SoftwareX 22 (2023) 101330

Contents lists available at ScienceDirect

SoftwareX

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

Original software publication

TopicTracker: A platform for topic trajectory identification and visualisation

Yong-Bin Kang [a,*], Timos Sellis [b,1]

[a] ARC Centre of Excellence for Automated Decision-Making and Society, Swinburne University of Technology, Victoria, Australia
[b] Archimedes Research Unit, Athena Research Center, Greece

ARTICLE INFO

Article history:
Received 1 March 2021
Received in revised form 24 November 2022
Accepted 8 December 2022

Keywords:
TopicTracker
Topic trajectory
Topic evolution
Topic tracking

ABSTRACT

Topic trajectory information provides crucial insight into the dynamics of topics and their evolutionary relationships over a given time. Also, this information can improve our understanding on how new topics have emerged or formed through a sequential or interrelated events of emergence, modification and integration of prior topics. Nevertheless, the implementation of the existing methods for topic trajectory identification is rarely available as usable software. In this paper, we present TopicTracker, a platform for topic trajectory identification and visualisation. The key of TopicTracker is that it can represent the three facets of information together, given two kinds of input: a time-stamped topic profile consisting of the set of the underlying topics over time, and the evolution strength matrix among them: evolutionary pathways of dynamic topics, evolution states of the topics, and topic importance. TopicTracker is a publicly available software implemented using the R software.

Code metadata

Current code version	V1.0
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-21-00046
Legal Code License	MIT License
Code versioning system used	git
Software code languages, tools, and services used	R
Compilation requirements, operating environments & dependencies	R version $\geq$ 3.6, and R packages (igraph, hash, plotrix)
If available Link to developer documentation/manual	https://github.com/Yongbinkang/topicTracker/README
Support email for questions	ykang@swin.edu.au, yongbin.kang@gmail.com

1. Motivation and significance

Topic trajectory identification is a research area that has attracted significant attention from scientific institutions and innovation-industry sectors. In this area, a fundamental is the use of *topic modelling* to discover latent thematic topics (or concepts) from a document collection, where each topic consists of its representative terms extracted from the collection [1]. More recently, dynamic topic modelling has also been utilised to identify topics and their evolution over a time period [2–4]. Identifying topic trajectories provides precious insights into the dynamics of topics over time. For example, in scientific and patented innovation domains, such trajectories can significantly help to distinguish outstanding research or technological topics, and discover their evolutionary pathways reflecting how new topics have been emerged or formed through a sequential of the events of the emergence, modification and integration of past topics [5–8]. We view a *trajectory of topics* as a main stream or an evolutionary pathway of topics over time.[2] Also, we define an *evolutionary pathway* as a series of evolutionary relationships between older topics and newer topics. Consequently, due to this

* Corresponding author.
E-mail addresses: ykang@swin.edu.au (Yong-Bin Kang), timos@athenarc.gr (Timos Sellis).

[1] Work done while at Swinburne University of Technology.

[2] To simplify the presentation, we do not distinguish between 'trajectory' and 'evolutionary pathways' and use them interchangeably.

https://doi.org/10.1016/j.softx.2023.101330

Table 1
Previous studies on topic trajectory identification and software availability.

Study	Software availability	Domain
He et al. (2009) [9]	✗	Scientific literature (topic: technology/knowledge concept)
Jo et al. (2011) [10]	✗	
Song et al. (2014) [6]	✗	
Zhou et al. (2017) [5]	✗	
Zhang et al. (2017) [8]	✗	
Jung et al. (2020) [11]	✗	
Yoon et al. (2011) [7]	✗	Patents (topic: technology/knowledge concept)
Zhong et al. (2016) [12]	✗	
Lee et al. (2017) [1]	✗	
Park et al. (2017) [13]	✗	
Triulzi et al. (2020) [14]	✗	
Huang et al. (2020) [15]	✗	
Qiu et al. (2020) [16]	✗	
Zhang et al. (2015) [2]	✗	E-commerce (topic: market brand)
Greene et al. (2016) [3]	✗	Politics (topic: political agenda)
Song et al. (2016) [4]	✗	History (topic: historical event)
Gaul et al. (2017) [17]	✗	Online news (topic: online news)

merit, many studies were conducted in recent years for designing different methods for topic trajectory identification (see Table 1).

Unfortunately, the implementation of these methods to encourage their use by the wider scientific community that are interested in topic trajectory identification has been still remained limited. Primarily, a large body of the methods is not available as readily usable software as seen in Table 1. In topic modelling and its application areas, existing software tools were developed with little consideration for preparing and formatting the data for topic trajectory identification in a simple and easy way to use. This requires the users to directly implement their algorithms for topic trajectory identification from the results of topic modelling. Further, this leads to unnecessary time spent for data preparation and limits effective comparison of results produced by different topic trajectory identification models as well.

To address these issues, we have developed TopicTracker, a platform for topic trajectory identification and visualisation. TopicTracker is a software implemented using the R software.

2. Software description

We present the architecture of TopicTracker and its main functionalities.

2.1. Software architecture

The architecture of TopicTracker's code design is depicted in Fig. 1. It is designed to distinguish the inferential module for building a *Topic Evolution Tree* (TET) of topics (Phase 1) from the code for visualising their topic trajectories (Phase 2). One key idea in designing TopicTracker is to discover evolutionary pathways (i.e., topic trajectories) between non-contemporary topics by constructing their most likely genealogy tree (i.e., TET) over a given time. Because of this flexibility, one can easily customise the code to modify or define a new shape of a TET.

The following summarises the workflow within the architecture:

- First, two kinds of input data must be provided: a *temporal topic profile* comprising of the profile of the underlying time-stamped topics, and a $N \times N$ *Topic Evolution Strength* (TES) matrix, where N denotes the number of the topics in the profile.
- Second, TopicTracker infers the TET of the topics in the profile with a user-specified parameter, min_tes (Phase 1), the minimum TES among the topics to find their possible ancestors. The topics with TESs less than min_tes will not be considered in finding their ancestors. Given a topic v in the TET, its pathway to the most ancestral topic indicates the trajectory of v. Thus, TET is a backbone reflecting evolutionary pathways of topics.
- Third, TopicTracker visualises the inferred TET with the five central modules (Phase 2) using three parameters: min_reborn is the minimum time period of topics which has been elapsed between the moment of their emergence until their evolved topics appear; min_dead is the minimum time period of topics unobserved; and min_tes.

An example TET is presented in Fig. 2, where each node denotes a topic. The evolutionary relationship between two topics is denoted by a directed edge, and its TES is represented by a different edge colour. The evolution state of each node is marked with a different node colour. The y-axis shows the importance of topics normalised in [0,1]. We emphasise that the integration of *three information facets* in TET provides insight about topic trajectories: the evolutionary pathways of topics with their TESs over time, evolution states of topics over time, and topic importance. Below we elaborate the construction process of a TET.

2.2. Software functionalities

We now present the detailed descriptions about the input data and the two phases embodied in TopicTracker.

2.2.1. Input data format

The two kinds of input data[3] must be given to run TopicTracker: a temporal topic profile $\mathcal{P}$, and a TES matrix $\mathbf{M}$. $\mathcal{P}$ contains the descriptions about the target topics whose trajectories will be generated. An example of $\mathcal{P}$ is given in Table 2, and it has the following fields for each topic v:

- `id` is the unique identifier of v.
- `index` is the unique integer index of v (starting from 0).
- `weight` is the weight of v given each year.
- `year` (in format yyyy) is the year that v was generated.
- `words` are the top-N words representing v. The information about `weight`, `year`, and `words` can be generated by a topic model.[4]

[3] Discussion of generating this input data is beyond the scope of this paper.

[4] For example, a dynamic topic model [18] can generate such information. Discussion about generating such information is beyond the scope of this paper.

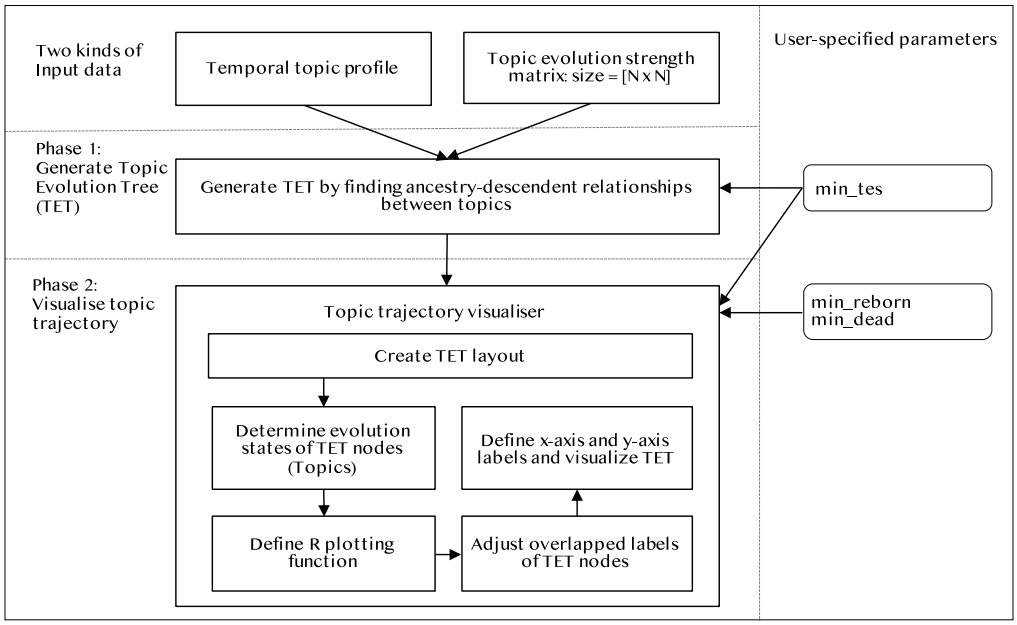


Fig. 1. The architecture of TopicTracker.

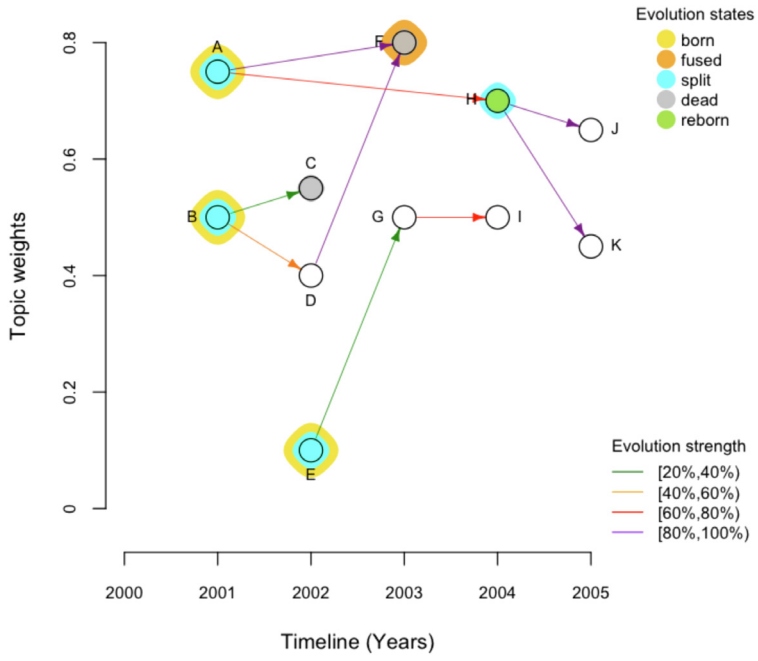


Fig. 2. An example TET generated by TopicTracker. The parameters are set as: min_reborn = 2 years, min_dead = 1 year, and min_tes = 0.2. The explanation about this example is provided in Section 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The **M** matrix is a $N \times N$ matrix, where N is the number of the topics presented in $\mathcal{P}$, where:

- The ith row and jth column of **M** represents the TES of the ith row topic (old) towards the jth column topic (new).
- A TES only exists between a pair of two *non-contemporary* topics as we only estimate the TES between topics on different time slots. Thus, we set TESs between *contemporary topics* as 0. Also, by default, we set the diagonal entries to be 1.
- All the topics are sorted in ascending order according to their time slots. Thus, the first is the most ancient topic and the last the most recent topic.
- **M** is a non-symmetric matrix, where all entries below the main diagonal do not hold any values, as we are only interested in the calculation of the TES between two topics x and y, if only if $time(x) < time(y)$, where *time* is the function returning the time slot of a given topic.
- All entries in **M** are normalised in [0,1], where the higher the more important.

An example of **M** is given in Table 3. To estimate TESs in **M**, the concept of similarity has been widely used in most related works [1,3,5,9,11,16]. The fundamental is to formulate a similarity measure between x and y using the aggregated similarities between x's top-k descriptive terms and y's top-k descriptive terms (often $K = 10$) [3,19].

2.2.2. Creating TET layout

We construct a TET from the given $\mathcal{P}$ and **M**. The goal of constructing a TET is to identify the most likely genealogy of topics over a given time period. Technically, this problem is formulated as finding an optimum branching in a directed tree, where each direct edge connects a topic and the topics evolved from it. Let TET $= (V, E)$ be a directed, weighted tree, where $V = \{v_1, \ldots, v_n\}$ is the set of nodes that correspond to the n topics in $\mathcal{P}$. These topics are time-stamped meaning that these are collected according to the k different time slots, where k is the carnality of the years observed in $\mathcal{P}$. $E \subseteq V \times V$ represents the set of directed edges that reflect evolutionary relationships in V, such that $(v_i, v_j) \in E$ if only if $time(v_i) < time(t_j)$, where the *time* function specifies the time slot of a given topic. An edge (v_i, v_j) is an ordered pair of two nodes v_i and v_j, and is interpreted as there is a directed dependency from an ancestry topic v_i to a descendent topic v_j. Each edge is associated with its TES, drawn from **M**. A TES reflects how strong each ancestry-descendent relationship is between two topics.

The TET algorithm is implemented in the function `buildTES()` in `topicTracker.R`, and has the following bases:

1. We assume that there is the dummy *root* node in a TET to merely make a tree. Thus, the most ancient topic for each topic is connected to the root node.
2. Ancestry topics always precede their descendant topics in time.
3. A topic v can have no ancestor, meaning that v is newly emerged. On this occasion, v is connected to the root node.
4. Each topic v can have multiple ancestors as more than one topics in the past can be assembled together into the emergence of v in a later time.
5. In a TET, evolutionary transitivity relationships are logically inferred by navigating the edges between past and new topics. Suppose that there are two evolutionary relationships: an older topic '`wireless technology`' influences on the generation of a recent topic '`mobile devices`'; and a topic '`mobile devices`' influences on the generation of a more recent topic '`Android phone`'. Then, we infer the relationship that '`wireless technology`' can also influence on the generation of '`Android phone`'.
6. A complexity of a TET is determined by the number of edges connected in the TET. This number is determined by min_tes. That is, we only include all edges whose TESs are equal to and greater than min_tes. Also, in a TET, cycles are not observed as ancestries cannot go back in time.
7. An evolutionary pathway of a topic v is defined as a sequence of connected topics from v to the root node. Given the same evolutionary pathway, among all possible parents Par(v) of v, some are more likely than others. A parent is the immediate ancestor of a topic. This likelihood is estimated based on the TESs, from **M**, between v and Par(v). We choose the parent with the highest TES with v. To illustrate, in Fig. 3(b), given a topic F, there are three possible evolutionary paths: (1) {(B, F)}, (2) {(B,C), (C,F)}, and (3) {(B,D), (D,F)}. Given the paths (1) and (2), there are two possible parents for F, B and C, as these exist on the same pathway. In this case, we choose C as its TES with F is higher than that of B. Thus, we do not connect F to B. The TESs are indicated by different edge colours as presented in the legend. In the same manner, given the paths (1) and (3), we choose D not B as the parent of F. The result is given in Fig. 3(b).

2.2.3. Topic trajectory visualiser

Topic trajectory visualiser is implemented in the function `visualiseTET()` in `topicTracker.R`. The function has the following modules:

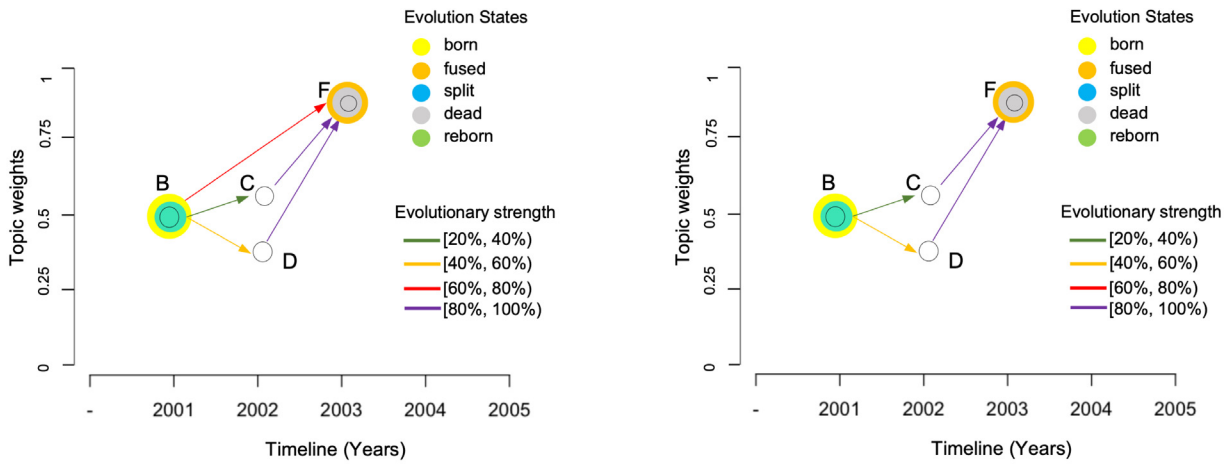
- *Create TET layout*: We create a TET using `layout_as_tree()` (the Reingold-Tilford graph layout algorithm) in the R `igraph` package. We also define the positions of topics by their weights given by `weight` in $\mathcal{P}$ and time information given by `year` in $\mathcal{P}$ within the layout.
- *Determine evolution states of TET nodes*: We determine the evolution states of topics in the TET, and mark them using different colours (see below).
- *Define R plotting function*: We define the `plot()` function: the properties of nodes and edges within the defined `layout_as_tree()` to visualise them in the TET.
- *Adjust overlapped labels of TET nodes*: The overlapped node labels in the TET are separated to avoid their overlaps using `thigmophobe.labels`[5] in the package `plotrix`. This function is used to automatically place topic labels avoiding partially or fully overlapping labels. The meaning of the name, 'thigmophobe' is 'one who fears being touched'.
- *Define x-axis and y-axis labels and visualise TET*: We finally define the ticks of x-axis and y-axis and draw their labels in the TET.

TopicTracker can identify the trajectories of the underlying topics, detecting their evolution states: (a) when and how is a new topic born? (b) how can a topic influence on the generation of newer topics? (c) what is a topic continually flourishing? (d) how can a new topic emerge implanted by what past topics? (e) what are the dead topics no longer observed? and (f) what are reborn topics distinguishable again in later time? To address them, we measure the evolution states of underlying topics using the TESs given in **M**:

- *born*: Topics born are the topics newly emerging. These are the topics emerging without any ancestors.
- *split*: Topics split are the topics split into more than two topics in the next generations.
- *fused*: Topics fused are the topics whose emergence has been made by more than two topics in the previous generations.
- *reborn*: Topics reborn are the topics that re-emerge after a user-specific time period, min_reborn. This state indicates that the topic reborn has attention after min_reborn, while during min_reborn it had been unnoticed.
- *dead*: Topics dead are the topics that go into unobserved (being unpopular) during min_dead and do not contain any descendent topics. During min_dead, if a topic rarely influences the generation of topics in the next generation, we regard it as dead topics.
- *flourishing*: Topics flourishing are the topics continually and actively being used or influencing on the generation of some topics in the next generations.

Note also that each topic can be marked by two evolutionary states: its emergence reason from the past generation, and its influence on the next generation. The former state is called 'emerging-state' as it captures how it emerges. The latter is called 'evolving-state' as it reveals its evolving state for topics in the next generations. The evolution states are integrated with the TET, and this unified knowledge identifies the trajectories of the underlying topics. Finally, the underlying topics are positioned considering their weights, given by the `weight` field in $\mathcal{P}$, on the y-axis (see also Fig. 1). This helps us to identify which topics more important than the others.

[5] https://www.rdocumentation.org/packages/plotrix/versions/3.8-2/topics/thigmophobe.labels



(a) Before choosing the best parent of F

(b) After choosing the best parent of F

Fig. 3. An example of the trajectories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
An example temporal topic profile $\mathcal{P}$. See also `softwarex_example_topic_profile.csv` under the `data` directory of the Github URL.

Id	Index	Label	Weight	Year	Words
t1	0	A	0.75	2001	['opinion', 'computer', 'lab', 'user', 'human']
t2	1	B	0.5	2001	['time', 'lab', 'opinion', 'human', 'computer']
t3	2	C	0.55	2002	['time', 'application', 'interface', 'user', 'computer']
t4	3	D	0.4	2002	['lab', 'user', 'abc', 'interface', 'application']
t5	4	E	0.1	2002	['application', 'interface', 'abc', 'opinion', 'computer']
t6	5	F	0.8	2003	['interface', 'computer', 'application', 'response', 'system']
t7	6	G	0.5	2003	['lab', 'abc', 'survey', 'opinion', 'time']
t8	7	H	0.7	2004	['machine', 'lab', 'system', 'response', 'human']
t9	8	I	0.5	2004	['opinion', 'interface', 'time', 'application', 'lab']
t10	9	J	0.65	2005	['opinion', 'human', 'user', 'survey', 'system']
t11	10	K	0.45	2005	['application', 'interface', 'system', 'survey', 'abc']

Table 3
An example TES **M**. The labels A-K are provided for the illustration only. In the actual input TES, only a $N \times N$ matrix without the labels should be given, where N is the number of the topics in $\mathcal{P}$. See also `softwarex_example_tes_matrix.csv` under the `data` directory of the Github URL.

	A	B	C	D	E	F	G	H	I	J	K
A	1	0	0.1	0.1	0.1	0.9	0.1	0.7	0.1	0.1	0.1
B		1	0.3	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1
C			1	0	0	0.1	0.1	0.1	0.1	0.1	0.1
D				1	0	0.9	0.1	0.1	0.1	0.1	0.1
E					1	0.2	0.3	0.1	0.1	0.1	0.1
F						1	0	0.1	0.1	0.1	0.1
G							1	0.1	0.75	0.1	0.1
H								1	0	0.9	0.9
I									1	0.1	0.1
J										1	0
K											1

3. Illustrative example

We present an illustrative example of TopicTracker which can also be reproduced in the provided Github URL. Another larger example is also provided in Appendix. Assume that the following temporal topic profile $\mathcal{P}$ in Table 2 and the TES matrix **M** in Table 3 are given to TopicTracker. In $\mathcal{P}$, suppose that the `words` field contains the top-4 words for each topic.

As seen in Table A.4–Table 3, there are 11 topics whose time slots are from 2001 to 2005. Given $\mathcal{P}$ and **M**, TopicTracker creates the TES shown in Fig. 2. We now discuss what insight we observe from this TET:

- The label of each topic comes from the `label` string in Table 2. If the user wants to display `id` and/or `words` as a label, the user can modify the code in `visualiseTES()`.[6]
- The topics at the same time slot are aligned together by the y-axis. The timeline is split by years 2001–2005 assuming that topics are captured and identified using a year.
- The TESs are represented by the 'Evolutionary strength' legend. For example, A influences the generation of both F and H, evidenced by A → F and A → H. The TES of A associated with F and H are assigned to the corresponding edges, respectively, using different colours. There is no meaning of the length of the edges.
- The evolutionary states of the topics are represented by different colours in the 'Evolution states' legend. Each topic has two evolution states: emerging-state and evolving-state. For example, the emerging-state of A is born as it newly emerges at 2001 without any ancestors. Its evolving-state is split as it influences the generation of both F and H as indicated by the directed edges. H's emerging-state is reborn, with min_reborn = 2 years, meaning that H had been unnoticed during min_reborn from 2001 to 2003, but noticed again by being influenced by A after min_reborn at 2004. Its evolving-state is split as it influences on the generation of both J and K.

  Uncoloured topics represent flourishing. Both emerging-states and evolving-states of topics D, G, I, J and K are flourishing. D and G are influencing the generation of other topics in the next generations: D → F, G → I. I, J and K are

[6] We have commented which block needs to be modified in the code

under the incubation period which is less than min_reborn from the latest observed time slot, 2005.
The emerging-state of C is flourishing as generated by the influence of B. Its evolving-state is dead as its influence had not been observed for the past 3 years from the latest time slot 2005, where the 3 years are greater than min_dead. The emerging-state of F is fused as generated by the co-influence of both A and D. Its evolving-state is dead with the same reason of that of C.

- The *y*-axis shows relative importance of topics. Referring to B → D, we see a declining trend of B (the weight decreases from 0.5 to 0.4) during 2001–2002, but that topic is getting more popular during 2002–2003 as indicated by D → F (the weight increases from 0.4 to 0.8).

4. Impact

Through a TET, we can observe the integrated, precious information about topic trajectories: evolutionary pathways of dynamic topics with evolution strengths indicated by the directed edges with different edge colours, evolution states of the topics indicated by different node colours, and the topic importance indicated by topic weight on the *y*-axis. The goal of building a TET is to build the most likely genealogy of non-contemporary topics over time. This TET models topic trajectory information: the evolution pathways between non-contemporary topics based on their evolution strengths.

TopicTracker enlightens credible evolutionary relationships between non-contemporary topics. For example, in science and technology domains, TopicTracker could contribute to uncovering how technological or knowledge topics can change and influence the generation of newer topics through a series of evolution events over time. In e-commerce domains, TopicTracker could also help to track the evolution of market-competitive product-related topics in a product market over time.

TopicTracker is designed to generate topic trajectory identification and visualisation with few parameters. It can run with two simple data formats as explained in Section 2: a temporal topic profile, and a TES matrix. For any models that can provide these formats, their ability to identify topic trajectories can be easily analysed and visualised by TopicTracker. Note that TopicTracker differs from existing graph visualisation tools (denoted as GVTs for the sake of explanation below) such as Gephi[7] in that (1) GVTs aim to visualise relationships between nodes, not incorporating temporal aspects of nodes, while TopicTracker aims to visualise evolutionary relationships between nodes (i.e., topics) considering their temporal property; (2) Accordingly, GVTs cannot detect evolution states of nodes, while TopicTracker has been designed to enlighten such information; and (3) GVTs usually incorporate numerous visualisation layouts, however, none of them visualise nodes and edges based on the temporal and weights of the nodes as TopicTracker does.

We believe that TopicTracker could help the user to allow a greater focus on methodological developments of their evolution strength matrix between time-stamped topics, rather than their implementation for topic trajectory identification. To the best of our knowledge, TopicTracker is the first generation of a framework that can be used in broader communities interested in identifying and visualising topic trajectory information.

[7] https://gephi.org/

5. Conclusions

In this paper, we presented a platform, TopicTracker, that can identify and visualise topic trajectory information. TopicTracker is equipped with the capability of addressing the issues still remained in the information retrieval community: how to identify trajectories of evolving topics over time? how to represent evolution states of the underlying topics at a particular time slot? and how to visualise topic trajectory information? As the backbone of topic trajectory information, we presented that TopicTracker uses TET which aims to induce a most likely genealogy tree for evolving topics. We presented the formal definition of TET, its constituent elements, and detailed descriptions about how to construct a TET from two kinds of input data: a temporal topic profile and a TES matrix showing the evolution strengths among the topics in the profile. Another key strength of TopicTracker is its ability to visualise three facets of useful trajectory information in a TET together: the evolutionary pathways of dynamic topics with their inter-evolution strengths, their evolution states at a particular time, and their relative importance. We believe that TopicTracker can provide a platform for topic trajectory analysis that can promote ease-of-use. As future work, we plan to extend TopicTracker in the way that enables a user to analyse TET properties and structure interactively that improves the capability of TopicTracker. Also, a usability test would improve the design and effectiveness of TopicTracker.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared my code and data in the mentioned URL in the paper

Acknowledgements

This work was supported by the LP170100416 project, funded by an Australian Research Council's Linkage grant. We thank Professor Beth Webster, Director of the Centre for Transformative Innovation at Swinburne University of Technology, for providing construct advice and support in developing TopicTracker. We also thank both Dr. Don Klinkenberg and Dr. Thibaut Jombart who are the creators of the 'phybreak' and 'seqtrack' packages, respectively, for inspiring the foundation of TopicTracker. Special thanks to Dr. Don Klinkenberg for providing valuable suggestions for implementing TopicTracker.

Appendix. TopicTracker example

In this Appendix, we provide an example of a topic evolution visualisation using 4 topics. These topics were extracted from the international patent corpora from United States Patent and Trademark Office (USPTO) data using topic modelling.[8] Patent data have been one of the obvious choices for analysis of topic trajectory identification. In the patent domain, topics are seen as technological or knowledge concepts and topic trajectory is focused on discovering evolutionary pathways of such topics over time. The USPTO patent data during 2000 to 2018 were collected with their CPC (Cooperative Patent Classification) data. The CPC

[8] Presenting details of the adopted topic modelling and how the temporal topic profile and the TES matrix is out of scope of this paper.

Table A.4
An example temporal topic profile $\mathcal{P}$. See also `softwarex_example_A61K_topic_profile.csv` under the `data` directory of the Github URL.

Id	Index	Label	Weight	Year	Words
t1	0	T2	0.4	2000	[antigen, adjuvant, epitope, immune]
t2	1	T3	0.45	2000	[tumor, carcinoma, lung, metastasis, breast]
t3	2	T4	0.7	2000	[virus, influenza, hepatitis, immunodeficiency, herpes]
t4	3	T1	0.2	2002	[immunoglobulin, region, light, mrna]
t5	4	T2	0.3	2002	[antigen, adjuvant, epitope, immune]
t6	5	T3	0.5	2002	[tumor, carcinoma, lung, metastasis, breast]
t7	6	T4	0.55	2002	[virus, influenza, hepatitis, immunodeficiency, herpes]
t8	7	T2	0.33	2004	[antigen, adjuvant, epitope, immune]
t9	8	T3	0.45	2004	[tumor, carcinoma, lung, metastasis, breast]
t10	9	T4	0.4	2004	[virus, influenza, hepatitis, immunodeficiency, herpes]
t11	10	T1	0.25	2006	[immunoglobulin, region, light, mrna]
t12	11	T2	0.6	2006	[antigen, adjuvant, epitope, immune]
t13	12	T3	0.4	2006	[tumor, carcinoma, lung, metastasis, breast]
t14	13	T4	0.3	2006	[virus, influenza, hepatitis, immunodeficiency, herpes]
t15	14	T1	0.3	2008	[immunoglobulin, region, light, mrna]
t16	15	T2	0.67	2008	[antigen, adjuvant, epitope, immune]
t17	16	T3	0.55	2008	[tumor, carcinoma, lung, metastasis, breast]
t18	17	T4	0.73	2008	[virus, influenza, hepatitis, immunodeficiency, herpes]
t19	18	T1	0.45	2010	[immunoglobulin, region, light, mrna]
t20	19	T2	0.55	2010	[antigen, adjuvant, epitope, immune]
t21	20	T3	0.65	2010	[tumor, carcinoma, lung, metastasis, breast]
t22	21	T4	0.6	2010	[virus, influenza, hepatitis, immunodeficiency, herpes]
t23	22	T2	0.77	2012	[antigen, adjuvant, epitope, immune]
t24	23	T3	0.65	2012	[tumor, carcinoma, lung, metastasis, breast]
t25	24	T4	0.55	2012	[virus, influenza, hepatitis, immunodeficiency, herpes]
t26	25	T2	0.4	2014	[antigen, adjuvant, epitope, immune]
t27	26	T3	0.7	2014	[tumor, carcinoma, lung, metastasis, breast]
t28	27	T4	0.6	2014	[virus, influenza, hepatitis, immunodeficiency, herpes]
t29	28	T1	0.65	2016	[immunoglobulin, region, light, mrna]
t30	29	T2	0.55	2016	[antigen, adjuvant, epitope, immune]
t31	30	T3	0.85	2016	[tumor, carcinoma, lung, metastasis, breast]
t32	31	T4	0.5	2016	[virus, influenza, hepatitis, immunodeficiency, herpes]
t33	32	T2	0.5	2018	[antigen, adjuvant, epitope, immune]
t34	33	T3	0.65	2018	[tumor, carcinoma, lung, metastasis, breast]

Table A.5
A 34 × 24 TES matrix. See also `softwarex_example_A61K_matrix.csv` under the `data` directory of the Github URL.

	T2	T3	T4	T1	T2	T3	T4	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T2	T3	T4	T1	T2	T3	T4	T2	T3
T2	1.00	0.20	0.06	0.75	0.87	0.16	0.42	0.35	0.41	0.27	0.05	0.45	0.09	0.34	0.13	0.36	0.18	0.05	0.17	0.05	0.23	0.37	0.16	0.22	0.06	0.36	0.38	0.20	0.19	0.36	0.33	0.36	0.38	0.12
T3		1.00	0.15	0.01	0.41	0.70	0.02	0.42	0.09	0.24	0.20	0.42	0.23	0.34	0.20	0.28	0.22	0.37	0.36	0.34	0.24	0.42	0.35	0.90	0.04	0.70	0.49	0.38	0.37	0.40	0.16	0.40	0.43	0.49
T4			1.00	0.24	0.48	0.47	0.92	0.02	0.04	0.30	0.43	0.11	0.10	0.43	0.10	0.00	0.02	0.37	0.07	0.39	0.34	0.01	0.44	0.35	0.12	0.14	0.09	0.19	0.08	0.40	0.47	0.46	0.34	0.12
T1				1.00	0.30	0.22	0.21	0.40	0.47	0.27	0.85	0.49	0.08	0.35	0.34	0.11	0.40	0.44	0.41	0.32	0.29	0.45	0.46	0.10	0.31	0.48	0.10	0.16	0.41	0.38	0.09	0.34	0.07	0.18
T2					1.00	0.12	0.16	0.90	0.25	0.49	0.09	0.05	0.15	0.04	0.12	0.15	0.41	0.15	0.41	0.26	0.45	0.41	0.35	0.38	0.09	0.41	0.42	0.48	0.34	0.29	0.29	0.07	0.48	0.18
T3						1.00	0.01	0.23	0.95	0.21	0.43	0.27	0.18	0.23	0.01	0.29	0.33	0.20	0.44	0.29	0.06	0.48	0.22	0.28	0.40	0.19	0.26	0.36	0.09	0.01	0.40	0.49	0.16	0.17
T4							1.00	0.27	0.22	0.93	0.25	0.18	0.41	0.90	0.19	0.05	0.16	0.33	0.05	0.29	0.08	0.15	0.45	0.02	0.11	0.03	0.49	0.01	0.11	0.26	0.11	0.13	0.48	0.15
T2								1.00	0.35	0.46	0.36	0.90	0.13	0.03	0.47	0.16	0.07	0.21	0.30	0.30	0.28	0.17	0.47	0.43	0.42	0.11	0.15	0.16	0.34	0.35	0.10	0.17	0.10	0.44
T3									1.00	0.09	0.45	0.35	0.90	0.05	0.25	0.00	0.43	0.24	0.36	0.21	0.22	0.06	0.20	0.33	0.39	0.18	0.20	0.48	0.40	0.23	0.30	0.06	0.10	0.23
T4										1.00	0.15	0.05	0.34	0.95	0.42	0.25	0.10	0.31	0.48	0.48	0.42	0.25	0.10	0.21	0.11	0.09	0.15	0.40	0.25	0.10	0.40	0.25	0.19	0.15
T1											1.00	0.44	0.44	0.45	0.90	0.00	0.49	0.01	0.29	0.38	0.46	0.31	0.42	0.44	0.02	0.40	0.10	0.28	0.38	0.27	0.28	0.03	0.09	0.35
T2												1.00	0.25	0.44	0.29	0.90	0.02	0.49	0.06	0.48	0.34	0.37	0.34	0.05	0.27	0.15	0.11	0.29	0.46	0.45	0.32	0.22	0.06	0.27
T3													1.00	0.45	0.31	0.50	0.90	0.06	0.15	0.07	0.27	0.00	0.06	0.20	0.29	0.04	0.22	0.04	0.36	0.43	0.31	0.47	0.07	0.09
T4														1.00	0.04	0.32	0.44	0.90	0.09	0.20	0.19	0.45	0.48	0.20	0.02	0.02	0.30	0.41	0.43	0.03	0.12	0.34	0.18	0.44
T1															1.00	0.36	0.33	0.18	0.65	0.25	0.01	0.39	0.19	0.23	0.41	0.13	0.11	0.22	0.30	0.07	0.25	0.38	0.06	0.15
T2																1.00	0.19	0.33	0.14	0.85	0.03	0.15	0.38	0.21	0.25	0.23	0.44	0.41	0.13	0.07	0.27	0.38	0.38	0.08
T3																	1.00	0.24	0.30	0.23	0.95	0.29	0.01	0.26	0.26	0.10	0.23	0.29	0.46	0.45	0.43	0.45	0.38	0.17
T4																		1.00	0.34	0.48	0.46	0.90	0.41	0.08	0.11	0.39	0.19	0.12	0.13	0.46	0.08	0.26	0.26	0.35
T1																			1.00	0.09	0.38	0.10	0.44	0.17	0.13	0.21	0.42	0.08	0.36	0.15	0.16	0.27	0.13	0.49
T2																				1.00	0.20	0.03	0.90	0.22	0.30	0.70	0.04	0.00	0.43	0.21	0.14	0.44	0.03	0.44
T3																					1.00	0.12	0.05	0.90	0.00	0.13	0.18	0.16	0.33	0.19	0.48	0.23	0.01	0.32
T4																						1.00	0.26	0.42	0.90	0.04	0.07	0.20	0.12	0.48	0.07	0.24	0.31	0.11
T2																							1.00	0.13	0.25	0.90	0.30	0.39	0.80	0.49	0.19	0.26	0.24	0.20
T3																								1.00	0.35	0.46	0.90	0.48	0.07	0.23	0.35	0.14	0.15	0.00
T4																									1.00	0.16	0.27	0.90	0.25	0.09	0.13	0.02	0.27	0.38
T2																										1.00	0.00	0.36	0.43	0.90	0.44	0.39	0.16	0.43
T3																											1.00	0.40	0.10	0.30	0.90	0.04	0.19	0.30
T4																												1.00	0.48	0.31	0.35	0.90	0.20	0.26
T1																													1.00	0.22	0.23	0.28	0.18	0.16
T2																														1.00	0.22	0.19	0.85	0.24
T3																															1.00	0.27	0.39	0.90
T4																																1.00	0.09	0.48
T2																																	1.00	0.16
T3																																		1.00

code data have bee effectively used to index patents. The CPC code data is divided into nine major sections, A-H and Y, which in turn are further divided into classes, sub-classes, groups and subgroups. The topics used in this example were indexed with 'A61K' (PREPARATIONS FOR MEDICAL, DENTAL, OR TOILET PURPOSES: *devices or methods specially adapted for bringing pharmaceutical products into particular physical or administering forms A61J 3/00; chemical aspects of, or use of materials for deodorisation of air, for disinfection or sterilisation, or for bandages, dressings, absorbent pads or surgical articles A61L; soap compositions C11D)*).

The temporal topic profile of the 4 topics are presented in Table A.4, where we observe that these are the labelled as T1,

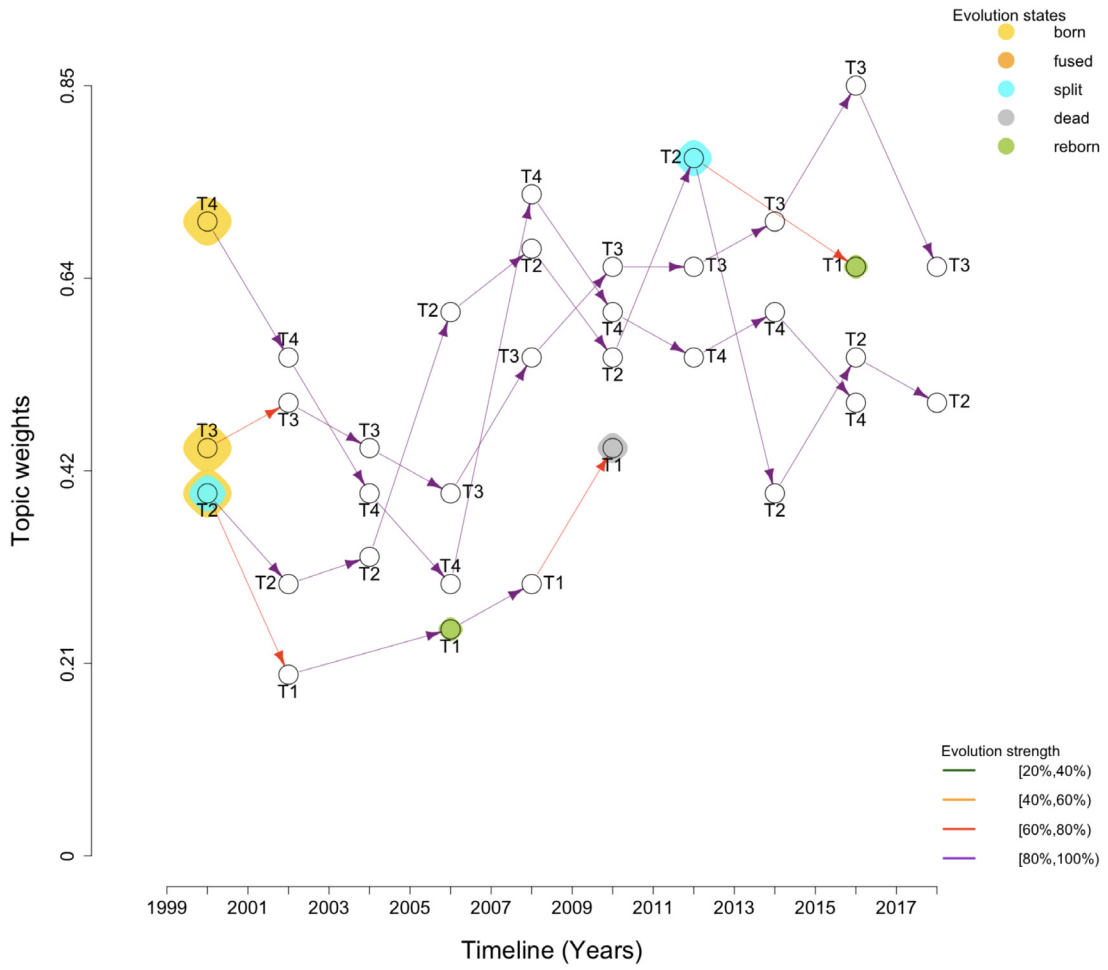


Fig. A.4. An example TET generated by TopicTracker. The parameters are set as: min_reborn = 2 years, min_dead = 1 year, and min_tes = 0.5.

T2, T3 and T4, their top-5 words, and weights per year during 2000 and 2018. Table A.5 shows the topic evolution strength (TES) matrix used to create an topic evolution tree (TET).

Given the above topic profile and TES, the following TET in Fig. A.4 is generation using TopicTracker. We can observer the following from the TET:

- As the colours of the edges are indicated, this example includes the evolution strengths which are equal to or greater than min_tes = 0.5.
- There are 3 newly born topics T2, T3, and T4 on 2000.
- There is one dead topic T1 on 2010 which does not have any descendent topics since then.
- The topic T1 was reborn on 2006 whose ancestor is topic T1 on 2022, and also reborn on 2016 affected by T2 on 2022.
- The topic T2 was split on 2000 and 2014. T2 on 2000 influenced strongly (as the edge strengths indicated) the generation of both T2 and T1 on 2002. That topic also influenced generation of T2 on 2014 and T1 on 2016. This means that sometimes some knowledge and techniques (concept) can be somehow used or adopted to create some other innovations.
- We see that most of the topics were flourishing throughout the timeline (2000–2018) except T1.

References

[1] Lee WS, Sohn SY. Identifying emerging trends of financial business method patents. Sustainability 2017;9(9):1–21, URL https://ideas.repec.org/a/gam/jsusta/v9y2017i9p1670-d112597.html.

[2] Zhang H, Kim G, Xing EP. Dynamic topic modeling for monitoring market competition from online text and image data. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2015, p. 1425–34. http://dx.doi.org/10.1145/2783258.2783293.

[3] Greene D, Cross JP. Exploring the political agenda of the European parliament using a dynamic topic modeling approach. 2016, CoRR abs/1607.03055 URL http://dblp.uni-trier.de/db/journals/corr/corr1607.html#GreeneC16.

[4] Song J, Huang Y, Qi X, Li Y, Li F, Fu K, et al. Discovering hierarchical topic evolution in time-stamped documents. J Assoc Inform Sci Technol 2016;67(4):915–27. http://dx.doi.org/10.1002/asi.23439, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23439.

[5] Zhou H-k, Yu H-m, Hu R. Topic discovery and evolution in scientific literature based on content and citations. Front Inf Technol Electron Eng 2017;18(10):1511–24. http://dx.doi.org/10.1631/fitee.1601125, URL https://app.dimensions.ai/details/publication/pub.1099692062.

[6] Song M, Heo GE, Kim SY. Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP. Scientometrics 2014;101(1):397–428. http://dx.doi.org/10.1007/s11192-014-1246-2.

[7] Yoon J, Kim K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics 2011;88(1):213–28. http://dx.doi.org/10.1007/s11192-011-0383-0.

[8] Zhang Y, Zhang G, Zhu D, Lu J. Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. J Assoc Inform Sci Technol 2017;68(8):1925–39. http://dx.doi.org/10.1002/asi.23814, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23814.

[9] He Q, Chen B, Pei J, Qiu B, Mitra P, Giles L. Detecting topic evolution in scientific literature: How can citations help? In: Proceedings of the 18th ACM conference on information and knowledge management. New York, NY, USA: Association for Computing Machinery; 2009, p. 957–66. http://dx.doi.org/10.1145/1645953.1646076.

[10] Jo Y, Hopcroft JE, Lagoze C. The web of topics: Discovering the topology of topic evolution in a corpus. In: Proceedings of the 20th international conference on world wide web. New York, NY, USA: Association for Computing Machinery; 2011, p. 257–66. http://dx.doi.org/10.1145/1963405.1963444.

[11] Jung S, Yoon WC. An alternative topic model based on common interest authors for topic evolution analysis. J Informetr 2020;14(3):101040. http://dx.doi.org/10.1016/j.joi.2020.101040, URL http://www.sciencedirect.com/science/article/pii/S1751157719303517.

[12] Zhong S, Verspagen B. The role of technological trajectories in catching-up-based development: An application to energy efficiency technologies. MERIT Working Papers 013, United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT); 2016, URL https://ideas.repec.org/p/unm/unumer/2016013.html.

[13] Park H, Magee CL. Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. PLOS ONE 2017;12(1):1–18. http://dx.doi.org/10.1371/journal.pone.0170895.

[14] Triulzi G, Alstott J, Magee CL. Estimating technology performance improvement rates by mining patent data. Technol Forecast Soc Change 2020;158:120100. http://dx.doi.org/10.1016/j.techfore.2020.120100, URL http://www.sciencedirect.com/science/article/pii/S0040162520309264.

[15] Huang Y, Zhu F, Porter AL, Zhang Y, Zhu D, Guo Y. Exploring technology evolution pathways to facilitate technology management: From a technology life cycle perspective. IEEE Trans Eng Manage 2020:1–13.

[16] Qiu Z, Wang Z. Technology forecasting based on semantic and citation analysis of patents: A case of robotics domain. IEEE Trans Eng Manage 2020;1–21.

[17] Gaul W, Vincent D. Evaluation of the evolution of relationships between topics over time. Adv Data Anal Classif 2017;11(1):159–78. http://dx.doi.org/10.1007/s11634-016-0241-2.

[18] Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning. New York, NY, USA: Association for Computing Machinery; 2006, p. 113–20. http://dx.doi.org/10.1145/1143844.1143859.

[19] O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 2015;42(13):5645–57.