Optimizing Research Progress Trajectories with Semantic Power Graphs

G.S. Mahalakshmi[1] and S. Sendhilkumar[2]

[1] Department of Computer Science and Engineering, Anna University
[2] Department of Information Science and Technology, Anna University
{gsmaha,thamaraikumar}@annauniv.edu

Abstract. Any researcher who is taking up a new research work must explore the works done in the past. For this we propose an idea to track the possible work progresses of a particular research article through semantic based approaches. In addition we analyze the co-citations and cross-citations among research works to avoid leaving out any significant works of the past. Finally we attempt to represent the citation networks and related meta-info using power graphs. This technique reduces the overhead of huge dimensions of citation networks thereby providing optimized trajectory representations which leads to finding significant research progress trajectories.

Keywords: citation, co-citation, trajectory, semantics, H-index, power graph, main path, key-route path, backward ideal path.

1 Introduction

An Integrated Approach for Tracking Work Progress is a bibliometric method capable of tracking the most significant paths of work progress and is also used to trace the developing trajectory of a research paper. The study provides valuable information for academic researchers and scholars to gain insights into development trajectory of the given research paper. It incorporates the integrated main path analysis approach, with which it helps to view the true path in which the research work has progressed. The outcome of the study shows the most significant work progress graph of the research paper. It also captures the works that are carried out in parallel and country contribution.

Many researchers have applied the main path analysis method since then to investigate the developmental trajectories of various science and technology domains using bibliographical citation data, patent citation data, or both [2, 4-8, 11]. Hummon and Doreian [9] addressed the same issue by introducing main path analysis. The method helps to escape from the maze by offering the most significant trajectories the main paths of a citation network. Methodologically, Batagelj [1] proposed a major advancement of the method, enhancing main path analysis by offering efficient algorithms for determining various versions of the significance index.

An earlier work on patent data [4] integratse social network analysis and main path analysis to explore the technological development trajectories of automatic

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 708–713, 2013.

identification techniques. The method, however, has some limitations, as it either offers a complex "network of main paths," as suggested in Hummon et al. [10], or one and only the most significant development path. For a large citation network, the "network of main paths" does not achieve the goal of simplifying the citation network. On the other hand, the one and only path, which limits our view to the target science or technology domain, is not satisfactory for explorers who are looking for more than the most significant development path. Further, the "priority first search" algorithm traces the most significant route at each juncture when several new ideas are competing. The result obtained may not be the path with the largest overall impact. Indeed, many close-second significant routes also can be neglected in the process.

In this paper, we attempt to apply power graphs over the semantically built citation network through which the 'bigness' of citation networks is reduced to a greater extent. We analyse the citation networks to evolve technological work progress which are semantically appealing unlike the earlier methods. We have suggested three different optimization approaches in finding main path analysis over semantic citation networks.

2 Forming Work Progress Trajectories

The citations of the seed paper are denoted as N, at the first generation. All the immediate first level citation papers are retrieved for the seed paper. The procedure is iterated to retrieve all the citation papers at all generations till 2012. Next to this, the corpus is subjected to preprocessing.. After these steps similarity between the text document are found using cosine similarity and topics are found using LDA algorithm. The cosine similarity value is determined for comparing seed paper and the citation papers. The average value of every level is also determined. Papers with similarity score greater than the average are only retained and the rest are eliminated. This dynamic threshold value is used to eliminate citation papers that do not semantically match with the seed paper. Citation papers that are semantically related to the seed paper only survive after this phase. Rest of the papers is filtered. The set of citation papers that are preserved after elimination is notated by N_p, where $N_p \leq N$. This is done for all levels.

Next we proceed to find out the co-citations of papers and trace all possible co-citations. To find the co-citations we used Cite Seer. This enables us to find the co-citations and compute the bibliographic coupling strength. For all the citation papers we determine the co-citations which is notated as M_{ij}, where i is a co-citation to j, i= 1 to n, j=1 to k where k is the no of co-citations in N_i. For all the co-citation papers, we perform the semantic analysis, get their concept maps and compare with the seed paper to generate a semantic index score and use a dynamic threshold value to determine the semantically matched citation papers. The set of co-citation papers that are preserved after elimination is notated by $N_{co\text{-}p}$.

Graph Node blending and Feature preservation are done in redundancy elimination phase. Level 2 Graph (G2) is an input to the redundancy elimination phase. The graph formed in level 2 will have redundancies. The graph may contain duplicate nodes which have to be removed. The graph may or may not contain duplicates. If duplicates are present Graph Node Blending has to be done. The papers which are

present in N and not present in N_p and has again appeared in $N_{co\text{-}p}$ whose features has to be preserved. Has a Output of this module we get Reduced (N_p,$N_{co\text{-}p}$). After redundancy elimination we get Reduced (N_p,$N_{co\text{-}p}$) set of papers for which cross citation analyses have to be performed (refer section: Results). For instance, if a paper occurs in citation levels,L2,L4,L7 in graph(G1) and the same paper occurs in co-citation levels CO1 and CO3 in the graph(G2),then the node is blended with all these information as L2,L4,L7,CO1,CO3 in G2.

We further proceed to include cross citations between papers in our corpus across the entire timeline. Cross citation can occur across levels as well. We use k-means clustering and link analysis for finding the cross citations. Clustering is done to group the papers according to the topics dealt in the papers. In each cluster the new citation links between the papers are found and results in increase of edges in the graph (G2), which leads to the graph (G3).

We construct a graph by considering the seed paper as the root node with the citation papers N_p as the child nodes with directed edges indicating the citation relationship with the seed paper. This is done at all generations. We form an acyclic graph that has directed edges and is devoid of cycles, initially to start the graph formation process. Nodes represent the papers and the edges represent the citation relationship between them. In G1, the nodes purely represent the citation papers whereas in G2, the nodes include both the citation and co-citation papers. In G2,there is an increase of nodes and edges present whereas in G3, no new nodes are added, but results in increase of edges.

1. Directed Acyclic Graphs (DAG's): We obtain the first graph G1 after semantic analysis with filtration using DAG. The citations of the seed paper at all generations are represented using directed acyclic graphs. Constructing DAGs ensures there are no cycles within the graphs. Consider C1, C2, C3 to be the citation papers of the seed paper, S. Let C1 be present at levels L1,L3 and C2 to be present in levels L4 and L5.Let C3 belong to level L2.These are represented in DAG using 5 citation nodes and a seed paper node. We call this graph, say G1.

2. Power Graphs: Power Graphs are used to reduce the dimensionality in graphs. C1 node can be blended with the information that it occurs in levels L1, L3 in a single powered node. Similarly the other citation papers C2 and C3 are represented using power nodes. Now the graph G1 can be represented using 3 citation nodes and a seed paper node, thus reducing dimensionality. The two conditions to be satisfied for a power graph are:
    - Power node hierarchy condition: Any two power nodes are either disjointed or one is included in the other.
    - Power edge disjointness condition: There is an onto mapping from edges of the original graph to power edges

Since we are representing the nodes as power nodes, they have information about which citation and co-citation levels it occurs. So each node is represented using multi colors, which denote the levels it occurs. The co-citations for the graph G1 is also brought into the graph representation. Since this involves increase in number of

nodes, the co-citation information is also blended within the power nodes. If M1 and M2 are co-citations of C1, 2 co-citation nodes are to be added to G1. But this co-citation information is blended in C1 power node thus bringing in co-citation information without adding further nodes. We call this graph as G2. The citations that are present between various clusters are analyzed in cross citation analysis. Suppose if C1 belongs to network cluster and C2 in software engineering cluster. The citations between them are analyzed after clustering the citation papers. After bringing in cross citation information in G2, we call the final graph as G3.

We use four approaches to track the work progress of a paper in the integrated approach. They are namely: Global Main Path, Backward Local Main Path, Multiple Main Path, and Key Route Main Path. The global main path searches forward from the source to the sinks and is traced using Priority First Search. This uses the Dijkstra's Algorithm, which finds the path with the lowest cost. The Backward Local Main Path is found using Reverse Priority First Search Algorithm. We trace the path backwards according to semantic scores from a recent paper until we reach the seed. For the nodes that are retained through optimization approach 1, we find the paths from leaf node towards seed node. Multiple Main Path uses Semantic Index Relaxation which basically relaxes the search constraint. The constraints are relaxed by bringing the next longest path also. Using the Key route Algorithm we take the most significant link and begin a search from the key rather than from the source or sink. This key is the research paper with a very close semantic score compared to the seed. It guarantees that this key route is included in the main path.

3 Optimising Work Progress Trajectories

The Paths obtained are ranked based on path length, node popularity, and relevancy with the seed paper. More Length indicates that more frequently the continuing research works have been carried out. Generally any significant research idea attracts more citations, hence is an indicative of good work progress (fetched through high citation count). The citation counts for all the nodes in a path are added and the average is taken as popularity score. Average of similarity scores of nodes in a path with respect to the seed paper, gives relevancy score which indicates more relevant work progress. The Paths found in the graph (G3) is quite large in number. Hence in order to improve the quality of the paths found we propose three optimization approaches and the results are discussed.

Approach 1: Matching Paths by Semantic score Average Filtration Level wise
In this approach, the semantic score for each node in the graph is determined from the cosine similarity algorithm in comparison with the seed paper. The nodes in each level are considered separately and their semantic score average is calculated for each level and the nodes whose semantic score are above each level's average are only considered and retained. Later paths are determined for the filtered nodes that offers better semantic match between the papers in the path and reduction in the number of paths between any nodes.

Approach 2: Using Semantic score Average Filtration Level wise
In this approach semantic score is considered for tracing the path from the seed till the leaf. It is a very narrow approach in which we consider only the nodes that survive the threshold based semantic filtration of first level. Next level nodes are considered only for the nodes that have survived the previous level filtration. Since the corresponding nodes of prior filtered nodes are alone considered, the number of nodes under optimization is reduced largely in number. It proceeds till the latest level generation, which is fourth level. The difference between approach 1 and approach 2 is that, in the first approach the filtration was done level wise before and then the link was found whereas in the second approach only the nodes which have survived the first level filtration are considered for the second level and hence forth. The path obtained thus obtained is narrow.

Approach 3: Using Semantic score and Popularity Average Filtration Level wise
In this approach we use two criteria's namely semantic score and popularity of the paper for narrowing the graph. The similarity between all papers of the first level is determined using cosine similarity algorithm. The average semantic score for the first level alone is taken into account and the papers that are above the average are retained. For the remaining levels we use the popularity score as criteria for reducing the paths. We take the average of the popularity for the second level and for all the remaining levels. The papers above average are retained and thus results in a fewer number of paths than before optimization.

Approach 4: Best Semantic Match Path
In order to find the best semantic match path, for all the nodes in the graph we find the cosine similarity score with the seed paper. Then we find the nodes with maximum semantic score at each levels. The best semantic path is the path which contains these maximum semantic score nodes.

4 Results and Discussions

The papers across seven generations are taken and a link analysis is made across papers. The seed paper [3] has totally 2706 and 2038 citations at the first and second generations. After the semantic scores are computed a dynamic threshold based filtration is performed. The optimized research progress paths are listed in Table 1.

Table 1. Optimization Approaches

Approach	Paths	From Node – To Node
Optimization I	26 Paths	000 - 56
Optimization II	000-174-167	000 - 180
Optimization III	000-174-167	000 - 167
Best	000-209-102-141-168	000 - 168

5 Conclusion

We have proposed semantics based approach to track the work progress of a research paper across timeline. We include co-citations and cross-citations to improvise the graph at every stage to obtain an efficient graph. We have also proposed different optimization methods to get better results. In the future, the download can be automated.

References

1. Batagelj, V.: Efficient algorithms for citation network analysis. Prep. Ser. Univ. Ljubljana, Inst. Math. 41(897), 1–29 (2003)
2. Carlero-Medina, C., Noyons, E.C.M.: 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 (2008)
3. Hirsch, J.E.: 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 (2005)
4. Lu, L.Y.Y., Lan, Y.L., Liu, J.S.: A novel approach for exploring technological development trajectories. In: 2012 IEEE International Conference on Management of Innovation and Technology (ICMIT), pp. 504–509 (2012)
5. Lu, L.Y.Y., Lin, B.J.Y., Liu, J.S., Yu, C.Y.: Ethics in nanotechnology: What's being done? What's missing? Journal of Business Ethics 109(4), 583–598 (2011)
6. Lucio-Arias, D., Leydesdorff, L.: Knowledge Emergence in Scientific Communication: From "Fullerenes" to "Nanotubes". Scientometrics 70(3), 603–632 (2007)
7. Mina, A.R., Ramlogan, G., Tampubolon, J.S.: Metcalfe,'Mapping Evolutionary Trajectories: Applications to the Growth and Transformation of Medical Knowledge'. Research Policy 36(5), 789–806 (2007)
8. Moore, S., Haines, V., Hawe, P., Shiell, A.: Lost in translation: A genealogy of the "social capital" concept in public health. Journal of Epidemiology and Community Health 60, 729–734 (2006)
9. Hummon, N., Doreian, P.: Computational methods for social network analysis. Social Networks 12, 273–288 (1990)
10. Hummon, N.P., Doreian, P., Freeman, L.C.: Analyzing the Structure of the Centrality-Productivity Literature Created Between 1948 and 1979. Science Communication 11(4), 459–480 (1990)
11. Verspagen, B.: Mapping Technological Trajectories as Patent Citation Networks: A Study on the History of Fuel Cell Research. Advances in Complex Systems 10, 93–115 (2007)