Chapter 6 Pathfinder Associative Network

The Pathfinder associative network (*PFNET*) was originally designed to assist researchers with psychological analysis based on a proximity data set (Schvaneveldt et al., 1989). It is a structural and procedural modeling technique that extracts underlying connection patterns in proximity data and represents them spatially in a class of networks (Cooke et al., 1996).

The power of the Pathfinder associative network rests on its ability to discard insignificant links in the original network while it reserves the salient semantic structure of the network. The simplified network still maintains the proximity connections and fundamental characteristics of the original network. *PFNET* can be used to visualize semantic relations of related nodes in a more effective and meaningful way. The Pathfinder associative network can handle data with both an ordinal and ratio nature.

The triangle inequality principle which is centered in the Pathfinder associative network algorithm is applied to simplifying an original network. The triangle inequality is used to identify paths with the lowest weights in the network, eliminate redundant ones, and make the network more economical. In the Euclidean space, the triangle inequality can be easily interpreted and illustrated. Given three points (*A, B, and C*) in the Euclidean two-dimensional plane, the distance between *AB* is always smaller than or equal to distances of *AC* and *CB*(See Fig. 6.1). When *C* is situated on the line determined by *AB*, the distance between *AB* is equal to distances of *AC* and *CB*. In other words, *AB* is always the shortest path in the network. If there is a network consisting of multiple connected points and the network is pruned in a way that all shortest paths are preserved and redundant paths are discarded, the final pruned network would be a Pathfinder network. The main idea of the Pathfinder associative network is to discard the redundant paths and keep the significant ones in a network.

The principle of the triangle inequality can be extended to an abstract space. In that case, connection proximity between two points may be measured in other forms such as invisible semantic similarity between two objects rather than distance.

The Pathfinder associative networks can be applied to many different fields of study, such as cognitive science, artificial intelligence, psychological analysis, information retrieval, knowledge organization, and information visualization as well.

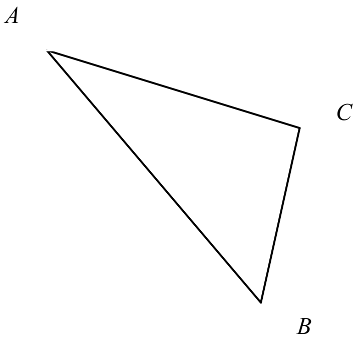

Fig. 6.1. Display of three points in the Euclidean space

6.1 Pathfinder associative network properties and descriptions

6.1.1 Definitions of concepts and explanations

A graph can be defined as *G(V, E)*. *V* is a set of vertices (or nodes) {$N_1, N_2, ..., N_n$ }and *E* is a set of edges in which an edge is connected by a pair of vertices (nodes) in *V*. $|V|=n$ is defined as the number of nodes in *V*. In a Euclidean plane, a graph can be depicted with vertices as points and edges as segments linking these vertices. A graph *G* is also called a network.

Connections and relationships of all edges in *E* can be described in an adjacent $n\times n$ matrix *EG* (See Eq. (6.1)). Headings of both the column and row are nodes and orders of these nodes in both column and row are exactly the same. The matrix *EG* is represented as:

$$EG = \begin{pmatrix} e_{11} & \cdots & \cdots & e_{1n} \\ \cdots & \cdots & e_{ij} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ e_{n1} & \cdots & \cdots & e_{nn} \end{pmatrix}_{n\times n} \quad 1 \le i, j \le n \tag{6.1}$$

Where e_{ij} is defined as an edge from node N_i to node N_j. If there is an edge between N_i to N_j, then the corresponding e_{ij} is equal to 1, otherwise e_{ij} is equal to 0. We define $e_{ii}=0$, assuming that a node is not linked to itself. It suggests that the diagonal elements of the matrix are always equal to zero. Constant n is the number of nodes in a graph. If a graph is undirected, then we have $e_{ij}=e_{ji}$. Therefore, the matrix EG is symmetric against its diagonal. If a graph is directed, the equation ($e_{ij}=e_{ji}$) may not hold. Thus the corresponding matrix EG is asymmetric against its diagonal.

Parallel to the matrix EG, the weight matrix W (See Eq. (6.2)) defines a weight w_{ij} that is associated with an edge e_{ij} in a graph. In other words, w_{ij} is the weight assigned to e_{ij}.

$$W=\begin{pmatrix} w_{11} & \cdots & \cdots & w_{1n} \\ \cdots & \cdots & w_{ij} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & \cdots & \cdots & w_{nn} \end{pmatrix}_{n\times n}, 1\le i,j\le n \tag{6.2}$$

Similar to e_{ii}, w_{ii} is always equal to 0. W and EG have the same matrix structure but different contents and meanings. It is clear that if $e_{ij}=0$, then $w_{ij}=0$. That is, if there is no link between two nodes, the weight is zero.

As we know, the Pathfinder associative network is a simplified network. It always has the same nodes as the original network but possesses fewer edges than the original network. Therefore, the Pathfinder associative network can also be defined as a matrix PF where p_{ij} is a weight assigned to the edge e_{ij}.

$$PF=\begin{pmatrix} p_{11} & \cdots & \cdots & p_{1n} \\ \cdots & \cdots & p_{ij} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & \cdots & \cdots & p_{nn} \end{pmatrix}_{n\times n}, 1\le i,j\le n \tag{6.3}$$

$$PF \subseteq W \tag{6.4}$$

A path in the graph/network is comprised of several connected edges. For instance, path $P=\{e_{ab}, e_{bc}, e_{cd}\}$ is a path consisting of three edges e_{ab}, e_{bc}, and e_{cd}. The weight of a path is calculated by the Minkowski r-metric (See Eq. (6.5)):

$$W(Path)=\left(\sum_{i=1}^{k} w_i^r\right)^{1/r}, \quad r=1,\ldots,\infty. \tag{6.5}$$

In the above equation, w_i is the weight of edge i and $Path(e_1, e_2, \ldots, e_k)$ is a path and ($w_1, w_2, \ldots, w_k$) are weights of the edges on the path. The legitimate value of the parameter r in Eq. (6.5) can range from 1 to ∞. The parameter r affects the path weight significantly. When r is equal to 1, the path weight is the sum of all edge weighs along the path; when r is equal to 2, the path weight is the Euclidean distance calculation of the path weight; and when r is equal to ∞, the path weight is equal to the maximum edge weight among all involved edge weights.

Path length is defined as the number of edges along a path. For instance, the length of *Path*(e_1, e_2, ..., e_k) is *k*.

$$L(Path) = k \tag{6.6}$$

Notice that the concept of the path length is quite different from that of the path weight even though they have a very close relationship. The path length is not dependent on the edge weights along the path whereas the path weight is calculated based on these edge weights.

A graph is *q*-triangular with the Minkowski *r*-metric if and only if all possible weights of these paths in a network, whose path lengths are smaller than and equal to the parameter *q,* meet the triangle inequality (See Eq. (6.7)):

$$w_{ag} \le (\overbrace{w_{ab}^{r} + w_{bc}^{r} + ... + w_{fg}^{r}}^{m})^{1/r}, m = 1,2,3,...,q \tag{6.7}$$

$$m = L((e_{ab},...,e_{fg})) \tag{6.8}$$

In *G(V, E*), the valid value of *q* can range from 1 to *n*-1. The associated weights of e_{ab}, e_{bc}, ..., e_{fg} are w_{ab}, w_{bc}, …, w_{fg} respectively. Parameter *m* is the path length.

The two parameters *q* and *r* can determine a family of similar Pathfinder associative networks respectively. The Pathfinder associative network family is also called isomorphic Pathfinder associative networks.

EG^i is a *path-length-i* matrix. In the matrix, if there is a path from node *l* to node *k* with path length *i,* then the element e_{1k}^{i} is equal to 1; otherwise 0.

Now let us define another very important concept: *path-length-i* minimum weight matrix which contains the most economical weights for a certain path length in a network. For the definition, see Eqs. (6.9) and (6.10):

$$W^{i+1} = W^1 \otimes W^i \tag{6.9}$$

$$w_{jk}^{i+1} = MIN\left\{ \left((w_{j1})^r + (w_{1k}^{i})^r \right)^{1/r},..., \left((w_{jm})^r + (w_{mk}^{i})^r \right)^{1/r},..., \left((w_{jn})^r + (w_{nk}^{i})^r \right)^{1/r} \right\}$$

$$for \quad w_{mk}, \quad m <> k, \quad for \quad w_{jm}, \quad m <> j, \quad 1 \le m \le n, \quad 1 \le i \le n-1 \tag{6.10}$$

W^1 is the original weight matrix *W*. Parameter *n* is the number of all nodes in a network. The above two equations are used to calculate the weight of a path when the path length increases by 1. Observe that if path growth in a network happens, it should consider all possibilities of path growths and select the most economical one from all possible paths. For instance, an existing path with path length *i* will increases by 1, that is, convert W^i to W^{i+1}. It first should consult W^1 to determine all possibilities for path growth. For weight w_{jk}^{i}, the possible paths with path length *i+1* are e_{j1}^{1} and e_{1k}^{i}, e_{j2}^{1} and e_{2k}^{i}, …, e_{jn}^{1} and e_{nk}^{i} are considered if the corresponding e_{jm}^{1} exists for the path increase. The next step is to use the Minkowski *r*-metric to calculate new path weights for all newly generated paths with path length *i+1*. And the final step is to select the best (the lowest weight) path from the all newly calculated path weights. The reason that for the weight w_{mk}, *m* can not be equal to *k*, and for the weight w_{jm}, *m* cannot be equal to *j*, is that adding either w_{kk} or w_{jj} can not result in an increase in the path length. In other words, the path length from a node to itself is defined as 0.

In the *path-length-i* minimum weight matrix, the meaning of an element w_{jk}^{i} is defined as the lowest weight of a path whose path length is exactly equal to *i* that starts from node *j* and ends in node *k* in a network. The *path-length-i* minimum weight matrix W^i (1<=*i*<=*n*) is introduced to calculate the *path-length-i* complete minimum weight matrix D^i (See Eqs. (6.11) and (6.12)).

$$D^{i} = \begin{pmatrix} d_{11}^{i} & \dots & \dots & d_{1n}^{i} \\ \dots & \dots & d_{lk}^{i} & \dots \\ \dots & \dots & \dots & \dots \\ d_{1n}^{i} & \dots & \dots & d_{nn}^{i} \end{pmatrix}_{n \times n} ,1 \le l,k \le n, \qquad 1 \le i \le n-1 \tag{6.11}$$

$$d_{lk}^{i} = MIN(w_{lk}^{1}, w_{lk}^{2}, \dots.., \quad w_{lk}^{i}), \quad l <> k \tag{6.12}$$

D^i is also a square matrix like W^i. The *path-length-i* complete minimum weight matrix D^i is different from W^i. The former is generated based upon the latter. The element d_{lk}^{i} means the weight of a path that meets two conditions: it comes from one of a group of paths whose path are lengths equal to 1, 2, 3,…, *i*, respectively, and its weight is the lowest among weights of these paths. Notice that when value of *i* increases, values of the elements in D^i may decrease. That is because the number of the paths from any node *A* to another *B* increases due to increase of *i*. As a result, the possibility of finding a lower path weight increases. If it happens, according the algorithm, the path with the lower path weight would replace the old path in D^i, which leads to lower values of elements in D^i. When *i* is equal to *n*-1, it reaches its maximum because linking a node to itself in a network does not construct a valid edge.

6.1.2 The algorithm description

The Pathfinder associative network (*PFNET(r, q))* generation algorithm is described as follows. *PFNET(r, q)* means the produced Pathfinder associative network is q-triangular with the Minkowski r-metric. W is an input original weight matrix, *PF* matrix is an output matrix of the generation algorithm, and p_{ij} is an element of *PF* ($1 \le i,\ j \le n$). The algorithm is adapted from the original Pathfinder algorithm (Dearholt and Schvaneveldt, 1990). The algorithm can handle both symmetric and asymmetric matrices.

```
L1    Begin
L2    Initialize PF matrix;
L3    Input parameters r, q, and the proximity matrix W;
L4    For m=1 To q-1 Step 1
L5        For k=1 To n Step 1
L6            For l=1 To n Step 1
L7
        $w_{lk}^{m+1} = MIN\left(\left(\left(w_{l1}^{1}\right)^{r} + \left(w_{1k}^{m}\right)^{r}\right)^{1/r},\ldots,\ \left(\left(w_{\ln}^{1}\right)^{r} + \left(w_{nk}^{m}\right)^{r}\right)^{1/r}\ \right)^{r}$;
L8            Next l;
L9        Next k;
L10   Next m;
L11   For k=1 To n Step 1
L12       For l=1 To n Step 1
L13       $d_{lk}^{q}$ =MIN( $w_{lk}^{1}, w_{lk}^{2},\ldots,\ w_{lk}^{q}$ );
L14       Next l;
L15   Next k;
L16   For k=1 To n Step 1
L17       For l=1 To n Step 1
L18           If $w_{lk}$= $d_{lk}^{q}$ Then let $p_{lk}$ =$w_{lk}$;
L19           EndIf;
L20       Next l;
L21   Next k;
L22   End.
```

In the algorithm, from lines *L2* to *L3* all variables are initialized, and parameters and proximity matrix are received. Lines *L4* to *L10* calculate a group of the *path-length-i* minimum weight matrices W^i, which will be used as inputs for calculation of the *path-length-q* complete minimum weight matrices D^q. Lines *L11* to *L15* compute the *path-length-q* complete minimum weight matrices D^q. Lines *L16* to *L21* examine whether each of the edges in the matrix D^q meets the condition. If so, these edges are moved from the matrix D^q to the final *PF* matrix. The condition is that the weight of an edge from the matrix D^q is equal to that of the corresponding weight from W^1. If the condition is satisfied, it suggests that the edge from D^q has the same weight as the weight of the corresponding edge in W^1 but different path lengths. Notice that as the value of parameter q increases, fewer elements in the matrix D^q may be qualified for the condition according to the algorithm because more paths whose path lengths are larger than 1 and path weights are lower than those in *W* may be found. If so, they replace the old ones. Consequently fewer edges in *W* have chances to be added to *PF.*

The inputs for this generation algorithm are the two parameters *r, q,* and matrix *W*, which describes the proximity among objects. Here *n* is the number of all nodes in the network. The output of the algorithm is the Pathfinder matrix *PF* which may be employed to draw a Pathfinder associative network graph in the visual space. Parameter *l, k,* and *m* are control variables. We have 1=< *q* =< *n*-1, and 1=< *r* =<∞. Notice that this presented algorithm does not have an edge labeling feature. If there are multiple paths that have the same lowest path weight and the same path length, all of these paths are included in the final Pathfinder associative network.

Observe that the Pathfinder algorithm generates a new network with the same nodes as the original matrix *W* but a sub-set of the edges of the original matrix *W.* In any case, the edges with the lowest weight in the original matrix *W* will be included to *PF* because no other paths in the network can have path weights which are lower than these edges and the nodes linking the edges must be in the final result network. If an edge links two independent sub-graphs and it is the only edge to the two graphs, then the edge is included in the final result network regardless of its weight.

Now we give an example to illustrate the generation process of the Pathfinder associative network. The given example proximity matrix *W* is symmetric (See Eq. (6.13)). All diagonal elements of the matrix are 0. The corresponding graph sees Fig. 6.2.

$$W = \begin{pmatrix} 0 & 1 & 7 & 3 & 7 & 5 \\ 1 & 0 & 1 & 7 & 8 & 8 \\ 7 & 1 & 0 & 2 & 7 & 7 \\ 3 & 7 & 2 & 0 & 2 & 7 \\ 7 & 8 & 7 & 2 & 0 & 1 \\ 5 & 8 & 7 & 7 & 1 & 0 \end{pmatrix} \tag{6.13}$$

For simplicity, the two parameters *r* and *q* for the Pathfinder associative network are set to ∞ and 2 or *PFNET (∞, 2)*. First, the algorithm needs to calculate *path-length-2* minimum weight matrix W^2. For instance, we can calculate w_{12}^2 as follows.

$$w_{12}^2 = MIN\{MAX(w_{13,}w_{32}), MAX(w_{14,}w_{42}), MAX(w_{15,}w_{52}), MAX(w_{16,}w_{62})\}$$

$$w_{12}^2 = MIN(7, \quad 7, \quad 8, \;\; 8)$$

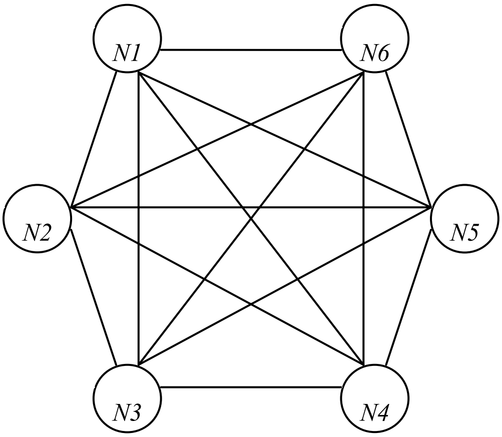

Fig. 6.2. Original network display of an example

$$w_{12}^2 = 7 \tag{6.14}$$

Following a similar calculation procedure, we can calculate the rests of elements in W^2. The final result sees Eq. (6.15).

$$W^2 = \begin{pmatrix} 0 & 7 & 1 & 7 & 3 & 7 \\ 7 & 0 & 7 & 2 & 7 & 5 \\ 1 & 7 & 0 & 7 & 2 & 7 \\ 7 & 2 & 7 & 0 & 7 & 2 \\ 3 & 7 & 2 & 7 & 0 & 7 \\ 7 & 5 & 7 & 2 & 7 & 0 \end{pmatrix} \tag{6.15}$$

The next step is to calculate the *path-length-2* complete minimum weight matrix D^2 based on both W and W^2: Compare w_{ij} and w_{ij}^2 in both W and W^2, find a minimum weight, and put the minimum weight into d_{ij}^2. For results of the calculations, see Eq. (6.16). D^1 is equal to W, so we don't need to calculate it.

$$D^2 = \begin{pmatrix} 0 & 1 & 1 & 3 & 3 & 5 \\ 1 & 0 & 1 & 2 & 7 & 5 \\ 1 & 1 & 0 & 2 & 2 & 7 \\ 3 & 2 & 2 & 0 & 2 & 2 \\ 3 & 7 & 2 & 2 & 0 & 1 \\ 5 & 5 & 7 & 2 & 1 & 0 \end{pmatrix} \tag{6.16}$$

The final step is to compare D^2 and *W*, identify the edges which satisfy the conditions $d_{ij}^2 = w_{ij}$, then put the satisfied edges into the *PF* matrix. It is clear that edges e_{12}, e_{14}, e_{16}, e_{23}, e_{34}, e_{36}, e_{45}, and e_{56} meet the condition and should be added to the *PF* matrix (See Eq. 6.17). As a final result, the final Pathfinder associative network is shown in Fig. 6.3. This network demonstrates two characteristics of the triangle inequality: No link violates the triangle inequality within path length 2 in terms of Minkowski ∞-metric, and there may be some links violating the triangle inequality when their path lengths are longer than 2.

$$PF = \begin{pmatrix} 0 & 1 & 0 & 3 & 0 & 5 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 7 \\ 3 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 \\ 5 & 0 & 7 & 0 & 1 & 0 \end{pmatrix} \tag{6.17}$$

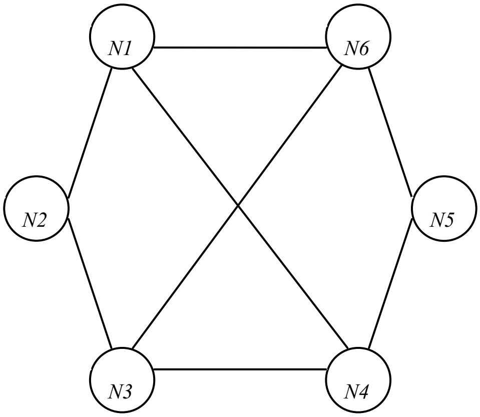

Fig. 6.3. Final display of *PFNET*(∞, 2)

6.1.3 Graph layout method

Unlike other information visualization approaches such as the self-organizing maps, *DARE, TOFIR* and so on, the Pathfinder associative network has a unique problem of graph drawing in a visual space. The problem is raised because logic relationships of nodes in *PFNET* are separate from physical relationships of the *PFNET* nodes in the visual space. Logic relationships of nodes in *PFNET* are described in the matrix (*PF*). But *PF* does not illustrate how these nodes are projected onto the visual space. The physical relationships of these nodes refer to nodes' positions and locations, and edges linking these nodes in a visual space. For instance, nodes *A* and *B* are linked in *PF*, *A* and *B* can be positioned anywhere in a visual space as long as they are connected in the visual space. Graph drawing, an independent research field, addresses how to effectively arrange connected nodes in a low visual space while preserving logic connections and relationships of the nodes. Issues regarding aesthetics for drawing an undirected graph include graph symmetry, minimal edge crossing, bending of edges, uniform edge length, reflection of inherent symmetry, conformation to the frame, and uniform vertex distribution (Battista et al., 1994; Fruchterman and Reingold, 1991).

A spring model for graph drawing was introduced by Kamada and Kawaii (1989). The model simulates a dynamic spring system where an edge in a graph stands for a spring, and a ring for a node in the graph. Two springs are linked by a ring in the system. When new springs are added to the system (or existing springs are deleted from the system), or an external force is imposed upon the system, the previous balance of the spring system is no longer maintained. The system reaches new equilibrium when the energies of all springs are released to the minimum status. This optimistic status is used to draw the graph in the visual space.

The energy of a spring is given in the following equation:

$$E = \frac{1}{2} K \times X^2 \tag{6.18}$$

Where X is the spring length from the position of its free status to the position of its stretching status, and K is the force constant of the spring which is primarily determined by its material quality.

Eq. (6.19) can be extended to a multiple spring system. Given a dynamic spring system in which n nodes are mutually linked by springs. Denote p_i (i=1, 2, 3,…, n) a node in a graph. The energy of the whole system is defined as:

$$E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{2} k_{ij} \times \left(\left| p_i - p_j \right| - l_{ij} \right) \tag{6.19}$$

Where l_{ij} is the original length of the free spring determined by p_i and p_j, k_{ij} is the force constant of the corresponding spring, and | p_i - p_j | is the distance between p_i and p_j in the graph. In order to achieve the equilibrium status in which lengths of all springs are the shortest, E must reach its minimum value.

Finally, use the Newton-Raphson method (Rowe et al., 1987) to find out solutions to all variables in the equation which determine positions of all involved nodes in the graph.

6.2 Implications on information retrieval

Application of a *PFNET* to a domain problem requires identifying two basic, necessary, and indispensable elements from application domain: the first is the objects which are used as nodes in the network; and the second is the proximity relationship between the two objects which is used to form a link between the two objects. For certain types of objects, it may correspond to multiple methods to define their proximity relationship between the two objects. Different types of objects may have different proximities. Proximity can be procured by either a human-interference method or an automatic computation method. It is not surprising that different objects and proximity methods can lead to different Pathfinder associative networks.

Clearly defining objects and the proximity method are essential to construction of a Pathfinder associative network.

6.2.1 Author co-citation analysis

Author co-citation refers to the phenomena occurring when the authors of two different papers both co-cite the same paper(s) in their works. The concept is also called bibliographic coupling. Usually papers are cited to demonstrate previous related research works, or support the author's arguments. It is a very common and natural phenomenon that two authors cite the same paper(s) if they address the same topic or a related topic. As a supplement to subject analysis, author co-citation analysis is unique and important because the cited documents have a close semantic relationship with a citing document. Views, themes, ideas, concepts, theories, issues, problems, trends, approaches, and people from the cited documents are naturally embedded in the contexts of the citing paper. It is believed that the concepts and conceptual relations based on cited documents have an advantage over concepts and conceptual relations created from conventional co-term analysis (Rees-Potter, 1989).

Author co-citation analysis uses co-citation data to structure and summarize a scientific field which can be depicted in a co-citation network or a collaboration graph. It is apparent that in this case the objects, one of the two basic elements for the Pathfinder associative network construction, are documents which cite or are cited by each other in author co-citation analysis. There are various approaches to define proximity relationship between documents in co-citation analysis. Proximity relationship between two documents is used to produce a document-document proximity matrix providing the input for the Pathfinder associative network generation algorithm.

The first approach, the cosine similarity measure, was described in Eq. (6.20) (Chen and Morris, 2003).

$$S(d_i,d_j)=\frac{cocit(d_i,d_j)}{\left(cit(d_i)\times cit(d_j)\right)^{1/2}} \tag{6.20}$$

In Eq. (6.20), *cit(x)* denotes the number of all citations of a document *x* and *cocit(x, y)* stands for the number of co-citations that both document *x* and document *y* cite. Here both d_i and d_j are two citing documents in a document collection. The equation suggests that the proximity or similarity between two documents increases if the number of citations that the two documents co-cite grows, and vice versa. And the proximity or similarity between two documents increases if the number of all citations for either of the documents decreases and the co-cited documents stay the same, and vice versa.

The second proximity approach is called the Jaccard or Tanimoto similarity measure. See Eq. (6.21). It was used in co-citation analysis study (Schneider and Borland, 2004; Schneider, 2005). Definitions of *cit(x)* and *cocit(x, y)* in Eq. (6.21) are the same as these in Eq. (6.20). The difference between the two equations reflects in their denominators, or the way that they normalize co-citations.

$$S(d_i,d_j)=\frac{cocit(d_i,d_j)}{\left(cit(d_i)+cit(d_j)-cocit(d_i,d_j)\right)} \tag{6.21}$$

The author co-citation Pathfinder network can be used to visualize progress in knowledge domain (Chen, 2004). In order to illustrate the progress, a time interval was divided into a number of meaningful time slices (saying one year, or five years), and an individual co-citation Pathfinder network was derived from each time slice. The final time series of co-citation networks was generated when all time slices were connected according to their time sequences. That is, time slices consisted of a continuous time series. In the time series of co-citation Pathfinder associative networks, salient changes between neighboring time slices were identified and scientific research evolution was visualized and analyzed. In the co-citation Pathfinder associative networks, a node size, and width and length of a link were proportional to the number of the citations of a document, co-citation similarity value respectively. And nodes in the network were classified as landmark nodes which had significant attribute values; Hub nodes that were the widely co-cited documents; and pivot nodes that were joints between different sub-networks.

The third proximity approach is the Pearson *r* correlation co-efficient. It is widely recognized and used in author co-citation analysis. It is an easily understood concept. Many commercial statistical packages support the Pearson *r* correlation co-efficient. Author co-citation matrices can serve as input to principal component analysis as well as multidimensional scaling and hierarchical clustering routing. The Pearson *r* co-efficient method can produce highly intelligible results (White, 2003).

Correlation analysis addresses measuring the association degree between two variables. In the Pearson *r* correlation co-efficient (or Pearson product moment correlation co-efficient), the two variables should have a linear relationship, and either of the variables is normally distributed. They should be interval or ratio. The Pearson *r* can be computed from Eq. (6.22).

$$r = \frac{n\sum XY - \sum X \times \sum Y}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right) \times \left(n\sum Y^2 - \left(\sum Y\right)^2\right)}} \tag{6.22}$$

Where *n* is the number of observations, *X* and *Y* are two variables.

Result *r* ranges from -1 to 1. If the result *r* is larger than 0, it suggests that there is a positive relationship between the two variables. If the result *r* is smaller than 0, it suggests that there is a negative relationship between the two variables. If the result *r* is equal to 0, it suggests that there is no relationship between them. For instance, result *r* (from 0.9 to 1) indicates very high correlation, from 0.7 to 0.9 high correlation, from 0.5 to 0.7 moderate correlation, from 0.3 to 0.5 low correlation, and from 0 to 0.3 little correlation.

However, there are debates over application of the Pearson correlation coefficient approach to co-citation analysis (White, 2003). The issues include: Pearson *r* becomes unstable when smaller co-citation count matrices are combined; The treatment of diagonal in matrices from which measures like *r* are produced remains a problem; Pearson's *r* is supposed to handle data with normal distribution while author co-citation data is highly skewed; The standard significance test for *r* assumes random sampling of independent observations from population; And so on.

6.2.2 Term associative network

Term co-occurrence analysis addresses term co-occurrence behavior in a full-text document. Keywords appearing together in a predefined length of text in the same document are regarded as the co-occurrence terms.

Term co-occurrence information can be utilized to produce the Pathfinder associative term network which may be utilized to explore and discover related terms in a domain that users are not familiar with. For instance, the idea, the so-called "term seeding" method (Buzdlowski et al., 2001), is that a user starts with a seed term as a starting point, then it can trigger other associative terms that most frequently co-occur with the seed term. Documents including the seed term are systematically examined to return co-occurred terms.

Term Pathfinder associative networks are also expected to help users to better formulate their queries.

It is clear that objects of term Pathfinder associative networks are terms and the proximity is the relationship among terms in the full-text contexts. There are several options to define such proximity in the full-text contexts.

The first proximity method is based on term adjacency information. In analyzing provided texts, all stop words are filtered by a predefined stop word list, and the remaining words are stemmed. Term pair proximity or similarity is calculated as the sum of values added when they are adjacent, or occur in the same sentence, paragraph, or documents. For each of the term pairs, similarity is increased by 5 if they are adjacent in the same sentence, 4 for a nonadjacent term pair in the same sentence, 3 for a nonadjacent term pair in the same paragraph, 2 for a nonadjacent term pair in the same section/chapter, and 1 for a term pair in the same document. The results of this processing lead to a final term-term proximity matrix that is used for construction of a term *PFNET* (Fowler and Dearholt, 1990). The term co-occurrence matrix is organized as follows. Both the columns and rows are defined as terms respectively. The order of terms in both the column and row are the same. The interaction of a column and a row in the matrix is the term proximity value between two terms.

The second proximity method is based on term probability in a full-text. Association between two terms in a full text can be calculated by the equivalent index (Turner et al., 1988; Schneider and Borland, 2004), see Eq. (6.23).

$$S(t_i, t_j) = \frac{f_{ij}^2}{f_i \times f_j} \tag{6.23}$$

Where f_{ij} is the number of co-occurrences of term t_i and term t_j in citation contexts, both f_j and f_i are occurrence of term t_i and term t_j respectively. $S(t_i, t_j)$ indicates the probability of term t_i (t_j) appearing simultaneously in a set of the citation context with term t_j (t_i). $S(t_i, t_j)$ is also called a coefficient of mutual inclusion because of this reason. Eq. (6.23) can be used to produce the term-term proximity matrix.

6.2.3 Hyperlink

The *PFNET* technique can be applied to Internet information representation. In this case, the objects of the Pathfinder associative network are Web pages and proximity is strength of a hyperlink which connects Web pages.

In fact, there are two approaches to construct a *PFNET* based on hyperlink strength. The first one is similar to the author co-citation analysis method. The number of co-cited hyperlinks can be used to measure the similarity between two Web pages. The cell value in the webpage co-citation matrix is defined as the number of the same Web pages that two Web pages co-cite. The webpage co-citation matrix should be symmetric. This is because if webpage *A* cites webpage *C* and webpage *B* also cites webpage *C*, the direction of a citation does not play any role. Therefore, the final *PFNET* is an un-directed graph.

The second approach is based on hyperlink connections between two pages (Chen, 1997). In this case, a webpage connection matrix is defined as follows. The cell value of the webpage connection matrix is defined as the number of hyperlinks that a webpage cites another webpage. It is apparent that the webpage connection matrix is asymmetric. That is because if webpage *A* cites webpage *B*, it does not necessarily mean that webpage *B* also cites webpage *A*, therefore, the final *PFNET* is a directed graph.

6.2.4 Search in pathfinder associative networks

Users are allowed to search an established Pathfinder associative network (Chen, 1999). After a query is submitted to the network, the relevance between a query and a document is calculated by the Pearson correlation coefficient. Then, search results can be demonstrated or highlighted on the network. The relevance magnitude of a search query and a document is indicated by the height of a raising spike from the document sphere. The longer a spike is, the more relevant it is to the document; and vice versa. Users can also browse the Pathfinder associative network at will. Clicking a document sphere, users can view its contents in detail. Documents on the center ring in a Pathfinder associative network appear to be more generic than leaf-documents on a branch.

Query search in a Pathfinder associative network can be a different scenario (Fowler et al., 1991; Fowler and Dearholt, 1990) where both a query and a document are converted into two Pathfinder associative networks respectively. The similarity between a query and a document hinges on the similarity between the two Pathfinder networks. The query process can begin with the user's entry of a natural language request for information. Query revision can be accomplished by deleting nodes, entering more texts, or dragging any terms that the system display to the query Pathfinder associative network. Keyword adjacency information in both a natural language based query and a full-text based document can be employed to generate a query Pathfinder network and a document Pathfinder network respectively. Since both a query and a document are presented in the *PFNET* form, the match technique between a query and a document is a little different from traditional ones. The proximity algorithm for a query network structure and a document network structure consists of two parts. The first part is defined as the ratio of the number of common terms in both a query and a document to the number of all terms in the query. It is clear that the first part only measures the term relevance between the query and the document. The second part is supposed to measure the network structure similarity between the query network and a document network. The value of this part increases when nodes (terms) connected in the query network also appear closely connected in the document network. For instance, the similarity value of two network structures increases by 2 when two terms appear in both the query and the document, and they are directly linked in both networks; the similarity value increases by 1 when two terms appear in both the query and the document, but are indirectly linked in both networks. All network similarity values are summed up and the total is divided by 2 times the number of links in the query to normalize the structure similarity between 0 and 1. Finally, the two parts are weighted and integrated into a final similarity value which is used to make a decision on whether the document is relevant to the query or not.

6.3 Summary

In a Pathfinder associative network *PFNET(r, q)*, the triangle inequality is always satisfied in terms of a path weight calculated by the Minkowski *r*-metric within path length *q*. Characteristics of a Pathfinder associative network are determined by the two important parameters, *r* and *q*. The weight of a path is affected by the Minkowsk*i* *r*-metric while path length is affected the parameter *q*. *PFNET* can have systematic variations when the two parameters *q* and *r* are varied. The changes of these two parameters can impact the Pathfinder associative network complexity. The complexity of the network decreases as either or both of these two parameters increases. In other words, when the parameters *r* and *q* are equal to their maximum values ∞ and *n*-1 respectively (*n* is the number of all nodes in a network), the *PFNET* is the simplest and most economical network. However, increase of *q* would result in an increase of computational complexity.

The strength of *PFNET* lies in reveling accurate, detailed, and specific connections of nodes in a network.

The weaknesses of the Pathfinder associative network include its computational complexity, which may prevent *PFNET* from not only visualizing a large dataset, but also dynamically modifying a *PFNET* caused by interactions between users and the network. The *PFNET* generation algorithm requires many large intermediate matrices to yield the final result. This may lead to occupying a large amount of memory to support the generation of these matrices. Another disadvantage of *PFNETs* in the present state of development is that people have no way of knowing the features upon which similarity judgments are made, which results in that the semantic content of links is not easily discernible (Dearholt and Schvaneveldt, 1990). It is clear that *PFNET* cannot generate a local visual configuration based users' individual information needs and it only produces a global overview for a data collection.

Since the logic relationships of nodes in a Pathfinder associative network are separate from its physical relationships of the nodes, the logical relations are not directly assigned to a coordination system of the visual space. It leads to a graph drawing problem when the Pathfinder associative networks are projected onto a *2D* or *3D* visual space. Fortunately, people have found an effective solution to the problem.

The Pathfinder network technique is very effective and efficient for display of complex relationships among objects such as sophisticated semantic networks. As an information visualization means, it can be applied to a wide spectrum of information retrieval environments, ranging from information searches, author co-citation analysis, term co-occurrence analysis, thesaurus construction, to the Internet information representation.