

Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs

Journal of Information Science
XX(X):2–35

©The Author(s) 2020
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/



Mauro D L Tosi¹ and Julio Cesar dos Reis¹

Abstract

The process of tracking the evolution of a scientific field is arduous. It allows researchers to understand trends in areas of science and predict how they may evolve. Nowadays, most of the automated mechanisms developed to assist researchers in this process do not consider the content of articles to identify changes in its structure, only the articles metadata. These methods are not suited to easily assist researchers to study the concepts that compose an area and its evolution. In this article, we propose a method to track the evolution of a scientific field at a concept-level. Our method structures a scientific field using two knowledge graphs, representing distinct periods of the studied field. Then, it clusters them and identifies correspondent clusters between the knowledge graphs, representing the same sub-areas in distinct time periods. Our solution enables to compare the corresponding clusters, tracking their evolution. We apply and experiment our method in two case studies concerning the Artificial Intelligence and the Biotechnology fields. Findings indicate befitting results regarding the way their evolution can be assessed with our implemented software tool. From our analyses, we perceived evolution in broader sub-areas of a scientific field, as the growth of the “Convolutional Neural Network” area from 2006; to specific ones, as the decrease of researches using mice to study BRAF-mutation lung cancer from 2018. This work contributes with the development of a web application with interactive user interfaces to assist researchers in representing, analyzing, and tracking the evolution of scientific fields at a concept-level.

Keywords

Knowledge Representation, Knowledge Graphs, Scientific Evolution

Introduction

The process of studying a scientific field or to do a literature review is arduous for both newcomers and expert researchers in a specific area.

¹Institute of Computing, University of Campinas, Campinas - SP, Brazil

The increasing amount of publications (1) and, consequently, the coming era of big scholarly data have aggravated this issue, making it more laborious for researchers to search for relevant documents related to their studies (2). Therefore, they usually go after review articles to help them understanding how the field under study is organized. After familiarizing themselves with the state of the art of the studied field, researchers must maintain themselves updated reading novel findings published in prestigious conferences and journals.

Both the processes of studying a new scientific field, or staying up-to-date with fresh discoveries in a previously known field are time-consuming. A common reason for this problem is that survey articles may be outdated or even do not exist for a desired area. Consequently, researchers have to spend a significant part of their research-time reading content, sometimes irrelevant to their studies, until finding the right articles related to their investigations. When keeping themselves updated with fresh discoveries, researchers do not have time to read all newly published articles in their areas.

Recent investigations have developed mechanisms to understand the structure and the evolution of scientific fields. Most of those approaches make inferences based on manually annotated data or metadata information from articles that belong to the studied field(3)(4). However, these characteristics limit the understanding of a field, as they are unfeasible on large-scale or do not consider the semantics of content from its articles. In this context, Tosi and dos Reis (5) proposed SciKGraph, a framework to structure a scientific field as a knowledge graph (KG) at a concept level, considering a KG a structure that integrates information into a knowledge base and applies a reasoner to generate new knowledge from it. Instead of representing a scientific field using only metadata information, SciKGraph uses the concepts extracted from texts on academic articles to represent and structure it. SciKGraph clusters the constructed knowledge graph by identifying the main sub-areas of the studied field, their relation and the concepts belonging to them. Still, their framework only represents a fixed time-period of the studied field and does not allow us to analyze its evolution over different periods of time.

It is not a trivial task the adequate evolution tracking of a scientific field considering how it evolved semantically, examining its concepts, not just its metadata. The study of the evolution of a scientific field at

a concept level requires investigating how to track changes in its structure. We describe the structure of a scientific field as the set of its sub-areas and their relations, containing the concepts that belong to it along with how they are correlated and connected among themselves. Therefore, the evolution tracking of a scientific field becomes more challenging than only detecting and representing inserted or deleted concepts. Considering that not just its structure, but all concepts that shape it can be modified, it is hard to determine if some sub-area of the structure changed, was replaced, or suffered from noise data. For example, one could determine that a sub-area a_1 did not evolve to a_2 because they have only 30% of their concepts in common. However, 90% of the time, a_1 is referred to by one of those common concepts. Thus, the other 70% of not similar concepts, despite being part of a_1 , caused a miss classification and should have been considered noise.

In this article, we propose an approach to track the evolution of a scientific field at a concept level. We first define two scientific field periods from which we shall track the evolution. Then, we use SciKGraph (5) to construct knowledge graph representations of those two periods, clustered in their main sub-areas. Next, our solution calculates the similarity of all clusters from both knowledge graphs. Our method is suited to identify clusters from distinct knowledge graphs that represent the same sub-area of the scientific field. Finally, after determining which clusters represent the same sub-area, we compare them and track their evolution, identifying concepts added or excluded from the analyzed sub-area between the periods analyzed.

In addition, with the aim of computationally assisting researchers understanding how a scientific field organized at a concept level evolves, we provide an application software to facilitate this process. Our solution is a web application designed for researchers without programming skills or background knowledge in the area. Our application contains graphical interfaces to assist users to represent a scientific field as a knowledge graph, analyze its characteristics, and track its evolution. All these features allow graphical representations of the scientific field.

We evaluate the proposed method and application software in two case studies from distinct knowledge areas. The first one is the Artificial Intelligence scientific field, represented using 1,002 academic articles. Based on the proposed approach, we found the possibility of tracking

the evolution that occurred on the “Image Analysis” and the “Neural Network” sub-areas from 2006. As an example identified with our tool from this period, we observed a drop in “Image Analysis” articles using supervised learning; also, the increase of popularity of “Convolutional Neural Networks”. We also explore our tool to analyze the Biotechnology scientific field, represented by 8,964 abstracts. We structure the input textual documents in knowledge graphs and analyze their evolution by comparing articles published between 2014-2016 and 2018-2020. As an example, this analysis allowed us to detect via the software tool that researches using mice to study BRAF-mutation lung cancer decreased.

Our proposal and software application contributes to assist researchers in understanding how a scientific field is organized on a concept level. In addition, our solution enables to track and understand the evolution of the identified topics (represented as clusters) in different knowledge graphs. Our software application provides graphical visualizations for the analyzed fields, their sub-areas, and their relations.

The remaining of this article is organized as follows: Section “Background” introduces background studies. Section “Tracking the evolution of a scientific field” presents the proposed method to track the evolution of the scientific field. Section “Software tool for evolution analysis” fully describes our constructed and available to use application software. Section “Experimental Results” reports on our experimental analyses. Section “Discussion” discusses the obtained findings. Section “Conclusion” concludes this article with our final considerations.

Background

Scientific knowledge has been structured differently over the past years. Nowadays, publishers and researchers usually use one of four methods to this end — citation networks, manual assignment, classification techniques, or knowledge graphs. Citation networks study the topology of the citations among papers to determine how to structure scientific knowledge (3). Manual assignments depend on experts to determine how to segment the areas and sub-areas of a scientific-field, like the ones in (6). Classification techniques determine in which sub-area of a scientific field a document belongs (7), usually considering the sub-areas previously identified by citation networks, or manual assignments. Knowledge graphs

structure a scientific-field based only on the textual content of its articles, using them to segment the scientific-field sub-areas (5).

Researchers use some of those methods to identify how the structure of scientific-fields changes over time. Citation network methods are common for this end. Jung and Segev (8), for example, studied not only the evolution of a scientific-field, but inferred future changes in it. Hopcroft *et al.* (4) proposed to track evolving communities in citation networks. They identified groups of communities, also known as covers (9), from citation networks built with documents from distinct time-periods. Then, they classified if clusters from distinct covers were similar enough to be representing the same sub-area. If they were, they could compare those clusters and track sub-areas' evolution.

The use of the Normalized Mutual Information (NMI) metric has been used to train classification methods, or to identify if a clusterization algorithm effectively segmented a citation network in its main sub-areas. This allows comparing the cover generated by the analyzed algorithm with a baseline cover. NMI is a metric that varies from 0 to 1, which evaluates how close two covers are between themselves. The closer the NMI value is to 1, the more similar the covers are. Chakraborty and Chakraborty (10) adopted this approach in which they used a variant of the NMI metric, optimized for overlapping clusters, to evaluate their algorithm results identifying overlapping clusters in citation networks.

These and other citation network approaches study only the metadata of the articles. The tracking of changes concerning the structure of a scientific-field based on its content remains an open research challenge. If one wants to identify how the structure of a scientific-field evolved at a concept level, the usage of metadata-dependent-only techniques is not adequate. On the other hand, review articles (manual assignment method) could be a solution to this issue. Though many scientific fields do not have review articles, or they are out-dated.

Furthermore, Amancio *et al.* (11) semantically assessed approaches for citations in scientific papers using complex networks. They conclude that factors that may not consider the scientific merit are used in those approaches by researchers, and recommend a method that researchers may use to identify relevant references without bias. Therefore, they advocate for the development of less biased software tools to assist researchers in better understanding the literature. Also, Amancio, Oliveira

Jr, and Fontoura Costa (12) studied how specific metadata and non-metadata parameters (number of citations, date of publication, and content similarity) influence citation networks and further impact authors h-index. Similarly to Amancio *et al.* (11), they concluded that not only the content similarity between articles influence citations, but also the date of publication, number of citations, and the reputation of their authors and institutions. Considering this, we point out that the bias included in citation networks is another indication that semantically oriented methods must be used to improve literature understanding.

The usage of knowledge graph approaches to represent scholarly data and scientific knowledge is very recent in literature. The study conducted by Vahdati *et al.* (13) tackled the problem of knowledge discovery in scholarly knowledge graphs. They used a knowledge-driven framework able to unveil scholarly communities for the prediction of scholarly networks. Results observed from their evaluations suggested that exploiting semantics in scholarly knowledge graphs enables the identification of previously unknown relations between researchers. However, to the best of our knowledge, there is no work in literature aiming to analyze the evolution of knowledge graphs in the context of scientific knowledge.

Tosi and dos Reis (5) proposed SciKGraph, a knowledge graph framework to structure and represent a scientific-field. SciKGraph receives as input a compendium of textual documents used to represent a scientific-field. Then, SciKGraph identifies and disambiguates the concepts of those documents, and uses them as vertices of a co-occurrence knowledge graph. Next, by clustering the graph, the framework identifies the main sub-areas of the scientific-field. Finally, SciKGraph extracts key-concepts from the knowledge graph representing its topics. Tosi and dos Reis evaluated their framework using collections of documents from distinct fields of science, obtaining satisfactory results. This indicates that SciKgraph can be used to represent scientific knowledge independent from its field. Nevertheless, SciKGraph requires novel features to compare and compute similar clusters as well as to enable identifying change operations from one KG to other.

Our exploratory literature review indicates that there is no current method that can track how a scientific field structure evolves at a concept level. In this investigation, we propose a knowledge graph method to

track the evolution of a scientific field. We base our method on the SciKGraph framework (5), and combine techniques usually applied to citation networks to track the evolution of our knowledge graphs.

Tracking the evolution of a scientific field

This section describes how to structure a scientific field to track its evolution over time. The first subsection introduces how we use SciKGraph to represent a scientific field as a knowledge graph and extract its main sub-areas. The second subsection describes our approach to identify similar sub-areas in different knowledge graphs and how to use this information to track sub-areas' evolution.

Figure 1 presents our proposed methodology to track the evolution of a scientific field. First, it receives as input two sets of documents used to describe distinct periods of the same scientific field, 2010 and 2015 (as the example used in Figure 1). Next, it uses SciKGraph to construct a knowledge graph for each one of the input sets and extracts their main sub-areas (clusters). Finally, it compares the two knowledge graphs by identifying common sub-areas in both structures. The output refers to the dissimilarities between detected sub-areas in the different knowledge graphs. These dissimilarities represent what changed in the sub-areas during the time-window between both sets of documents (represented in bold and dotted elements in Figure 1). This enables describing concepts added or excluded from a sub-area during a time-window, which we describe here as the sub-area evolution.

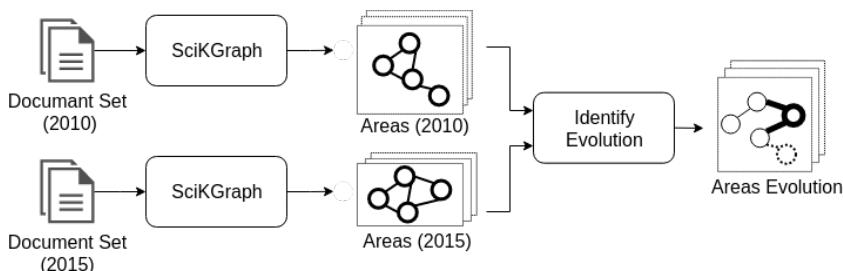


Figura 1. Methodology to track the evolution of a scientific field

Representing a Scientific Field as a Knowledge Graph

The representation of a scientific field as a knowledge graph enables the study of knowledge at a concept level. This allows users to understand concepts, their relations and structure. We represent scientific fields as knowledge graphs to track their evolution. To this end, we use the SciKGraph framework (5) (cf. Figure 2). First, SciKGraph receives as input a collection of scientific documents used to represent a scientific field. This is used to construct a knowledge graph that represents this collection. Second, SciKGraph clusters the constructed structure by identifying the scientific field's main sub-areas. Finally, SciKGraph extracts knowledge from the structures previously constructed.

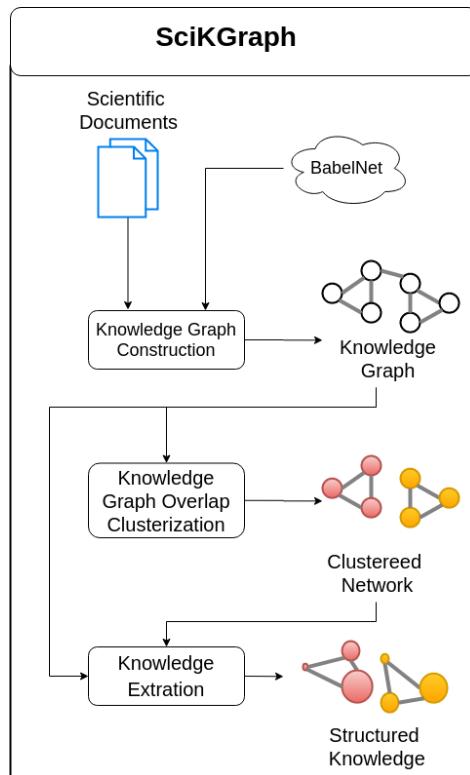


Figura 2. SciKGraph framework (5) to structure a scientific collection as a knowledge graph.

SciKGraph performs the “Knowledge Graph Construction” task, which takes as input a collection of scientific documents used to represent a scientific field, which in our case we use to represent a time-period

of this scientific field. This task parses those documents, excluding citations, images, and equations. Then, it uses Babelfy (14), a text disambiguation software to perform word sense disambiguation - the task to computationally find the contextual meaning of words (15) - and link corresponding concepts between our documents and BabelNet(16), a multilingual semantic network. This operation returns concept babel synsets, which are their unique identification codes.

Next, this task constructs the knowledge graph by using the identified corresponding concepts as vertices and their co-occurrence in the text as undirected edges, both weighted based on the number of times they appeared in the text. Figure 3 presents an example, by receiving as input the phrase “Knowledge graphs can structure knowledge [72].”. The solution parses it excluding a citation (pre-processing); identifies the “knowledge graph”, “structure”, and “knowledge” as correspondent concepts with BabelNet; constructs the knowledge graph taking the three concepts as vertices and their direct co-occurrence as edges, weighted based on the number of times they appeared in the whole collection. We note that if those concepts or their co-occurrences appear again in other documents of the collection, their weights in the knowledge graph are updated considering these events.

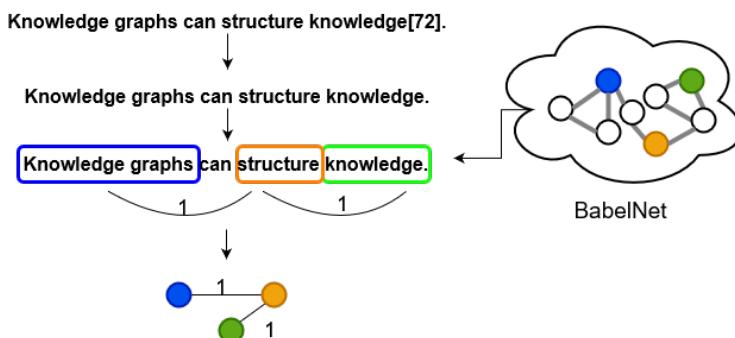


Figura 3. Example of generating a knowledge graph from an input sentence.

SciKGraph performs the “Knowledge Graph Overlap Clusterization” task (cf. Figure 2), which takes as input the knowledge graph previously constructed and clusters it, identifying overlapping sub-graphs, representing the scientific field sub-areas. The idea behind this task is that concepts that belong to the same sub-area co-occur more than those from

different ones because they tend to appear more times in the same context. Therefore, as clusterization techniques divide elements into groups that are more correlated with themselves than with others, SciKGraph states that when clustering a knowledge graph with edges weighted based on concepts co-occurrence, the clusters identified represent distinct sub-areas of the field the knowledge graph represents. Moreover, as a concept can belong to multiple sub-areas simultaneously, it is an intrinsic part of the problem to enable overlapping representations of sub-areas.

In the second step, to cluster the knowledge graph, SciKGraph uses the OClustR algorithm (17). It is a graph-based clusterization algorithm that automatically identifies the number of overlapping clusters on which it divides the knowledge graph. The user does not need to input the number of sub-areas to organize the scientific field because OClustR identifies it automatically. However, if the user desires to fix a number and reduces the number of clusters identified, SciKGraph proposes three agglomerative algorithms for this task (5), which we do not use to track the evolution of the identified sub-areas.

In the third step, the “Knowledge Extraction” task organizes and extracts knowledge from the knowledge graph, and its clusters (sub-graphs). It presents the sub-areas identified through the clusterization process and identifies their adequate labels using the C-Rank algorithm. C-Rank (18) is a non-supervised keyphrase extraction technique that identifies keyphrases from scientific documents receiving as input from the user only the document from which it has to extract the keyphrases. At this stage, keyphrases are not relevant for the identification of similar sub-areas in distinct knowledge graphs, nor to identify their evolution over time. Therefore, we do not further detail their extraction in this work.

Knowledge Graph parameters To optimize the clusterization, SciKGraph pre-processes the knowledge graph. This pre-processing contains three steps: (1) it removes edges weighted below a certain $threshold_{edges}$, reducing noise; (2) it excludes vertices with higher degree centrality that are inside a $threshold_{centrality}$, which eliminates general concepts that are relevant for the scientific field as a whole, but are too general to be relevant for a specific sub-area; (3) it discards small disjoint sub-graphs created after the previous steps.

In the performance of the pre-processing, two variables must be defined $threshold_{edges}$ and $threshold_{centrality}$; both of them depend

on the domain and the size of the input document collection. When one increases $threshold_{edges}$, the clusters produced are better defined and the clusterization has better accuracy, but contain less information, decreasing the number of vertices in the knowledge graph (5). By reducing $threshold_{centrality}$, the size of the identified clusters decreases, but the number of identified clusters increases, as less of their central nodes are being removed. To obtain optimal values, SciKGraph, based on the graph structure, recommends an interactive approach. This approach presents estimated values to the user, who can change them based on the observed generated output.

In this investigation, our objective is to track the evolution of a scientific field by comparing the structure of its knowledge graph representations in distinct periods. In this sense, the closer the knowledge graph representations are to each other, the less noise to identify the evolution of its sub-areas will be. Thus, we acknowledge the fact that the $threshold_{edges}$ is related to the number of vertices and edges in the knowledge graph, and the evidence that the number of vertices influences the structure of the identified clusters (5). On this basis, we shall set $threshold_{edges}$ values that result in knowledge graphs with a similar number of vertices.

In our analyses, we observed that representations of the studied scientific fields containing around 1,600 concepts held enough information to be further processed and generated clusters with high modularity. Therefore, we examined which $threshold_{edges}$ values we would need to apply to our representations to produce knowledge graphs with this amount of vertices. Based on these values, we identified a power series relation between the $threshold_{edges}$ and the number of edges in the knowledge graph ($|edges|$) (cf. Figure 4).

Figure 4 shows a chart with the relation between the number of edges of knowledge graphs ($|edges|$), and the $threshold_{edges}$ values we determined to pre-process them in order to obtain their representations with around 1,600 concepts each. This relation does not expect a constant value and is powered to 0.85. Moreover, setting the coefficient of the power series to 0.00014, we obtained the approximate number of vertices we expected. However, one can increase or decrease this coefficient to respectively raise or reduce the number of vertices of the representations in order to identify more specific or general sub-areas of the studied scientific field.

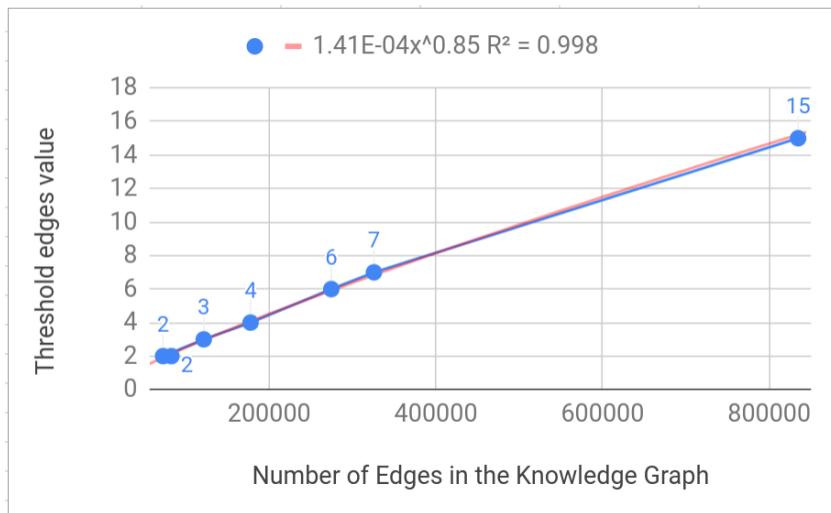


Figura 4. Power series relation between the number of edges and the $threshold_{edges}$ value to generate knowledge graphs with the same amount of vertices.

Therefore, considering this relation, when tracking the evolution of a scientific field, we use Equation 1 to determine $threshold_{edges}$ value, maintaining a similar amount of concepts in distinct scientific field representations. The equation coefficient is domain-dependent, and, as explained previously, can vary based on the scientific field analyzed and the number of articles used to represent it. In this sense, the equation coefficient follows the same trend of the $threshold_{edges}$ parameter, when one increases it, the clusterization algorithm has better accuracy. However, as it also removes more vertices, the knowledge graph loses more information.

$$threshold_{edges} = |edges|^{0.85} * 0.00014 \quad (1)$$

Identifying dissimilarities between Knowledge Graphs

We represent two time-periods of the same scientific field as distinct knowledge graphs. Our goal is to track the evolution of this scientific field during the time-window between both representations. The identification of concepts and connections that were added or excluded from the knowledge graphs during the analyzed time-window would give the researcher a shallow perspective of the evolution of the scientific field.

Therefore, to improve the researcher experience, instead of analyzing only the knowledge graph, we track the evolution of the scientific field by identifying the concepts added and/or excluded from each of the scientific field sub-areas (knowledge graph clusters).

In order to track the evolution of a cluster comparing distinct covers, it is necessary to identify the corresponding clusters in both covers. Figure 5 presents an example to illustrate the comparison of clusters from distinct knowledge graphs. Even though both covers have three clusters each, it would be inaccurate to compare the clusters labeled with the same numbers; this comparison would result in total dissimilarity between the covers. It requires further considering the meaning of concepts present in the clusters.

In our approach, we analyze the content of each cluster in both covers, so that we may identify that “Cluster #1” from “Cover 1” and “Cluster #3” from “Cover 2” are corresponding clusters in our example (corresponding clusters linked by dotted lines in Figure 5). By comparing the elements of the clusters, we can identify that both covers have the same structure, they only had their clusters sorted differently.

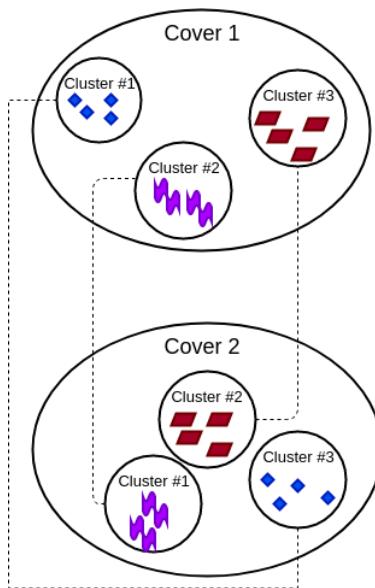


Figura 5. Example of covers comparison; correspondent clusters from distinct covers are linked by dotted lines.

We use a similarity measure to determine the correspondence ratio between two clusters c_1 and c_2 . It is based on the $match(c_1, c_2)$ equation used by Hopcroft *et. al* in (4) (cf. Equation 2), which quantifies how well two clusters match with each other, computing the number of intersection elements between the clusters; and normalizing it by the size of the bigger cluster. We chose this measure because its definition ensures that for clusters to have high similarity, they must roughly have the same size. Therefore, broader clusters will not have a high correspondence ratio with small ones just because they also contain almost all the elements that the small clusters have.

$$match(c_1, c_2) = \min\left(\frac{|c_1 \cap c_2|}{|c_1|}, \frac{|c_1 \cap c_2|}{|c_2|}\right) \quad (2)$$

However, instead of giving the same importance to all concepts in the clusters, our similarity measure weights the concepts based on their degree centralities in the clusters sub-graphs. Therefore, when analyzing a concept n , we consider its weight $weight(n)$ as the average of its degree centrality $central_c(n)$ in the sub-graphs c that it belongs (cf. Equation 3).

$$weight(n) = \begin{cases} central_{c1}(n), & \text{if } n \in c1, n \notin c2 \\ central_{c2}(n), & \text{if } n \notin c1, n \in c2 \\ \frac{central_{c1}(n)+centrality_{c2}(n)}{2}, & \text{if } n \in c1, n \in c2 \end{cases} \quad (3)$$

Furthermore, considering the weight of each concept to calculate the similarity between two clusters, our similarity measure $similarity(c_1, c_2)$ is defined as in Equation 4. That way, we reduce the relevance of miss classified noisy concepts by using a weighted function based on the centrality of the concepts in the sub-graphs to which they belong.

$$similarity(c_1, c_2) = \min\left(\frac{\sum_{n \in c_1 \cap c_2} weight(n)}{\sum_{n \in c_1} weight(n)}, \frac{\sum_{n \in c_1 \cap c_2} weight(n)}{\sum_{n \in c_2} weight(n)}\right) \quad (4)$$

After defining how to calculate the similarity measure between two clusters, we can identify the more similar ones. To this end, we calculate the similarity between every pair of clusters from different covers. In this procedure, each cluster from one cover has a certain similarity value with

all clusters from the other cover. At the final stage, we select the most similar pairs of clusters, which are those that have the similarity value above the $threshold_{similarity}$ value. This threshold value varies from 0 to 1 and determines to which extent the similarity between two clusters must be to consider them correspondent, or related.

After identifying correspondent clusters, representing the same sub-area in distinct time-periods, the researcher can directly analyze the concepts and relations added and excluded from them during the studied time-window. Considering this, one may visualize how this sub-area evolved at a concept-level, enabling researchers to, for example, observe how a specific concept impacted in the evolution of a sub-area. Practical examples and results obtained analyzing correspondent clusters in two case studies are presented in Section “Experimental Results”.

Figure 6 presents an example. Considering that all concepts in the example have the same degree centrality, “Cluster #1” from “Cover 1” has 0.75 of similarity with “Cluster #2” from “Cover 2”, and 0.25 with “Cluster #1” also from “Cover 2”. Therefore, defining a $threshold_{similarity}$ (0.5 for example), one may observe that “Cluster #1” from “Cover 1” and “Cluster #2” from “Cover 2” are correspondents. By reducing this $threshold_{similarity}$ number to 0.25, besides the correspondent clusters, one may also observe related ones, as the cluster from “Cover 1” and “Cluster #1” “from “Cover 2”. These relations, for example, can describe sub-areas that were merged, or split themselves in distinct ones or have overlapping concepts shared between themselves.

It is further possible to quantify the similarity of the knowledge graphs and their clusters as a whole. However, it is not simple to compare covers that allow overlapping clusters, and we cannot apply a procedure as used to compare just a pair of clusters. We use a metric proposed by McDaid, Greene, and Hurley (9) to quantify the similarity between the whole covers generated from both knowledge graphs. It is a Normalized Mutual Information metric used to evaluate overlapping clusterization algorithms. Usually, Mutual Information metrics calculate the similarity between a cover of clusters generated by the algorithm it is evaluating, and the control cover, containing correct segmented clusters. In this work, instead of using this metric to evaluate a clusterization algorithm, we use it to compare the covers obtained from clustering the two generated knowledge graphs. By calculating the amount of mutual information between the

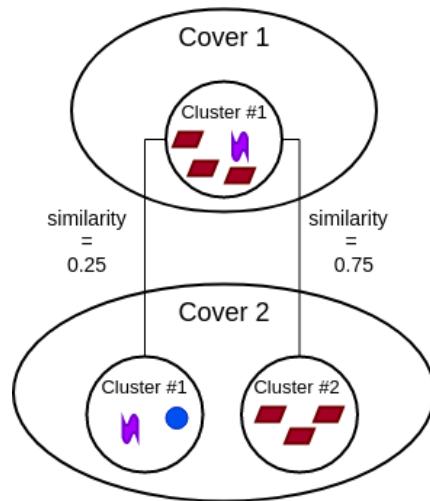


Figura 6. Example of similarities among clusters distinct covers.

covers, we can use this normalized value to inform the researcher of the similarity between the generated covers, or, in other terms, the similarity of the scientific field in the analyzed time-periods.

Software tool for evolution analysis

Our method proposed to track the evolution of a scientific field based on SciKGraph. It allows the analysis of the most diverse scientific areas, identifying concepts added and/or excluded from sub-areas of the studied scientific field. The Jupyter Notebook (19) interface proposed in SciKGraph limits its usability to researchers with at least a minimum programming knowledge. This limitation negatively impacts the usage of the framework for a significant part of the scientific community. At this stage, our goal is to enable the usage of our method by researchers non literate in programming. For this purpose, we developed a software tool with an user graphic interface. Section “Defined features” presents the software application, its features and interfaces; Section “Technical details” specifics the technologies applied to its construction.

Defined features

Our developed software tool enables the user to examine the evolution of the desired learning field by structuring and analyzing scientific

knowledge as proposed by SciKGraph. In this sense, we segment the functionalities into three primary actions in the software.

1. “Create”, in which the user creates the knowledge graph representation of the desired scientific field and identifies its main sub-areas.
2. “Analyze”, devised to enable the user to obtain quantitative metrics of the previously generated structure and extract knowledge from it.
3. “Track Evolution”, in which the user can compare previously generated scientific field representations and track the evolution and similarities between those (the main contribution of this article).

The “Create” interface (cf. Figure 7) assists researchers to represent a scientific field as a knowledge graph. To do so, the user input a collection of textual documents that represent the desired scientific field, the language of the documents, and a Babelfy key - obtained by registering in the Babelfy website*. In the sequence, the software builds and plots the constructed knowledge graph. In order to better understand the topology, the user can pre-process and cluster it, identifying the main sub-areas of the analyzed scientific field. The pre-processing step consists of choosing the most generic vertices of the graph through a threshold or a list. After defining that, the software can cluster the knowledge graph, identifying and plotting the main sub-areas of the scientific field.

The “Analyze” interface (cf. Figure 8) enables the user study the sub-areas previously identified. First, if the researcher assumes that the application identified too many sub-areas, (s)he can choose the number of sub-areas (s)he would consider optimal. Then, our software application uses an agglomerative method (5) to merge the sub-areas until it reaches the number chosen by the user. Moreover, to evaluate how well-segmented are the defined sub-areas, the user can calculate the modularities of the knowledge graph and its sub-areas. Our tool lists the keyphrases of the knowledge graph and the identified sub-areas to assist in the scientific field analysis. At last, the researcher can plot a cluster relation graph to understand how those sub-areas are related among themselves.

The “Track Evolution” interface (cf. Figure 9) enables researchers to analyze how sub-areas of a scientific field have modified over time by

*<http://babelfy.org/login>

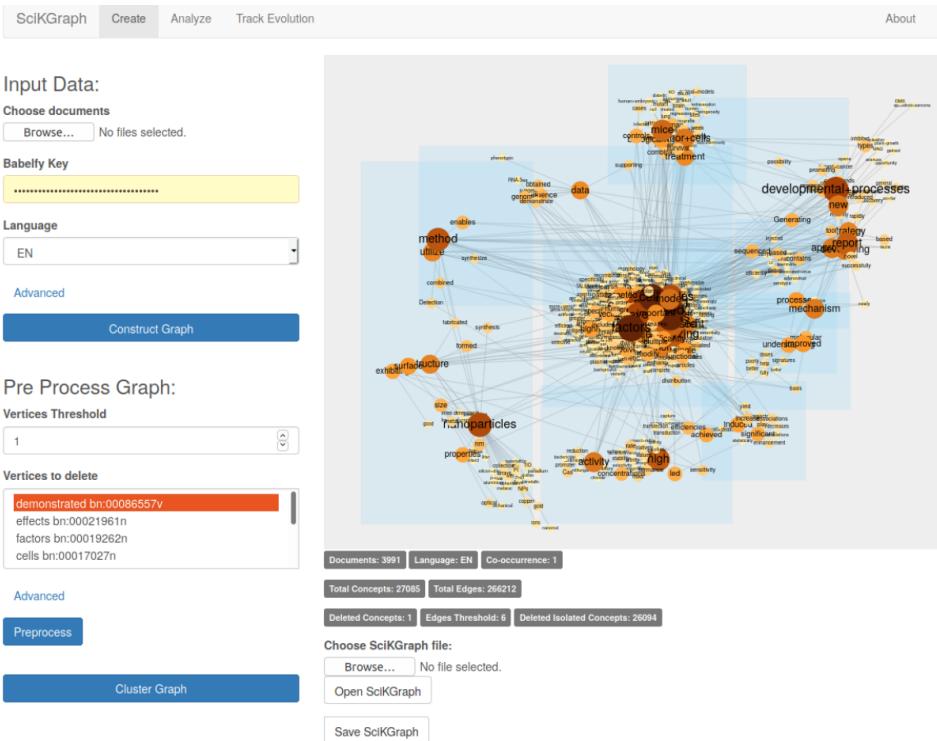


Figura 7. “Create” interface used to represent a scientific field as a knowledge graph. This allows to cluster the graph by extracting its main sub-areas.

identifying concepts added and/or excluded from them. The user needs to input two knowledge graphs representing the scientific field time-periods for comparison. In the tool, those knowledge graphs are generated by the “Create” feature by inputting only academic articles inside the time-period to be represented. The tool supports the importing of already generated knowledge graphs. As an example, to track the evolution of a Biotechnology area between 2015 and 2019, the user creates a knowledge graph to represent each of those years. One knowledge graph is created by receiving as input biotechnology articles published in 2015, and the other one biotechnology articles published in 2019.

After the knowledge graphs are available, the researcher can compare the similarity between the whole covers, identifying the amount of concepts added and/or excluded from the scientific field clusters in the time-window between the time-periods analyzed. Our software tool

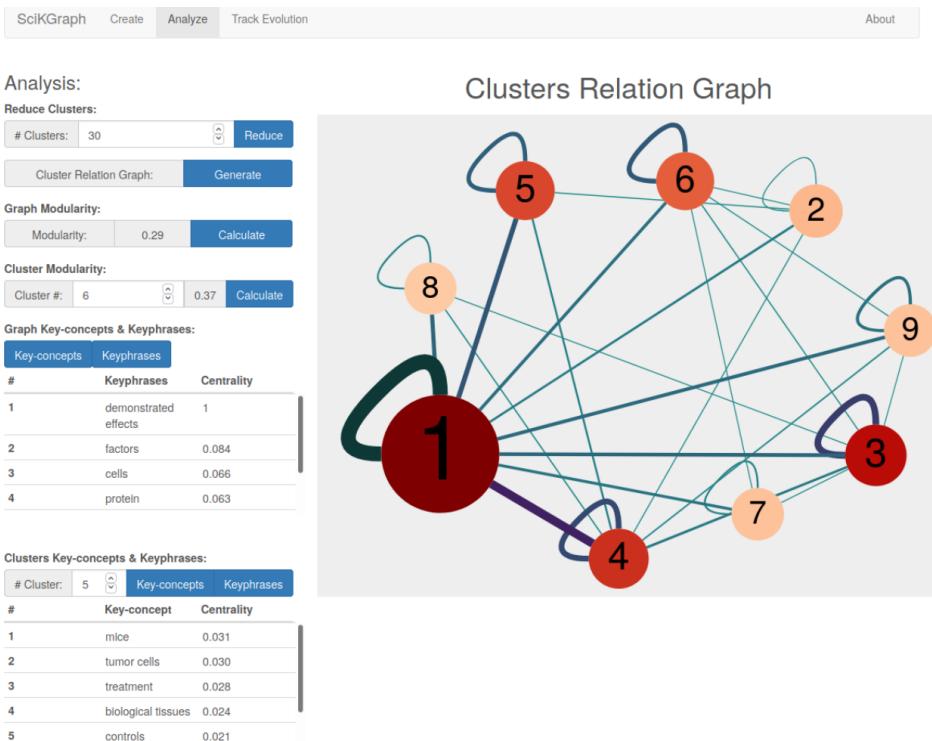


Figura 8. “Analyze” interface in our tool used to extract knowledge and presents quantitative metrics from the scientific field previously structured.

can identify correspondent clusters from distinct covers, based on the “Similarity Threshold” configured by the user. Those correspondent clusters are essential because they represent the same sub-areas in different time-periods. Consequently, comparing those, the application can identify how the sub-areas evolved, presenting to the user concepts that were removed, added, or maintained in a specific sub-area. The application provides a graph visualization of this evolution, as presented in the screenshot of the “Track Evolution” functionality (cf. Figure 9).

Technical details

The software application is available online[†]. It is a web application with its back-end developed in Python 3.7, web server interface in flask

[†]<https://github.com/maurodlt/SciKGraph>

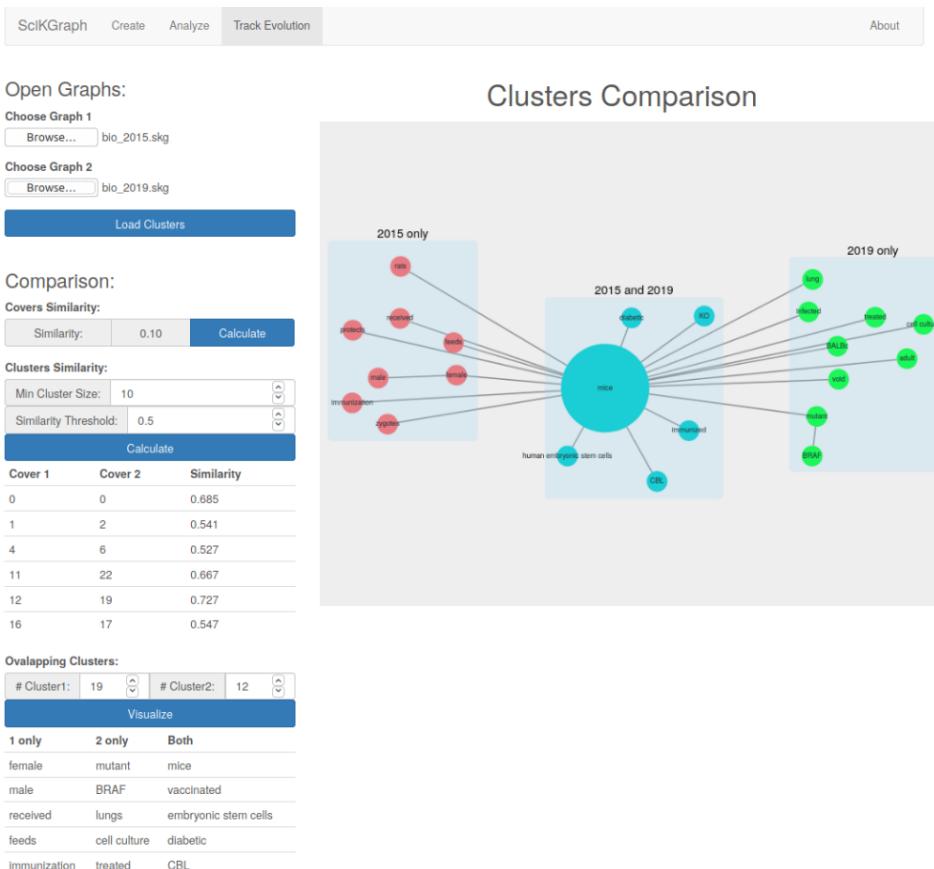


Figura 9. “Evolve” interface in our tool used to track the evolution of a scientific field and its sub-areas.

1.1.1, and front-end in HTML 5, CSS 3, Bootstrap 3.3.7, and javascript 6. Furthermore, to plot the constructed knowledge graphs, we used Cytoscape (20), which is an open-source software to visualize complex networks. It has the py2cytoscape library (21) that allows us to link it with our back-end solution. In addition, it has the cytoscape.js library (21) that uses javascript to link py2cytoscape and our web pages. For the construction of the SciKGraph knowledge graphs, we use the pybabelfy

library [‡], which utilizes the Babelfy HTTP API [§] to link the corresponding concepts between the input documents and BabelNet.

Our application software uses cytoscape.js to plot the knowledge graph. In the current version, the application software will not perform well when rendering tens of thousands of vertices simultaneously. Nevertheless, this should not be an issue because scientific fields have their own lexicon, which consists of only a fraction of the words of a language. In both of our study cases, the knowledge graphs, with around 40,000 concepts each, after the pre-processing step were reduced to less than 3,000 concepts. These concepts can be plotted in a couple of minutes. In addition, we highlight that the higher the number of articles represented by the knowledge graph, the lower is the probability of new concepts to be added when inserting articles from the same scientific field. This occurs because the number of vertices tends to stabilize as most of the concepts of the scientific field have been previously inserted in the knowledge graph. Therefore, the number of articles inserted in the application should not impair its visualization.

Regarding the used algorithms, we adopted the SciKGraph framework (5) to represent a scientific field as a knowledge graph. Therefore, if the user requires to reduce the number of clusters to n , we use the agglomerative technique recommended in SciKGraph, which selects the n bigger clusters and simulates to merge them with the smallest cluster of the cover. Then, the algorithm permanently merges the smallest cluster with the one that achieved the higher modularity in the simulation. This algorithm repeats this process until the number of clusters is equal to n . The modularity metric used in this algorithm, and over other analyses in our work, must consider the overlapping cluster. To this end, as in SciKGraph, we adopted the Q_{ov}^L metric, suggested by Chen and Szymanski (22). Our software tool application uses the same algorithms as SciKGraph to cluster the knowledge graph and to extract its keyphrases, which are the OClustR (17) and C-Rank (18) algorithms, respectively. At last, to identify the similarity of distinct covers, we used the NMI metric proposed by McDaid, Greene, and Hurley (9). Our tool implements a procedure considering the Equation 4 to identify the similarity of sub-areas from distinct covers.

[‡]<https://github.com/aghie/pybabelfy>

[§]<http://babelfy.org/guide>

Other than the developed graphical user interface (GUI) software application, our investigation contributed in providing a python library for those who wants to automatize its usage (available online[¶]).

Experimental Results

We present the application of our solution in different scenarios. Subsection “Datasets” introduces two datasets used to represent the Artificial Intelligence (AI) and the Biotechnology (BIO) scientific fields. Then, we show how the proposed application structures and illustrates them. Finally, we demonstrate how our method tracks a scientific field evolution and how a researcher can analyze those results, extracting knowledge from them.

Datasets

This investigation explores the AI and the BIO datasets to, respectively, represent the Artificial Intelligence and the Biotechnology scientific fields. We chose those datasets because they represent distinct branches of science. Our goal is to understand to which extent our devised methods and implemented tool can correctly track the evolution of scientific fields independently of their areas.

Tosi and dos Reis (5) constructed the Artificial Intelligence dataset (AI) to experimentally evaluate the SciKGraph. It contains 1,002 academic documents, and their publishing dates, from the Artificial Intelligence area. Those documents were crawled from the IEEE Xplorer website^{||} based on a search using “Artificial Intelligence” as a keyphrase. The results were sorted based on their number of citations by other papers, and the most cited documents were downloaded. At last, our solution parsed the downloaded documents from PDF to text, generating the final dataset with full academic documents, in text format, representing the AI scientific field. To our evolution analyses, we organized AI dataset into two parts: the first one containing 481 documents published before 2006; and the second one containing 521 documents published from 2006 (2006 included).

[¶]<https://github.com/maurodlt/SciKGraph>

^{||}<https://ieeexplore.ieee.org/Xplore/home.jsp>

We constructed the Biotechnology dataset (BIO) to understand how our method would perform tracking evolution in a different science branch. We developed a crawler to download the content of the dataset. Instead of downloading full documents, we downloaded only their abstract and publishing date. In this procedure, we aim to compare the obtained results taking as input full academic articles, from AI, and only abstracts, from the BIO dataset.

Our crawler automatically searched for articles in the nature website** selecting only “research” or “reviews” documents having the “biotechnology” subject, sorted by relevance. In order to analyze the evolution of the Biotechnology scientific area in a specific period, we downloaded 3,991 articles abstracts published between 2013 and 2015 and 4,973 articles abstracts published between 2018 and 2020. This allows use to analyze the evolution that occurred between 2015 and 2018 in the Biotechnology scientific field.

Case study results

We illustrate the generation and evolution tracking analysis with the use of our software application based on the two cases studies by representing the Artificial Intelligence and the Biotechnology scientific fields. Both case studies are available online^{††}

Artificial Intelligence: Using all 1,002 AI textual documents, our software application can structure and cluster the Artificial Intelligence scientific field (cf. Figure 10). Figure 10 shows the constructed knowledge graph, which researchers can use to understand how concepts within the same or from sub-areas (clusters) are related. Figure 10 highlights a portion of the knowledge graph to facilitate its visualization, which can be accomplished by using the zooming mechanism in the software.

In Figure 10, the knowledge graph originally contained 40,343 concepts that the pre-processing step reduced to 2,899 before the clusterization, which identified 18 distinct sub-areas containing at least 10 concepts each. We observed that most concepts are in the same central region of the figure. This occurs with the most generic ones. On the other

**<https://www.nature.com/search/advanced>

††<https://github.com/maurodlr/SciKGraph>

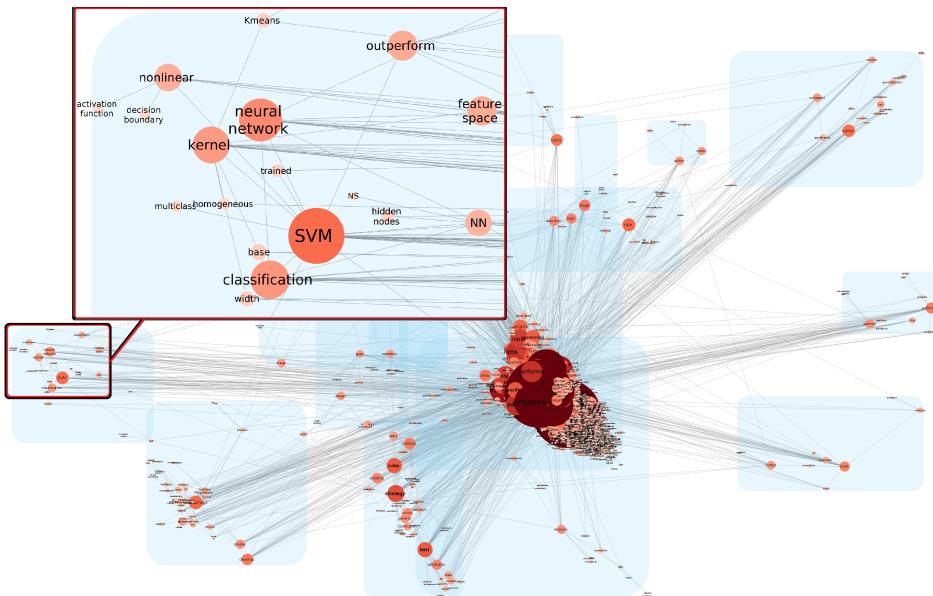


Figura 10. Knowledge graph illustrating the Artificial Intelligence scientific field. It highlights a region of peripheral vertices representing specific concepts related to machine learning and classification.

hand, more specific concepts, the ones that differentiate distinct sub-areas, are positioned at the edge of the knowledge graph. Therefore, this type of visualization automatically segments more specific concepts; this occurs with the highlighted sub-area in Figure 10. It shows specific concepts related to classification and machine learning techniques, which are central to the AI scientific field. For example, if a researcher wants to understand more about “neural networks”, using the visualization proposed, (s)he can observe that it is linked to “classification”, “NN” (neural networks), and “trained” concepts. Moreover, it belongs to the same sub-area as “kernel”, “hidden nodes”, and “feature space”, indicating that those concepts are related among themselves and used in similar problems.

By applying our tool, we exemplify how to track the evolution in Artificial Intelligence. We analyzed how its sub-areas change comparing two time-fixed covers of the AI field, one constructed based on articles published before 2006 and the other based on articles published from 2006. Our analysis consists first in identifying the similarities between

the clusters from different covers (cf. Table 1). To exhibit only the most similar clusters, we set a similarity threshold and display only those clusters with more than 40% of similarity between themselves. Also, clusters with less than 10 concepts are not displayed because those contain insufficient information and would disturb the user analysis. Table 1 shows 13 clusters from different covers that have similarities above 40%, indicating that they can represent the same sub-areas in distinct time-periods.

Tabela 1. Similarities between Artificial Intelligence clusters from two distinct periods.

Artificial Intelligence		
# Cluster (Before 2006)	# Cluster (From 2006)	Similarity
0	0	0.586
4	1	0.458
10	10	0.414
11	3	0.415
11	6	0.557
11	8	0.401
12	4	0.666
14	9	0.416
24	32	0.815
26	33	0.450
28	16	0.427
28	37	0.493
33	38	0.514

Figure 11 represents clusters 4 and 1, from covers 1 and 2, respectively. It shows that the sub-area the clusters represent is centered around the “image” concept. We assume here that it represents the Image Analysis sub-area. All the concepts that both clusters have in common are related to metrics, characteristics, and domains in which Image Analysis is used, as “grayscale image”, “low-resolution”, and “iris”. Moreover, a researcher analyzing this representation can observe not only concepts of a sub-area, but concepts removed or added to it. For example, before 2006, concepts as “training set”, “labeled”, and “test image” were part of the analyzed sub-area; and from 2006, concepts as “videos” and “facial” were added to it. By interpreting this information, one may understand that supervised

learning - which depends on training sets, labeled information, and test sets - has been less used recently in this context. On the other hand, new domains of problems acquired more attention in the most recent years, video processing, and facial-related analysis, for example.

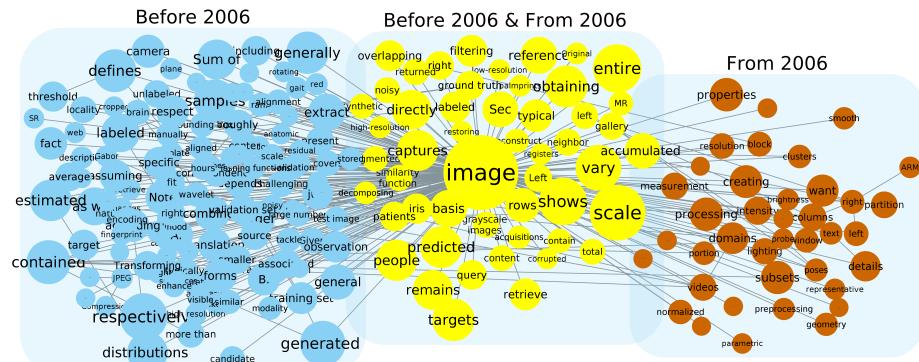


Figura 11. Evolution of the “Image Analysis” sub-area comparing its representation by cluster 4 from cover 1 (before 2006) and by cluster 1 from cover 2 (from 2006).

Other than that, it is also possible to analyze how two sub-areas evolved together to form a third one. Figure 12 illustrates this, in which the clusters 14 and 19 from cover 1 evolved to form cluster 5 from cover 2. We note that in this case, the similarity between the cluster from cover 1 and the cluster from cover 2 is smaller than the 40% we determined before. This occurs because the third cluster is formed not only by concepts from the other two clusters, but by newer concepts added to it. In this sense, cluster 5 from cover 2 has 35% of similarity with cluster 14 and 33% of similarity with cluster 19, both from cover 1. Therefore, Figure 12 illustrates three groups, clusters 14 and 19 from cover 1, and cluster 5 from cover 2. Those we represent by orange, blue, and yellow concepts, and we assume that they represent “Neural Networks”, “Neural Networks Structures”, and “Convolutional Neural Networks” sub-areas, respectively. We assumed this to facilitate the overall comprehension of the example. This assumption relies on the concepts that belong to each of those clusters.

Figure 12 representation allows researchers to understand how two clusters evolved into a single one. It shows all the concepts from the three

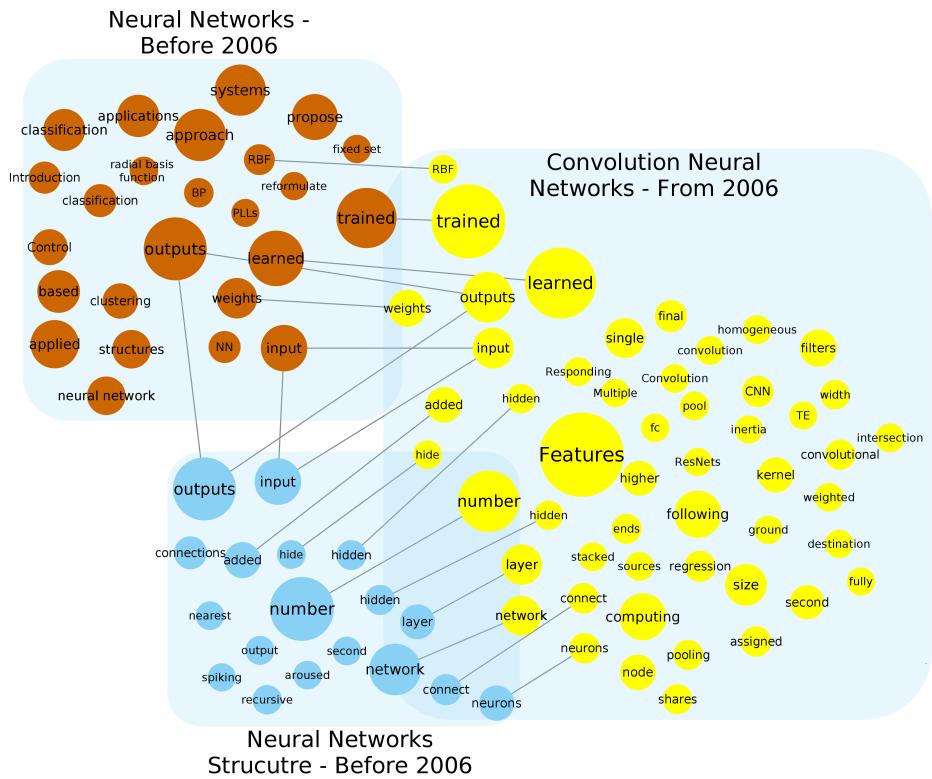


Figura 12. Evolution of the “Convolution Neural Networks” sub-area comparing its representation by cluster 14 and 19 from cover 1 (before 2006) and cluster 5 from cover 2 (from 2006).

clusters and links the corresponding ones. A researcher can observe which concepts the clusters share in common. For example, we mentioned that the similarity between the orange and the yellow clusters is higher than the blue and the yellow one. Though, this visualization allows us to observe that the last pair of clusters have more concepts in common between themselves. In addition, it is interesting to note how the “Convolutional Neural Networks” sub-area was created. Most of its concepts are not in the two clusters from cover 1, not because they were miss classified, but because they were not relevant enough to be part of the scientific field

representation. From 2006, concepts related to neural networks and its structure were used several times together with concepts as “convolution” and “CNN”, showing their relevance for the Artificial Intelligence field. Therefore, both sub-areas from cover 1, before 2006, were merged with other newer concepts, creating the “Convolutional Neural Networks” sub-area.

Biotechnology: We used our developed software tool to analyze, structure and cluster the Biotechnology field based on the 8,964 abstracts from the BIO dataset (cf. Figure 13). It shows the most relevant concepts of the biotechnology field and how they are correlated. We zoomed in a region of Figure 13 to highlight how correlated concepts are plotted close to each other in this visualization.

Similar to the structure of the Artificial Intelligence knowledge graph representation, Figure 13 exhibits a central group of vertices containing generic concepts and peripheral vertices representing more specific ones. We observe this pattern in the zoomed part of Figure 13, in which biotechnology specific concepts related to microbiology, as “Bacterial”, “strains”, “genome”, and “virus” are close to each other in the knowledge graph. Therefore, if a biologist would like to understand more about bacterial related researches, for example, (s)he could study other concepts connected to the “bacterial” and those that are in its surroundings.

In order to study the evolution of the Biotechnology field, similar to what we performed to the AI knowledge graph, we analyzed the evolution of its sub-areas comparing their representations in two distinct time-periods. In this case, our first representation contains 3,991 abstracts published between 2013 and 2015; and the second containing 4,973 abstracts published between 2018 and 2020. We are tracking the evolution of the Biotechnology field that occurred between these two time-windows. Accordingly, we represent the two time-periods as knowledge graphs, cluster both of them, and identify the similarities among their clusters. Afterwards, we set a threshold similarity to display only the most similar clusters, which was defined as 50%, a higher value compared to the one used in the Artificial Intelligence analysis, increased to reduce the number of sub-areas displayed to the user (cf. Section “Tracking the evolution of a scientific field”). Table 2 presents the clusters that have more than 50% of similarity with each other, and therefore, we assume that they represent the same sub-area of the analyzed scientific field.

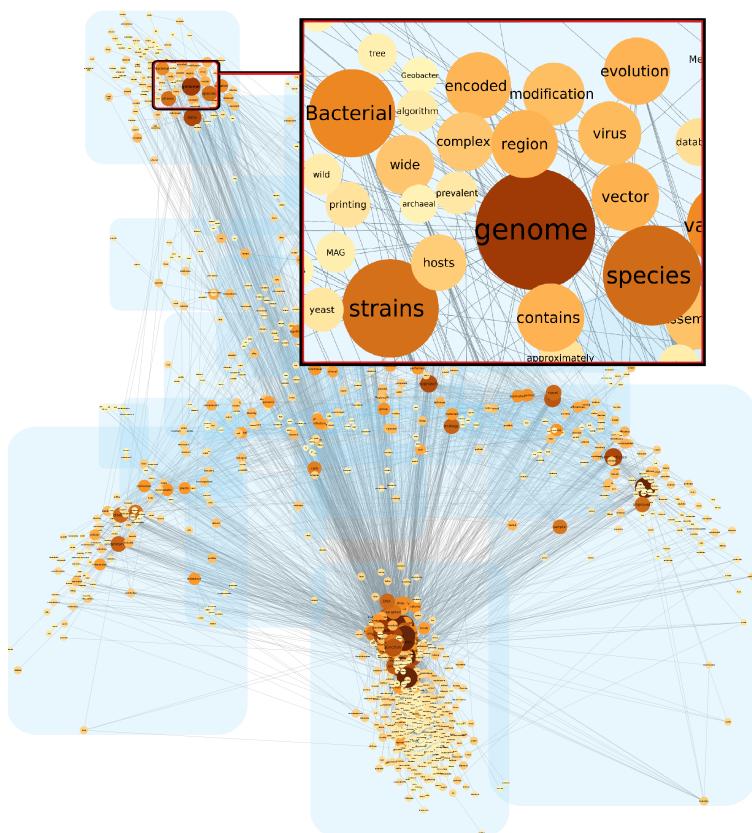


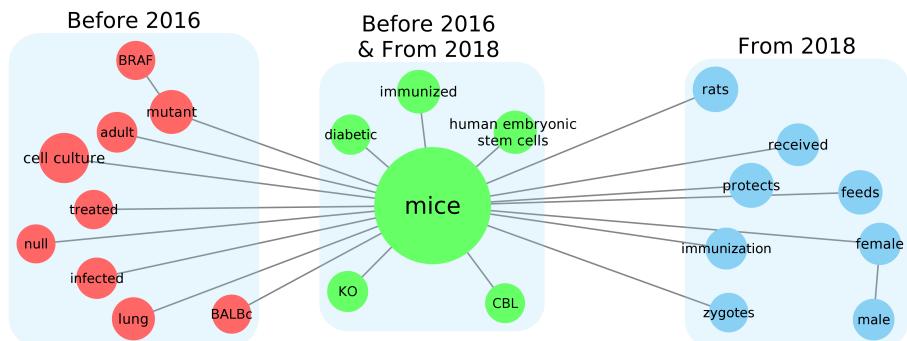
Figura 13. Knowledge graph illustrating the Biotechnology scientific field. It highlights a region of peripheral vertices representing specific concepts related to the microbiology sub-area.

We further analyzed a sub-area, represented by a pair of clusters listed in Table 2 with the aim of tracking and understanding the evolution of its concepts. Figure 14 illustrates how cluster 12 from cover 1 changed into cluster 19 from cover 2. Figure 14 shows the concepts belonging only to cluster 12 colored in orange; the concepts belonging only to cluster 19 colored in blue; and belonging to both of them colored in yellow. We observed that both of these clusters are centered around the “mice” concept. We assume that they represent a sub-area of “Researches using mice”, in two distinct periods. On this basis, we notice that concepts as “mutant”, “BRAF” (cancer-related gene), and “lung” are present only on

Tabela 2. Similarities between Biotechnology clusters from distinct periods.

Biotechnology		
# Cluster (Before 2016)	# Cluster (From 2018)	Similarity
0	0	0.685
1	1	0.541
4	6	0.527
11	22	0.667
12	19	0.727
16	17	0.547

the representation of the field before 2016, which indicates that researches using mice focused on studying BRAF-mutation lung cancer decreased. Our method to track the changes of a scientific field can be used by researchers to understand better how its sub-areas evolved at a concept level.

**Figura 14.** Evolution of the “Researches using mice” sub-area comparing its representation as cluster 4 in cover 1 (before 2016) and cluster 6 from cover 2 (from 2018).

Discussion

We studied how to track the evolution of a scientific field and its sub-areas. Fulfilling this objective, we can enable researchers understand how an academic field changes over time. To this end, we proposed a method

to track the evolution that occurred in a scientific field between two time-periods.

We represented both time-periods using SciKGraph (5). Our solution compares the obtained representations and identifies what changed between them. Our implemented software tool outputs all concepts and relations that differed from the representations of both time-periods to identify how its sub-areas changed over time and better track the evolution of a scientific field.

Tracking the evolution of sub-areas represented as clusters of distinct covers remains a challenging issue. This complexity happens because before comparing two clusters, it is necessary to identify which clusters, from distinct covers, represent the same sub-area. For this purpose, we employed a similarity metric to determine if two clusters are similar enough to be considered correspondents from distinct covers. We set our similarity metric inspired by the metric proposed by Hopcroft in (4). In our solution, we did not consider that all concepts have the same weight for the similarity value. Instead, we considered the relevance of each concept in the knowledge graph for similarity computation, favoring clusters formed around the same relevant concepts to be considered correspondents. Consequently, the importance of less relevant concepts for the similarity value is low, reducing the noise they generate in the identification of the correspondent clusters.

By using the similarity metric proposed, we successfully identified representations of the same sub-areas in distinct time-periods. Their comparison was effective to support researcher to observe what changed in a particular sub-area, identifying, for example, sub-areas merged, segmented, and concepts added or excluded.

Analyze how a scientific field evolves at a concept level is a novel proposition. Our research contributed to a method and a software tool in this direction. To the best of our knowledge, there is no gold standard available for objective evaluations and we found no similar methods from which we could directly compare our solution. Therefore, we conducted experimental case studies, from which we found that our studied method is valuable to track sub-areas represented in different periods and to compare them, which by our analyses, occurred successfully.

In particular, we analyzed the evolution of the Artificial Intelligence and the Biotechnology fields. Their representation and the evolution observed

in both scientific fields were befitting with the studied fields. The analysis of our solution occurred in distinct knowledge areas by assessing the results in a case with full academic articles and other considering only abstracts.

An relevant aspect to further consider in our approach is to help end-users to deal with noise data and with the configuration of the required parameters in the software tool. In addition, we plan to study other representations to express semantics in our approach. For instance, investigate to which extent word embedding (23) and semantic similarity (24) techniques can be combined to advance our proposal.

Conclusion

It is of utmost relevance the proposal of methods and tools to help researchers to better understand how scientific sub-areas evolved at a concept level. Textual documents as research papers provide rich information, but they are not structured. In this article, we proposed an approach to track changes in a scientific field, identifying how it evolves at a concept level. Our approach used knowledge graphs to structure the periods of a scientific field. It considered relations among concepts of a studied field to identify its main clusters, representing their relevant sub-areas to track their evolution. We used a similarity metric, based on the relevance of the concepts within the clusters, to determine if two clusters from distinct periods represent the same sub-area. Obtained results attested that our approach is domain-independent and, despite being further effective based on full articles, can structure a scientific field using only academic abstracts. In future work, we plan to study further interactive mechanisms to navigate over the graph can help users to explore the concepts in the detected clusters.

Declaration of Conflicting Interests

The Author(s) declare(s) that there is no conflict of interest.

Funding

This work was financially supported by the São Paulo Research Foundation (FAPESP) (grants #2017/02325-5 and #2013/08293-7)^{†‡} and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Referências

- [1] Bornmann L and Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 2015; 66(11): 2215–2222.
- [2] Xia F, Liu H, Lee I et al. Scientific article recommendation: Exploiting common author relations and historical preferences. *IEEE Transactions on Big Data* 2016; 2(2): 101–112.
- [3] Silva FN, Amancio DR, Bardosova M et al. Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics* 2016; 10(2): 487–502.
- [4] Hopcroft J, Khan O, Kulis B et al. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences* 2004; 101(suppl 1): 5249–5253.
- [5] Tosi MDL. *Constructing knowledge graphs from textual documents for scientific literature analysis: Construindo grafos de conhecimento utilizando documentos textuais para análise de literatura científica*. Master's Thesis, University of Campinas, Institute of Computing, Campinas - SP, 2020.
- [6] Rous B. Major update to acm's computing classification system. *Communications of the ACM* 2012; 55(11): 12–12.
- [7] Kowsari K, Jafari Meimandi K, Heidarysafa M et al. Text classification algorithms: A survey. *Information* 2019; 10(4): 150.
- [8] Jung S and Segev A. Analyzing future communities in growing citation networks. *Knowledge-Based Systems* 2014; 69: 34–44.
- [9] McDaid AF, Greene D and Hurley N. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:11102515* 2011; .
- [10] Chakraborty T and Chakraborty A. Overcite: Finding overlapping communities in citation network. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, pp. 1124–1131.
- [11] Amancio DR, Nunes MdGV, Oliveira O et al. Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics* 2012; 91(3): 827–842.
- [12] Amancio DR, Oliveira Jr ON and da Fontoura Costa L. Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index. *Journal of informetrics* 2012; 6(3): 427–434.

^{†‡}The opinions expressed in here are not necessarily shared by the financial support agency.

-
- [13] Vahdati S, Palma G, Nath RJ et al. Unveiling scholarly communities over knowledge graphs. In Méndez E, Crestani F, Ribeiro C et al. (eds.) *Digital Libraries for Open Knowledge*. Cham: Springer International Publishing. ISBN 978-3-030-00066-0, pp. 103–115.
 - [14] Moro A, Cecconi F and Navigli R. Multilingual word sense disambiguation and entity linking for everybody. In *International Semantic Web Conference (Posters & Demos)*. pp. 25–28.
 - [15] Navigli R. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)* 2009; 41(2): 1–69.
 - [16] Navigli R and Ponzetto SP. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pp. 216–225.
 - [17] Pérez-Suárez A, Martínez-Trinidad JF, Carrasco-Ochoa JA et al. Oclustr: A new graph-based algorithm for overlapping clustering. *Neurocomputing* 2013; 121: 234–247.
 - [18] Tosi MDL and dos Reis JC. C-rank: A concept linking approach to unsupervised keyphrase extraction. In *Research Conference on Metadata and Semantics Research*. Springer, pp. 236–247.
 - [19] Kluyver T, Ragan-Kelley B, Pérez F et al. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pp. 87–90.
 - [20] Shannon P, Markiel A, Ozier O et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 2003; 13(11): 2498–2504.
 - [21] Otasek D, Morris JH, Bouças J et al. Cytoscape automation: empowering workflow-based network analysis. *Genome biology* 2019; 20(1): 1–15.
 - [22] Chen M and Szymanski BK. Fuzzy overlapping community quality metrics. *Social Network Analysis and Mining* 2015; 5(1): 40.
 - [23] Naili M, Chaibi AH and Ghezala HHB. Comparative study of word embedding methods in topic segmentation. *Procedia computer science* 2017; 112: 340–349.
 - [24] Cilibrasi RL and Vitanyi PM. The google similarity distance. *IEEE Transactions on knowledge and data engineering* 2007; 19(3): 370–383.