Tracing the Evolution of Approaches to Semantic Similarity Analysis

Weronika T. Adrian[a], Sebastian Skoczeń[b], Szymon Majkut,
Krzysztof Kluza[c] and Antoni Ligęza[d]
AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland

Keywords: Knowledge Representation, Knowledge Metrics, Semantic Similarity, Knowledge Graphs, Literature Review, Knowledge Engineering, Knowledge Visualization.

Abstract: Capturing the essence of semantic similarity of words or concepts in order to quantify it and measure has been an inspiring challenge for the last decades. From corpus-based statistics to metrics based on structured knowledge bases, a plethora of methods has been proposed in several branches of Artificial Intelligence. Recently, with the advent of knowledge graphs, a renewed interest in similarity metrics can be observed. Choosing appropriate metrics that will work best in a given situation is not a trivial task. To help navigate through the semantic similarity algorithms and understand the characteristics of them, we have analyzed the fundamental proposals in this domain and the evolution of them over the years. In this paper, we present a review of the approaches to measuring semantic similarity of entities in knowledge bases. We organize the findings into a taxonomy and analyze the relations between and within the identified categories. To complement the research with a practical solution, we present a new tool that supports the literature review process with graph-based and temporal visualizations.

1 INTRODUCTION

We live in an information society, where such an abstract concept as knowledge may have bigger value than any physical resource. Not only we became *information-driven* and *information-oriented*, but also a general tendency towards automatization of information processing can be observed. Knowledge bases such as WordNet, Sensus, Gene Ontology or Generalized Upper Model are visible examples of increasing need for a databases that contain, along the information, also its meaning. Ability to process complex data in an intelligent way opens up new possibilities in pattern discovery, recognition and analysis, and therefore leads to take the full advantage of the gargantuan amount of data that is produced every second.

Recently, various knowledge-rich resources gain increasing attention, offering flexibility of the data and knowledge representation and allowing to represent complex relations that better reflect the reality around us. Knowledge graphs, not only encyclopedic-like, such as Wikipedia, DBPedia, Wikidata or BabelNet, but also lexical such as WordNet, or taxonomical such as domain ontologies, are invaluable resources for comprehending the meaning of words and phrases, and also real world objects and categories. Exploiting these knowledge bases lead to results that are not only universal, but also interpretable.

Semantic similarity analysis has been considered for many years, and the graph-based knowledge representation has always played an important role in it. Similarity may be considered at different levels: from word senses, through words, phrases, sentences up to whole documents. For each of these levels, numerous methods have been proposed over the years and still new metrics appear every year. The reason for that is two-fold: on the one hand we have new resources, machine learning methods and application areas that come with new datasets and input formats; on the other hand, the methods are still not satisfactory on more challenging and domain-specific cases. Thus, we state the following research questions:

1. How to measure semantic similarity of entities about which we have some (taxonomical, statistical or graph-based) knowledge?
2. What base methods are there, how have they influenced one another and in which domains have they been used?

[a] https://orcid.org/0000-0002-1860-6989
[b] https://orcid.org/0000-0003-0242-2373
[c] https://orcid.org/0000-0003-1876-9603
[d] https://orcid.org/0000-0002-6573-4246

Adrian, W., Skoczeń, S., Majkut, S., Kluza, K. and Ligęza, A.
Tracing the Evolution of Approaches to Semantic Similarity Analysis.
DOI: 10.5220/0010108401570164
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, pages 157-164
ISBN: 978-989-758-474-9

3. What methods gain attention recently and why?

To address the questions listed above, we have conducted a literature review of the semantic similarity metrics and analyzed their characteristics and inter-dependencies. Then we used methods from the knowledge engineering domain to better internalize and capture the findings. In particular, we have developed an ontology of semantic similarity metrics that organizes them into classes and captures other attributes (such as application domain) and relations among them (such as influence). To support the analysis of our findings, we have developed a simple tool that provides useful visualizations based on graph-based knowledge representation. Thus, the paper contributes in the following ways:

- we provide a review of semantic similarity metrics for concepts, objects and words that use different aspects of knowledge about the entities;
- we present a classification and analysis of the relations between different similarity metrics that can guide those who are starting their journey with the semantic world, and we enhance it with bibliographic analysis of their citations over the years;
- we propose a graph-based tool supporting literature review process and we demonstrate its usage on the case of semantic similarity metrics.

The rest of the paper is organized as follows: We put forward our motivation and give some context in Section 2. Then we present the review of the approaches to semantic similarity illustrating their relations with a simple taxonomy and temporal and functional dependencies in Section 3. In Section 4, we present our tool that proved useful when analyzing the *state-of-the-art*, we explain its design and implementation together with some directions for usage also beyond this work. We conclude our paper in Section 5 outlining the future development of our research.

2 CONTEXT AND MOTIVATION

Assessing similarity has multiple practical applications and the metrics provided for some domains may prove useful in another one. Whether in recommendation engines that propose similar objects based on the ones liked by a user or natural language translators that suggest synonyms, assessing (semantic) similarity is a crucial phase. Because of a rich mathematical and lexical background of knowledge graphs and ontologies, there are multiple applications that exploit their characteristics, from measuring semantic similarity (Agirre et al., 2010) or relatedness (Agirre et al., 2015) between the abstract concepts, up to more sophisticated problems that build on the previously mentioned tasks, such as named entity disambiguation (Zhu and Iglesias, 2018), entity set expansion (Adrian and Manna, 2018) or case-based reasoning (Zbroja and Ligęza, 2001).

Semantic similarity methods may be also useful for determining similarity between graph-based models, such as e.g. business process models. As companies usually own many business processes and store their models in several versions, it may cause misunderstandings, errors and delays, especially when two departments that work together use similar, but not identical models. Thanks to the comparing algorithms, it is possible to find similar processes and standardize the procedures in a company. There exist many algorithms for comparing business process models, mostly based on element labels, syntax (element types and model structure), and model behaviour (Dumas et al., 2009). However, in practice, it is hard to compare and evaluate them, because each algorithm has its specific context of application, and they may give different results depending on the features of the models (Cayoglu U. et al., 2014; Antunes G. et al., 2015).

The main objective of this paper is to give an overview of fundamental semantic similarity metrics and additionally grasp their influence on each other, providing an extended perspective on their principles. We believe that it is a firm starting point that can lead to a better understanding of different interpretations of semantic similarity, the resulting metrics and what each of this methods "brings to the table".

3 EVOLUTION AND ANALYSIS OF SIMILARITY METRICS

By definition, similarity is *a state of being almost the same*, what leads to a variety of possible interpretations – and therefore becomes a concept very hard to standardize by any single measure (see Table 1). In this section, we outline the directions in which the metrics have been developed and analyze the relations among them.

3.1 Evolution of Approaches

Apart from purely experimental attempts to discover the universal notion of similarity, the works such as "Dimensions of Similarity" (Attneave, 1950) began the period of associating similarity with a geometrical representation of the concepts characteristics, using the mathematical distance measuring methods to quantify the result. On the other hand, in 1977 Tversky published an article that can be considered as one of the first contributions to the modern discussion about the similarity metrics, starting a feature-based class of methods (Tversky, 1977). The paper questioned the geometric approach towards similarity, and presented a novel "Contrast Model" method of assessing similarity based on the features of two concepts, treating them as simple sets, taking into consideration both their common attributes and their differences.

A different approach to this problem was then presented in 1989, when the semantic similarity was defined as the aggregate of the interconnections between the concepts (Rada et al., 1989). The paper introduced an edge-based method leveraging the tree structure of the graphs. Six years later, one of the most significant works in this fields was published by Philip Resnik, who presented a novel approach that used "Information Content" to calculate the similarity between two concepts (Resnik, 1995). That seminal paper started a node-based group of methods that uses a text corpus to calculate the IC metric (estimating probability of a term's occurrence) and influenced later on both edge-based and hybrid approaches.

A hybrid class of methods attempts to combine advantages of both node- and edge-based approach, for example incorporating knowledge from a particular domain while calculating similarity (Knappe et al., 2003). This class of methods has been intensively developed especially in 2007 and 2008 and was also inspired by the edge-based similarity measures (e.g. Jiang and Conrath method influenced the Othman et al. measure) and the node-based ones (e.g. Zhou et al. similarity measure was inspired by the Wu and Palmer's work) (Jiang and Conrath, 1997; Othman et al., 2008; Zhou et al., 2008; Wu and Palmer, 1994).

Ultimately, a new category of semantic-based measures emerged in 2016 along with the Fähndrich et al. work (Fähndrich et al., 2016). They described the similarity methodology that decomposes the concepts into semantic "primes" and then applies marker passing, counting the activations that occurs and normalising them by the number of initial activation to obtain the semantic distance.

Nowadays, the concepts from different approaches are being mixed, such as in one of the newest feature-based metrics, the Sigmoid similarity (Likavec et al., 2019) that can take into account the underlying structure of the ontology describing the analyzed concepts.

3.2 Ontological View of the Approaches

The metrics can be thus organized into categories defined by which characteristics of the *description* of the considered entities are taken into consideration: in graph-based knowledge bases the entities are described by attributes and relations with other entities, and thus we call the metrics either Node or Edge Based (see Figure 1). Node Based is a class including metrics based on node analysis, which use internal issues (such as link density, number of children, etc.) and external ones such as shared annotations or information content measuring how specific and informative a particular term is. Edge Based metrics include those focusing on relationship analysis and often use structural measures such as shared path or distance. Metrics that are based on both node-specific information and edge-based measures are called Hybrid. Moreover, Feature Based and Semantic methods are considered separately.

Each method classified in our ontology has its attributes, such as the year in which it was developed or an application domain for which it was proposed, and relations with other methods on which it builds. These different aspects can be represented visually as we can see in Fig. 1. Multiple aspects of the research landscape of semantic similarity analysis contain:

- on the timeline, we can see when certain methods were developed;
- the classes of methods are represented by swimlanes;
- the methods' influences on each other are represented by arrows, and
- the domain that the method was developed for (note that this does not necessarily limit the usage of the method to this domain) are marked with different colours and referenced below the graph.

Some methods demonstrate unique characteristics, such as the one in (Rodríguez and Egenhofer, 2003), where the feature-based model allows to compare terms across different ontologies. Three years later, the X-Similarity metric, that was built upon it, improved the correlation with the human notion of similarity reaching 84% which can be considered quite high score for this class (Petrakis et al., 2006). The closest to human guess of similarity from all the metrics compared in this article was reached by a semantic-based MP metric with the correlation of 88.2% (Fähndrich et al., 2016). Another approach called *Align, Disambiguate and Walk* or ADW for short, presented in 2013, is until now considered to present a state-of-the-art performance in textual, word and sense similarity (Pilehvar et al., 2013). Some of

Table 1: Various approaches to measuring similarity grouped in classes. The concepts used in the above formulas: c_1, c_2 – compared concepts; $IC = -log p(c)$ – Information Content; C_{MICA} – most informative common ancestor; $p(C_A)$ – probability of c occurring in a specific corpus, estimated by frequency of annotation; CDA – Common disjunctive ancestor; W_{k1}, W_{k2} – fuzzy membership matrix of graph G; $Pr[c_k] = \frac{\sum_{t_j \in C}(W_{kj} \cdot |t_j|)}{|U|}$; $Pr[c_i|c_k] = \frac{\sum_{t_j \in C}(\min(W_{ij}, W_{kj}) \cdot |t_j|)}{\sum_{t_j \in C}(W_{kj} \cdot |t_j|)}$; r_i^j – denote the rank of sense $s_i \in S$ in signature j; α – variable representing possible asymmetric similarity relation; $S_{neighb}(c_1, c_2) = max_{i \in R} \frac{|c_{1i} \cap c_{2i}|}{|c_{1i} \cup c_{2i}|}$; $S_{descr}(c_1, c_2) = \frac{|c_1 \cap c_2|}{|c_1 \cup c_2|}$; S_w, S_u, S_z are respectively the measure of the similarity between synonym sets, features and semantic neighbourhoods among classes c_1 of ontology p and classes c_2 of ontology q; SP – shortest path relating concepts; $\vec{P}$ – vector representation of measured concepts; $\delta(a, b)$ – number of edges on the shortest path between a and b. l – shortest path between concepts;; h – depth of the subsumer in the hierarchy; α, β – parameters scaling the contribution of l and h; Depth – the depth of the taxonomy; C,k – constants derived throughout experiments; d – the number of changes of direction in the path that relates c_1 and c_2; N1,N2 – the distances that separates c1 and c2 from the root node; N – the distance between closest common ancestor of C1 and c2 from the root node; $PF(c_1, c_2) = (1 - \lambda)(Min(N1, N2) - N) + \lambda(|N1 - N2| + 1)^{-1}$ and λ is a boolean coefficient; $SV(c) = \sum_{t \in T_c} S_c(t)$; $Ans(C1), Ans(C2)$ – description sets of terms C1 and C2 respectively.

Class	Method	Formula
Node Based	1. Resnik (Resnik, 1995)	$Sim_{Res}(c_1, c_2) = IC(c_{MICA})$
	2. Lin (Lin et al., 1998)	$Sim_L(c_1, c_2) = \frac{2 \cdot IC(c_{MICA})}{IC(c_1) + IC(c_2)}$
	3. Jiang (Jiang and Conrath, 1997)	$Sim_{JC}(c_1, c_2) = 1 - IC(c_1) + IC(c2) - 2 \cdot IC(c_{MICA})$
	4. Schlicker (Schlicker et al., 2006)	$Sim_{Rel}(c_1, c_2) = Sim_L(c_1, c_2) \cdot (1 - p(c_A))$
	5. GraSM (Couto et al., 2005)	$Sim_G(c_1, c_2) = \overline{\{IC(a) \mid a \in CDA(c_1, c_2)\}}$
	6. ADW (Pilehvar et al., 2013)	$Sim_{ADW} = \frac{\sum_{i=1}^{\lvert S \rvert} (r_i^1 + r_i^2)^{-1}}{\sum_{i=1}^{\lvert S \rvert} (2i)^{-1}}$
	7. Maguitman (Maguitman et al., 2005)	$Sim_M(c_1, c_2) = \max_k \frac{2 \cdot \min(W_{k1}, W_{k2}) \cdot \log Pr[c_k]}{\log(Pr[c_1 \mid c_k] \cdot Pr[c_k]) + \log(Pr[c_2 \mid c_k] \cdot Pr[c_k])}$
Feature	8. Tversky (Tversky, 1977)	$Sim_T(c_1, c_2) = \frac{\lvert c_1 \cap c_2 \rvert}{\lvert c_1 \cap c_2 \rvert + \alpha \lvert c_1 - c_2 \rvert + (\alpha - 1) \lvert c_2 - c_1 \rvert}$
	9. X-similarity (Petrakis et al., 2006)	$Sim_x(c_1, c_2) = \begin{cases} 1 & \text{if } S_{syns} > 0 \\ maxS_{neig}(c_1, c_2), S_{desc}(c_1, c_2) & \text{if } S_{syns} = 0 \end{cases}$
	10. Rodriguez (Rodríguez and Egenhofer, 2003)	$Sim_R(c_1^p, c_2^q) = W_w S_w(c_1^p, c_2^q) + W_u S_u(c_1^p, c_2^q) + W_n S_n(c_1^p, c_2^q)$
Edge Based	11. Rada (Rada et al., 1989)	$Sim_{SP}(c_1, c_2) = 2 \cdot Max(c_1, c_2) - SP$
	12. Pozo (Del Pozo et al., 2008)	$Sim(GO_i, GO_j) = cos(\vec{P}_i, \vec{P}_j) = \frac{\vec{P}_i * \vec{P}_j}{\lvert \vec{P}_i \rvert \lvert \vec{P}_j \rvert}$
	13. Wu et al. (Wu and Palmer, 1994)	$Sim_{WU}(L_s, L_t) = \max_{L_s \in V_s, L_t \in V_t} \left\{ \begin{array}{c} \text{the number of common} \\ \text{terms between} L_s \text{and} L_t \end{array} \right\}$
	14. Pekar (Pekar and Staab, 2002)	$Sim_{PS}(c_1, c_2) = \frac{\delta(c_a, root)}{\delta(c_a, root) + \delta(c_1, c_a) + \delta(c_2, c_a)}$
	15. Richardson (Richardson et al., 1994)	$Sim_{Rich}(c_1, c_2) = \max_{c_i} log \frac{1}{P(c_i)}$
	16. Li et al. (Li et al., 2003)	$Sim_{Li}(c_1, c_2) = e^{-\alpha l} \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} if c_1 \neq c_2$
	17. IntelliGO (Benabderrahmane et al., 2010)	$Sim_{Int}(c_1, c_2) = \frac{\vec{(c_1)} * \vec{(c_2)}}{\sqrt{\vec{(c_1)} * \vec{(c_1)}}^{\sqrt{\vec{(c_2)} * \vec{(c_2)}}}}$
	18. Leacock (Leacock, 1994)	$Sim_{L\&C}(c_1, c_2) = -log \frac{SP}{2 \cdot Depth}$
	19. HSO (Hirst et al., 1998)	$Sim_{HSO}(c_1, c_2) = C - SP - k \cdot d$
	20. Wu (Wu et al., 2005)	$Sim_{wup}(c_1, c_2) = \frac{2 \cdot N}{N1 + N2 + 2 \cdot N}$
	21. TBK (Slimani et al., 2006)	$Sim_{TBK}(c_1, c_2) = \frac{2 \cdot N}{N1 + N2} \cdot PF(c_1, c_2)$
Hybrid	22. Othman (Othman et al., 2008)	$Sim_O(c_1, c_2) = 1 - \min\{1, \frac{dist(c_1, c_2)}{\max IC(c)}\}$
	23. Wang (Wang et al., 2007)	$Sim_W(c_1, c_2) = \frac{\sum_{t \in T_{c_1} \cap T_{c_2}} (S_{c_1}(t) + S_{c_2}(t))}{SV(c_1) + SV(c_2)}$
	24. Knappe (Knappe et al., 2003)	$Sim_K(c_1, c_2) = p \cdot \frac{\lvert Ans(c_1) \cap Ans(c_2) \rvert}{\lvert Ans(c_1) \rvert} + (1 - p) \cdot \frac{\lvert Ans(c_1) \cap Ans(c_2) \rvert}{\lvert Ans(c_2) \rvert}$
	25. Zhou (Zhou et al., 2008)	$Sim_Z(c_1, c_2) = 1 - k(\frac{ln(len(c_1, c_2) + 1}{ln(2 \cdot (deep_{max} - 1))})$ $-(1 - k) \cdot ((IC(c_1) + IC(c_2) - 2 \cdot IC(\frac{lso(c_1, c_2))}{2})$
Semantic	26. MP (Fähndrich et al., 2016)	$Sim_{MP}(c_1, c_2) = \frac{\sum_{t=0}^{t_{max}} \sum_{x \in V} \phi(\hat{a}_t^*(x), c_1, c_2)}{\sum_{\forall w \in V} a^0(w)}$

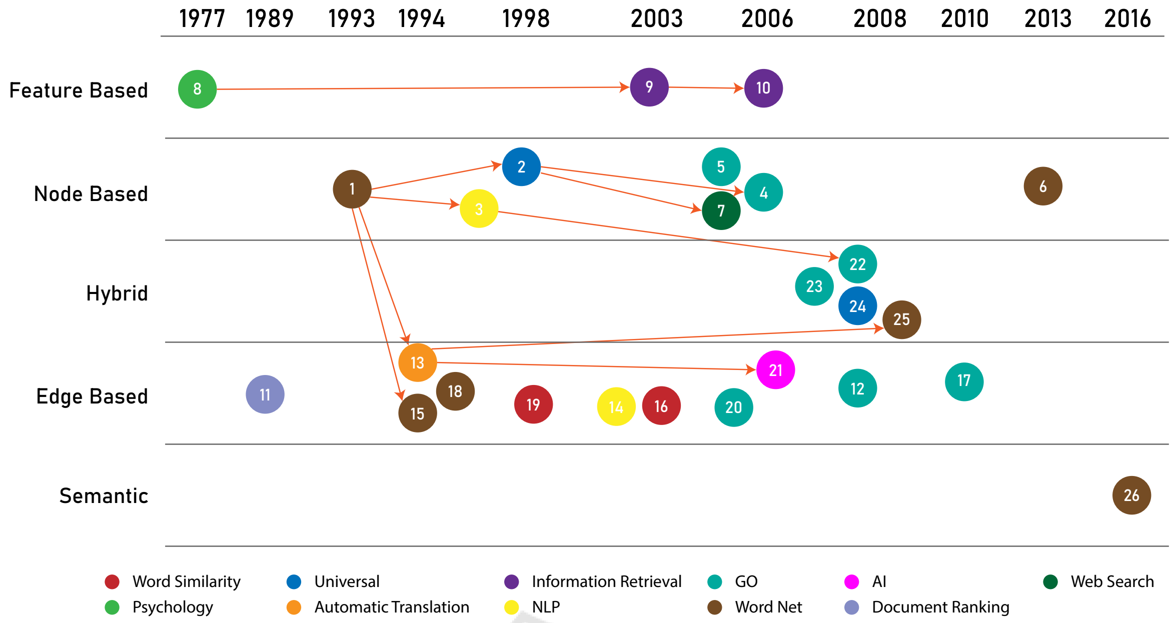


Figure 1: Evolution and influence relations among the classified methods.

the methods however perform well only in ideal conditions where quality of data is very good – an example of such a method is the one presented in (Schlicker et al., 2006) – or by definition contain a significant bias of symmetry. A good example of such bias is the Li et al. measure where the asymmetric nature of the similarity relation is consciously not considered (Li et al., 2003). Such features should be thus taken into consideration when selecting a method.

3.3 Bibliometric Analysis

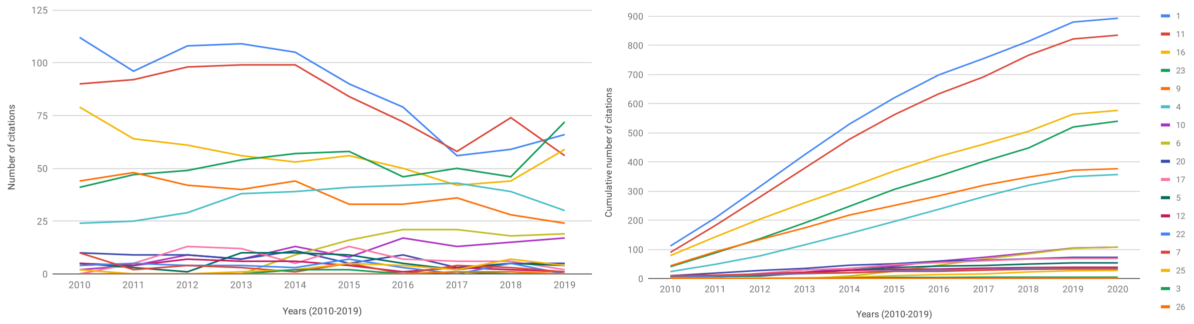
For the works analyzed in this paper, based on the data from the Scopus bibliometric database, in Figure 2, we present two charts with the distribution of citations to these works. It is easy to notice that two works (Wang et al., 2007; Li et al., 2003) have been recently increasingly cited. They fall into the categories of edge-based and hybrid approaches. On the other hand, from the cumulative citation chart one can observe that apart from the two mentioned works there are other pairs of highly cited papers – the classic edge-based (Rada et al., 1989) and node-based (Resnik, 1995) approaches from 90', which still gain attention as the base for the newly developed methods, as well as the feature-based (Rodríguez and Egenhofer, 2003) and node-based (Schlicker et al., 2006) methods, which provided foundations for modern semantic similarity measures.

4 AN INTERACTIVE HISTORICAL ATLAS FOR RESEARCH METHODS

To facilitate the literature review, we propose to use a simple tool based on a concept of a historical atlas. Management and visualization of historical data that concern multiple actors, events and references is a specific problem that can be used to alleviate the acquisition of large collections of knowledge. One of the main intentions of the tool is to keep the data model general enough to be analyzed from different points of views and used for different visualizations (e.g., chronology, spatial map or dependency graph). We adopted an assumption that the tool should be intuitive even for a non-technical researcher and the performance should allow real-time work with data. Although the architecture of the tool uses a web browser, the tool can also work offline.

4.1 Representation of the Methods as Historical Literature

Similarity methods and papers about them can be considered as a part of literature history which could be easily visualized with interactive historical atlas described earlier. Methods, articles and authors have been modelled as part of a historical atlas data model where methods and authors are recognized as specific



(a) in each year

(b) cummulative

Figure 2: Number of citations of the state-of-the-art papers in the field (based on the data from the Scopus database).

types of the same general "event" concept and articles are "references" for methods. This approach allows to connect methods through influence relationship and attach a number of papers to methods and authors.

Data prepared according to this model can be then visualized in our application. The front-end of the tool uses data from back-end endpoint provided by user in online mode or from uploaded file with data encoded in JSON format. For our research, we have prepared data instances describing the analyzed methods (the data is available on the tool website). Below we present a snippet with Node Based – Information Content method and example of influence relationship with indeterminate "test" method.

```
{ "nodes": [{
        "label": "Author",
        "name": "Philip Resnik",
        "id": "author-resnik"
     },{
        "label": "Method",
        "name": "Node-based -
            Information Content",
        "description": "The method uses
            shared information content...",
        "id": "method-IC"
     },{
       "label": "Reference",
       "title": "Semantic Similarity in a
           Taxonomy: An...",
       "id": "resnik1999semantic"
     },
  "edges": [{
       "from": "author-resnik",
       "to": "method-IC",
       "label": "AUTHOR"
     },{
       "from": "method-IC",
       "to": "method-test",
       "label": "INFLUENCED"
     }]}
```

For chronology view, all the method's details and the related articles' titles should be stored in one array, and the information about the authors can be omitted. Below the same example method encoded with additional start and end dates of method, where the start date is the year of publication and the end date is the year of the publication of the latest influenced method.

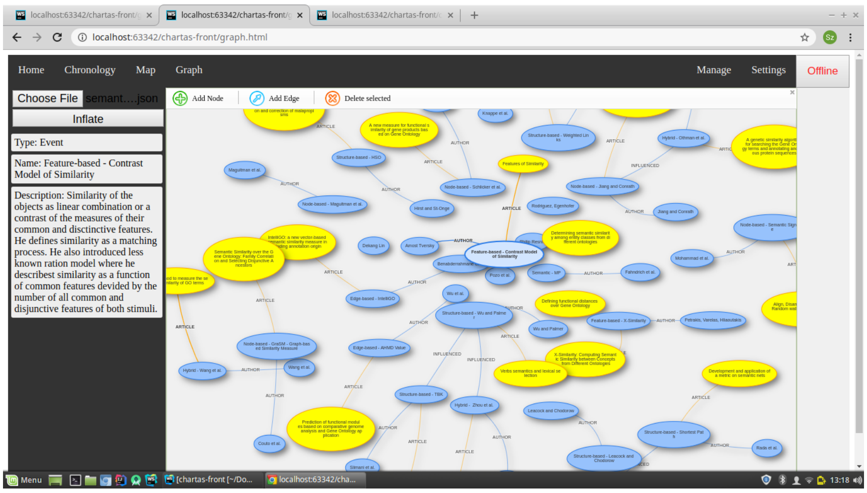
```
{ "events": [{
      "id": "method-IC",
      "content": "Node-based
      - Information Content",
      "start": "1990-01-01",
      "end": "2006-01-01",
      "description": "The method
      uses shared information...",
      "references" : ["Semantic
      Similarity in a Taxonomy..."]
}]}
```

4.2 Implementation of the Atlas

The front-end of the application has been developed as an interactive website with scripts implemented in JavaScript. The project uses two third-party libraries: vis-network and vis-timeline, dual licensed under The Apache 2.0 and MIT License. For dependency management npm Software Registry was used. The source code of the application and the sample data is available at: https://anonymous.4open.science/r/95f844e7-afb8-4876-b27e-1e48d56907a6/.

The application allows users to upload files with data encoded in format presented earlier or use an online mode in which it is possible to set the already deployed endpoint with data and then receive it from the server. Navigating to graph or chronology page, the user can see the data in a selected view (see Fig. 3, 4).

As the natural relationships between papers, authors and methods can be modelled and visualized as a graph or using a simplified chronology, we believe that the proposed tool can be useful for researchers in various domains. The flexible model, based on an "event" entity easily captures any phenomena occurring in time and the interactive visualizations help in analysis of the state-of-the-art.


localhost:63342/chartas-front/graph.html
Home
Chronology
Map
Graph
Manage
Settings
Offline
Add Node
Add Edge
Delete selected
Feature based - Contrast Model of Similarity
Structure-based - Wu and Palmer
Node-based - Jiang and Conrath
Structure-based - Shortest Path
Edge-based - AHMD Value
ARTICLE
AUTHOR
INFLUENCED
Menu
13:18

Choose File semant….json

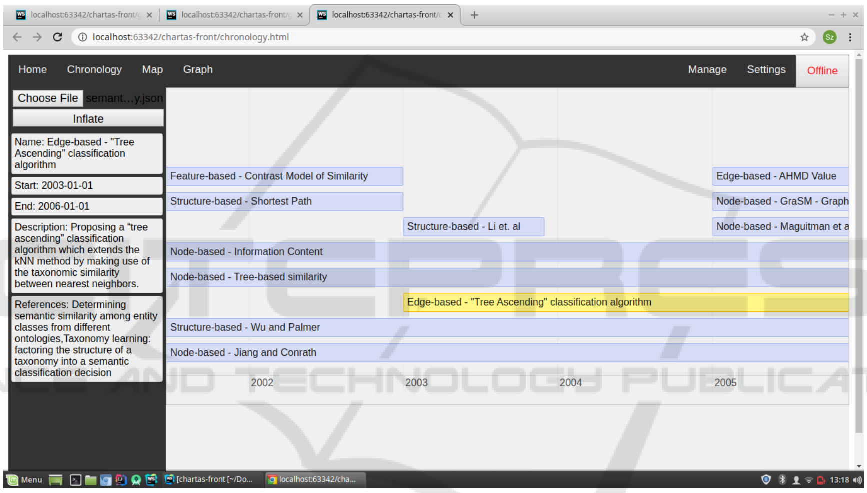
Inflate

Type: Event

Name: Feature-based - Contrast Model of Similarity

Description: Similarity of the objects as linear combination or a contrast of the measures of their common and disctinctive features. He defines similarity as a matching process. He also introduced less known ration model where he describest similarity as a function of common features devided by the number of all common and disjunctive features of both stimuli.

Figure 3: Screenshot of semantic similarity methods graph visualization.



Choose File semant…y.json

Inflate

Name: Edge-based - "Tree Ascending" classification algorithm

Start: 2003-01-01

End: 2006-01-01

Description: Proposing a "tree ascending" classification algorithm which extends the kNN method by making use of the taxonomic similarity between nearest neighbors.

References: Determining semantic similarity among entity classes from different ontologies,Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision

Feature-based - Contrast Model of Similarity

Structure-based - Shortest Path

Structure-based - Li et. al

Edge-based - AHMD Value

Node-based - GraSM - Graph

Node-based - Maguitman et a

Node-based - Information Content

Node-based - Tree-based similarity

Edge-based - "Tree Ascending" classification algorithm

Structure-based - Wu and Palmer

Node-based - Jiang and Conrath

Figure 4: Screenshot of semantic similarity methods chronology visualization.

5 CONCLUSION

Semantic similarity of concepts, objects or words, described with some degree of formalization can be quantified in different ways. The semantics itself may be defined based on features or geometrical properties of the underlying knowledge base. In this paper, we have reviewed existing approaches to semantic similarity analysis and presented different metrics within a simple ontology. We have analyzed how the approaches evolved in time and in which application domains they have been used. We formalized their interrelatedness with a graph-based model, and provided a tool that can facilitate literature review of any topic. For future, we plan to further enrich the methods' ontology with new instance and possibly relations, and extend the tool with more analytical capabilities.

REFERENCES

Adrian, W. T. and Manna, M. (2018). Navigating online semantic resources for entity set expansion. In Calimeri, F., Hamlen, K., and Leone, N., editors, *Practical Aspects of Declarative Languages*, pages 170–185, Cham. Springer International Publishing.

Agirre, E., Barrena, A., and Soroa, A. (2015). Studying the wikipedia hyperlink graph for relatedness and disambiguation. *arXiv preprint arXiv:1503.01655*.

Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring knowledge bases for similarity. In *LREC*.

Antunes G. et al. (2015). The process model matching contest 2015. In *Proc. of the 6th Intl. Workshop on Enterprise Modelling and Information Systems Architectures, September 3-4, 2015 Innsbruck, Austria*, volume 248, pages 127–155, Bonn.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, 63(4):516–556.

Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., and Devignes, M.-D. (2010). Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC bioinformatics*, 11(1):588.

Cayoglu U. et al. (2014). Report: The process model matching contest 2013. In Lohmann, N., Song, M., and Wohed, P., editors, *Business Process Management Workshops*, pages 442–463, Cham. Springer International Publishing.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005). Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In *Proc. of the 14th ACM int. conf. on Information and knowledge management*, pages 343–344.

Del Pozo, A., Pazos, F., and Valencia, A. (2008). Defining functional distances over gene ontology. *BMC bioinformatics*, 9(1):50.

Dumas, M., García-Bañuelos, L., and Dijkman, R. M. (2009). Similarity search of business process models. *IEEE Data Eng. Bull.*, 32(3):23–28.

Fähndrich, J., Weber, S., and Ahrndt, S. (2016). Design and use of a semantic similarity measure for interoperability among agents. In *German Conference on Multiagent System Technologies*, pages 41–57. Springer.

Hirst, G., St-Onge, D., et al. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Knappe, R., Bulskov, H., Andreasen, T., and Kaynak, O. (2003). On similarity measures for content-based querying. In *10th International Fuzzy Systems Association World Congress, IFSA*, pages 400–403. Citeseer.

Leacock, C. (1994). Filling in a sparse training space for word sense identification. *Ph. D. thesis, Macquarie University*.

Li, Y., Bandar, Z. A., and McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882.

Likavec, S., Lombardi, I., and Cena, F. (2019). Sigmoid similarity-a new feature-based similarity measure. *Information Sciences*, 481:203–218.

Lin, D. et al. (1998). An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304.

Maguitman, A. G., Menczer, F., Roinestad, H., and Vespignani, A. (2005). Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, pages 107–116.

Othman, R. M., Deris, S., and Illias, R. M. (2008). A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of biomedical informatics*, 41(1):65–81.

Pekar, V. and Staab, S. (2002). Taxonomy learning-factoring the structure of a taxonomy into a semantic classification decision. In *COLING 2002: The 19th Int. Conference on Computational Linguistics*.

Petrakis, E. G., Varelas, G., Hliaoutakis, A., and Raftopoulou, P. (2006). X-similarity: Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management*, 4(4).

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 1)*, pages 1341–1351.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Richardson, R., Smeaton, A., and Murphy, J. (1994). Using wordnet as a knowledge base for measuring semantic similarity between words.

Rodríguez, M. A. and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE transactions on knowledge and data engineering*, 15(2):442–456.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1):302.

Slimani, T., Yaghlane, B. B., and Mellouli, K. (2006). A new similarity measure based on edge counting. *Proceedings of the World Academy of Science, Engineering and Technology*, 17:3.

Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4):327.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281.

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic acids research*, 33(9):2822–2837.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Zbroja, S. and Ligęza, A. (2001). Case-based reasoning within tabular systems. extended structural data representation and partial matching. In *Flexible Query Answering Systems*, pages 230–239. Springer.

Zhou, Z., Wang, Y., and Gu, J. (2008). New model of semantic similarity measuring in wordnet. In *2008 3rd Int. Conference on Intelligent System and Knowledge Engineering*, volume 1, pages 256–261. IEEE.

Zhu, G. and Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, 101:8–24.