

Representing Scientific Literature Evolution via Temporal Knowledge Graphs

Anderson Rossanez¹, Julio Cesar dos Reis¹, and Ricardo da Silva Torres²

¹ Institute of Computing, University of Campinas, Campinas - SP, Brazil
{anderson.rossanez, jreis}@ic.unicamp.br

² Department of ICT and Natural Sciences, Norwegian University of Science and Technology, Ålesund, Norway
ricardo.torres@ntnu.no

Abstract. Scientific publications register the current knowledge in a specific domain. As new researches are conducted, knowledge evolves, getting documented in dissertations, theses and articles. In this article, we introduce new methods that exploit Temporal Knowledge Graphs (TKGs) to model temporal knowledge evolution in corpora of unstructured texts. In our approach, complex network measurements are applied over TKGs to determine the relevance of concepts dealt with in the corpora under analysis. We demonstrate the effectiveness of our method by conducting experimental analyses on TKGs constructed from a corpus of scientific papers extracted from different editions of the International Semantic Web Conference (ISWC). The results demonstrate the effectiveness of the method in representing and tracking the knowledge evolution over time.

Keywords: Temporal Knowledge Graphs · Information Retrieval · Knowledge Evolution · Complex Network Measurements

1 Introduction

Scientific publications describe studies that contribute to advancements in the state of the art of a given research domain. Such publications document new methodologies and findings, as well as new data, from which new knowledge is produced. As researches are continuously conducted, an overwhelming amount of scientific studies become available on a timely basis, documenting the evolution of knowledge in a field of study.

To further illustrate it, one may refer to scientific articles published from time to time in journals, focusing on studies from a specific field, or in the proceedings of important conferences released every year. For example, the proceedings of the International Semantic Web Conference (ISWC)³ is a relevant corpora source representing the evolution of the state of the art in the Semantic Web domain.

The amount of papers submitted and published in journals and conferences has increased drastically over the years. Proper reading and understanding, as well as tracking new knowledge from such amount of publications is now a very hard task to be accomplished without proper help. Our claim is that computational methods are a relevant

³ <https://link.springer.com/conference/semweb> (As of Aug. 2020).

solution to help scientists understand the trends and needs of their research domain. This requires investigations to adequately extract information and represent knowledge conveyed in the scientific literature, as well as tracking the evolution of knowledge in a temporal manner from texts.

Building knowledge representations requires the determination of facts conveyed in a portion of a scientific text. Facts are described by relations between entities in the text. For instance, let us consider a text in an initial time frame, containing the sentence *The Scholarly Ontology documents research processes*. The entities *Scholarly Ontology* and *research processes* relate to each other through the verb *documents*. The knowledge representation for the sentence, therefore, relies on those entities and relation. Furthermore, one may notice that, *Scholarly Ontology* is a sub-type of *Ontology*, and *research processes* is a sub-type of *processes*. If we consider another text, in a subsequent time frame, containing another sentence, e.g., *The Gene Ontology describes the function of genes*. There is a new knowledge representation, that relies on the relation (*describes*), between two entities (*Gene Ontology*, and *function of genes*). Similarly, *Gene Ontology* is a sub-type of *Ontology*, and *function of genes* is a *function*, pertaining to *genes*. Considering both time frames, one may observe that there are two sub-types of *Ontology*: *Scholarly Ontology*, and *Gene Ontology*, both having their specific relations with other entities, which are, in turn, sub-types of other entities. In this sense, knowledge regarding the *ontology* concept evolves across the considered time frames. In the first time frame, we knew the existence of *Scholarly Ontology* and the relations in which such concepts take part. In the second time frame, we had the *Gene Ontology* concept and its relations aggregated to the overall knowledge. Such concepts and relations were not known in the initial time frame. This illustrates the relevance of modeling temporal aspects of knowledge.

In this research work, we address the following research questions: (1) How to represent knowledge conveyed in temporal corpora of natural language texts? (2) How to characterize the knowledge evolution in such temporal corpora of natural language texts?

Knowledge Graphs (KGs) [9] are computational tools that model knowledge by means of the interrelations of real-world entities in facts, through a graph format. Linking graph entities and properties from a KG with computational ontologies, which describe concepts from a specific domain, helps in the comparison of different KGs that are mapped to the same, or mapped, ontologies. We address the concept of Temporal Knowledge Graphs (TKGs) as a way to benefit from the graph format to represent temporal aspects from texts in KGs.

In this article, we investigate the use of TKGs to represent a corpus of unstructured scientific texts, and how they evolve over time. The generation of KGs from scientific literature is still an open research problem. In our previous work [11], we investigated how to generate KGs from unstructured texts via the KGen tool. However, how to analyze and compare several generated KGs in a temporal perspective requires further investigation. The way of representing and characterizing the evolution of knowledge over time via KGs is the main novelty aspect in this work.

In summary, our objectives are: (1) enhance the KGen method and tool introduced in our previous work [11], originally conceived considering biomedical texts, to semi-

automatically generate ontology-linked TKGs for corpora of unstructured texts in the Computer Science domain; and (2) propose a method to analyze and compare the generated TKGs using centrality metrics obtained from complex networks to characterize the evolution of knowledge over time. In our evaluation, we use a set of temporal texts obtained from articles in the proceedings of previous editions of the ISWC conference. The Computer Science Ontology [12] was used to determine and link the entities from the TKGs.

The remaining of this paper is organized as follows: Section 2 discusses related work; Section 3 presents our method; Section 4 shows the evaluation with TKGs generated from temporal sets of scientific papers; Section 5 discusses our findings. Finally, Section 6 closes the paper presenting conclusions and future work.

2 Background and Related Work

Studies that deal with KGs consider a different concept of TKGs from the one we consider in this work. Han *et al.* [3] considered TKGs as a graph that captures the moment in which a represented fact happened. Such facts are represented by quadruples (subject, predicate, object, and time), rather than common triples. Their work proposed a neural network approach to predict future events that should be added to the graph. Similarly, a survey on KGs [4] considered TKGs as graphs that incorporate the time aspect through quadruple structures.

Trivedi *et al.* [14] proposed a deep learning method to infer the moment in which some events in the graph occur, based on events in which the time information is explicit. All these studies considered a single, but rather large, graph to perform the proposed inferences. Liu *et al.* [6] proposed a framework to derive a new TKG that predicts future events, considering the weights of graph vertices. Their work employed a quadruple representation to model the temporal facts.

As for graph metrics, Park *et al.* [8] used centrality measurements, combined with a neural model, to determine the most important nodes of a KG. Shi *et al.* [13], in turn, presented a method to extract key-phrases from texts by means of strategies that are based on degree, closeness, betweenness, and eigenvector centrality measures taken for the nodes of a KG that represent the text.

Gengchen *et al.* [7] described an approach to infer the weights of KG nodes by means of centrality measures, such as degree, eigenvector, and betweenness centralities. Such measures are fed into a neural model, and used to determine the centrality of nodes from secondary graphs.

Existing literature dealing with TKGs often considered a quadruple representation that embeds temporal information, rather than the traditional representation in form of triples. The quadruple representation allows for encoding distinct temporal information in the same graph. In our work, triples that constitute each graph generated for the same temporal set of texts, represent facts from the same temporal time frame, *e.g.*, a KG generated for the proceedings from a given year has facts that hold true for that year. For this reason, we consider the traditional triple format in our TKG approach. Merging our TKGs into a single graph, considering the quadruple format, would result in a TKG as defined in literature. To the best of our knowledge, there is no study in

literature considering centrality measurements in TKGs to characterize the evolution of knowledge in a temporal manner.

3 Construction and Analysis of TKGs

This section describes our method to represent time-evolving scientific texts by means of ontology-linked TKGs. In our approach, we explore analyses based on complex networks' centrality measurements in KGs to determine how the represented knowledge evolves over time. Figure 1 presents an overview of our method.

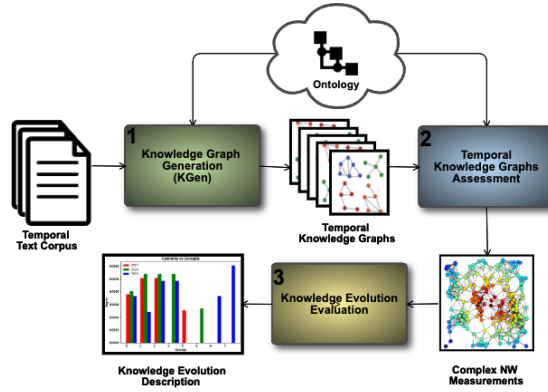


Fig. 1: **Method overview.** KGen [11] is depicted in step 1. Steps 2 and 3 are the contributions of this work.

A Knowledge Graph \mathcal{KG} is formally described as a set of vertices \mathcal{V} and edges \mathcal{E} , *i.e.*, $\mathcal{KG} = (\mathcal{V}, \mathcal{E})$, where the vertices represent entities, and the edges represent the relations among such entities. A KG can also be defined as a set of RDF triples, composed of subject, predicate, and object defined as $t = (s, p, o)$. The subjects and objects are vertices, and the predicates are the edges in a KG, resulting that a $\mathcal{KG} = \{t_0, t_1, \dots, t_n\}$, where $t_0 = (s_0, p_0, o_0), t_1 = (s_1, p_1, o_1), \dots, t_n = (s_n, p_n, o_n)$. An ontology \mathcal{O} describes a real-world domain in terms of concepts, represented by classes $\mathcal{C}_{\mathcal{O}}$ interrelated by directed relations \mathcal{R} , and a set of attributes $\mathcal{A}_{\mathcal{O}}$, *i.e.*, $\mathcal{O} = (\mathcal{C}_{\mathcal{O}}, \mathcal{R}_{\mathcal{O}}, \mathcal{A}_{\mathcal{O}})$. Subjects and objects from triples in a KG may be linked to an ontology, denoting that they are instances of those particular ontology's entities.

In this work, we consider a Temporal Knowledge Graph as a \mathcal{KG} in which the triples represent facts that are available in a determined time frame i , *i.e.*, $\mathcal{KG}^i = \{t_0^i, t_1^i, \dots, t_n^i\}$. In this sense, if we consider KGs in two different time frames (i and $i + 1$), *i.e.*, $\mathcal{KG}^i = \{t_0^i, t_1^i, \dots, t_n^i\}$, and $\mathcal{KG}^{i+1} = \{t_0^{i+1}, t_1^{i+1}, \dots, t_n^{i+1}\}$, we may have triples like $t_a \in \mathcal{KG}^i$ and $t_a \in \mathcal{KG}^{i+1}$, meaning that the fact described by t_a is available in both time frames. We may also have triples like $t_b \in \mathcal{KG}^i$ and $t_b \notin \mathcal{KG}^{i+1}$, *i.e.*, the fact described by the t_b is available only in the time frame i , and not in $i + 1$. The triples' constituents in the TKG may also be linked to ontologies. Linking nodes of

the TKGs to ontologies, are a key step to enable the comparison in between TKGs, as nodes linked to the same ontology’s classes, attributes, or properties, denote an instance of the same concept, although in different time frames.

To generate TKGs, we evolved the KGen tool [11] to consider determining named entities from other domains in the text for triples extraction. The extracted triples’ constituents are matched and linked to an ontology. KGen was enhanced to allow batch processing for generating TKGs taking multiple unstructured texts as input. A set of texts from a determined time frame produce a single KG as output, *i.e.*, the TKG for that specific time frame. A set of TKGs are the output of multiple sets of unstructured texts from different time frames.

Once the set of TKGs is computed, $\mathcal{KG}^0, \mathcal{KG}^1, \dots, \mathcal{KG}^m$, we analyze and determine triples and entities available in different subsets of TKGs. Each generated KG, though, may contain an overwhelmingly large amount of triples, whose constituents may be not linked to the target ontology, making it difficult to determine if nodes from different TKGs are instances of the same concepts. For this reason, we consider the graph nodes that are linked to the target ontology in the analysis.

Due to the possibly huge amount of edges (nodes) and vertices that such KGs may contain, our analyses consider only the most relevant nodes in the KG. For this purpose, we take centrality measurements for complex networks [1] in each node in order to determine its relevance in the graph. In particular, we consider three measurements in our method validation (*cf.* Section 4): degree centrality [5], eigenvector centrality [2], and betweenness centrality [1]. Such measures denote, respectively, the amount of neighbouring nodes linked to a particular node, the number of nodes in the overall graph that have connections to a particular node, and the amount of shortest paths that a particular node is part of. In our solution, ontology-linked nodes with the highest centrality values for all the graphs enables evaluating how knowledge evolves. The analysis of these measurements change in between the temporal KGs, and thus, over time.

Regarding implementation aspects,⁴ our temporal analyses on TKGs were implemented using the Python programming language. The TKGs are handled using RDFLib.⁵ The centrality measurements are computed using APIs from the NetworkX⁶ library.

4 Evaluation

We applied our solution to generate and assess TKGs by using abstracts of the research track papers, obtained from the proceedings of three subsequent editions of the ISWC conference (2017 to 2019). The sets of papers for each year were used as input to our solution generating three ontology-linked TKGs (available online⁷). The Computer Science Ontology (CSO) [12] was used to assist the method in determining named entities from the computer science domain in the text for triples extraction.

⁴ <https://github.com/rossanez/TKGAnalyzer> (As of Aug. 2020).

⁵ <https://github.com/RDFLib> (As of Aug. 2020).

⁶ <https://networkx.github.io/> (As of Aug. 2020).

⁷ <https://github.com/rossanez/KGen> (As of Aug. 2020).

Aiming to identify the most relevant concepts dealt within each edition of the ISWC conference, we conducted three analyses based on distinct centrality measurements, which are widely used in literature, in each KG: degree centrality, eigenvector centrality, and betweenness centrality. Figure 2 summarizes the top values for each measurement.

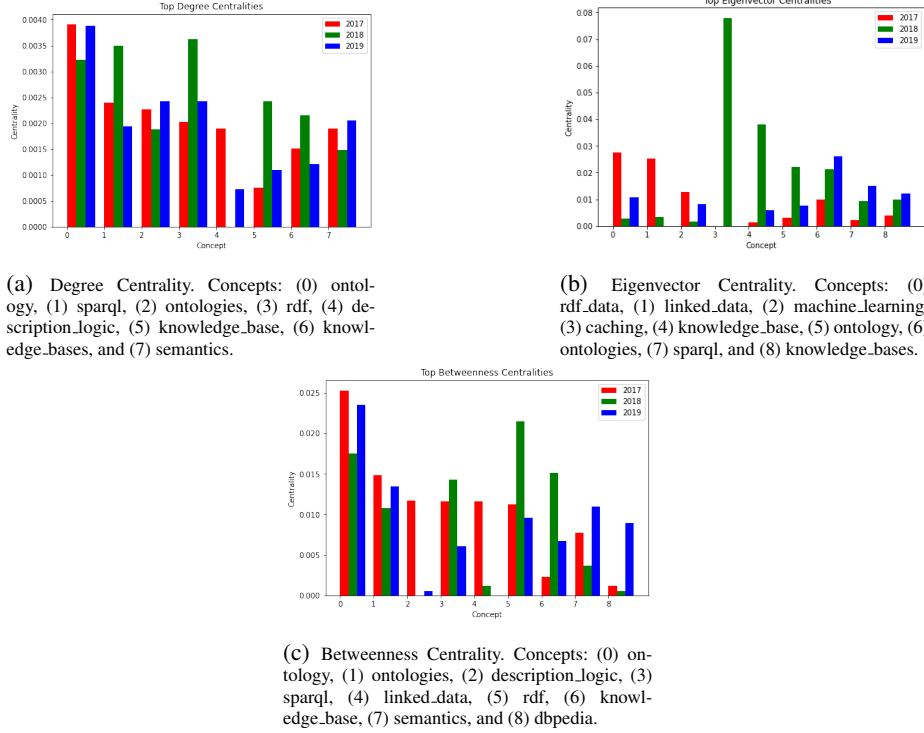


Fig. 2: **Top centrality values and their evolution.** Values are normalized, and concepts are preceded by [https://cso.kmi.open.ac.uk/topics/].

Degree centrality. The degree centrality of a node denotes the amount of nodes that are directly connected to it (*i.e.*, its immediate neighbors). Let $G = (V, E)$ be a graph, represented by a $|V| \times |V|$ adjacency matrix A with elements a_{ij} . The degree centrality is defined as $c_i^D = \sum_{j=1}^N \phi_{ij}$, $\phi_{ij} = 1$, if $a_{ij} > 0$, and 0, otherwise ($i = 1, 2, \dots, N$); where c_i^D is the degree of node i and N is the number of nodes. The normalized degree centrality is computed by $nc_i = \frac{c_i^D}{N-1}$. Figure 2a shows the concepts from the CSO ontology represented by the nodes with the highest degree centrality for each ISWC edition (the complete list, along with sub-graph examples, is available online⁸).

Considering our TKGs, the nodes with the highest degree centrality values represent concepts from the CSO ontology that mostly occur among all abstracts of that particular

⁸ <https://github.com/rossanez/TKGAnalyzer> (As of Aug. 2020).

edition. In 2017, *ontology* has the top value; in 2018, *RDF*; and in 2019, *ontology* has the top value. Figure 3 presents the immediate neighbors for the node representing the *ontology* concept in the TKGs for the ISWC editions considered in the analysis. We observe that the amount of immediate neighbors vary in between the editions: in 2017, it has 31 neighbor nodes; in 2018, 22 neighbor nodes; and in 2019, it has 32 neighbor nodes.

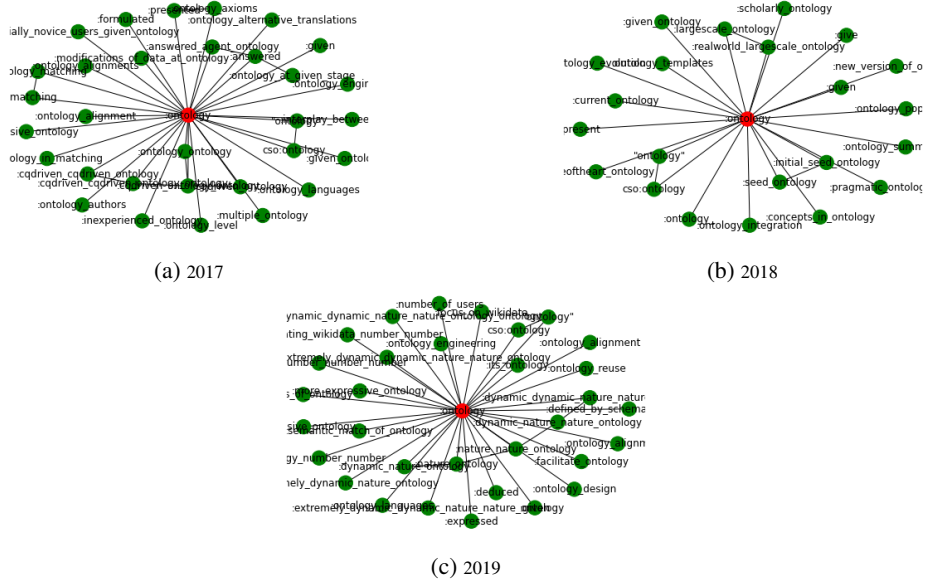


Fig. 3: **Sub-TKGs for each conference edition.** This figure presents a subgraph from our TKGs generated (from the three corpus under study) considering the ontology concept. *Ontology* node is shown in red, and its immediate neighbors are in green.

Figure 2a shows that the degree centrality of concepts vary over time. *SPARQL*, for instance, increases from 2017 to 2018, and then decreases in 2019; *Ontologies* decreases from 2017 to 2018, and increases to its highest value in 2019; *Description logic* is available in 2017, not available in 2018, and reappears with a lower value in 2019.

Eigenvector centrality. The eigenvector centrality of a node denotes the amount of nodes that are directly connected to it, extending to nodes directly connected to those, and so forth throughout the network. It is a measurement of the influence that a node has on the network, relative to the number of connections that this node has to all the other nodes of the entire network. Formally, $\lambda c_i^E = \sum_j a_{ij} c_j^E$ ($i = 1, 2, \dots, N$), $\lambda c = \mathbf{A}c$; where c_i^E is the eigenvector centrality of node i , c is an N -dimensional vector whose entry i represents the centrality score of node i . This formulation leads to the problem of finding the eigenvalues (λ) and the eigenvectors (c) of the adjacency matrix \mathbf{A} . The eigenvector centrality is associated with the dominant eigenvalue found.

Figure 2b presents the concepts from the CSO ontology represented by the nodes with the highest eigenvector centrality values over the considered ISWC editions.

The interpretation of the eigenvector centrality in our TKGs is that nodes with the highest values, connected to most of the other nodes, are highly related to the overall context of the conference at that edition. In 2017, *RDF data* has the highest eigenvector centrality value; in 2018, *caching* has the top value; and in 2019, it is *ontologies*.

The analysis shows that not all nodes with high degree centrality (*cf.* Figure 2a) are the same that have high eigenvector centrality (*cf.* Figure 2b). This particularly occurs to nodes with high degree centrality, whose neighbors have low eigenvector centrality, *i.e.*, the eigenvector centrality of a node highly depends on the eigenvector centrality of its neighbors.

Betweenness centrality. The betweenness centrality denotes how many times a node lies between the shortest path between two different nodes. A node with high betweenness centrality serves as a bridge in many shortest paths between different nodes, being either in a more centered position of the network, or in an important cluster. It can be computed as follows: $c_i^B = \sum_{j=1, j \neq i}^N \sum_{k=1, k \neq i, j}^N \frac{\eta_{jk}(i)}{\eta_{jk}}$ ($i = 1, 2, \dots, N$); where N is the number of nodes, η_{jk} is the number of the shortest paths from node j to node k and $\eta_{jk}(i)$ is the number of the shortest paths from j to k that contain node i . Figure 2c shows the concepts from the CSO ontology represented by the nodes with the highest betweenness centrality values for the considered ISWC editions.

Considering our TKGs, nodes with higher betweenness values are those representing concepts that are related to most of the other concepts represented in the graphs, as they lie between the paths to most of the other concepts' nodes. It is the case of *ontology*, in 2017; *RDF*, in 2018; and *ontology*, in 2019. Considering the previous centrality measurement, a node may have a high betweenness centrality, but a low eigenvector centrality value, if it links nodes that are disconnected from the overall network. For this reason, not all the concepts with the top values from Figure 2c (Betweenness) are the same as Figure 2b (Eigenvector).

5 Discussion

We proposed the generation of ontology-linked TKGs to represent the knowledge conveyed in sets of temporal texts. Our approach explored complex networks' centrality measurements to characterize the evolution of knowledge in between specific time frames. We applied our solution to the Computer science domain by analyzing a set of papers from distinct editions of the ISWC.

We found that our proposal was successful to represent knowledge conveyed in temporal corpora of natural language texts. The use of the distinct complex networks' centrality measurements was suited to help analyzing and characterizing the knowledge evolution in such temporal corpora of natural language texts. We judge our proposal of TKGs appropriate as means to conduct the analyses because KGs are well suited to represent knowledge in terms of facts. In addition, our generated TKGs can be retrieved via SPARQL queries whenever required.

The language employed in scientific papers from the computer science domain poses some extra challenges in the generation of TKGs. One of such aspects is the

use of LaTeX mathematical expressions (even on abstracts), that, when converted to plain text format, generate incorrect information (especially due to characters such as \$, %, {, }, etc.). Despite such challenge, KGen was able to run to completion, but some meaningless triples are generated. For this reason, and taking advantage of the fact that KGen is a semi-automated method, some manual interventions were performed. Therefore, the preprocessing of the text requires further improvements for smoothly batch processing corpora of unstructured texts.

Regarding performance aspects, the generation of KGs using KGen is indeed a time-consuming procedure, even for relatively small texts such as abstracts. This is mainly due to the NLP tools and techniques that are employed, which take a considerable amount of time to process texts. As for the centrality measurements, although reading a KG in turtle format, and converting it to a complex network object does not take a great amount of time, some measurements, especially the betweenness centrality, take a considerable amount of time to complete. Computing the shortest paths for all possible pairs of nodes is indeed a time-consuming procedure, especially true for large KGs.

As for the centrality measurements, the aspect of considering nodes that represent entities that match concepts from the CSO ontology opens opportunities of improvements in our method. Certainly, there will be nodes that have high centrality values, that are left out of the analysis. Possibly, considering multiple ontologies to match and link concepts, and taking advantage of mappings between ontologies could be a solution to tackle this improvement opportunity. Furthermore, extra centrality measurements, besides the three considered in this work, could be incorporated to the analysis, to further enrich it.

We plan to address a better strategy to describe temporal evolution of the concepts aiming to add value in closing the analysis. Bar plots can be confusing to visualize and understand. A possible future work therefore would be to study the feasibility of using a graph-based visual structure to represent such evolution [10]. We also plan to further extend and enhance our results by considering all available proceedings of the ISWC conference. Additional metrics and analyses opportunities shall be considered, and possibly considering predictions of the most important concepts on future editions of the conference, based on the evolution of the previous editions.

Finally, it is worth mentioning that results obtained in the evaluation are entirely based on the abstracts of the papers from the ISWC proceedings. We could achieve different results if we consider the entire text from papers. Furthermore, it is important to enforce that different centrality measurements possibly lead to different most important concepts.

6 Conclusion

The way of representing and analyzing temporal knowledge is valuable to understand the evolution of key domain concepts. This article presented a method to characterize the evolution of knowledge from temporal, unstructured texts in scientific literature, based on analyses conducted via as we named TKGs. We evaluated our proposal by building TKGs from abstracts of three subsequent editions of the ISWC conference.

The conducted complex networks centrality measurements analyses show promising results in determining how the knowledge evolved in between the conference editions. Future work mainly involves the evaluation of analyses using different complex network measurements. Also, we plan to define suitable visual structures to represent temporal knowledge evolution based on network measurements.

Acknowledgment

We thank the São Paulo Research Foundation (FAPESP) (Grant #2017/02325-5).⁹

References

1. Barthelemy, M.: Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter* **38**, 163–168 (2004)
2. Bonacich, P.: Some unique properties of eigenvector centrality. *Social Networks* **29**(4), 555 – 564 (2007)
3. Han, Z., Ma, Y., Wang, Y., Günemann, S., Tresp, V.: Graph hawkes neural network for forecasting on temporal knowledge graphs (2020)
4. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A survey on knowledge graphs: Representation, acquisition and applications (2020)
5. Kapoor, K., Sharma, D., Srivastava, J.: Weighted node degree centrality for hypergraphs. In: 2013 IEEE 2nd Network Science Workshop. pp. 152–155 (2013)
6. Liu, J., Zhang, Q., Fu, L., Wang, X., Lu, S.: Evolving knowledge graphs. In: IEEE Conference on Computer Communications. pp. 2260–2268 (2019)
7. Mai, G., Janowicz, K., Yan, B.: Support and centrality: Learning weights for knowledge graph embedding models. In: Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds.) *Knowledge Engineering and Knowledge Management*. pp. 212–227 (2018)
8. Park, N., Kan, A., Dong, X.L., Zhao, T., Faloutsos, C.: Estimating node importance in knowledge graphs using graph neural networks. In: *Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining*. p. 596–606 (2019)
9. Reinanda, R., Meij, E., de Rijke, M.: Knowledge graphs: An information retrieval perspective. *Foundations and Trends in Information Retrieval* **14**, 1–158 (2020)
10. Rodrigues, D.C.U.M., Moura, F.A., Cunha, S.A., da S. Torres, R.: Graph visual rhythms in temporal network analyses. *Graphical Models* **103**, 101021 (2019)
11. Rossanez, A., Dos Reis, J.C.: Generating knowledge graphs from scientific literature of degenerative diseases. In: *Proceedings of the 4th International Workshop on Semantics-Powered Data Mining and Analytics*. pp. 12–23 (2019)
12. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: A large-scale taxonomy of research areas. In: *ISWC 2018: The Semantic Web*. pp. 187–205 (2018)
13. Shi, W., Zheng, W., Yu, J., Cheng, H., Zou, L.: Keyphrase extraction using knowledge graphs. In: *Data Science and Engineering*. pp. 132–148 (2017)
14. Trivedi, R., Dai, H., Wang, Y., Song, L.: Know-evolve: Deep temporal reasoning for dynamic knowledge graphs (2017)

⁹ The opinions expressed in this work do not necessarily reflect those of the funding agencies.