

Binary Pathfinder: An improvement to the Pathfinder algorithm

Vicente P. Guerrero-Bote ^{a,*}, Felipe Zapico-Alonso ^a,
María Eugenia Espinosa-Calvo ^a, Rocío Gómez Crisóstomo ^a,
Félix de Moya-Anegón ^b

^a *Library and Information Science Faculty, University of Extremadura, Alcazaba de Badajoz (Antiguo Hospital Militar), Plaza Ibn Marwan s/n, 06071 Badajoz, Spain*

^b *Library and Information Science Faculty, University of Granada, Campus Cartuja, Colegio Máximo, 18071 Granada, Spain*

Received 16 March 2006; accepted 16 March 2006

Abstract

The Pathfinder algorithm is widely used to prune social networks. The pruning maintains the geodesic distances between nodes. It has shown itself to be very useful in the analysis of, amongst others, citations in BIS (bibliometrics, informetrics, and scientometrics). It has even been proposed for the online display of the search results in an information retrieval system. However, its great time and space complexity limits its use in real-time applications and in networks of any considerable size.

The present work describes an improved algorithm with considerably reduced time and space complexity. Its lower execution costs thus increase its applicability both in real time and to large networks.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: PFNETs; Social networks; Citation analysis; Information visualization

1. Introduction

Network analysis centres on seeking network structures in the representation of entities, such as nodes, and their relationships, such as links (Börner, Chen, & Boyack, 2003). In certain types of problems, however, the tangle of links is so complex that it does not allow the principal relationships to be seen. Associated with the links, there usually exists a number that indicates the distance between nodes or the strength of the relationship. This number can be used to prune the less significant links. However, such a pruning process is far from having a trivial solution, since links that may not be important for a given structure may be so for another.

The Pathfinder algorithm was developed in cognitive science to determine the most important links in a network (Schvaneveldt, 1990). The resulting pruned networks are known as Pathfinder networks or PFNETs.

* Corresponding author. Tel.: +34 924289300; fax: +34 924286401.

E-mail addresses: guerrero@unex.es (V.P. Guerrero-Bote), fzapalo@alcazaba.unex.es (F. Zapico-Alonso), meespcal@alcazaba.unex.es (M.E. Espinosa-Calvo), mrgomcrl@alcazaba.unex.es (R. Gómez Crisóstomo), felix@ugr.es (F. de Moya-Anegón).

The pruning carried out by this algorithm is based on eliminating those links which violate the triangle inequality. This inequality states that the direct distance between two points must be less than or equal to the distance between them passing through an intermediate point. This inequality always holds in Euclidean spaces, but there are other domains in which it does not. Dearholt and Schvaneveldt (1990) give the example where the similarity or the distance is objectively measured by means of intersections of sets. In particular, if one knows the intersection between sets A and B on the one hand, and between B and C on the other, one can deduce nothing about the intersection between A and C. In these cases, since the transitive relationship is not satisfied, neither can one expect the triangle inequality to be.

However, it is always attractive to be able to use geodesic distances, where the distance between a pair of nodes is that corresponding to the shortest path connecting them. Indeed, this is the definition of distance usually employed in graph theory. In this sense, a link that does not satisfy the triangle inequality will never form part of the shortest path between two nodes, because there will always be a better alternative. For this reason, as indicated by Schvaneveldt, Dearholt, and Durso (1988), eliminating the links which violate the triangle inequality preserves the geodesic distance between nodes while simplifying the structure of the network, thus clarifying the subsequent analysis.

The triangle inequality, which is of course defined on the basis of a triangle, must also hold for higher-order polygons. This is equivalent to saying that the direct distance between two nodes will never be greater than the distance between them passing through different intermediate nodes. This gives rise to the first of the algorithm's parameters, the maximum number of intermediate links that will be considered.

One also needs a procedure for the calculation of the distance between two nodes along a path that is not direct. On a road map, one simply sums the intermediate distances, but this may not carry over to all problems. The Pathfinder algorithm calculates the said distance using the Minkowski r -metric. For $r = 2$, this reduces to the Euclidean distance, and for $r = 1$ to the road-map distance, i.e., the sum of the lengths of the segments on the map and for $r = \infty$, it is the greatest of the distances between successive intermediate nodes. The Minkowski r -metric parameter is the second of the algorithm's parameters.

Although the algorithm is defined for networks where each link is weighted by a distance, a dissimilarity, it is readily generalizable to networks where each link is weighted by a similarity between its nodes.

As Dearholt and Schvaneveldt (1990) indicate, this type of network has the capacity to model asymmetric relationships directly, which is more difficult with other techniques such as multidimensional scaling (MDS). Local relationships are represented more precisely than with MDS, which is based on minimizing the overall error, the hierarchical constraints in most cluster analysis techniques do not apply to PFNETs, which bring out the most "salient" relationships present in the data. In sum, they represent a novel quantitative methodological framework in which to study classification models, being applicable to networks that have been intuitively designed or based on qualitative information.

Pathfinder network scaling is a procedural modeling algorithm originally developed by cognitive psychologists to capture salient relationships between concepts. The strengths of such relationships are typically measured by human experts' subjective ratings of how similar those concepts are. Initial studies exclusively used Pathfinder networks to represent interrelations between concepts or keywords. Other work has extended the use of Pathfinder networks to a far richer range of applications, especially cocitation networks (Chen, 1998; Chen & Paul, 2001).

Chen (1999) was the first to apply the technique to citation analysis, adapting it to form an integral part of a structuring and visualization framework. Unlike the classical ACA (author co-citation analysis) (Lin, White, & Buzydowski, 2003; White, 2003) which uses the Pearson correlation, and MDS, what matters with this method is not the location itself at which an author is represented, per se, but the links between the nodes which are represented by means of lines (Buzydowski, 2002).

The networks that are constructed in the analysis of citations, co-citations, bibliographic coupling, or co-word analysis, whether of authors, journals, ISI categories, etc., usually have a weight associated with each link. This weight has the meaning of similarity, the number of times that two authors are co-cited, the references they have in common, etc. Since one usually employs $r = \infty$ and $q = n - 1$, this gives rise to associating with each path formed by a set of links the similarity of the link of least similarity. In analogy with a plumbing network, or with Internet access, the rate of flow or bandwidth reaching a given point will depend on the narrowest bottleneck in the path. Therefore, by also using $q = n - 1$, no link will be left that does not satisfy the

triangle inequality, and only the least cost paths will be retained, which in this case will be those of maximum flow rate or bandwidth.

With these parameters, some nodes appear linked to only one other node, whereas some are linked to many. In the ACA case, for example, an author needs to be highly cited in order to be highly co-cited, and hence be linked with others (White, 2003). These nodes appear as stars, connected to many others, and with a great centrality. In terms of author co-citation, one says that these authors dominate their research speciality.

One of the great advantages of PFNETs is that alone without the aid of other methods, they are able to show what is happening in a field or discipline through the links that have been retained (White, 2003). The dominant authors, and the links starting out from them, define specialities. If a field lacks any major figures, it will be seen as relatively disconnected, with no hierarchical structure. In contrast, classical ACA techniques require various methods to extract the same information (Pearson correlation, MDS, PCA, cluster analysis, etc.).

The result of the procedure consists of networks with all the original nodes and the most salient links. In effect, a totally or highly connected network will have been pruned to eliminate much of its complexity and visual noise.

This type of network has also been proposed for the real-time display of search results (Lin et al., 2003), in combination with which it could be used to represent visually the profile of an author, a research team, or a discipline.

The computational cost of the algorithm, however, increases rapidly with increasing numbers of nodes and links. This is a limitation against its application in real time or to large networks. It is in this sense that we here propose to obtain another algorithm whose time and space complexity is less than that of the original algorithm. It will thus be applicable to large networks, including real-time application to networks of considerable size.

2. Description of the original algorithm

Pathfinder is based mainly on two elements: the Minkowski distance, and an extension of the triangle inequality (Schvaneveldt, 1990; Schvaneveldt et al., 1988).

The Minkowski distance is used to calculate the distance between two points along a path through several links. It is defined by means of a parametric equation:

$$D = \left(\sum_i d_i^r \right)^{1/r}$$

This equation reduces to the sum of distances for $r = 1$ (r is the parameter), to the Euclidean distance for $r = 2$, and allows r to tend to infinity, in which case the equivalent is finding the greatest of the intermediate distances.

The second element that comes into play is the triangle inequality, indicating that the distance of a direct path between two points can never be greater than that corresponding to another path that passes through one or more intermediate points. This holds in Euclidean spaces, but not necessarily in other spaces. Here it will be applied by eliminating all those links that have an associated distance which is greater than that of another path between the same two points but passing through other intermediate nodes. The distance through these intermediate nodes is determined by means of the Minkowski equation.

The Pathfinder algorithm thus has two parameters r (associated with the Minkowski distance being employed) and q (associated with the length in number of links of the paths being compared, or, equivalently, the number of intermediate points). The links which violate the triangle inequality will be eliminated, since they have an associated distance which is greater than another path between the same points consisting of up to q links, and with the overall distance of this second path calculated by means of the Minkowski equation with parameter r . The maximum possible value of q is $n - 1$, where n is the number of nodes.

In the description of the algorithm, Dearholt and Schvaneveldt (1990) use the following definitions:

- A PFNET has n nodes, denoted N_1, N_2, \dots, N_n (or N_a, N_b, \dots).
- A *link* is an association between a pair of nodes which can be either undirected or directed. A *directed link* is called an *arc*, and an *undirected link* is called an *edge*.

- Links are labelled e_{ij} , for the edge between N_i and N_j (or for the arc from N_i to N_j). N_i and N_j are *end nodes* of the link e_{ij} .
- The distance from node N_i to N_j (along the link e_{ij}) is the weight w_{ij} , and these weights are often written in matrix form as an $n \times n$ matrix W .
- $W^{i+1} = W \Theta W^i$ is computed as follows:

$$w_{jk}^{i+1} = \text{MIN}((w_{jm}^i)^r + (w_{mk}^i)^r)^{1/r} \quad \text{for } 1 \leq m \leq n$$

where $w_{jm} \geq 0$ and $w_{mk}^i \geq 0$.

- The minimum-distance matrix for paths not exceeding i links is denoted D^i , and its elements are computed as follows:

$$d_{jk}^i = \text{MIN}(w_{jk}^1, w_{jk}^2, \dots, w_{jk}^i) \quad \text{for } j \neq k$$

The procedure for calculating the PFNETs is therefore the following:

1. Compute $W^2, D^2, W^3, D^3, \dots, W^q, D^q$;
2. Comparing elements of D^q and W^1 , wherever $d_{ij} = w_{ij}$, mark e_{ij} as a link in the PFNET.

As is expressed in the algorithm, one has to create q matrices W^i and D^i , each of which has n^2 weights or distances. The resulting space is thus of complexity $O(qn^2)$. To calculate each weight one has to make n comparisons, so that the algorithm has time complexity $O(qn^3)$, and since q can take a maximum value of $n - 1$, one can say that in this maximum case the algorithm is of time complexity $O(n^4)$.

3. Binary Pathfinder

Our modest contribution will be based on two aspects:

- It is only necessary to attain D^q for the comparison with the initial weight matrix. It is unnecessary to generate the rest of the D^i .
- Similarly to how the W^i are calculated, and with the same complexity, the matrices D^i can be calculated directly from another two distance matrices, so that $D^{i+j} = D^i \Theta D^j$.

The distance matrices would then be calculated in the following form:

$$d_{kl}^{i+j} = \text{MIN}\left\{d_{kl}^i, d_{kl}^j, ((d_{km}^i)^r + (d_{ml}^j)^r)^{1/r}\right\} \quad \text{for } 1 \leq m \leq n$$

where $d_{kl}^1 = w_{kl}$.

The idea is to ensure that in the matrix D^i there are the minimum distances between two nodes using up to i links. In the original algorithm this is done by calculating all the matrices $W^1, W^2, W^3, \dots, W^q$, which contain the minimum distances using exactly i links, and from them calculates the matrices D^i by taking the first i matrices $W^1, W^2, W^3, \dots, W^i$. Nevertheless, all the parts into which a minimum path of up to i links can be decomposed will also be minima, at least among the paths with fewer than i links. This means that if one starts from D^i and D^j , any minimum distance path considered in D^{i+j} has necessarily to be either one already considered in D^i or D^j , or a combination of two links – one in each origin matrix. I.e., if l is the number of links of any minimum path considered in D^{i+j} :

- If $l \leq i$, then that path has to be considered in D^i .
- If $l \leq j$, then that path has to be considered in D^j .
- Otherwise, as $l \leq (i + j)$, one will always be able to make a decomposition into one of length i which has to be considered in D^i and another of length $l - i$ which has to be considered in D^j , because, as we said above, any of the parts of a path considered as optimal for D^{i+j} has to be optimal among the paths with fewer than $i + j$ links, and therefore of up to j links.

With this, it is shown that the distance matrix D^{i+j} can be calculated from D^i and D^j , obtaining the same result as with the method described by Dearholt and Schvaneveldt (1990).

One can thus make larger steps. One begins by calculating $D^1, D^2, D^4, D^8, \dots$, and the algorithm can be expressed as follows:

```

i = 1
nq = 0
Generate  $D^1 = W$ ;  $D^q = \infty$ 
If  $(q \bmod 2) = 1$ 
    Compute  $D^q = D^q \Theta D^1$ 
nq = 1
While  $(i * 2) \leq q$ 
    Compute  $D^{i*2} = D^i \Theta D^i$ 
    If  $((q - nq) \bmod (4 * i)) > 0$ 
        Compute  $D^q = D^q \Theta D^{i*2}$ 
    nq = nq + (2 * i)
    i = i * 2

```

Comparing elements of D^q and W^1 , wherever $d_{ij} = w_{ij}$, then mark e_{ij} as a link in the PFNET.

Here, the operator “mod” means modulus, i.e., the remainder of a division by an integer.

The first and greater part of the algorithm (everything except the last line) concerns the calculation of the matrix D^i . The last line is exactly the same as the traditional method.

One observes that the procedure is similar to that of transforming a number to the binary system, so that the principal loop has a time complexity of $O(\log(q))$ instead of $O(q)$.

This means that the algorithm’s overall time complexity is $O(\log(q) \cdot n^3)$, instead of $O(qn^3)$. And, in the maximum case, the time complexity would be $O(\log(n) \cdot n^3)$, instead of $O(n^4)$. While this may in principle seem only a slight difference, as ever larger graphs are dealt with the difference grows immensely.

The space complexity is also reduced to $O(n^2)$.

4. Results

In order to test the difference in execution time of the two algorithms, we conducted a trial with networks of different numbers of nodes. We took networks of the type usually employed in Information Science, generated from author co-citation (Buzydlowski, 2002; Lin et al., 2003; White, 2003), JCR category co-citation (Moya-Aneón et al., 2004, 2005), etc. In these networks, the weights associated with the links are not distances but similarities.

We used $q = n - 1$ and $r = \infty$ for the parameters, since these are the customary choices in this type of study.

The two algorithms were implemented in the C programming language, and compiled using DJGPP, an acronym for DJ’s GNU Programming Platform, a project which brings the GNU development tools to the MS-DOS and MS-Windows systems. Its originator and principal maintainer is DJ Delorie, which is where the “DJ” in DJGPP comes from. They were run on two machines under MS-Windows XP, one with a Pentium processor at 2.80 GHz, and the other with a Centrino at 2.13 GHz. In defence of the Pentium, we must note that the corresponding disk ran at 4200 r.p.m., while the Centrino machine’s disk ran at 5400 r.p.m. This was especially noticeable in the execution of the first example which should have been the fastest.

One can see from Table 1 that the results showed hardly any variation from one machine to the other.

As was expected, however, the trend of the times of the two algorithms differed markedly. For a small number of nodes, the difference was almost non-existent, and one could even say that the original Pathfinder algorithm was superior. But as the number of nodes increased, the execution time of the original algorithm increased much faster than that of the Binary Pathfinder. Both, of course, yielded the same result.

The execution time was not monotonously increasing with number of nodes because it also depends on the number of links between the nodes. This explains the result obtained with the file e217.net, for example.

Table 1
Times (in seconds) employed by the Pathfinder and Binary Pathfinder algorithms, using two different computers

File	N	Binary Pathfinder		Pathfinder	
		Centrino (2, 13)	Pentium (2, 8)	Centrino (2, 13)	Pentium (2, 8)
e8.net	8	00.05	00.14	00.05	00.11
e16.net	16	00.05	00.07	00.03	00.06
e17.net	17	00.04	00.06	00.04	00.07
e18.net	18	00.08	00.06	00.05	00.06
e24.net	24	00.05	00.06	00.04	00.06
e26.net	26	00.05	00.06	00.05	00.06
e34.net	34	00.05	00.06	00.05	00.07
e53.net	53	00.06	00.08	00.11	00.13
e89.net	89	00.16	00.18	00.36	00.38
e108.net	108	00.20	00.21	00.46	00.45
e109.net	109	00.31	00.32	00.52	00.49
e126.net	126	00.28	00.30	00.78	00.78
e207.net	207	02.09	02.28	12.19	12.72
e216.net	216	02.65	02.91	15.52	15.65
e217.net	217	01.94	02.08	07.75	07.26
e241.net	241	03.30	03.66	24.25	24.72

5. Conclusions

As we noted at the beginning of the work, PFNETs are of great interest in the study of all types of networks. For the particular case of citation analysis, they are found to be extremely useful in studying advancing frontiers of research, disciplines, and even the profiles of authors, research teams, or institutions.

This algorithm has also been used in online systems of information retrieval for the display of search results.

Their application has been quite limited, however, by their complexity, which makes their costs of execution increase rapidly as the numbers of nodes and of links in the networks grow.

The improved performance of the Binary Pathfinder is based on two aspects:

- It is only necessary to reach the final distance matrix in order to make the comparisons, with the previous matrices being dispensable.
- The matrices can be generated additively, thereby reducing the number of times that new matrices need to be generated.

As a result, one succeeds in reducing the space complexity as well as the time complexity, achieving very significant differences in execution time for medium- and large-sized networks, and greatly reducing the limitations for the pruning procedure.

Acknowledgments

This work was financed by the Junta de Extremadura-Consejería de Educación Ciencia & Tecnología and the Fondo Social Europeo, as part of research project 2PR02A041.

References

- Börner, K., Chen, C., & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179–255.
- Buzydowski, J. W. (2002). A comparison of self-organizing maps and pathfinder networks for the mapping of co-cited authors, Ph.D. thesis. Drexel University.
- Chen, C. (1998). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers*, 10(2), 107–128.
- Chen, C. (1999). *Information visualization and virtual environments*. Berlin, Germany: Springer.

- Chen, C., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer*, 34(3), 65–71.
- Dearholt, D. W., & Schvaneveldt, R. W. (1990). Properties of pathfinder networks. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: studies in knowledge organization* (pp. 1–30). Norwood, NJ: Ablex.
- Lin, X., White, H. D., & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *Information Processing and Management*, 39(5), 689–706.
- Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Herrero-Solana, V., & Muñoz-Fernández, F. J. (2005). Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. *Information Processing and Management*, 41(6), 1520–1533.
- Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Muñoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129–145.
- Schvaneveldt, R. W. (1990). *Pathfinder associative networks*. Norwood, NJ: Ablex.
- Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of pathfinder networks. *Computers and Mathematics with Applications*, 15(4), 337–345.
- White, H. D. (2003). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423–434.