Journal of Informetrics 16 (2022) 101281

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A semantic main path analysis method to identify multiple developmental trajectories

Liang Chen[a], Shuo Xu[b,*], Lijun Zhu[a], Jing Zhang[a], Haiyun Xu[c], Guancan Yang[d]

[a] Institute of Scientific and Technical Information of China, Beijing 100038, PR China
[b] College of Economics and Management, Beijing University of Technology, Beijing 100124, PR China
[c] Business School, Shandong University of Technology, Zibo 255000, PR China
[d] School of Information Resource Management, Renmin University of China, Beijing 100872, PR China

ARTICLE INFO

Keywords:
Main path analysis
Developmental trajectory
Patent mining
Topic coherence
Lithium-ion battery

ABSTRACT

Main Path Analysis (MPA) is widely used to trace the developmental trajectory of a technological field through a citation network. The citation-based traversal weight is usually utilized to cherry-pick the most significant path. However, the theme of documents along a main path may not be so coherent, and it is very possible to miss the main paths of significant sub-fields overall in a domain. Furthermore, the global path search algorithm in conventional MPA also suffers from high space complexity due to the exhaustive strategy. To address these limitations, a new method, named as semantic MPA (sMPA), is proposed by leveraging semantic information in two steps of candidate path generation and main path selection. In the meanwhile, the resulting source code can be freely accessed. To demonstrate the advantages of our method, extensive experiments are conducted on a patent dataset pertaining to lithium-ion battery in electric vehicle. Experimental results show that our sMPA is capable of discovering more knowledge flows from important sub-fields, and improving the topical coherence of candidate paths as well.

1. Introduction

Historical reconstructions of scientific and technological developments are important tools for scheming future R&D plan, capturing significant opportunities, forecasting the trends of future technology evolution, etc. Especially to beginning researchers, they can serve as effective means to gain a proper understanding of the constantly updating field (Xu, Hao, An, Pang & Li, 2020). In early empirical studies, qualitative methods are mainly used, such as literature review, document reading and interview to obtain a discipline's development (Kim & Shin, 2018). However, such methods suffer from subject bias, bounded knowledge (Kim & Shin, 2018) and time consuming given the increasing volumes of scientific publications.

With the drastic improvement in computational techniques, a series of automatic methods (Hummon & Doreian, 1989; Leydesdorff, Bornmann, Marx & Milojević, 2014; Tan et al., 2016) have been proposed for discovering developmental trajectories of an interested field. Among these methods, the main path analysis (MPA), pioneered by Hummon and Doreian (1989), is the most well-known. The MPA takes a citation network to represent the diffusion relations among the pieces of knowledge in individual documents, and citation-based traversal counts (Hummon & Doreian, 1989; Batagelj 2003) as a "proxy" for knowledge significance. Then, a skeleton structure, consisting of documents which play an important role in the development of a field, is extracted from the citation network. This skeleton structure is usually called as the main paths, through which one can gain valuable insights into the main advancements in a specific field.

[*] Corresponding author.
E-mail address: xushuo@bjut.edu.cn (S. Xu).

https://doi.org/10.1016/j.joi.2022.101281
Received 27 July 2021; Received in revised form 10 March 2022; Accepted 17 March 2022
Available online 6 April 2022

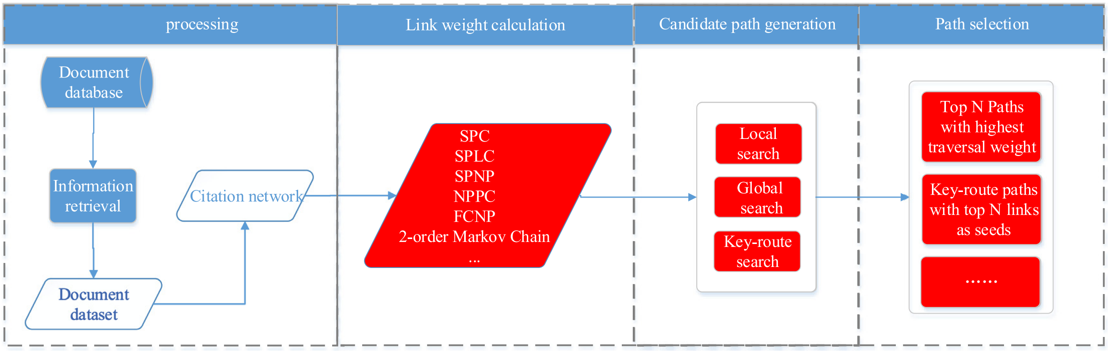


Fig. 1. The basic procedure of conventional MPA.

The software *Pajek* (Batagelj & Mrvar, 2004) has implemented several MPA-related functionalities, which makes the analysis accessible for the academia. Since then, the MPA has been successfully applied to multiple domains (Liu & Lu, 2012; Liu, Lu, Lu & Lin, 2013; Huang et al., 2017; Xu et al., 2020; Yu & Sheng, 2020). It is well known that the development in one technological field is often driven by the advancements of multiple sub-fields at the same time. But, since the MPA mainly cherry-picks one or several main paths on the basis of citation-based traversal counts, it is very possible that the evolutionary pathways of some sub-fields are missed. According to our observations (cf. Section 4), the paths with high traversal weights tend to focus on a certain theme. In other words, the top paths with highest traversal weight are not guaranteed to cover the main paths of significant sub-fields overall in a domain.

In addition, due to exhaustive search strategy for main paths, the MPA does not scale directly to a very large citation network. Hence, current empirical studies in the literature are just limited to the small-scale networks. In the meanwhile, the MPA uses some topological indicator to weight citation relations while ignoring semantic information attached to each vertex, which leads to topical inconsistency for documents along a same main path and reduces the interpretability of the resulting path (Yeo, Kim, Lee & Kang, 2014; Tu & Hsu, 2016). Last but not least, until now it is still very difficult to find an open-source implementation version of the MPA, which is not conducive to follow-up further studies and improvements.

For the purpose of dealing with these limitations, a semantic main path analysis (sMPA for short) approach is proposed in this study to identify simultaneously multiple developmental trajectories in a target field. Each trajectory corresponds to the resulting evolutionary pathway of a significant sub-field. In more details, the topological structure based link weights are combined with the semantic information based ones to weight each link in the first place. Thus, topical coherence of documents along a same path can be improved dramatically. Then, to pick up multiple trajectories of significant sub-fields, a density-based clustering algorithm is utilized to divide the relevant paths into several groups after all paths are enumerated effectively with a dynamic programming based search algorithm. The following summarizes main contributions of this work:

- A semantic main path analysis approach is put forward to identify simultaneously multiple developmental trajectories in a target field.
- To improve topical coherence of documents along a same trajectory, the conventional link weights are armed with the semantic information based ones.
- After all paths are enumerated effectively with a dynamic programming based search algorithm, a density-based clustering method is used to divide them into several groups.
- The source codes in Python language can be freely accessed at the GitHub[1] and PyPi[2] with detailed API documentation[3], thus to promote the related studies.

The organization of the rest of this paper is as follows. After related work is briefly reviewed in Section 2, a framework utilizing semantic information attached to each vertex of the citation network to improve MPA is put forward in Section 3 and core modules are also described in more details in this section. Section 3 conducts extensive experiments on the field of lithium-ion battery in electric vehicle, in which the conventional MPA is taken as a baseline to demonstrate the advantages of our methodology. The last section concludes this contribution with future directions.

2. Related works

Given a citation network, the basic procedure of MPA is illustrated in Fig. 1. After the weight of each link is calculated, candidate paths between source vertices and sink ones are generated with a search algorithm. By source vertex, we mean the vertex that is

[1] https://github.com/awesome-patent-mining/sMPA-documentation.

[2] https://pypi.org/project/sMPA/1.0.

[3] https://awesome-patent-mining.github.io/sMPA-documentation.

cited while referring to no other vertices and the sink vertex is the reverse. The last step of MPA is to select the sequence with the largest sum of link weights or multiple sequences meeting certain criteria as the final main path(s). Before delving into more specifies, several key modules in Fig. 1, as shown in red box, will be briefly reviewed in the following subsections.

2.1. Link weight calculation

As early as Hummon and Doreian (1989) developed three traversal counts to weight the resulting links in their seminal work: search path link count (SPLC), search path node pair (SPNP) and node pair projection count (NPPC). Later, Batagelj (2003) proposed an efficient algorithm to compute these traversal counts, and another traversal count, search path count (SPC), was raised. Due to the following nice properties (Batagelj, 2003), the SPC has become the first choice of many scholars (Tu & Hsu, 2016; Huang et al., 2017; Kim & Shin, 2018; Yu & Pan, 2021): (1) The SPC takes into all direct and indirect citation-links considerations, and (2) compared to the other traversal counts, the SPC has the lowest time complexity.

To understand the difference between these traversal counts, through a delicate analogy of messenger and tollway, Liu, Lu & Hu (2019, 2020) found that the SPLC should be the closest to the real-world knowledge diffusion scenario. They argued that the intermediates should not only pass knowledge flow through but also add new knowledge into it, while in the SPC the intermediates just passed the knowledge flow and the SPNP considered intermediates to be knowledge depositories. As for the NPPC, it is rarely used in actual applications due to its high computational complexity.

In this way, more and more scholars turn to the SPLC for calculating link weights, such as Huang et al. (2021), Xu et al. (2020), and Lai, Chen, Chang, Kumar, & Bhatt (2021), though it has been shown that the MPAs based on different traversal counts will produce almost the same results (Batagelj, 2003; Martinelli, 2012). Recently, the following concerns on these traversal counts emerge in the literature: (1) they cannot reflect the information loss as the knowledge flows through the citation network (Liu & Kuan, 2016), and (2) the documents tend to involve mixed topics along a same path (J. Kim & Shin, 2018; M. Kim, Baek, & Min, 2018). For the sake of alleviating these concerns, various alternatives of link weights are proposed, such as forward citation node pair (FCNP) (Choi & Park, 2009), search-path arithmetic decay (SPAD), search-path geometric decay (SPGD) and search-path harmonic decay (SPHD) (Liu & Kuan, 2016). On the other hand, Kim, Baek & Min (2018) measured the link weight by topical similarity between a citing document and a cited one as well as the relative importance of the documents.

2.2. Candidate path generation

Once the weight of each link is ready, the next step is to generate candidate paths between source vertices and sink ones. To facilitate the narrative, this step is referred to as candidate path generation hereinafter. In the literature, there are two main strategies for path search, namely greedy search and exhaustive search. The former starts at a source/sink vertex and searches a citation sequence (path) using a greedy strategy (Yeo et al., 2014), while the latter exhaustively runs through all possible paths in a citation network to find the sequence that is the highest in term of sum of link weights (Verspagen, 2007).

Since greedy search does not guarantee to result in a globally optimal path, it is named as local search strategy. Similarly, the exhaustive search is named as global search strategy. According to the search direction (from source vertex to sink one or reverse), these strategies can be further divided into forward and backward local/global search (Liu et al., 2013). Liu et al. (2013) observed that the main paths from both the local and global search strategies might not include the citations with the highest link weight. Thereupon, they suggested an alternative strategy: key-route search. This search strategy begins with several seed citation-links, viz. the citations with top link weights, and then searches forward until a sink vertex is hit and backward until a source vertex is hit.

In addition, Yeo et al. (2014) found that prior studies heavily depended on the current information at the moment of selecting the next vertex. This tends to yield main paths with mixed themes. Thereupon, Yeo et al. (2014) suggested a path search algorithm on the basis of 2nd-order Markov chains. Tu & Hsu (2016) merged the papers on the main path with similar topics to form a conceptual path. Along this direction, Kim et al. (2018) further integrated PageRank (Page, Brin, Motwani & Winograd, 1999) with the Citation Influence Model (CIM) (Dietz, Bickel & Scheffer, 2007; Xu, Hao, An, Yang & Wang., 2019) to identify multiple evolutionary pathways in the field of protein p53 with improved topical inheritance.

2.3. Path selection

The early studies of MPA simply selected the path with the longest in term of length or the highest in term of sum of link weights as the main path in a focal citation network (Yeo et al., 2014). The constraint of a single path limits the explorers from expanding the perspective in exploring a research domain (Liu & Lu, 2012), and the path suffers from missing some important vertices, citations and paths (Kim & Shin, 2018). To loosen this limitation, Verspagen (2007) retrieved the top paths with the highest overall traversal weight to construct a network. Fontana, Nuvolari, & Verspagen (2009) further extended the main path network by merging the second- and third-highest paths. Since the network not only contains the path with the highest link weight sum, but also those with a link weight sum falling within certain criteria, Liu and Lu (2012) named the new method as multiple MPA.

Similar to forward and backward search strategies in the MPA, the citations with the highest link weight may not be included in the main paths from multiple MPA. As a remedy, Xiao, Lu, Liu and Zhou (2014) introduced the key-route path search algorithm in multiple MPA. In more details, the links with the top weights are retrieved as the seed links, then key-route path search is conducted for each seed link and the resulting paths are combined together to generate multiple main paths. Since multiple MPA with the

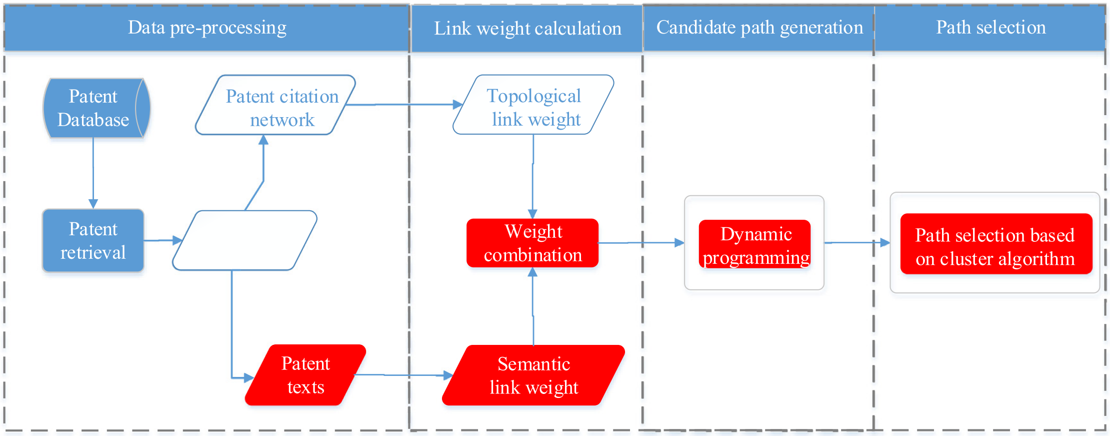


Fig. 2. The procedure of sMPA to identify multiple developmental trajectories.

key-route search strategy is capable of revealing more details of the progress of a scientific or technological domain, it has received widespread attention (Xiao, Lu, Liu, & Zhou, 2014; Yu & Pan, 2021; Xu et al., 2020).

On the other hand, Kim & Shin (2018) and Yu & Pan (2021) focused on technology junctures where important derivative paths emanating from a main path. More specifically, the citation network is first divided into several sub-networks with a community detection algorithm. Then, the main sub-paths from each sub-network are retrieved with the conventional MPA, and linked to the overall main path for juncture analysis. Another strategy was adopted in Martinelli (2012) to obtain main sub-paths by fixing the starting year while changing the ending year for each time period. Thus, the significant developments in different periods can be observed.

2.4. General remarks

Although the MPA is a powerful tool to identify developmental trajectory in an interested field, it suffers from a fundamental limitation that overemphasizing the topological structure of a citation network while ignoring topical coherence between two linked documents. Consequently, the main path tends to involve mixed topics and fails to show what topic evolves along a path, which is an essential part for knowledge flow identification (Kim et al., 2018). Recently a series of methods are proposed to alleviate the problem, such as 2nd-order Markov chains and topic model, but they need frequent intervention of experts to select main paths from candidate ones. In other words, until now an objective and comprehensive method is still missing for identifying multiple developmental trajectories with improved topical coherence.

To this end, we propose a new MPA framework which integrates semantic similarities with link weights to improve the topical coherence on a same path. Then, a cluster algorithm is applied to identify the resulting main paths that best represent the developmental trajectories of different subareas from a global perspective. To scale to a large-scale citation network, we re-implement global search algorithm by combining broad-first search algorithm and dynamic programming strategy. In this way, the space complexity for the citation network with n vertices is reduced from $\mathcal{O}(n^2)$ in the conventional MPA to $\mathcal{O}(n\log n)$ in our framework. For the accessibility and application of this framework, a corresponding suite is developed in Python language and open sourced in Github.com. To the best of our knowledge, this is the first suite of the MPA with released source codes.

3. Research framework and methodology

Fig. 2 is our research framework for identifying multiple developmental trajectories in an interested domain. Similar to the conventional MPA in Fig. 1, the whole procedure consists of four phases: data pre-processing, link weight calculation, candidate path generation, and path selection. However, different from the conventional MPA, we induce semantic information into the link weight calculation to improve the topical coherence of documents along a same path. Apart from this, the path search algorithm and main path selection strategy are optimized as well. The modules with red box in Fig. 2 will be detailed in the following subsections.

3.1. Link weight calculation

To improve the topical coherence of documents along a same path, the conventional link weights on the basis of topological structure (cf. Section 2.1) are combined with semantic similarity between the documents. Formally, the weight for the link between the vertices i and j is defined as follows.

$$weight\left(link_{i,j}\right) = \alpha * weight_s\left(link_{i,j}\right) + (1-\alpha) * weight_t\left(link_{i,j}\right) \tag{1}$$

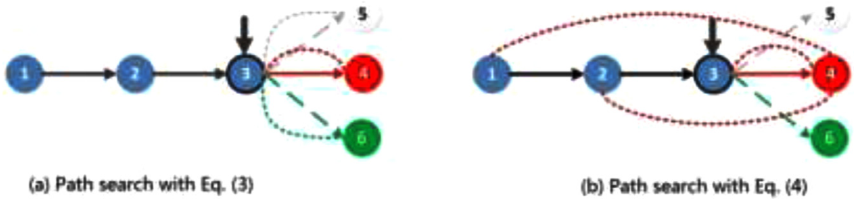


Fig. 3. Path search with different equations.

From Eq. (1), one can see that the weight of $link_{i,j}$ is composed of two parts, namely semantic link weight $weight_s(link_{i,j})$ and topological link weight $weight_t(link_{i,j})$. The former is obtained by computing the semantic similarity between documents attached to the vertices i and j. To be specific, the dense vectors generated by Latent Semantic Index (LSI) (Foltz, 1990) are used to represent documents to avoid the sparsity problem caused by one-hot representation. Then, the cosine similarity is utilized to measure semantic closeness between two documents due to its superior performance in measuring textual similarity and low calculation workload (Boyack et al., 2011). Any link weight calculation method in Section 2.1 can serve directly as the later. Though, the SPC is chosen in this study as the topological link weight. Here, α is a hyper-parameter ranging from 0 to 1 to adjust the relative importance between semantic and topological link weights.

Correspondingly, the calculation of path traversal weight is adapted as well. The traditional MPA calculates the traversal weight of a path by adding up the weights of all links along it, as shown in Eq. (2). When applied to combined weight [cf. Eq. (1)], the weight of a path can be obtained with Eq. (3). However, according to our observations, Eq. (3) still may cause semantic drift problem in path search process, viz. the topic of sink vertex rapidly drifts from that of the source one. In our opinion, the main reason is that Eq. (3) only considers direct citation-links while omitting indirect ones. Let us take the toy path in Fig. 3 (a) as an example. When path search algorithm determines the successor of vertex 3, only semantic similarities between vertex 3 and its forward vertices of 4, 5 and 6, as marked in dotted line, are utilized to maximize the path weight. However, the successor, vertex 4 in this case, cannot guarantee its topical coherence to vertices 1 and 2. In other words, even though the textual content between vertices 1 and 2, 2 and 3, 3 and 4 are very relevant to each other, it is very possible that the vertices 1 and 4 discuss different topics.

Therefore, to consider semantic similarities of both direct and indirect citation-links for the path search, we change Eqs. (3), (4) in this study. In this way, the topological weight of a path is calculated just same as in Eq. (3). But, for the semantic weight of a path, all possible vertex pairs are enumerated on the path and then their semantic similarities are added up together. When a new vertex, such as vertex 4, is added to the path in Fig. 3 (b), the similarities of this vertex and every other vertex, as shown in red dotted line, are also taken into consideration to guarantee that any two documents on the path are closely related to each other.

$$W_{path,\,t} = \sum_{link_{i,j} \in path} weight_t\left(link_{i,j}\right) \tag{2}$$

$$W_{path,c} = \sum_{link_{i,j} \in path} \alpha * weight_s\left(link_{i,j}\right) + \sum_{link_{i,j} \in path} (1-\alpha) * weight_t\left(link_{i,j}\right) \tag{3}$$

$$W_{path,c} = \sum_{vertex_i \in path} \sum_{\substack{vertex_j \in path \\ i>j}} \alpha * weight_s\left(vertex_i, vertex_j\right) + \sum_{link_{i,j} \in path} (1-\alpha) * weight_t\left(link_{i,j}\right) \tag{4}$$

Since topological weight and semantic weight follow different calculation rules in Eq. (4), their values may be in different magnitudes. Hence, before combining them, these two weights are linearly normalized with Eqs. (5)–(7) to the interval [0, 1], where $\max(W_{path,\cdot})$ and $\min(W_{path,\cdot})$ indicate the maximum and minimum of semantic or topological weights. To highlight its distinction from Eq. (3), the hyper-parameter α is also replaced by another symbol β to control the tradeoff between normalized topological and semantic weights.

$$W_{path,c} = \beta * \frac{W_{path,\,s} - \min\left(W_{path,s}\right)}{\max\left(W_{path,s}\right) - \min\left(W_{path,s}\right)} + (1-\beta) * \frac{W_{path,t} - \min\left(W_{path,t}\right)}{\max\left(W_{path,t}\right) - \min\left(W_{path,t}\right)} \tag{5}$$

$$W_{path,s} = \sum_{vertex_i \in path} \sum_{\substack{vertex_j \in path \\ i>j}} weight_s\left(vertex_i, vertex_j\right) \tag{6}$$

$$W_{path,t} = \sum_{link_{i,j} \in path} weight_t\left(link_{i,j}\right) \tag{7}$$

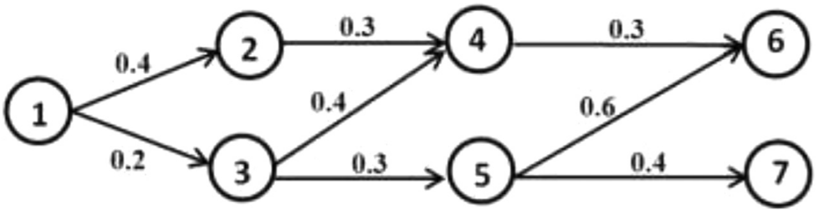


Fig. 4. A running example of citation network.

Note that we suggest a smaller value for β in the sMPA, since the topological structure of a citation network embeds the trajectories of technological developments. This is an implicit basic assumption of the MPA for identifying developmental trajectories. That is, the importance of a path should be mainly measured by its topological weight rather than semantic weight. The semantic weight just aims at improving the path's topical coherence.

3.2. Candidate path generation

Though prior studies seldom disclosed the algorithmic details of the MPA, it is not difficult to infer from graph traversal algorithms that the main process consists of two stages, (1) obtaining all the source/sink vertices, and (2) for each source/sink vertex, searching candidate paths with the greedy or exhaustive strategy, in which the former achieves the locally optimal path and the latter the globally optimal path. Thus, for a citation network with n vertices, the space complexity is $\mathcal{O}(n^2)$ for the exhaustive strategy. This means that out of memory problem may occur when identifying globally optimal path in a large-scale network.

To overcome such a limitation, a new algorithm for candidate path generation is developed here by combining dynamic programming strategy with breadth first search algorithm (hereinafter referred to as DP-BFS). Similar to conventional procedure, all source/sink vertices in a citation network are collected in the first place, and then the DP-BFS is applied to each source/sink vertex. Since the dynamic programming is adopted for global search instead of the exhaustive strategy, the space complexity is reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\log n)$. Please refer to Appendix 1 for more details on the derivative of space complexity. The pseudo-code of DP-BFS algorithm is shown as follows.

Input: A citation network G and a source vertex s of G.

Output: The path led by s with the highest sum in link weights.

```
(1) The procedure DP-BFS ($G$, $s$)
(2)   let $Q$ be a queue, $D$ be a dictionary
(3)   label $s$ as discovered
(4)   $Q$.enqueue ($s$), $D$.setKeyAndValue ($s$, $s$)
(5)   while $Q$ is not empty do
(6)     $v$:= $Q$.dequeue ()
(7)     for all $u$ in $G$.adjacentVertices ($v$) do
(8)       if $u$ is found in $D$.keys () then
(9)         $p$:= $D$.getValueByKey ($u$)
(10)        $p\_pre$:= D.getValueByKey ($v$)
(11)        if $W_p < W_{p\_pre} + W_{v \to u}$ then
(12)          append $v \to u$ to $p\_pre$
(13)          $D$.setKeyAndValue ($u$, $p\_pre$)
(14)          for $w$ in $D$.keys () do
(15)            if $u$ in $D$.getValueByKey ($w$) then
(16)              D.updateValueByKey($u$)
(17)      else
(18)        $Q$.enqueue ($u$)
(19)        $p$:= D.getValueByKey ($v$)
(20)        append arc $v \to u$ to $p$
(21)        $D$.setKeyAndValue ($u$, $p$)
(22)  return path with the highest traversal weight in $D$.values()
```

For ease of understanding our DP-BFS algorithm, Fig. 5 details the involving steps on the citation network in Fig. 4. In the meanwhile, to keep things simple, only topological weight, as a label attached to the resulting link in Fig. 4, is used here. It is worth noting that one can easily switch back to the full weight version just by replacing Eq. (2) with Eq. (5) while keeping other things unchanged.

Step 1 (Line 2): To initialize an empty queue Q and an empty dictionary D;

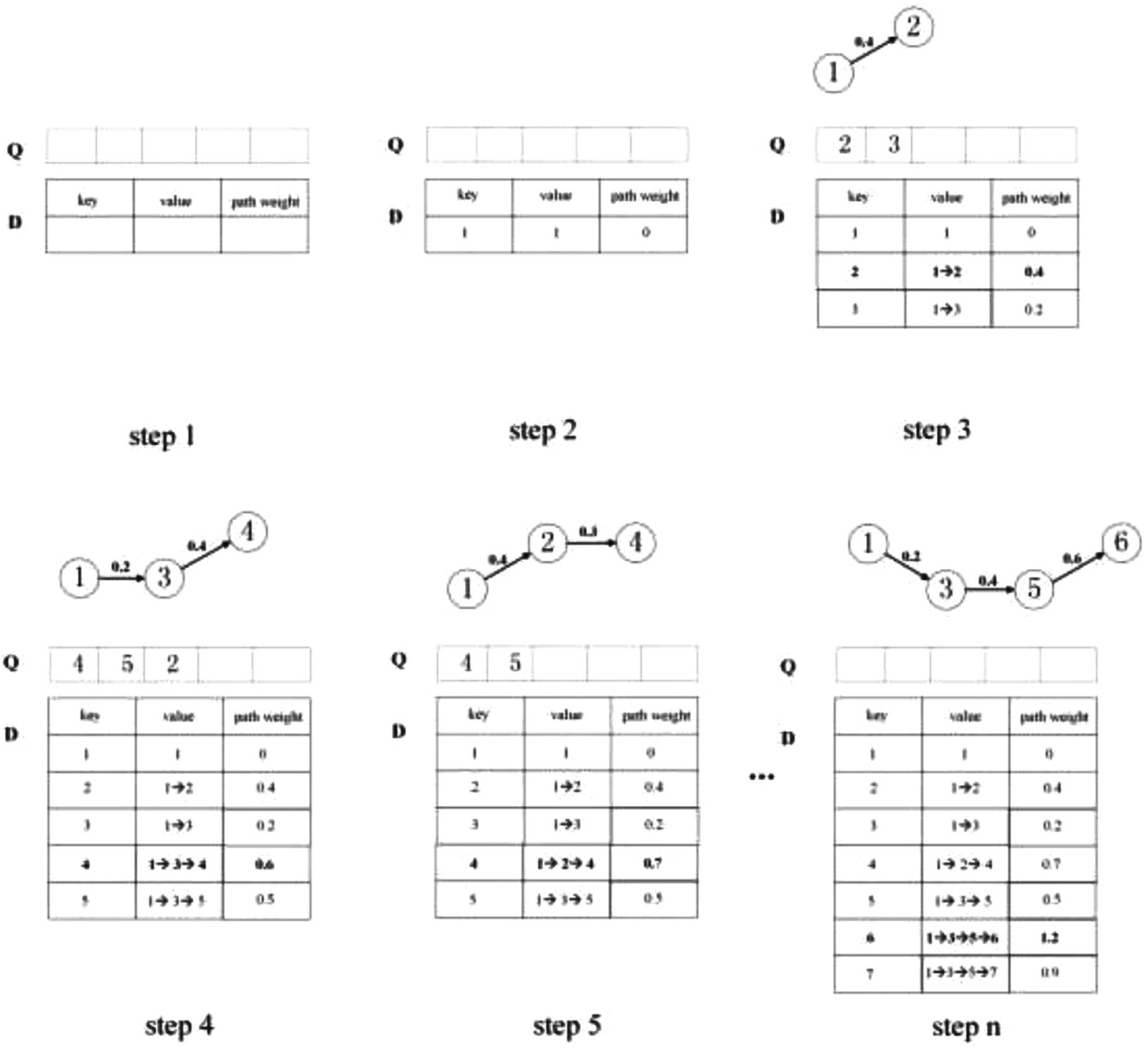


Fig. 5. The procedure of DP-BFS algorithm in the citation network in Fig. 4.

Step 2 (Line 3–7): To start traversing the citation network from vertex 1 and explore all of the neighbor vertices at the present depth, namely vertices 2 and 3;
Step 3 (Line 8, 17–21): Since vertices 2 and 3 are not in the key set of *D*, (2, 1→2), (3, 1→3) are put into *D*;
Step 4 (Line 5–8, 17–21): To move on to vertex 3 and explore its neighbor vertices 4 and 5. Then, (4, 1→3→4) and (5, 1→3→5) are put into *D*;
Step 5 (Line 5–16): To move on to vertex 2 and explores its neighbor vertices at the next depth level, namely vertex 4. As the vertex 4 has already been in *D* as a key, one has to compare the weight of current path 1→2→4 (0.7) to its counterpart 1→3→4 with weight 0.6 in *D*. Since the weight of the path corresponding to key 4 in *D* is lower than that of current path, the value of key 4 is updated by the new path 1→2→4.
Step n: To iterate the procedure above;

Output (line 22): To output the path with the highest sum in link weights, which turns out to be 1→3→5→6 with weight 1.1.

3.3. Main path selection

The final step is to select representative paths of each significant sub-field overall in the research area. Previous studies mainly choose the paths with top traversal weights (Hummon & Dereian, 1989; Liu & Lu, 2012; Batagelj, 2003), but such criterion is incapable of identifying the paths from different sub-fields. As a remedy, a cluster-based method for main path selection is proposed, which consists of three steps as follows: (1) Each candidate path is represented with a vector reflecting its semantic information; (2) A cluster algorithm is utilized on the candidate paths to group them into different sub-fields; (3) One can choose a path meeting certain criteria from each sub-field as its main path. It is worth mentioning that the method for main path selection is only a basic procedure. For a specific application, it is necessary to determine the appropriate techniques and criteria. In addition, the results of some clustering algorithms, such as *k*-means clustering (Hartigan and Wong, 1979) and Gaussian mixture model (Renolds, 2015), dependent on random initialization methods. To say it in another way, it is very possible to obtain different clustering results for multiple runs. So we suggest more robust clustering algorithms when using our sMPA, such as density peak clustering (Rodriguez & Laio, 2014). This algorithm assumes that the centroid of each cluster is surrounded by neighbors with lower local density and at a relatively large distance from any points with a higher local density. By computing local density and distance from higher density for each data point, the correct number of clusters and their centroids can be found automatically.

4. Experimental results and discussions

4.1. Dataset

For case study, a patent dataset pertaining to lithium-ion battery in electric vehicle is built via expert intervention. In more details, the search strategy in Zhang, Li and Wu (2017) is used here to retrieve patents related to electric vehicle batteries from Derwent Innovation Index database. Then, the patents on lithium-ion battery are filtered by domain experts and further extended through forward and backward citations. To merge the closely related patents derived from the same core technology but issued by authorities in different countries, the patent family is taken as the vertex of a citation network (Huang et al., 2017, Huang et al., 2021). In addition, we remove the patents published after April, 2010 to analyze the development of lithium-ion battery technique in its R&D stage (Yong, Ramachandaramurthy, Tan & Mithulananthan, 2015), since films in this stage strived to find a dominant design by introducing a number of alternative designs (Fernando & James, 1995). In the end, the number of patent families in our dataset is 13,401, and the publication times range from June 1973 to April 2010.

To calculate the semantic similarity between different vertices, the first published patent in each patent family, namely basic patent[4], is taken to represent the family and its abstract field to represent each vertex's textual content[5]. In addition, the Derwent Primary Accession Number of a patent family, viz. GA, is utilized to label the resulting vertex. Similar to Hummon and Dereian (1989) and Xu et al. (2020), the largest weakly connected component is taken here for semantic main path analysis. The largest weakly connected component in our dataset consists of 3603 patent families, in which there are 1248 source vertices, 1085 intermediate vertices and 1270 sink vertices. It can be readily seen that our largest weakly connected component includes a large share of source and sink vertices. We argue that main reason should be attributed to our extended operation through backward and forward citations during dataset construction. Fig. 6 illustrates the distribution of patent families with publication year. From Fig. 6, one can see that though the lithium-ion battery in electric vehicle appeared as early as 1975, it began to develop rapidly after 1990. This is consistent with the observation in Yong et al. (2015) that governments have implemented regulatory actions to reduce air emissions and promote electric and hybrid vehicles development since 1990's.

4.2. Hyper-parameter tuning and candidate path generation

For improvement of topical consistency, combined weight is used for candidate path generation. With the DP-BFS algorithm (cf. Section 3.2), 1248 candidate paths can be obtained from our citation network. Note that each path is attached with three types of path weights: topological path weight, semantic path weight and combined path weight in Eq. (5), where combined weight is yielded directly from DP-BFS algorithm, and the topological weight and semantic weight are calculated by Eqs. (6) and (7) correspondingly. Afterward, the normalized combined weight is obtained by Eq. (5). We still follow the assumption of prior MPA, that is, the topological weight of a path indicates the amount of knowledge passing through it. To enable the sMPA to reflect the main knowledge flow in the citation network, the topological weight of its main path should be close to that of the prior MPA.

To tune the hyper-parameter β, we use a step size of 0.01 to generate 101 candidate values for β in the interval [0.0, 1.0]. The trends of maximums and averages of topological, semantic, and combined weights of candidate paths are illustrated in Fig. 7. Several interesting phenomena can be observed:

(1) The maximal topological path weight stays unchanged for β in [0.0, 1.0] from Fig. 7 (a). This indicates that the corresponding paths are almost the same (cf. the three versions of path 1 in Fig. 11). On the other hand, the maximal semantic path weight shows similar behavior pattern, except two critical points appear in its curve where path weights vary significantly.

(2) As β increases, the average topological path weight declines gradually while the average semantic path weight rises gradually in Fig. 7 (b). When $\beta \geq 0.2$, two curves tend to be stable.

(3) In Fig. 7 (c), the curves of combined path weight are not so monotonic. More specifically, the maximal combined path weight equals to 1 when $\beta < 0.34$, then declines until reaching the bottom of 0.983 at $\beta = 0.44$ and rises back to 1 at $\beta = 1$. While the average combined path weight increases at the beginning until reaching a peak of 0.067 at $\beta = 0.04$, after that it gradually decreases to 0.021 at $\beta = 1$.

[4] https://dialog.com/derwent-world-patents-index/.

[5] The citation network can be accessed at the following link: https://github.com/awesome-patent-mining/sMPA-documentation/tree/master/experimental_data.

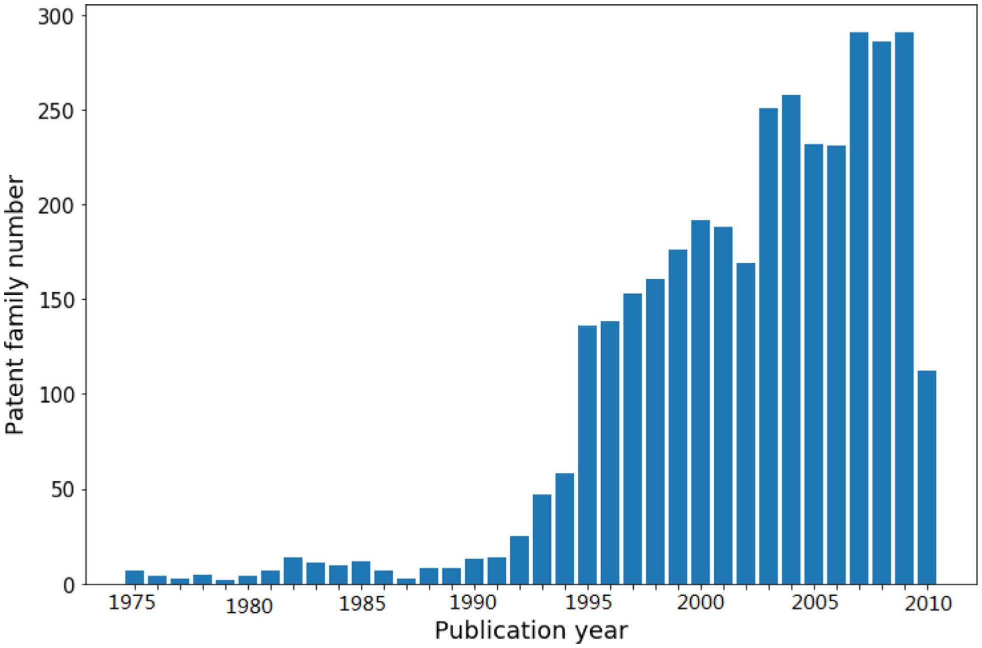
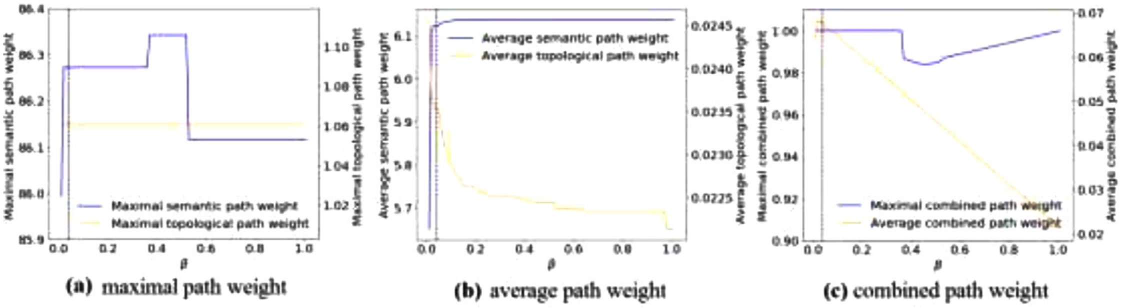


Fig. 6. The distribution of patent families with publication year.



(a) maximal path weight (b) average path weight (c) combined path weight

Fig. 7. The trend of average topological and semantic path weight.

Through comparison, three values for β (viz. 0, 0.05, 1) are adopted for further analysis, where 0 and 1 are the extreme cases that only topological weight or semantic weight are in action in candidate path generation. When β=0.05, as marked in dotted vertical line in Fig. 7, the change of weights for most types of path is relatively small. For purpose of obtaining valuable insights, the statistics of corresponding candidate paths for $\beta \in \{0, 0.05, 1\}$ are shown in Fig. 8. Obviously, semantic, topological and combined weights and path lengths of the candidate paths demonstrate a similar pattern for different values of β, which is conforming to power-law distribution. That is, the vast majority of candidate paths have lower weights and fewer vertices, but several candidate paths with higher weights and more vertices. For illustration, the case of $\beta = 0.05$ is taken as an example. Specifically, the semantic weights of candidate paths are distributed between 0.2 and 86.3, and only 10 paths with weights above 60 while paths with weights below 10 account for 84% of the total paths [cf. Fig. 8 (a)]. As for topological weights in Fig. 8 (b), not only the value range is much smaller than that of semantic weight, but also the weight distribution is more skewed. There are only 9 paths with weights above 0.6 while 1184 paths with weight below 0.1 occupy 94.9% of the total paths. A similar pattern can be observed for combined weights in Fig. 8 (c). As for path lengths, there are only 19 paths with length greater than 11, including 3 path with length 15 and 14, 5 paths with length 13 and 8 paths with length 12, while 88.4% candidate paths with length less than or equal to 6 [cf. Fig. 8 (d)].

4.3. Main path selection

With the characteristics of the dataset and the requirements of the downstream tasks in concern, we present an implementation of main path selection in Section 3.3 as follows. After each document in the citation network is vectorized with the Latent Semantic Index (LSI) (Foltz, 1990), the vectors of all documents along a candidate path are added up element-wise and normalized to represent the path. Then, the candidate paths are clustered into different sub-fields with the density peak clustering method

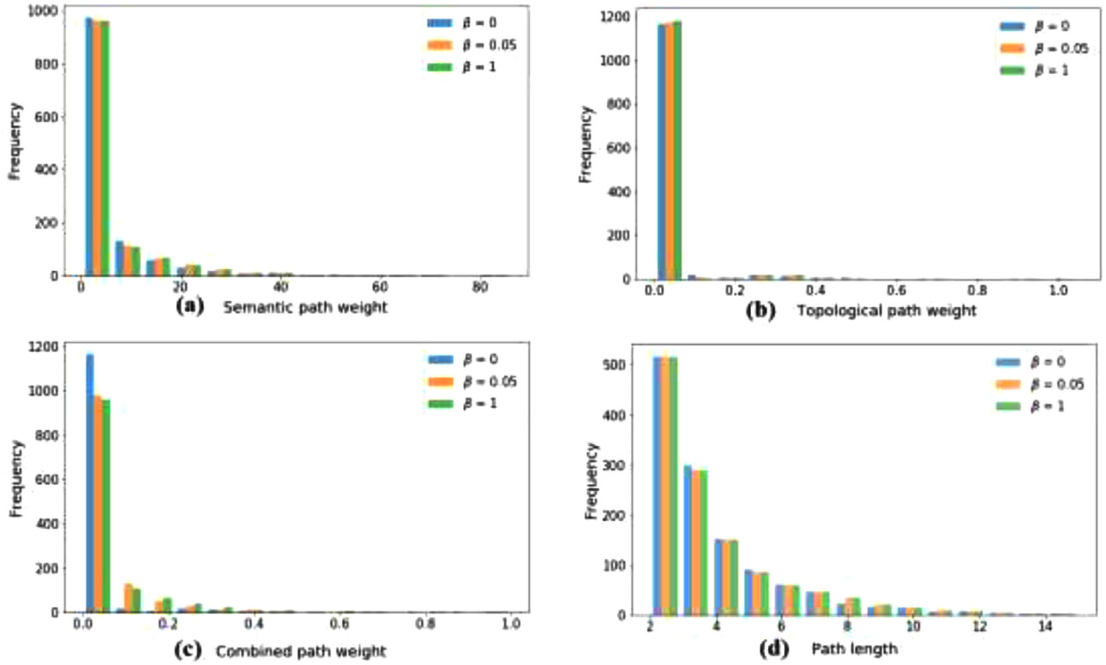


Fig. 8. Statistics of candidate paths for different βs.

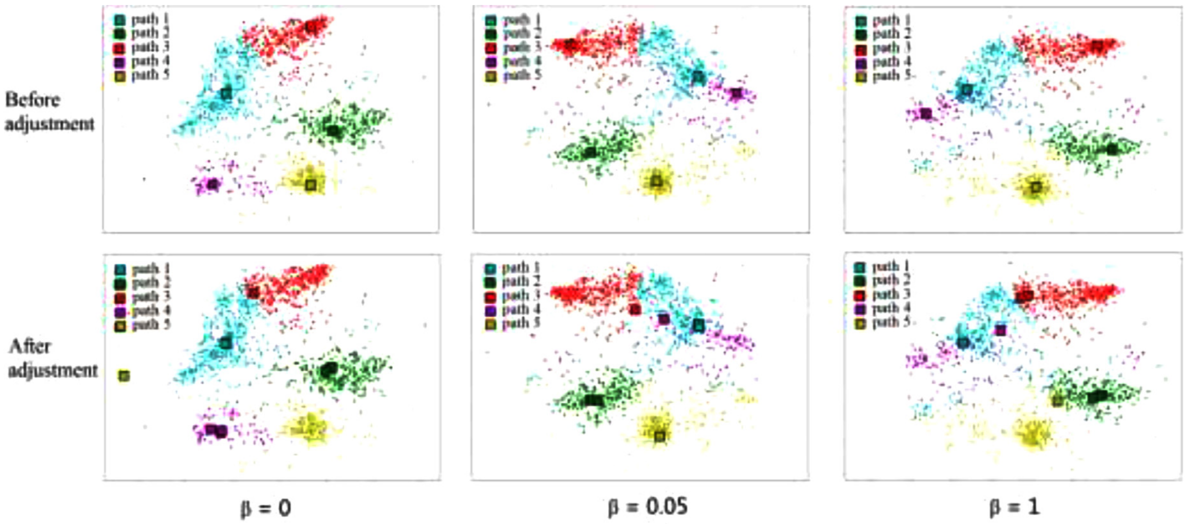


Fig. 9. The semantic distribution of candidate path in path clusters.

(Rodriguez & Laio, 2014). This density-based method can automatically find the correct number of clusters (Xu et al., 2016) and output the centroid of each cluster. To ensure the consistency of clustering results for different values of β, the thresholds of local density and distance from points of higher density are, respectively fixed to 32/0.38, 35.55/0.3, and 35.5/0.46 for $\beta \in \{0, 0.05, 1\}$ after trial and error.

Intuitively, the paths located at the cluster centers (cf. the 1st row in Fig. 9) can best represent the topics in each cluster. However, such paths are not suitable to be main paths due to their low topological weights (cf. the 2nd, 6th, and 10th rows in Table 1). In our opinion, the main reason lies in the extreme skewness of topological weight distribution [cf. Fig. 8 (b)]. Due to the overwhelming proportion of paths with low topological weights, they have more chance to be located at the cluster centers.

To address the problem, we take topological weight in concern and highlight topological path weight in different intervals with different colors and marks, as shown in Fig. 10. An interesting finding is that the paths with higher topological weights are always concentrated in a limited area. In fact, according to our observation, this phenomenon seems be independent from the hyper-

Table 1
The detailed information of main paths before and after adjustment.

β	Path ID		1	2	3	4	5
0.0	Before	topological weight	0.95	0.03	1.07×10^{-4}	1.28×10^{-3}	1.48×10^{-3}
	adjustment	path length	21	8	4	5	8
	After	topological weight	1.06	0.37	4.76×10^{-3}	1.85×10^{-3}	4.17×10^{-3}
	adjustment	path length	15	13	9	6	3
0.05	Before	topological weight	0.85	0.34	1.07×10^{-4}	0.02	2.85×10^{-4}
	adjustment	path length	13	13	4	6	5
	After	topological weight	1.06	0.34	7.82×10^{-3}	0.24	2.87×10^{-3}
	adjustment	path length	15	13	11	9	7
1.0	Before	topological weight	1.78×10^{-5}	4.33×10^{-3}	1.07×10^{-4}	0.02	2.85×10^{-4}
	adjustment	path length	2	6	4	6	5
	After	topological weight	1.06	0.34	4.76×10^{-3}	0.24	0.07
	adjustment	path length	15	13	9	9	11

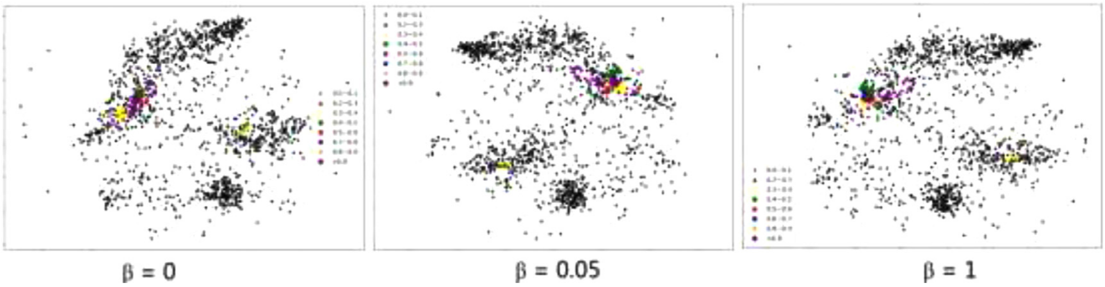


Fig. 10. The distribution of paths with topological weight in different intervals.

parameter β (cf. Fig. 10). It explains why prior studies (e.g., Verspagen 2007; Fontana, Nuvolari, & Verspagen, 2009; Liu et al. 2013; Xiao et al. 2014) are incapable of finding the developmental trajectories in different sub-fields. Hence, the paths with less topological weights can be used as supplement for global main path.

In this study, the path(s) with the largest topological weight in each cluster is chosen to represent the resulting cluster, viz. main path(s) as depicted in the 2nd row in Fig. 9. Due to the existence of multiple paths with equal topological weight, there are some clusters with more than one representative path. For example, three paths correspond to path 2 in case of $\beta = 0.05$, and all these paths are marked in Fig. 9. As we can see from the 4th, 8th, 12th rows in Table 1, the topological weight of the new representatives are greatly improved for all clusters and the corresponding path lengths also increase in most cases. The main paths in cases of $\beta \in \{0, 0.05, 1\}$ are presented in Fig. 11, where paths with equal topological weight are merged into one path. Due to the limited space, only detailed information of the main paths in case of $\beta = 0.05$ is provided in Tables A2–6 in Appendix section. As for the paths for $\beta = 0$ or 1 in Fig. 11, we refer readers to https://awesome-patent-mining.github.io/sMPA-paper for more details.

From Fig. 11, it is not difficult to observe several interesting phenomena:

(1) With the increase of β, the branches attaching to the main paths gradually decrease. The emerging of branch attributes to a considerable proportion of links with equal traversal weights in a citation network. However, when semantic similarity is integrated into the link weight calculation, the chance of two links with the same weights is reduced significantly, which turns out to eliminate branches from the main paths.

(2) The path with higher topological weight is less sensitive to β. For the paths with top 2 traversal weight sum, namely path 1 and path 2, their structures remain generally stable for $\beta \in \{0, 0.05, 1\}$. Especially for path 1, only branches are eliminated for β from 0 to 1 while the topological path weight keeps unchanged. In fact, path 1 in case of $\beta = 0$ is the resulting main path yielded by conventional MPA. In other words, the resulting path of conventional MPA (at least its trunk) is always identified as a sub-field trajectory by our sMPA. This is also applicable to the path 3. On the contrary, the path with less topological weight, such as path 4, changes dramatically when β switches from 0 to 0.05, although it keeps unchanged between the cases of $\beta = 0.05$ and 1. As for path 5, the change is even more drastic, as its structure is completely different for three different values of β.

(3) As β rises, the length of main path from each sub-field shows a generally increasing trend. In case of $\beta = 0$, even a path with length of 3 appears as the main path. But when β switches to 0.05 and 1, the shortest main path grows to 7 and 9, respectively. This indicates that the parameter β positively correlates the length of main paths. Note that the length of a main path cannot count the vertices on the branches in this study.

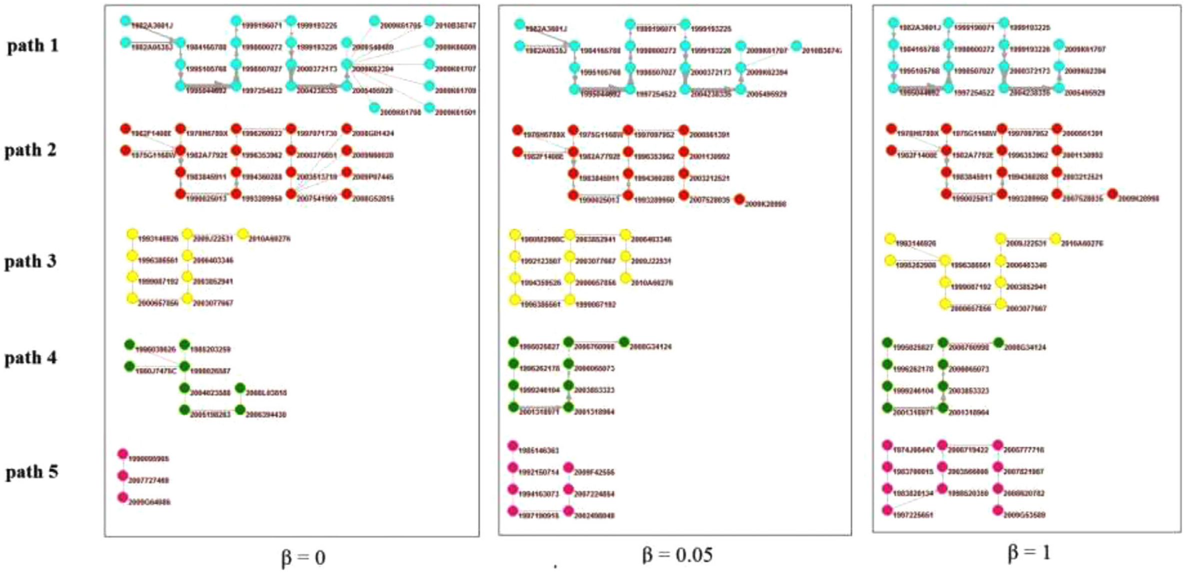


Fig. 11. The detailed information of main paths in different sub-fields.

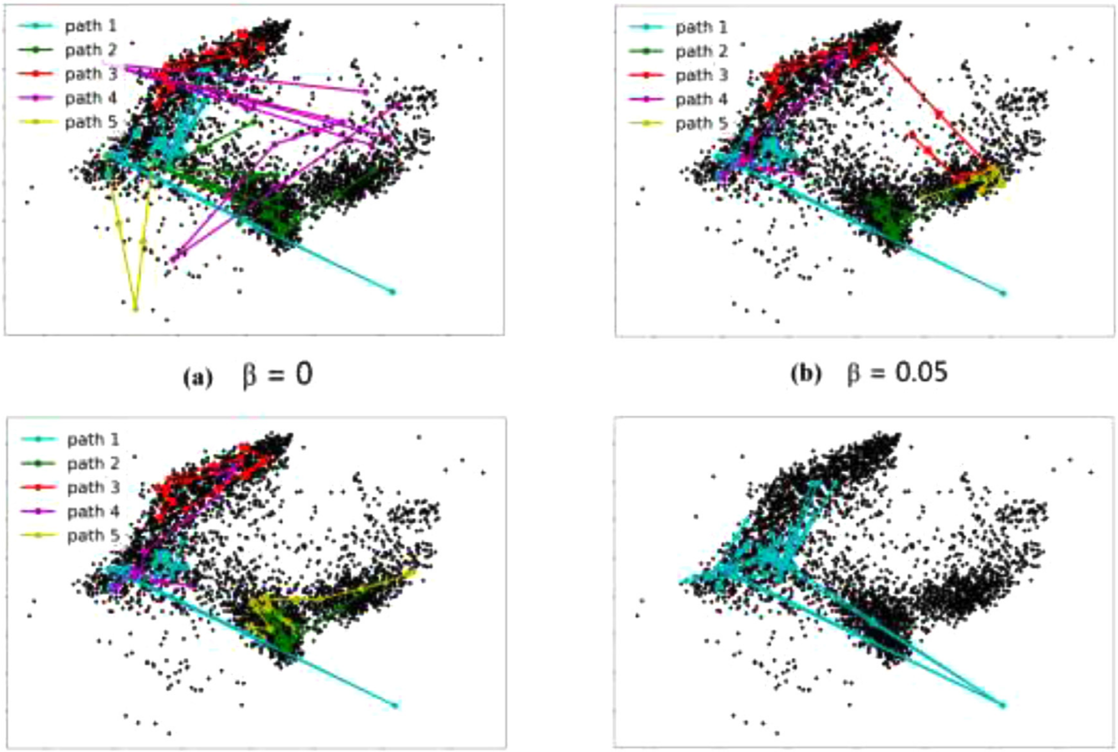
Table 2
The description of multiple main paths in different sub-fields.

Path ID	Path description
1	Battery cooling and structure design technology
2	Battery state control during charging and discharging process
3	Secondary battery technology for electrical and electronic equipment
4	Lithium secondary battery technology and battery pack structure design
5	Motor control technology for electric vehicle

5. Discussion

To obtain valuable insights about the sMPA on the topical coherence of documents along a same path, the experimental results are further analyzed at the macro and micro levels, respectively. At the macro level, the patent families represented by LSI vectors are projected to a two-dimensional plane by multidimensional scaling (MDS) (Xu et al., 2012). Afterward, the main paths in Fig. 11 are highlighted in Fig. 12 (a), (b) and (c). It is not difficult to see that main paths tend to span across different sub-fields with smaller value of β, especially path 2, 4 and 5 when β=0. But with the increasing of β, main paths are inclined to concentrate on certain sub-fields, which indicates the topical coherence of documents along a same path is improved. In the meanwhile, compared to the path from conventional multiple MPA when top $N = 10$ [cf. Fig. 12 (d)], our main paths cover most significant subareas in the whole field. It is worth noting that an outlier in path 1 is located at the lower right corner of each subplot in Fig. 12. This outlier corresponds to the patent family with GA = 1995044692. It seems that the outlier does not discuss the theme of path 1. Once it is removed from the whole citation network, the topological weight of path 1 would significantly drop from 1.06 to 0.438 in case of β=0. Though, this path has still the largest topological path weight among the candidate paths. This indicates that our main conclusions are not influenced by this outlier.

To dive into the micro level for more understanding of the sMPA, this study fixes β to 0.05 first, and the topics of corresponding main paths are summarized in Table 2 by reading the patent abstracts along each path. By reviewing the literature pertinent to electric vehicle, Jape &Thosar (2017) constructed the structure of electric vehicle in Fig. 13. Our sMPA identifies successfully most of the components relevant to battery, such as battery controller in path 1 and 2, battery design technology in path 3 and 4, motor controller in path 5. Furthermore, since the technological topic of each main path is quite distinguishable from each other, the sMPA provides an opportunity to observe various developmental trajectories concerning a same component. For example, even though path 1 and path 2 both discuss battery controller, path 1 describes the temperature control technology while path 2 concerns the voltage, current and residual capacity during charging and discharging process. According to the literature review (Hannan, Lipu, Hussain & Mohamed, 2017; Sanguesa, Vicente, Piedad, Francisco & Johann, 2021), these paths reflect the developmental trajectories of two critical components in battery management system (BMS), namely thermal management and charge control, whereas BMS is the key system in electric vehicle. On the other hand, the minimum topological weight of path 5 does not mean that the motor control is not the core technology but more independent of lithium-ion battery technology. Another example with small topological path weight is path 3, but the secondary battery technology has made a major advancement since 2010 (Zhang et al., 2017). This indicates the



(a) β = 0

(b) β = 0.05

(c) β = 1

(d) multiple MPA result when top N = 10

Fig. 12. The semantic distribution of candidate path in document clusters.

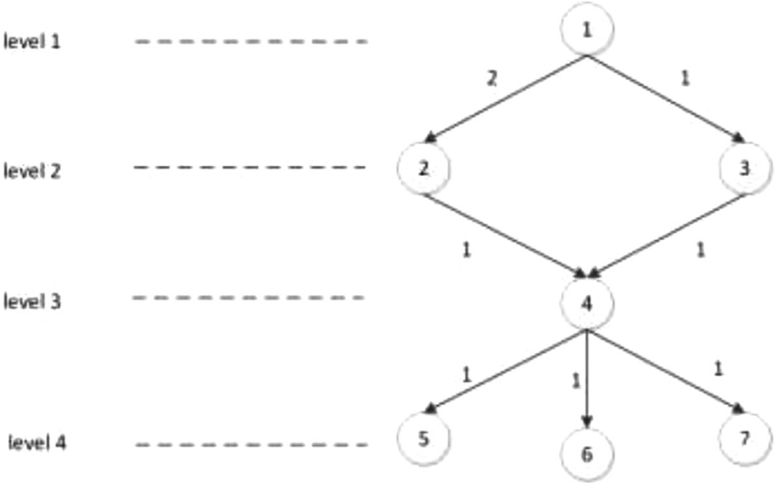


Fig. 13. The structure of electric vehicle (Jape & Thosar, 2017).

potential of our sMPA for forecasting technology developmental trajectories. As for battery pack in path 4 which is the most expensive component in any electric vehicle (Sanguesa et al., 2021), it has drawn much attention to optimize the structure of battery pack for thermal management with low cost (Arora, Kapoor & Shen, 2018; Wu et al., 2019; Zhuang, Liu, Su & Chen, 2021).

With the increase of β, not only the topical coherence of main paths is enhanced, but also the chance of picking up inappropriate path, such as the one deviating from the cluster center, with less length or topological weight, is reduced. Specifically, path 2 in case of $\beta = 0.05$ is much larger in term of semantic weight than its counterpart in case of $\beta = 0$. After reading the patents on the paths, we find that most of them discussed battery state control during charging and discharging process when $\beta = 0.05$. While for $\beta = 0$,

Table 3
The influence of our main paths in number of average forward citations and the resulting rank.

	path 1	path 2	path 3	path 4	path 5
avg. forward citations	57.5	71.4	39.1	24.8	28.9
rank	34	21	114	333	241
percentage	2.7	1.7	9.2	26.7	19.4

even through the vertices of path 2 is the same as the former in the first half, it has gone through a theme change since the vertex of 1996260922 and started to describe about data transmission and storage in battery control system. Path 3 is much alike to path 2 but the semantic distinction between different βs is relatively smaller. It is worth noting that path 4 is completely different for $\beta = 0$ and 0.05/1. The former refers to technology about permanent magnet type rotating electrical machine, while the latter describes lithium secondary battery technology and battery pack structure design. From the aspect of topological weight, the former of 1.85×10^{-3} is much less than the latter of 0.24, which indicates the main path about lithium secondary battery technology and battery pack structure design is a more important trajectory that should be cherry-picked. With a length of 3 and also deviating from the cluster centroid, path 5 in case of $\beta = 0$ is the last one to be main path. But with the increasing of β, the path has been improved in terms of path length and the distance to cluster centroid, and the path topic also transformed from battery mounting structure for $\beta = 0$ to motor control for $\beta = 0.05$, and eventually to battery control system for $\beta = 1$. It is very obvious that the latter two with path length of 7 and 11 are more suitable to serve as main paths.

Finally, we want to discuss the influence of the main semantic paths in the citation network. Inspired by Von Wartburg, Teichert and Rost (2005) and Lucio-Arias and Leydesdorf (2008), average forward citations received by vertices on a same path and the resulting rank among all candidate paths are utilized here to measure each path's influence. For this purpose, experimental dataset is extended to February 6th, 2022. The influence of our main paths is reported in Table 3. It is not difficult to see that four of five paths are in the top 20th percentile in average number of forward citation with the exception of path 4. Generally speaking, these main paths are still quite influential after even 12 years, and the patents on these paths are capable of receiving more citations than others in most cases. In addition, the ranks of the main paths in term of average forward citations are different from those in term of path weight. In more details, path 2 has surpassed path 1 and become the most influential technological trajectory, while path 5 has become the fourth influential main path. This may indicate more R&D activities have followed the directions of charge control, secondary battery and motor control technology during the 12 years.

6. Conclusions

Due to the ability of tracing the most significant developmental path of a field through a citation network, the MPA is widely used to find a set of papers that plays an important role in a specific area and identify a technology's main evolutionary pathway in the intellectual Property (IP) field. Especially when the citation network gets larger and more complicated, it significantly relieves researchers and technology managers from laborious, time consuming and cumbersome job in literature review. It is well known that there are usually multiple developmental trajectories corresponding to different sub-fields in an interested field. However, prior MPAs tend to focus on the developmental trace in one sub-field and ignore the others. In the meanwhile, high space complexity caused by exhaustive strategy limit this approach to be applicable to small-scale citation networks.

To address these limitations, a new method, named as semantic MPA (sMPA), is proposed in this work. Our approach is able to leverage semantic information to improve traditional MPA in steps of candidate path generation and main path selection. In more details, with the texts attached to vertices in a citation network, the textual similarity between two linked vertices is obtained and combined with topological weight to serve as a basis for topical consistency improvement in candidate path generation. Furthermore, to better represent the developmental trajectories in different sub-fields, the candidate paths are first clustered into several sub-fields, and then the paths with the highest topological weights in each cluster are chosen to represent the resulting sub-field.

To demonstrate the advantages of our method, extensive experiments are conducted on patent dataset pertaining to lithium-ion battery in electric vehicle. In the end, this study discovers that apart from the capability of identifying trajectories corresponding to different significant sub-fields, our sMPA can also improve topical coherence of documents along a same path and reduce the possibility of picking up inappropriate path, such as the one deviating from the cluster centroid, with less length or topological weight. Another nice property of the sMPA is that it always identifies the global main path (or with less branches) of conventional MPA as one of resulting multiple main paths. This indicates that without jeopardizing the ability of MPA to identify the most influential path from the entire network, our sMPA can also discover main paths from other underlying sub-fields as necessary complements to conventional MPA.

Though, this study is subject to the following limitations. (1) The sMPA selects the path with largest topological weight in each cluster as the resulting main path, but some paths may deviate from cluster centroid significantly, e.g. paths 4 and 5 in case of $\beta = 0$ in Fig. 9. Hence, it is necessary to construct a path selection strategy to make a tradeoff between semantic coherence and path topology. (2) Since topological path weight and semantic path weight follow different calculation rules, their values are in different magnitudes. By our observation, the semantic path weight is much larger than the topological path weight. In this case, a slight change of β may result in great impact on resulting paths from our sMPA. Several commonly used normalization methods, such as max-min scaling in this study, does not work very well, since these methods cannot weaken the impact of β. Hence, how to narrow the gap between magnitudes of the semantic path weight and topological path weight, and explore the impact of magnitude normalization on the resulting main paths deserve further study. (3) The Latent Semantic Index (LSI) is utilized here to vectorize the documents. As we all know, aside from LSI, more advanced distributed representations methods with better performance have been raised in the literature, such as Doc2vec (Le & Mikolov, 2014), XLnet (Yang, et al., 2019). To benefit from these advancements, our approach will be enhanced in the near future.

CRediT authorship contribution statement

Liang Chen: Conceived and designed the analysis, Collected the data, Performed the analysis, Wrote the paper. **Shuo Xu:** Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper. **Lijun Zhu:** Collected the data, Performed the analysis. **Jing Zhang:** Collected the data, Performed the analysis. **Haiyun Xu:** Contributed data or analysis tools, Performed the analysis. **Guancan Yang:** Collected the data, Contributed data or analysis tools.

Acknowledgment

This research received the financial support from National Natural Science Foundation of China under Grant Nos. of 71704169 and 72074014, respectively. Our gratitude also goes to Ms. Lili Chen in Institute of Scientific and Technical Information of China for her valuable contributions on the empirical study during paper revision and the anonymous reviewers for their valuable suggestions and comments.

Appendix 1. The inference of space complexity for sMPA

Since our sMPA embeds the dynamic programming technique, it does not need to store all paths in a citation network but the paths between source vertices and other vertices with maximal weights. For easy understanding, we take the network in Fig. A1 as an example, in which each link is labeled by its weight and the paths needing to be stored are reported in Table A1. Note that since the stored paths with maximal weights are from a source vertex to several intermediate/sink vertices, they can assemble a tree shown in Fig. A2.

It is well known that a citation network actually belongs to directed acyclic graph (DAG for short). Let us start with a special case of DAG, the tree with out-degree not greater than M, to learn how the amount of space for the paths scales with the size of a citation network, viz. space complexity. It is not difficult to infer the number of vertices on the second level of the tree is not more than M, that on the third level is not more than M^2, and so on. As a matter of fact, this property is also applicable for a general DAG when

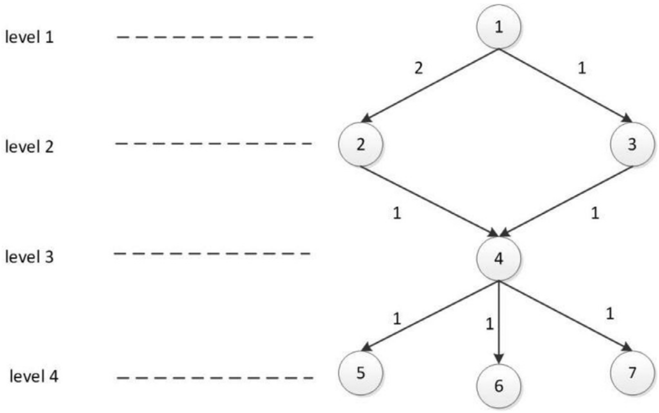


Fig. A1. An example of citation network.

Table A1
The paths need to be stored in running sMPA.

Source vertex	Tail vertex	Path with maximal weight in between	Weight
1	1	1	0
1	2	1→2	2
1	3	1→3	1
1	4	1→2→4	3
1	5	1→2→4→5	4
1	6	1→2→4→6	4
1	7	1→2→4→7	4

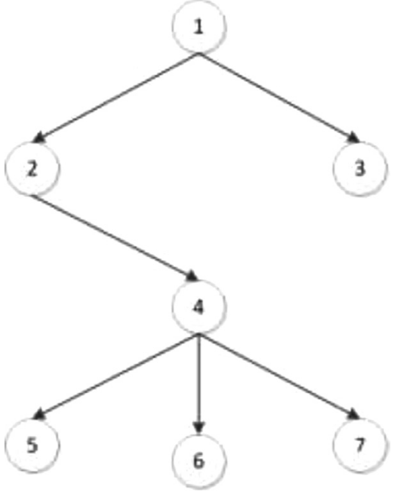


Fig. A2. The tree assembled by the stored paths in Table A1

M refers to the maximal out-degree, since DAG is stored as a series of paths which are capable of assembling a tree while running sMPA. Therefore, by deriving the space complexity of sMPA on a tree structure, we can naturally extend it to a general DAG.

Let h denote the height of the tree with n vertices, then we have

$$1 + M + M^2 + \cdots + M^{h-1} \geq n$$

$$\frac{M^{h-1} - 1}{M - 1} \leq n$$

$$h < \log_M(n * (M-1) + 1)$$

As the path length from source vertex to the ith level is i-1, the total number of vertices along these paths is not greater than $(i-1) * M^{i-1}$. Further, we can infer the number of vertices of paths from source vertex to all nodes on the tree, which is also the amount of space necessary for sMPA, is not greater than

$$1 + 1{*}M + 2{*}M^2 + \cdots + (h-1){*}M^{(h-1)}$$

$$1 + M{*}(1 + 2{*}M + \cdots + (h-1){*}M^{(h-2)}$$

$$1 + M{*}\left(M + M^2 + \cdots + M^{(h-1)}\right)'$$

$$1 + M{*}\left[\frac{M{*}\left(1 - M^{h-1}\right)}{1-M}\right]'$$

$$1 + \frac{M}{(1-M)^2}\left[1 - h{*}M^{h-1} + (h-1){*}M^h\right]$$

Recall that

$$\mathrm{h} < \log_M(n{*}(M-1) + 1)$$

Thus,

$$1 + \frac{M}{(1-M)^2}\left[1 - h{*}M^{h-1} + (h-1){*}M^h\right] <$$

$$1 + \frac{M}{(1-M)^2}\left[1 - (\log_M(n * (M-1) + 1)\right) * \frac{n * (M-1) + 1}{M} + (\log_M(n * (M-1) + 1) - 1) * (n * (M-1) + 1)$$

By omitting the constants such as M, $(M-1)$, the space complexity of $O(n \log n)$ is achieved.

Appendix 2

Tables A1–A5.

Table A2
A list of basic patents appearing along path 1 when $\beta = 0.05$.

GA	Title
1982A3601J	High-temperature battery with thermal insulation - has air cooling maximised due to positioning of inlet and outlet channels in cells of slide-in module
1982A0535J	High-temp. battery with cylindrical cells - held in springy rings inside insulating case allowing circulation of cooling air and separate renewal
1984165788	High temp. storage battery for electric vehicle - has cooling medium distributor plate between storage cells and double walled insulated housing
1995105768	Battery securing structure for electric vehicle - has lower inner spacer which has partitions divided into separate compartments in which batteries are disposed, and upper inner spacer member including pressing portions which hold batteries in position
1995044692	Electric vehicle traction battery exchange system including special lifter - hoists batteries in compact array with cabling to multipole plug connector in casing supported on underside of vehicle
1997254522	Battery structure of electric vehicle - comprises vent ports through which cooling air is caused to vertically flow in each compartment
1998507027	Battery assembly, e.g. for electric vehicle - has modules inserted into casing with fixed bulkheads and intermediate bulkheads
1998600272	Battery power supply unit for electric vehicles - has pair of end plates made of resin to which bar is inserted and molded
1999196071	Flat circuit harness for battery unit used in electric vehicle - has insulation sheet on its surface to cover flat conductor circuit that are connected to terminal of each battery of battery unit
1999193225	Connecting plate for electric vehicle battery holder
1999193226	Connecting plate for electric vehicle battery holder
2000372173	Connection plate for battery in electric vehicle, has wire protector provided in junction between bus bar and voltage detection terminal
2004238335	Battery connecting plate for electric vehicle, has upper and lower covers attached to plate body to fix terminal on body, where terminal has tolerance compensator compensating positional difference between two portions of terminal
2005495929	Lead member for secondary battery module useful in industrial instrument or electric cars, connects two adjacent cells and forms single series circuit through electrode terminals of cells
2009K62394	Battery system for use in e.g. electric vehicle, severs electrical connection between cells of battery packs locally in response to excessive impact force during over-current or over-temperature conditions
2009K61707	Electrochemical storage cell for lithium ion polymer battery system used in e.g. electric vehicle, has protection cover whose half portions are coupled with each other
2010B38747	End cover assembly of electrochemical cell of power lithium-ion polymer battery for e.g. electric vehicle, has sealing material positioned to extend beyond upper portion of scabbard and to wrap around scabbard to form protective flange

Table A3
A list of basic patents appearing along path 2 when $\beta = 0.05$.

GA	Title
1976H6789X	Electric vehicle drive control - with single regulator switched to either armature or field coils for wide range of torque control
1975G1168W	State-of-charge indicator for battery of electric vehicle - includes integrator and cct. for compensating discharge current
1982F1408E	Monitoring system for traction battery pack - has individual voltage monitors for each sub-pack and providing signal only when associated on-load voltage falls below preset limit
1982A7792E	Determining state of charge of accumulators - deriving battery voltage corrected for polarisation voltage and electrolyte-t electrolyte temp.
1983845911	Combined battery charge state evaluator and motor control - has microcomputer calculating used charge by integrating battery current data and remaining charge from voltage data
1990025013	Battery monitoring system for e.g. electric vehicle - calculates total possible discharge durations to indicate remaining discharge time available
1993289950	Operating parameter monitor for lead acid battery - has voltage, current and temperature sensors and periodically evaluates parameters of battery to determine state
1994360288	Rechargeable storage battery remaining capacity measuring appts - has switching device responsive to command signal from electronic control unit for setting battery in disconnected state to all loads except ECU
1996353962	Battery residual capacity meter for driving motor of electric vehicle - detects V-I characteristic by reading voltage and current when battery current is more than 0.75 C and increasing in high load state, and calculates residual capacity from measured V-I characteristic and stored relation
1997097952	Battery condition detecting system e.g for electric vehicle - includes regression line calculator which determines internal resistance and open-circuit voltage of battery based on detected values of voltage and current of battery stored in memory
2000561391	Charging state detector for battery used in e.g. electric vehicle, has battery ECU which corrects SOC electromotive voltage property based on detected supplement positive characteristic
2001130992	Charging state detector for battery of e.g. electric vehicle, enables correcting SOC electromotive voltage property based on battery voltage and battery current at predetermined period
2003212521	Memory effect detection method in secondary battery of vehicles, involves judging memory effect if ratio of variation in no-load voltage to variation in residual battery capacity exceeds predetermined threshold value
2007528035	Battery management system for use in vehicle, has current sensor for measuring amount of output current of battery, main switch turning on/off electricity provided from battery in response to control signal, and cooling fan
2009K28998	Vehicle electrical power system i.e. automotive electrical bus and charging system, controlling method, involves controlling current produced by alternator, where alternator current is equal to vehicle load current

Table A4
A list of basic patents appearing along path 3 when $\beta = 0.05$.

GA	Title
1980M2998C	Regenerative electric motor for battery powered vehicle - produces high back EMF which may be used to charge vehicle batteries in sequence with driving vehicle
1992123507	Electric vehicle integrated motor drive and recharge system - including bidirectional DC power source, two voltage-fed inverters, two induction motors and control unit
1994359526	Electric vehicle drive system - controls sharing of drive torque between efficient low speed motor and large capacity high speed motor
1996386561	Negative electrode for lithium secondary battery for electric automobile, motorbike, etc. - composing at least one cell of secondary battery of carbon material with particles carrying a metal forming alloy with lithium
1999087192	Lithium secondary battery for electrical and electronic equipment - uses positive electrode containing manganese content transition metal oxide and negative electrode containing carbon particles carrying metals forming or not forming alloy with lithium
2000657856	Long-life lithium secondary battery, e.g. for electric power storage systems and electric vehicles, has lithium-manganese complex oxide as positive electrode active material, and amorphous carbon as negative electrode active material
2003077667	Lithium secondary cell for electric car, has positive and negative electrodes made of metallic foil whose edges are joined with predetermined portions of positive/negative electrode collector
2003852941	Sealed rechargeable battery for e.g. electric vehicle, has electrode plate group accommodated in a case with liquid electrolyte, with its winding axis vertical relative to the open end of the case
2006403346	Closed type battery e.g. nickel hydrogen battery used in electric vehicle, has lower current collection plate welded to battery jar in range outside position corresponding to position below cap at upper portion of cover
2009J22531	Secondary battery e.g. lithium ion secondary battery for electric vehicle, has junction portion directly arranged in collector plate to uniformly distribute current to collector plate from electrode plate
2010A60276	Lithium ion secondary battery for electric vehicle, has steel shell as cathode, lower side edge of cathode is provided with copper foil, and copper foil is connected with steel shell by metal and/or non-metal conductive adhesive

Table A5
A list of basic patents appearing along path 4 when $\beta = 0.05$.

GA	Title
1995025827	Lithium secondary battery mfr. - comprises use of lithium as active material of negative electrode and active material of hydroxy or carboxyl group for positive electrode
1996262178	Sec. battery for use in electronic appts., memories, etc. - comprises positive electrode, negative electrode comprising carbon matter capable of absorbing and releasing lithium ions, etc.
1999246104	Lithium secondary battery has superior safety and high weight energy density
2001318971	Battery packs, e.g. for electric vehicle, has binding bands to bind battery modules and end plates which are arranged on both ends of battery modules
2001318964	Battery module, e.g. for electric vehicle, has projection of smaller height positioned opposite to partitions in battery case, than that of another projection
2003853323	Cooling device for battery pack, has cooling medium paths for allowing a cooling medium to flow in right and left directions between the longer side faces of the rechargeable batteries
2006065073	Vehicle-mounted power supply unit e.g. for electric vehicle has flat electronic component box arranged in contact with one of longitudinally extending side surface of battery pack casing
2006760998	Battery pack mounting structure mounted on electric vehicle, consists of direct current to direct current converter electrically connected by junction box and harness and provided at downward direction of driver's seat
2008G34124	Electrical-storage unit for motor vehicle, has reinforcement sections that are arranged at predetermined interval between electrical-storage cells of electrical-storage module along vertical direction

Table A6
A list of basic patents appearing along path 5 when $\beta = 0.05$.

GA	Title
1985146363	Turning speed control for electric vehicle with dual drive motors - automatically slows vehicle according to sharpness of turn and uses field excitation as function of steering angle
1992150714	Control system for electric motor vehicle - has right and left wheels driven by motors provided separately
1994163073	Motor current control device for electric automobile - has output torque cut commander for motor drive circuit to supply current to motor according to command value or to cut current supply by torque cut command
1997190915	Polyphase AC motor control appts used in electric vehicle - has second program which computes torque current command value from fixed exciting current command value and motor torque command value, while torque response is considered to be important
2002498048	Torque control method of permanent magnet synchronous motor, involves employing vector control technique to generate electrical control signals for adjusting frequency and magnitude of sinusoidal waveforms
2007224854	Alternating current motor driving apparatus for use in vehicle, has current sensor and motor phase provided to estimate motor current so that current detection is made from zero speed to high-speed rotations
2009F42556	Current detector unit for motor control device, has three-phase current detecting portion for detecting three-phase current, if judging portion judges that target time point belongs to period during which three-phase current is detected

References

Arora, S., Kapoor, A., & Shen, W. (2018). A novel thermal management system for improving discharge/charge performance of Li-ion battery packs under abuse. *Journal of Power Sources, 378*, 759–775.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv:*arXiv preprint* cs/0309023.

Batagelj, V., Mrvar, A., Jünger, M., & Mutzel, P. (2004). Pajek-analysis and visualization of large networks. In *Graph drawing software. Mathematics and visualization* (pp. 77–103). Springer.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS ONE, 6*(3), e18029.

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.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on machine learning* (pp. 233–240).

Fernando, F. S., & James, M. U. (1995). Dominant designs and the survival of firms. *Strategic Management Journal, 16*(6), 415–430.

Foltz, P. W. (1990). Using latent semantic indexing for information filtering. *ACM SIGOIS Bulletin, 11*(2-3), 40–47.

Hannan, M. A., Lipu, M. S. H., Hussain, A., & Mohamed, A. (2017). A review of lithium-ion battery state of charge estimation and management system in electric vehicle applications: Challenges and recommendations. *Renewable and Sustainable Energy Reviews*, (78), 834–854.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics), 28*(1), 100–108.

Huang, Y., Zhu, F., Porter, A. L., Zhang, Y., Zhu, D., & Guo, Y. (2021). Exploring technology evolution pathways to facilitate technology management: From a technology life cycle perspective. *IEEE Transactions on Engineering Management, 68*(5), 1347–1359.

Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A. L., Liu, Y., et al. (2017). A hybrid method to trace technology evolution pathways: A case study of 3D printing. *Scientometrics, 111*(1), 185–204.

Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39–63.

Jape, S. R., & Thosar, A. (2017). Comparison of electric motors for electric vehicle application. *International Journal of Research in Engineering and Technology, 6*(8), 12–17.

Kim, J., & Shin, J. (2018). Mapping extended technological trajectories: Integration of main path, derivative paths, and technology junctures. *Scientometrics, 116*(3), 1439–1459.

Kim, M., Baek, I., & Min, S. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *Journal of the Association for Information Science & Technology, 69*(2), 329–342.

Lai, K. K., Chen, H. C., Chang, Y. H., Kumar, V., & Bhatt, P. C. (2021). A structured MPA approach to explore technological core competence, knowledge flow, and technology development through social network patentometrics. *Journal of Knowledge Management, 25*(2), 402–432.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the international conference on machine learning* (pp. 1188–1196).

Leydesdorff, L., Bornmann, L., Marx, W., & Milojević, S. (2014). Referenced publication years spectroscopy applied to iMetrics: Scientometrics, journal of informetrics, and a relevant subset of JASIST. *Journal of Informetrics,, 8*(1), 162–174.

Liu, J. S., & Kuan, C. H. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology, 67*(2), 465–476.

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 American Society for Information Science and Technology, 63*(3), 528–542.

Liu, J. S., Lu, L. Y., & Ho, M. H. C. (2019). A few notes on main path analysis. *Scientometrics, 119*(1), 379–391.

Liu, J. S., Lu, L. Y., & Ho, M. H. C. (2020). A note on choosing traversal counts in main path analysis. *Scientometrics, 124*(1), 783–785.

Liu, J. S., Lu, L. Y., Lu, W. M., & Lin, B. J. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega, 41*(1), 3–15.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite$^{\mathrm{TM}}$-based historiograms. *Journal of the American Society for Information Science and Technology, 59*(12), 1948–1962.

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.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492–1496.

Sanguesa, A., Vicente, T., Piedad, G, Francisco, M., & Johann, M. (2021). A review on electric vehicles: Technologies and challenges. *Smart Cities, 4*(1), 372–404.

Tan, Z., Liu, C., Mao, Y., Guo, Y., Shen, J., & Wang, X. (2016). AceMap: A novel approach towards displaying relationship among academic literatures. In *Proceedings of the 25th international conference on world wide web* (pp. 437–442).

Tu, Y. N., & Hsu, S. L. (2016). Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology, 67*(8), 2016–2031.

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.

Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy, 34*(10), 1591–1607.

Wu, W., Wang, S., Wu, W., Chen, K., Hong, S., & Lai, Y. (2019). A critical review of battery thermal performance and liquid based battery thermal management. *Energy Conversion and Management, 182*, 262–281.

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

Xu, S., Qiao, X., Zhu, L., Zhang, Y., & Li, L. (2012). Fast but not bad initial configuration for metric multidimensional scaling. *Journal of Information & Computational Science, 9*(2), 257–265.

Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., & Li, L. (2016). Reviews on determining the number of clusters. *Applied Mathematics & Information Sciences, 10*(4), 1493–1520.

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*(5), Article 100983.

Xu, S., Hao, L., An, X., Pang, H., & Li, T. (2020). Review on emerging research topics with key-route main path analysis. *Scientometrics, 122*(1), 607–624.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach Hanna, Larochelle Hugo, Beygelzimer Alina, d'Alche-Buc Florence, Fox Edward, & Garnett Roman (Eds.), *Proceedings of the advances in neural information processing systems 32: Annual Confference on Neural Information Processing System 2019* (pp. 5753–5763). ACM.

Yeo, W., Kim, S., Lee, J. M., & Kang, J. (2014). Aggregative and stochastic model of main path identification: A case study on graphene. *Scientometrics, 98*(1), 633–655.

Yong, J. Y., Ramachandaramurthy, V. K., Tan, K. M., & Mithulananthan, N. (2015). A review on the state-of-the-art technologies of electric vehicle, its impacts and prospects. *Renewable and Sustainable Energy Reviews, 49*, 365–385.

Yu, D., & Pan, T. (2021). Tracing knowledge diffusion of TOPSIS: A historical perspective from citation network. *Expert Systems with Applications, 168*, 101–136.

Yu, D., & Sheng, L. (2020). Knowledge diffusion paths of blockchain domain: The main path analysis. *Scientometrics, 125*(1), 471–497.

Zhang, Q., Li, C., & Wu, Y. (2017). Analysis of research and development trend of the battery technology in electric vehicle with the perspective of patent. *Energy Procedia, 105*, 4274–4280.

Zhuang, W., Liu, Z., Su, H., & Chen, G. (2021). An intelligent thermal management system for optimized lithium-ion battery pack. *Applied Thermal Engineering*, (189), Article 116767.

Further reading

Fontana, R, Nuvolari, A, & Verspagen, B (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of innovation and new technology, 18*(4), 311–336.

Renolds, D (2015). Gaussian Mixture Models. In S Li, Z, & A Jain, K (Eds.), *Encyclopedia of Biometrics* (pp. 827–832). Springer.