1 Theoretical implication

Main path analysis is based on the connectivity of graphs which aims to find a simple and visible trajectory from the complex and large citation network to reveal the evolution process of the specific domain, which has become an important method of bibliometrics. In this paper, the concept of the influence flow is proposed based on the eigenvector centrality of the node in citation networks. The core idea of eigenvector centrality shown in the citation network is that the node cited by the influential node also has high influence because there will be influence flow from the citing nodes to the cited nodes through the edges between node pairs.

The main purpose of this paper is to construct the effective weighting algorithms for identifying the main path which is not based on the traversal counts. Based on the influence flow idea, the method of equally-distributed influence flow algorithm weighting for the edges generate the main path in the desalination field and the information security field respectively in chapter 4. The main paths based on PageRank and single traversal are totally the same, but the application scope of the two proposed algorithms is different. Single traversal algorithm is more suitable for large-scale citation networks, because its computational complexity is simpler and more efficient than PageRank-based algorithm, but the method of selecting main path based on PageRank algorithm can be applied not only to DAG such as citation networks, but also to other ring-structured networks, such as the web networks. Comparing the main path obtained from the equally-distributed influence flow traditional SPLC algorithm based on the traversal counts, the following four conclusions can be drawn:

1) From the morphological comparison, the main path obtained by the influence flow can be traced back to the earlier node in both two data sets, which can help the relevant researchers to analyze the specific field when analyzing evolution trends in specific fields to a certain extent.

2) From the comparative analysis of the cited frequency of the nodes in the main paths. The two methods of weighting the path by equally distributing the influence flow make up for the incompleteness of the nodes in the main path obtained by SPLC algorithm to a certain extent, and the two algorithms also supplement the key nodes that have not been searched by traditional traversal counts method.

3) Through the comparative analysis of node content in the main path, this paper finds that the results obtained by the influence flow algorithm in the information security field can effectively avoid the review literature nodes appearing in the main path compared to the main path extracted by the SPLC algorithm.

4) Through the analysis of the theme evolution content of the nodes in the main path, we can find that the technology or theme evolution process obtained by different algorithms will have different biases, and the evolution process of the main path obtained by using the same algorithm in different fields also will be different. Therefore, the algorithms based on influence flow idea can also provide another way to reveal the evolution of technologies or topics.

It can be judged from the above four points that the algorithm of the two equally-distributed influence flow weighting path can successfully identify the main path in the scientific citation network and the technological citation network compared with SPLC.

The influence flow algorithm based on the coupling strength proposed in this paper provides a new idea for revealing the close relationship between nodes in the main path. This method is used in the DNA network composed of 40 milestones events and get a different main path based on the traditional traversal counts algorithm. The extracted nodes searched not only takes into account the influence of nodes, but also the tightness between nodes. Analyzing the content of nodes in the main path provides a more refined evolutionary structure. The main path searched by this proposed method not only takes into account the influence of nodes, but also the tightness between nodes. It is for researchers. Analyzing the content of nodes in the main path provides a more refined evolutionary structure.

6.2 Limitations of the study

In the research experimental process in this paper, there are still the following limitations which are not considered.

Firstly, we compare the main paths obtained by different algorithms through the shape of the main path, the frequency of nodes being cited extracted, the content and the theme in the analysis of the experimental results. But these comparisons are not very comprehensive. For example, there are many indicators to evaluate the importance of a node, such as degree centrality, closeness centrality, betweeness centrality etc. This paper doesn’t compare and analyze the differences between different algorithms horizontally from more indicators to evaluate the importance of nodes. Similarly, in the patent citation network, the influence and importance of patent nodes are related to the commercial value of patents, patent age, patent litigation and other indicators expect the cited times examined in the experiments. Therefore, further research by researchers is needed for the influence of the searched nodes in the main path of science or technology. Secondly, thought the coupling strength method is used in DNA network and get the main path, this method is not used to verify the applicability of the method in large-scale citation network. Thirdly, this paper doesn’t combine the three proposed methods with traditional traversal counts method to search for more complete main paths in citation networks. The main path obtained by the influence flow algorithms the traditional main path algorithm is similar with a node number of about 10. Although the results show the most relevant path and also supplement the key nodes to the main path obtained by the traversal algorithm, there will still be other key nodes ignored if only this method is used to weight the path in the citation network. Lastly, the algorithms in Chapters 3 and 4 of this paper only use the global search method to search the main path. Although the local optimization problem is avoided, there may be some edges with the maximum weight value that are not searched in the main path.

6.3 Suggestions for future research

In view of the limitations of this paper and the current main path analysis research, this thesis will give four suggestions, including the applicability of main path analysis and the algorithms to weight the edges in citation networks.

1) The idea of algorithm needs to be more innovative and substantively expanded. In the current research, there are some problems such as single search target, immobilization of algorithms and so on, which will cause the intrinsic constraints to reveal the inaccurate evolution structure of the main path. The evolution process of technologies or topic is actually influenced by many factors in citation network, the main path search problem should be regarded as a multi-objective decision problem because the main path development of technologies or topic embodies the multi-objective optimization results of the influence of each stage. For example, the method of weighting the path based on the coupling strength is not only consider the influence of each edge but also consider the closeness between the nodes. Therefore, we should pay more attention to the different objectives in the evolution process when search the main path and study more on the model satisfying the multi-objective and the corresponding algorithm in the future.

2) Enhance the applicability and applicability of the main path algorithm. The traditional main path analysis method proposed by Hummon and Doreain is based on the exhaustive search of the path of network theory which will cause the high complex calculation. Therefore, it may take a long time to deal with the data set in large citation network when based on the exhaustive search idea.

3) Search for the main path more diverse. Most of the main paths searched by existing research methods are single main paths. Even the key-route search method can only ensure that the most important edge is extracted in the main paths. However, in addition to searching for the most important main path, relevant researchers may also want to trace the sub-critical paths as an aid to help them discover more elaborate disciplinary structures especially in the discipline with many sub-disciplines.

4) Combine the main path algorithm with other approaches. Because most existing main path methods will lose a lot of key information when searching the main path. Future main path analysis can be combined with other indicators which are used to evaluate the importance of nodes based on the improvement of their own algorithms to enrich and optimize the searched main path.

REFERENCES

Andersen B. (1998). The evolution of technological paths 1890- 1990. *Structural Change and Economic Dynamic*, 1998,4 (9):5-34.

Arrow K.J. (1972). Economic welfare and the allocation of resources for invention. In: Rowley C.K. (eds) *Readings in Industrial Economics*. Palgrave, London.

Asimov I. (1963). *The Genetic Code*. New York: New American Library.

Auber D. (2002). Using strahler numbers for real time visual exploration of huge graphs. *In: International Conference on Computer Vision and Graphics*, 10:1-3.

Batagelj V. (2003). Efficient algorithms for citation network analysis [J/OL]. arXiv, 2003: 0309023.

Bornmann L - Daniel H. (2006). What do citation counts measure? A review of studies onciting behavior. *Journal of Documentation*, 64(1):45-80.

Bornmann L. - Rüdiger Mutz. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11): 2215–2222.

Calero - Medina C - Noyons (2008). Combining mapping and citation network. analysis for a better understanding of the scientific development: the case of the absorptive capacity field. *Journal of Informetrics*, 2 (4): 272-279.

Carlero - Medina C – Noyons (2008). Combining mapping and citation. network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. *Journal of Informetrics*, 2(4): 272–279.

Carley KM - Hummon NP - Harty M. (1993). Scientific influence an analysis of the. main path structure in the journal of conflict resolution. *Science Communication*, 14(4): 417–447.

Chen C - Hicks D. (2004). Tracing knowledge diffusion. *Scientometrics*, 59(2): 199-211.

Chen L- Yang G.-Zhang J - Fan Y. (2015). Research on Multiple Main Paths Method Oriented to Analysis of Technological Evolution. *Library and Information Service*. 59(10): 124-130.

Choi C - Park Y. (2009). Monitoring the organic structure of technology based on the patent development paths. *Technological forecasting and social change*, 76(6): 754 - 768.

David Dreyfus - BalaIyer. (2005). Knowledge sharing and value flow in the software. industry: searching the patent citation network. Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii.

Delest M - Don A - Benois-Pineau J. (2006). Dag-based visual interfaces for navigation in indexed video content. *Multimedia Tools Appl,* 31(1):51–72.

Doreian Patrick. (1985). A measure of standing of journals in stratified networks. *Journal of the American Society for Information Science*, 8(5-6):341-363.

Dosi G. (1982). Technological paradigms and technological paths. *Research Policy*, 11(3): 147-162.

Duan Q.- Zhu D. - Wang X. (2012). Sorting method of citation. documents based on improved PageRank algorithm. *Information Studies: Theory & Application*, 1:115-119.

Diestel - Reinhard. (2005). “1.10 Other notions of graphs", Graph Theory (3rd. ed.) Springer, ISBN 978-3-540-26182-7.

Epicoco M. (2013). Knowledge patterns and sources of leadership: Mapping the. semiconductor miniaturization trajectory. *Research Policy*, 42(1):180–195.

Gagniuc PA. (2017). *Markov Chains: From Theory to Implementation and. Experimentation*. USA, NJ: John Wiley & Sons. pp. 1–235.

Garfield, E. Sher - R.J. Torpie. (1964). The use of citation data in writing the history of science. *Isis*, 56(4):487.

Garfield, E. (1965). Citation indexes for science. *Science*,123(3185): 61-21.

Garfield, E. (1970). Cition indexing for studying science. *Nature*, 227: 669-671.

Garfield, E. (1979). *Citation Indexing: Its theory and application in science, technology, and humanities*. PhiLAdELphia:Institute for Scientific Information Press.

Goffman W. (1966). Mathematical approach to the spread of scientific ideas-the. history of mast cell research. *Nature*, 212 (5061): 449.

Griliches Z. (1990). Patent statistics as economic indicators: a survey. *Journal of. Economic Literature*, 28: 1661-1707.

Gross P L - Gross E M. (1927). College libraries and chemical education. *Science*, 66(1713):385-389.

Halatechliyski - I.-Hecking T. - Goehnert, T.- Hoppe, H. U. (2014). Analyzing the Path. of Ideas and Activity of Contributors in an Open Learning Community. *Journal of Learning Analytics*, 1(2): 72–93.

Han Y - Jin B. (2012). A new perspective of citation network structure analysis. based on connectivity: main path analysis. *Studies in Science of Science*, 30(11):1634-640.

Hart C. (2018). *Doing a Literature Review: Releasing the Research Imagination*. SAGE Study Skills Series. SAGE. pp. xiii. ISBN 9781526423146.

Harris J.K. - Luke, D.A.- Zuckerman, R.B. - Shelton, S.C. (2009). Forty years of secondhand smoke research: The gap between discovery and delivery. *American Journal of Preventive Medicine*, 36(6): 538–548.

Hirsch J E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569-16572.

Ho MH - C - Lin VH - Liu JS. (2014). Exploring knowledge diffusion among nations: a study of core technologies in fuel cells. *Scientometrics*, 100(1):149–171.

Huenteler J - Schmidt TS - Ossenbrink J - Hoffmann VH. (2016). Technology life-cycles in the energy sector-Technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change*, 104:102–121.

Hu L. - Chen D. (2014). Visualization of Citation Analysis. *Journal of Information*, 11:78-81.

Hummon N.P. - Doreian P. (1989). Connectivity in a citation network: The. development of DNA theory. *Social networks*, 11(1): 39-63.

Hummon N.P. - Carley K. (1993). Social networks as normal science. *Social networks*, 15(1): 71-106.

Hummon N.P. - Doreian, P., - Freeman, L.C. (1990). Analyzing the structure of the centrality–productivity literature created between 1948 and 1979. *Science Communication*, 11(4), 459–480.

Hung SC - Liu JS - Lu LYY - Tseng YC. (2014). Technological change in lithium iron. phosphate battery: the key-route main path analysis. *Scientometric*, 100(1): 97-120.

Jahn M. (1972). *Changes with growth of the scientific literature of two biomedical. specialties*. Philadephia: Drexel University, MSthesis.

Jarneving B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of informetrics*, 1(4): 287-307.

John S. Liu - Louis Y. Y. Lu - Wen- Min Lu - Bruce J. Y. Lin. (2013). Data envelopment. analysis 1978–2010: A citation-based literature survey. *Omega*, 41(1): 3-15.

Jonathan C. Ho, Ewe - Chai Saw - Louis Y. Y. Lu - John S. Liu. (2014). Technological. barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Chang*, 82: 66-79.

John S. Liu-Louis Y.Y.Lu - Mei Hsiu-Ching Ho. (2019). A few notes on main path. analysis. *Scientometrics*, 119(1): 379-391.

James E-Kelley Jr - Morgan R - Walker. (1959). Critical-Path Planning and Scheduling. Proceedings of the Eastern Joint Computer Conference.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. A*merican Documentation*, 14, 10–25.

Liu, J.S. - Lu, L.Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the Association for Information Science and Technology*, 63(3): 528-542.

Liu-John S. - Chen- Hsiao-Hui-Ho- Mei Hsiu-Ching- Li- Yu-Chen. (2014). Citations with different levels of relevancy: Tracing the main paths of legal opinions. *Journal of the Association for Information Science and Technology*. 65(12): 2479–2488.

Lucio-Arias D. - Leydesdorff L. (2008). Main-path analysis and path-dependent. transitions in HistCite^{TM}-based historiograms. *Journal of the association for information science and technoogy*, 59(12): 1948-1962.

Ma R. - Zhang X. (2016). Discovering the Knowledge Communication Main Pathof a Domain Based on Pathfinder Algorithm. *Journal of the China Society for Scientific and Technological Information.* 8: 856-863.

Martinelli A - Nomaler O. (2014). Measuring knowledge persistence: a genetic. approach to patent citation networks. *Journal of Evolutionary Economics*, 24(3):623–652.

Martinelli A. (2012). An emerging paradigm or just another trajectory? Understanding. the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy*, 41(2): 414–429.

Martyn. (2019). What is a Literature Review? Retrieved April 1, 2019, from. https://explorable.com/what-is-a-literature-review

Marshakova I.V. (1973). A system of document connections based on references. *Scientific and Technological Information Serial of VINITI*, 6: 3–8.

Massimo Franceschet. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6):92-101.

Mei Hsiu - Ching Ho- Vincent H. Lin - John S. Liu. (2014). Exploring knowledge. diffusion among nations: a study of core technologies in fuel cells. Scientometrics, 100(1): 149-171.

McCain K W. (1991). Mapping economics through the journal literature: An. experiment in journal co-citation analysis. *Journal of the American Society for Information Science*, (42):290-296.

Mina A.- Ramlogan R., G. Tampubolon - J.S. Metcalfe. (2007). Mapping evolutionary trajectories: applications to the growth and transformation of medical knowledge. *Res. Policy*, 36 (5): 789–806.

Moed H F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature*, 415(6873):731-732.

Moore S- Haines V- Hawe P - Shiell A. (2006). Lost in translation: a genealogy of the。“social capital” concept in public health. *Journal of Epidemiology & Community Health*, 60 (8): 729-734.

Ninio F. (1976). A simple proof of the Perron-Frobenius theorem for positive symmetric matrices. *Journal of Physics A: Mathematical and General*, 9(8): 1281-1282.

National Institute of Science and Technology Policy (JA- PAN). (2007). Science map 2004 - study on hot research areas (1999-2004) by bibliometric method. NISTEP, REPORT No. 100 (NISTEP, REPORT No. 95 Follow-up).

Newman M. (2008). *The mathematics of networks*. In: Blume L, editor. SD, editors, The New Palgrave Encyclopedia of Economics. Basingstoke: Palgrave Macmillan. 2nd edition.

Page L - Brin S. The PageRank Citation Ranking: Bringing Order to the Web [EB/OL]. http://www.db.stanford.edu/~backub/PageR anksub.ps, 1998~2001.

Peritz B C. (1992). On the objectives of citation analysis: Problems of theory and. method. *Journal of the Association for Information Science & Technology*, 43(6):448-451.

Persson O. (1994). The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science*, 45(1):31-38.

Persson O. (2010). Identifying research themes with weighted direct citation links. Journal of Informetrics, 4(3): 415-422.

Price D J. (1965). Networks of scientific papers. *Science*, 149: 510 -515.

Qiu J- Dong K. (2013). Methods and Empirical Research on Deep Integration of Literature in Citation Network: Case Study on XML Research Literature from WOS. *Journal of Library Science in China*, 39(204):111-120.

Redner S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, 58(6): 49-54.

Renoust, B.-Claver, V. - Baffier, JF. (2017). Multiplex flows in citation networks. *Applied Network Science*, 2: 23. https://doi.org/10.1007/s41109-017-0035-2.

Small H. (1970). *Nuclear physics in the physical review 1927-1940*. Unpublished Report of a Study at the American Institute of Physics.

Small H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.

Small H. - Garffith B C. (1974). The structure of scientific literatures I:identifying. and graphing specialties. *Science Studies*, 4(1):17-40.

Trajtenberg M(Eds.). (2002). *Patents, Citations and Innovations-A.Window on the Knowledge Economy*. MIT Press, Cambridge, MA.

Verspagen B. (2007). Mapping technological trajectories as patent citation networks. A study on the history of Fuel Cell Research. *Advances in Complex Systems*, 10(1):93-115.

Virgo J A. (1977). A statistical procedure for evaluating the importance of scientific papers. *Library Quarterly*, 47(4):415-430.

Wang L -Zhang Q. (2014). Research on Mechanism of Knowledge Flow Based on Citation Network. *Journal of Harbin Institute of Technology (Social Sciences Edition)*, 16(1):110-116.

Wang Y. (2013). Research on identification of technological trajectories: With Main Path Analysis on Patent Citation Network as the Method. Wuhan University.

Wei L -Shu F. (2016). Review and Prospect of Research Progress on Main Path. of Citation Networks. *Information Studies: Theory & Application*, 9: 128-133.

White HD - Griffith BC. (1981). Author citation: A literature measure of intellectual. structure. *Journal of the Association for Information Science and Technology*, 32(3): 163-171.

Xiao Y- Lu Louis Y.Y.- Liu John S. - Zhou Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. *Journal of Informetrics*, 8(3): 594-605.

Yunwei Chen. (2016). Development of evolving citation network analysis, *Information Science*, 8:171-176.

Zhen Z- Li F- Li Q - Wang L- Cai W- Li X - Zhang H. (2016). State-of-the-art of R&D on seawater desalination technology. *Chinese Science Bulletin*, 21: 2344-2370.

Zhu S- Xue L- Xu Z. (2014). The Development History and An Analysis. of Status Quo of Desalination at Home and Abroad. *Technology of Water Treatment*, 7:12-15.

Zhuge H. (2006). Discovery of knowledge flow in science. *Communications of the. ACM*. 49(5): 101-107.

Zhu H- Yin X- Ma J - Hu W. (2016). Identifying the main paths of information diffusion in online social networks. *Physical A: Statistical Mechanics and its Application*, 452(15): 320-328.

APPENDIX 1

Event	Date	Authors	Discovery
1	1820	Braccont	Isolation of specific amino acids from protein
2	1860s	Mendel	Predictability of dominant and recessive traits in plants
3	1869	Miescher	Isolation of nucleic acid
4	1880	Flemming	Described replication of paired chromosomes within the cell nucleus
5	1886	Kossel	Study of purine and pyrimidine content of nucleic acid
6	1891	Fischer & Piloty	Isolation and synthesization of ribose as a freely occurring sugar
7	1900	DeVries	Concept that spontaneous alteration of the chromosome can lead to mutation
8	1900-10	Fischer	Demonstration of the peptide chemical linkage of amino acids forming protein
9	1909	Levene	Identified the 5carbon sugar ribose as a component of nucleic acid
10	1926	Muller	Produced altered genes and mutants with X-rays
11	1928	Griffith	Production of living capsulated bacteria from dead capsulated pneumococci
12	1929	Levene	Discovery that certain numcleic acids contain deoxyribose (DNA)
13	1931	Alloway	Proof that genetic material from a dead strain influences characteristics of a live strain
14	1935	Stanley	Isolated crystals of tobacco-mosaic virus
15	1935	Levene	Proposed formulae assigning linkages between the nucleotides
16	1936	Bawden & Pirie	Discovered the virus(cf.14) was also a nucleoprotein
17	1938	Caspersson & Schulz	RNA concentration is highest in cells where the rate of protein synthesis is highest

18	1941	Beadle & Talum	Via X-rays produced mutant molds requiring precise amino acid supplementation
19	1944	Martin & Synge	Development method of paper chromatographic separation of amino acids
20	1944	Avery et al.	Discovered DNA carried genetic information that can change a strain into another
21	1947	Chargaff	Purines and pyrimidines present in unequal quantites within nucleic acids
22	1950	Chargaff	Different nucleotides in the chain are in random order
23	1951	Pauling & Corey	Concept of polypeptides chains in a helical configuration
24	1953	Sanger	Determined the amino acid sequence of insulin
25	1952	Hershey & Chase	Nucleic acid portion of bacteriophage virus enters cell –not the protein shell
26	1953	Wilkins	Developed X-ray diffraction methods for studies of nucleic acid
27	1953	Watson & Crick	Constructed model of spatial molecular configuration of DNA (via method of 26)
28	1953	DuVigneaud	Extended 24 to determine amino acid sequence of oxytocin and vasopressin
29	1955	Todd	Confirmed Levene's (15) formulae through chemical synthesis
30	1953	Pallade	Discovered smaller particles associated with the microsomal fraction
31	1955	Fraenlal-Conrat	Separated nucleic acid and protein shell of tobacco-mosaic virus
32	1955	Ochoa	Isolated a bacterial enzyme producing polynucleotide strands of RNA
33	1956	Kornberg	Produced synthetic polynucleotides of RNA from an enzyme
34	1957-8	Hoagland	Demostration of transfer RNA as a triplet code
35	1960	Jacob & Monod	Discovered existence of second (Messenger) RNA

36	1961	Hurwitz	Manufactured Messenger RNA in test tube (from DNA, nucleptides,enzymes)
37	1961	Dintzis	Demonstrated concept of protein construction (in 34) was accurate
38	1961	Norvelli	Extended 36 via DNA nucleotides, ribosomes and amino acids
39	1962	Mirsky &Allbrey	Messenger RNA isolated from mammalian cells
40	1961	Nirenberg & Maltaei	Ultimate verification of triplet code (using method of 32)