Shen et al. Journal of Biomedical Semantics (2018) 9:20
https://doi.org/10.1186/s13326-018-0188-7

Journal of Biomedical Semantics

RESEARCH Open Access

EAPB: entropy-aware path-based metric for ontology quality

Ying Shen[1†], Daoyuan Chen[1†], Buzhou Tang[2], Min Yang[3] and Kai Lei[1*]

Abstract

Background: Entropy has become increasingly popular in computer science and information theory because it can be used to measure the predictability and redundancy of knowledge bases, especially ontologies. However, current entropy applications that evaluate ontologies consider only single-point connectivity rather than path connectivity, and they assign equal weights to each entity and path.

Results: We propose an **E**ntropy-**A**ware **P**ath-**B**ased (EAPB) metric for ontology quality by considering the path information between different vertices and textual information included in the path to calculate the connectivity path of the whole network and dynamic weights between different nodes. The information obtained from structure-based embedding and text-based embedding is multiplied by the connectivity matrix of the entropy computation. EAPB is analytically evaluated against the state-of-the-art criteria. We have performed empirical analysis on real-world medical ontologies and a synthetic ontology based on the following three aspects: ontology statistical information (data quantity), entropy evaluation (data quality), and a case study (ontology structure and text visualization). These aspects mutually demonstrate the reliability of the proposed metric. The experimental results show that the proposed EAPB can effectively evaluate ontologies, especially those in the medical informatics field.

Conclusions: We leverage path information and textual information to enrich the network representational learning and aid in entropy computation. The analytics and assessments of semantic web can benefit from the structure information but also the text information. We believe that EAPB is helpful for managing ontology development and evaluation projects. Our results are reproducible and we will release the source code and ontology of this work after publication. (Source code and ontology: https://github.com/AnonymousResearcher1/ontologyEvaluate).

Keywords: Ontology evaluation, Ontology modeling, Entropy-based metric, Knowledge representation, Big data and semantics

Background

The term ontology refers to "a representation and definition of the categories, properties, and relations of the concepts, data, and entities that substantiate one, many, or all domains." [1] Ontology has attracted increasing attention recently due to its broad applications such as information retrieval, relation extraction, and question answering. Significant progress has been made in the ontology construction [2]. However, the ontology evaluation is still a relatively new territory and under-explored. As a result, there are few commonly agreed-upon methodologies and metrics for ontology evaluation.

Considering each ontology as a graph or a network, entropy can be used as a measure of the complexity and redundancy of the graph. Ontologies may contain data and concepts redundancy that could be removed for the sake of consolidation and conciseness without changing the overall meaning. The information density is operationalized based on the normalized entropy measured between all concept pairs in the ontology [3]. States of lower entropy occur when ontology become organized. In the literature, the entropy evaluation of the lexical information included in the ontology has been studied in the last decade and have been proved to be helpful for ontology evaluation [4].

* Correspondence: leik@pkusz.edu.cn
†Ying Shen and Daoyuan Chen contributed equally to this work.
[1]Shenzhen Key Lab for Information Centric Networking & Blockchain Technology (ICNLAB), School of Electronics and Computer Engineering, Peking University Shenzhen Graduate School, 518055 Shenzhen, People's Republic of China
Full list of author information is available at the end of the article

Despite the effectiveness of previous studies, current entropy applications used to evaluate ontology have three limitations, in that they (1) Exclusively consider single point connectivity rather than paths [5], which neglects information pertaining to non-adjacent nodes. (2) Assign equal weights to edges and paths [6], which induces a loss of diversity. (3) Assume vertices are static, which ignores the various aspects of vertices when interacting with different neighboring vertices [7].

To address these three limitations, this article describes an Entropy-Aware Path-Eased quality metric for ontologies (EAPB) by comparing their information densities to those of other ontologies. We consider the path information between different vertices in ontology as well as the textual information included in the path to calculate the dynamic weight between different nodes and the connectivity path of the entire network. Specifically, we first apply CNN to learn the structure-based embedding and text-based embedding to capture both the ontology network structures and their encapsulated textual information. Subsequently, the information gain which is in the form of a matrix obtained by a cosine similarity calculation of the relevancy between nodes u and v, is multiplied by the connectivity matrix of entropy computations. Finally, we validate the effectiveness and robust superiority of our model on four real-world ontologies.

Three infectious disease-relevant ontologies, i.e. Infectious Disease Ontology[1] (IDO), Infectious Disease Ontology for Dengue[2] (IDODEN), and Disease Ontology[3] (DO) are adopted as baselines. Our material includes an in-house ontology that is used to develop an ontology-driven clinical decision support system for infectious disease diagnosis and antibiotic prescription (IDDAP) [8]. To demonstrate the applicability and generality of our quality metric for ontologies, we conduct evaluations on real-world ontologies with different structures and different textual information. To verify whether our quality metric can make a significant performance boost by incorporating textual information into the EAPB architecture, we assess ontologies with the same structures but different textual information, as well as report the ablation tests in terms of discarding the textual information. The textual attention visualization and ontology statistical information are used as references to evaluate the validity of the calculation.

To summarize, the core contributions of this study are as follows:

- To overcome the single point connectivity and equal weight problem, we consider the path information between different vertices in an ontology and the text information included in the path. The information gain obtained from these sources are used to adjust the connectivity matrix of entropy computation to provide a reliable metric for evaluating ontology redundancy.
- To solve the unreal assumption of traditional ontology computations, in which each vertex is represented as a static embedding vector, we consider that the nodes' interactions are dynamic by adapting mutual attention to emphasize those words that are focused by its neighbor vertices. The neural models are beneficial for representing ontology information.
- An ontology assessment is conducted on four real-world ontologies from three aspects: ontology statistical information (data quantity), entropy evaluation (data quality), and case study (ontology structure and text visualization), which mutually reflect the reliability of the proposed quality metric. The experimental results show that, compared with existing methods, EAPB can more effectively evaluate ontologies.

The remainder of this paper is organized as follows. "Related Work" section reviews the previous state-of-the-art ontology assessment methods. "Methods" section introduces the materials, i.e., ontologies related to infectious diseases. "Experiment and Results" section proposes the EAPB quality metric with information divergence by considering both text-based and structure-based embedding. "Discussion and conclusion" section presents the ontology statistical information, the evaluation experiments, and summarizes the evaluation results. Conclusions and possibilities for future work are outlined in "Methodology" section.

Related work

Ontology assessment with information density and entropy

Zaveri et al. [9] unified and formalized commonly used terminologies across papers related to data quality and provided a comprehensive list of the dimensions and metrics. In this list, these authors qualitatively and quantitatively selected 18 data dimensions, including accessibility dimensions, intrinsic dimensions, contextual dimensions, and representational dimensions, involving 69 metrics. Similarly, Färber et al. [10] used 34 evaluation indicators to perform statistics, analyses and comparisons of five renowned databases: DBpedia, Freebase, OpenCyc, Wikidata, and YAGO.

Ontology metrics can be divided into three main dimensions: structural, functional, and usability-profiling. The structural dimension of ontologies exploits the syntax and formal semantics of the ontologies represented as graphs. In this form, the topological, logical and meta-logical properties of an ontology are measured by means of a context-free metric [11]. The functional dimension is related to the intended use of the context of

the given ontology and its components, while the usability-profiling dimension employs the ontology annotations to address the communication context of an ontology [11]. Considering that redundancy is not only related to the subcharacteristics of structural dimension (e.g., a high structural redundancy indicates potential tangledness) but also to the subcharacteristics of the other two dimensions (e.g., a high structural redundancy indicates improper modularity, and a high textual redundancy harms the reusability since that the classes are easily confused) [12], we take into account these three metrics and conduct the entropy-based metric in deep neural network to interactively learn the structural information included in the ontology and the context information from entity annotations.

Entropy has become increasingly popular in computer science and medical informatics [13]. Calmet et al. [6] realized the distance measure using entropy and mutual information from information theory. They considered ontological structures and distances using centrality measures, such as the degree, closeness or betweenness. However, all the edges were assumed to have the same weight. Based on [6], Doran et al. [7] proposed a reformulation of the entropy metric to evaluate the amount of information carried by both the ontology structure and the semantics associated with the edges of the ontological graph. However, the assumption of [7] that all language level edges must be equal cannot be applied in most cases. Gurupur et al. [5] calculated the probability distributions and the information entropy of the knowledge base with a new metric that measures the node source's connectivity strength based on the number of unique paths from one node to another. Nevertheless, the study of [5] exclusively considers single-point connectivity rather than multiple connected paths.

Assessing ontology using network embedding

Network embedding aims to map vertices of a network (ontology) onto a low-dimensional space according to their structural roles in the network.

In recent years, a large number of network embedding models have been proposed to learn efficient vertex embeddings, including DeepWalk, LINE, and node2vec. However, these structure-only models do not consider the information that accompanies the vertices in networks. Therefore, compared with sophisticated deep learning architectures such as convolutional neural networks, these methods usually yield inferior results when applied to particular machine learning tasks.

Many studies have attempted to incorporate information into network embedding models or convolutional neural networks. For example, Yang et al. [14] proposed the text-associated DeepWalk model to incorporate textual features of vertices into network representation learning under a matrix factorization framework. Zhang et al. [15] proposed a content-enhanced network-based computational approach to jointly leverage the network structure and the content information. Tu et al. [16] proposed a max-margin Deep-Walk to enhance the discriminatory ability.

Most network embedding methods rely solely on network structure while ignoring the diverse roles of vertex interactions. The rich network content information used to describe the characteristics of a node is beneficial for evaluating entropy.

Methods

Ontologies for infectious disease

To verify the effectiveness of EAPB, four real-world ontologies related to infection diseases are adopted for comparison. Infection constitutes the invasion of an organism's body tissues by disease-causing agents, the multiplication of these agents, the reaction of host tissues to these organisms, and the toxins produced by these organisms [17, 18].

1) Disease Ontology (DO)

The Disease Ontology (DO) database represents a comprehensive knowledge base of 8043 inherited, developmental and acquired human diseases [19]. The DO database is divided into eight categories, including diseases by infectious agent, diseases of metabolism, diseases of mental health, and five others.

2) Infectious Disease Ontology (IDO)

The IDO [20] was designed as a set of interoperable ontologies that provide coverage of the infectious disease domain. At present, it is considered the most complete infectious disease ontology. However, the IDO lacks some infectious disease-relevant classes, such as "hemolytic-uremic syndrome."

3) Dengue ontology (IDODEN)

The IDODEN is an extension of the Infectious Disease Ontology (IDO) for dengue fever [21]. IDODEN reuses the Malaria Ontology IDOMAL [22], an existing infectious disease ontology. IDODEN contains a wide spectrum of ontological descriptions, from descriptions of the disease itself to descriptions of vector biology, virus biology, and epidemiology.

4) IDDAP ontology

For conciseness, the ontology applied in the IDDAP system we developed is called IDDAP ontology. The

ontology hierarchical conceptual schema of IDDAP ontology covers the following nine dimensions: (1) disease, (2) infection site, (3) bacteria, (4) animal, (5) symptom, (6) symptom type, (7) situation, (8) complication, and (9) antibiotic. Many infectious disease-relevant ontology resources were reused to construct the IDDAP ontology, including the DO and IDO mentioned above, as well as the NCBI organismal classification ontology,[4] the Human Phenotype Ontology,[5] DrugBank,[6] Antibiotic Guidelines (2015–2016),[7] the Antibacterial Spectrum Guide, and various websites (e.g., the U.S. National Library of Medicine (NLM)[8]). For example, the species included in the NCBI (including bacteria, viruses, fungi, and animals) were adopted to facilitate the completion of the IDDAP ontology knowledge base.

The constructed domain ontology contains 1,267,004 classes, 7,608,725 axioms, and 1,266,993 members of "SubClassOf" that pertain to infectious diseases, bacteria, syndromes, anti-bacterial drugs and other relevant components. The system includes 507 infectious diseases and their therapy methods in combination with 332 different infection sites; 936 relevant symptoms of the digestive, reproductive, neurological and other systems; 371 types of complications; 838,407 types of bacteria; 341 types of antibiotics; 1504 pairs of reaction rates (antibacterial spectrum) between antibiotics and bacteria; 431 pairs of drug interaction relationships; and 86 pairs of antibiotic-specific population contraindicated relationships. The amoxicillin class of IDDAP ontology is presented as an example (Fig. 1).

A graph G(V, E) is a math concept [23] that represents a set of vertices V and edges E. An ontology is an undirected graph with edges that connect unordered pairs of vertices. The IDDAP ontology and its three baselines (IDODEN, IDO, and DO), which contain no loops or multiple edges, can be considered as simple graphs.

Comparison of ontologies

The studied ontologies are compared by assessing the numbers of triples, classes, entities and relationships among the IDO, IDODEN, DO and IDDAP ontologies.

Triples

The quantity of triples included in different ontologies is calculated.

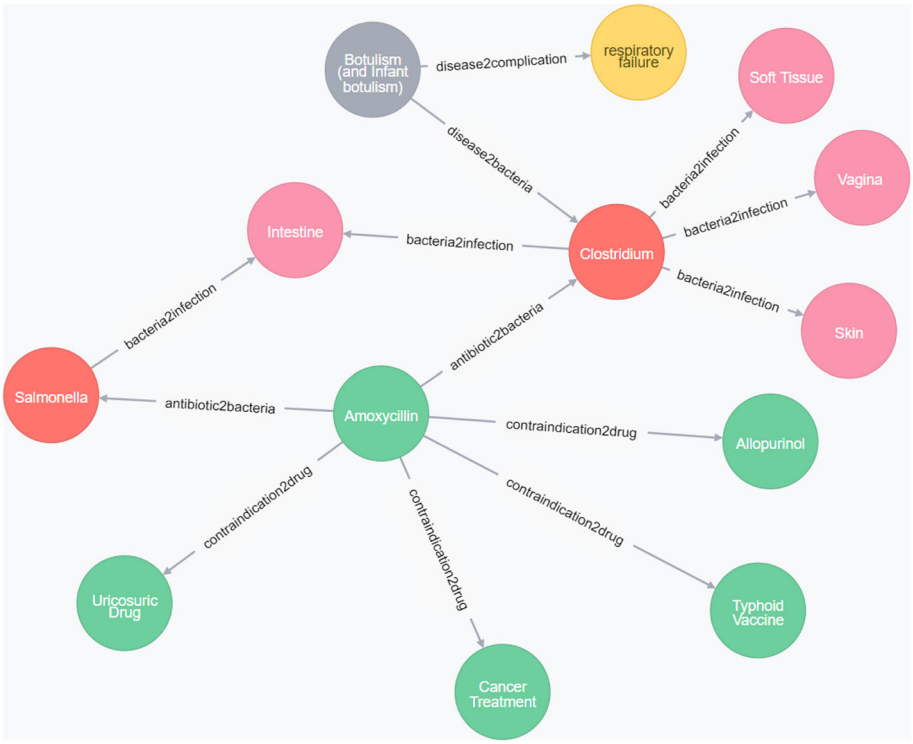

Fig. 1 IDDAP ontology—Amoxicillin. The green nodes represent contraindications between Amoxicillin and other drugs (relation_contraindication_drug). The red nodes identify bacteria that can be treated by Amoxicillin (relation_antibiotics_bacteria). The pink nodes indicate the relationships between infection sites and bacteria (relation_infs_bacteria). The gray nodes specify the relationships between diseases and bacteria (relation_disease_bacteria). Finally, the yellow nodes show the relationships between diseases and complications (relation_disease_ complication)

Classes and Entities

The classes of IDO, IDODEN and DO are recognized via the labels `owl:Class` and `rdfs:subClassOf`, whereas the class of IDDAP is identified via `Class IRI`. The statistics of the class quantities included in the ontologies is conducted using Protégé.

Relationships

The number of relationships is calculated via the labels `owl:ObjectProperty`, `owl:AnnotationProperty` and `owl:DataProperty`.

Granularity

Ontology granularity refers to the levels of semantic detail carried by an entity and the structural abstraction of entities. In this study, we estimate the ontology granularity by counting the number of properties with the same label. For example, DBpedia has two types of relationships for the word "creator", `dbo:author` and `dbo:director`, whereas YAGO only has only one relevant relationship, `yago:created`. In this case, DBpedia has finer granularity than YAGO.

Experiment and results

The ontology assessment is conducted based on the following three aspects: ontological (statistical) information (data quantity), entropy evaluation (data quality), and case study (ontology structure and text visualization).

Ontology information statistics

Table 1 shows the statistical information from IDO, IDODEN, DO and IDDAP. We can observe that the number of triplets and classes increases while the number of relations decreases. IDDAP has fewer relationships which are predefined, including the relationships of disease_complication, disease_symptom, disease_bacteria, and symptom_type. In Tables 1 and 3, the entries with largest values are italicized for facilitating the reader's reading.

With different relationship types in their schema, IDO, IDODEN and DO differ in the granularity. As Fig. 2 presents, for the "bearer of" relationship, no such relationship is observed in DO; three sub-relationships are observed in IDODEN; and four sub-relationships are observed in IDO, including "has disposition," "has function," "has quality" and "has role." Since semantic granularity addresses the different levels of specification of an entity, we can conclude that IDO has the finest relative granularity according to the statistics and example.

Table 1 Summary statistics of ontologies

	IDO	IDODEN	DO	IDDAP
Triples with annotation/data	3901	23,657	129,670	*3,807,709*
Triples without annotation/data	960	5845	10,060	*1,272,502*
Class/Instance/Entity	507	5007	11,088	*1,267,005*
Subclassof	582	5834	10,008	*1,266,996*
Equivalent classes	81	0	*46*	*1100*
Disjoint classes	17	11	*6*	*23*
Object property	*39*	25	20	8
Annotation property	*63*	*63*	33	1

Ontology visualization is adopted to intuitively reflect the density of ontology information. The vertex classification experiment consists of comparing the two-dimensional visualization created from the embedding. We employ the network embedding projector (t-SNE) to further visualize the low-dimensional node representations learned by the embedding models. The node representations of four ontologies (IDO, IDODEN, DO and IDDAP) are presented in Fig. 3.

In Fig. 3(a), we can see that IDO ontology is constituted by 51 type of colors which indicate the connected components. Its simple structure and clear node definition leads to the lowest entropy value. The IDODEN and IDDAP ontology, although with complicated structure, is only composed by 27 and 19 types of colors respectively (Fig. 3(b) and (d)). The limited valuable information and loose structure of these two ontologies result in a certain degree of redundancy. In the DO ontology (Fig. 3(c)), there is 81 types of colors, which means this ontology contain the most connected components. The DO ontology is discrete, sparse, and unevenly distributed. There is overlap between the nodes, indicating that some of the information is redundant. The unwanted redundancy can be reduced or eliminated by data compression [24].

Entropy evaluation performance

As ontological redundancy is mainly manifested in loose structures and lengthy textual information, we conducted experiments based on two aspects: ontologies with different structures and different textual information; and ontologies with unchanged structures but different textual information. The textual attention visualization and ontology statistical information is used as a reference to evaluate the validity of the calculations.

Data preprocessing

To conduct text embedding, the class definition in ontology is considered the text description. For a DO class that has no property or other text description, class labels are used as literal queries to extract no more than 5 sentences from the Wikipedia terminology introduction as the text description. For classes with no text description matching from Wikipedia, we have built the lineage of classes by performing a bottom-up extraction that copies text descriptions from their superclass nodes. For a DO class that

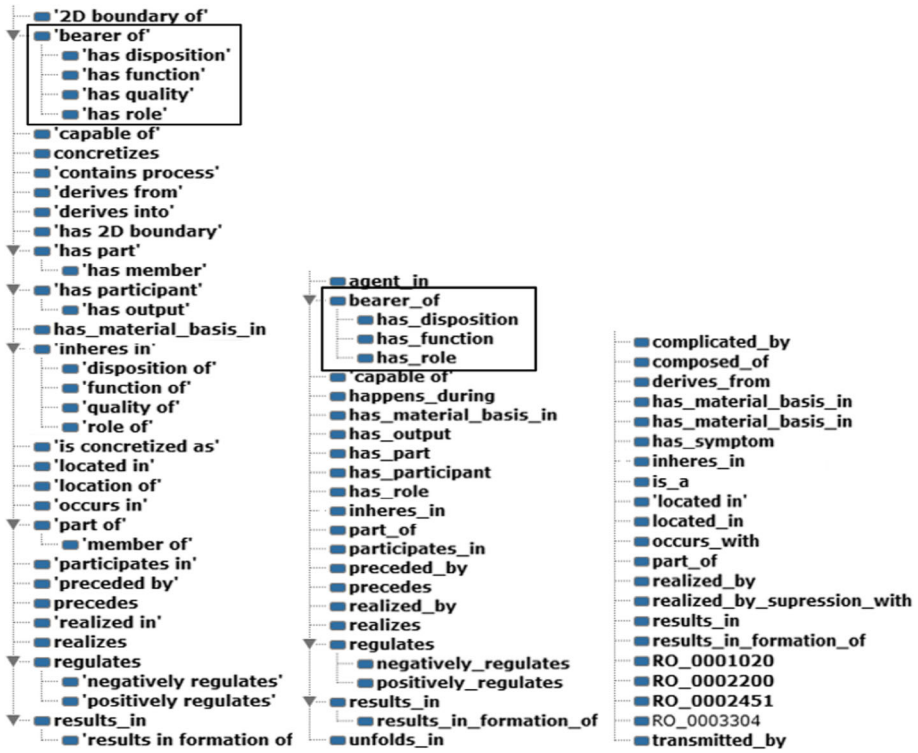

Fig. 2 Granularity observation in the IDO, IDODEN and DO ontologies. From left to right: the granularity decreases due to the increase of sparseness of relationship types in ontology schema

has neither text description nor superclass description, the label is adopted as the text description.

Impact of textual information

To verify the effectiveness of incorporating textual information into the EAPB architecture, we randomly choose one entity (dengue virus maturation) as well as its causal and inheritance relationship from IDODEN ontology for the experiment. Dengue virus maturation is a biological process about the disease.

It can be observed from Fig. 4 that, the class "dengue virus maturation" and "virus maturation" are directly connected, where the latter is a superclass of the former and their text information is thereby nearly the same. For another two indirectly connected classes, "dengue virus maturation" can cause "dengue virion". Their descriptive text is different from each other. Consistently, the information gain between word pairs (dengue virus maturation, virus maturation) and (dengue virus maturation, dengue virion) is 6.7545 and 2.2228 respectively. This is within our expectation that the information gain between the former two classes is higher than that of the latter, which proves that the information gain obtained by our model can help to identify the textual information, so as to evaluate the ontology redundancy.

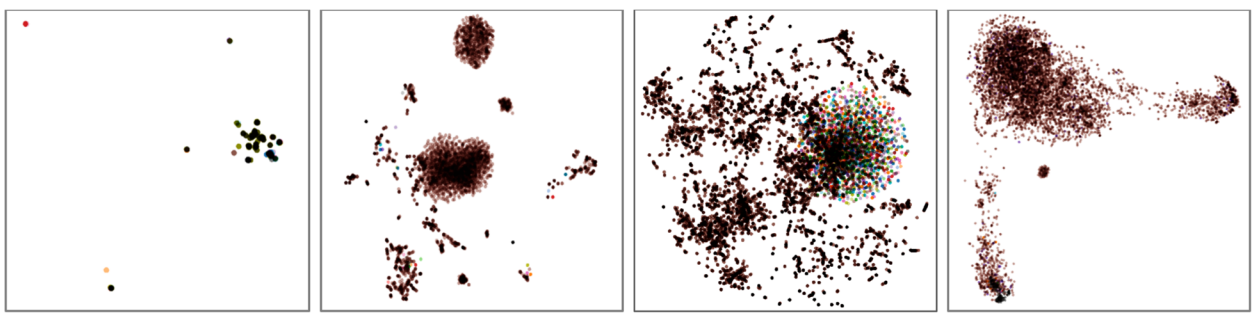

Fig. 3 Ontology visualization for the node representations of IDO, IDODEN, DO and IDDAP. From left to right: the redundancy increases due to the structure sparseness, relation incompleteness, and unclear node definition

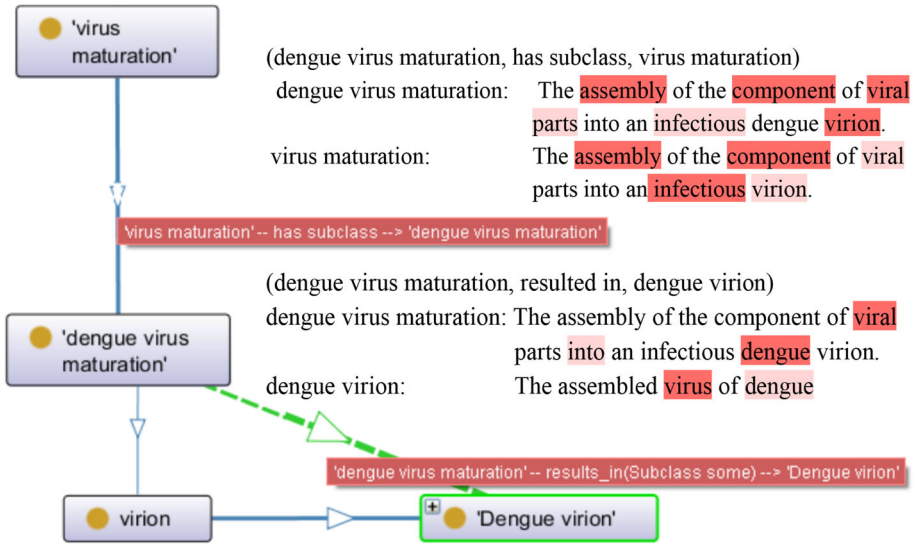

Fig. 4 An example of the structural and textual information of IDODEN ontology. The higher the EAPB calculated information gain is, the more information the attention model highlights

We visualize the attention scores predicted by EAPB in Fig. 4. The color depth indicates the importance degree of the words, the darker the more important. We can observe that, under the same ontology structure, a difference in text information contained in nodes will lead to differences in ontology redundancy; thus, the computation of entropy is useful in an ontology evaluation.

Furthermore, we employ four entire ontologies (IDO, IDODEN, DO and IDDAP) for the ablation tests of entropy evaluation in terms of discarding the textual information (w/o text). As one may expect, the consideration of textual information actually adjusts the network information density and affects the value of redundancy (see Table 2).

Performance comparison

According to the result reported in Table 2, Pearson correlation coefficient and Spearman rank correlation coefficient are adopted to evaluate the model performance. The IDO ontology, which is a well-recognized formal ontology for human infectious disease, is used as a base when calculating Pearson and Spearman rank correlation coefficients. Three other ontologies are compared with IDO from three aspects: the number of triples, the number of classes and entities, and the number of relations. The comparison results in three scores ranging from 0 to 1, obtained using the Softmax normalization method. To reduce the bias, we consider the average of these three scores as the final score. The geometric mean is employed to calculate the average, that is, the greater the difference, the smaller the mean; and the smaller the difference, the greater the mean.

Three aforementioned state-of-the-art baselines [5–7] are adopted for comparison (see Table 3). The experimental results on four ontologies are summarized in Table 3.

According to the entropy evaluation result, the Pearson and Spearman rank correlation coefficients indicate that the EAPB is strongly correlated with the ontology statistical information (the numbers of triples, classes and entities, and relations summarized in Table 1) and Ontology visualization (see Fig. 3).

Table 2 Entropy evaluation result on IDO, IDODEN, DO and IDDAP (with ablation study)

	IDO	IDODEN	DO	IDDAP
EAPB entropy evaluation	5.0554	4.4507	8.2320	10.4234
w/o text	5.7469	8.4408	9.0656	14.5873

Table 3 Performance comparison result on IDO, IDODEN, DO and IDDAP

	Calmet [6]	Doran [7]	Gurupur [5]	EAPB
Entropy evaluation - Pearson coefficient				
IDO	0.1791	0.2352	0.2488	*0.3528*
IDODEN	0.2229	0.2494	0.2862	*0.5973*
DO	0.3387	0.3620	0.3962	*0.6857*
IDDAP	0.3729	0.3936	0.4215	*0.7121*
Entropy evaluation - Spearman rank correlation				
IDO	0.1727	0.2121	0.2131	*0.2913*
IDODEN	0.2254	0.2367	0.2581	*0.5257*
DO	0.3173	0.3616	0.3844	*0.5793*
IDDAP	0.3238	0.3857	0.4023	*0.6231*

The result also shows that the IDDAP ontology scores obtained by calculating Pearson and Spearman rank correlation coefficients are higher than that of other ontologies. Although IDDAP has the largest numbers of triplets and classes/entities, it has the smallest number of relationships. Therefore, its loose structure and unevenly distributed nodes result in the highest redundancy. The IDODEN and DO ontologies have more numbers of triplets and classes/entities than IDO, leading to a certain degree of redundancy. The results of the performance comparison further validate the importance of evaluating the information redundancy by considering both structural information and textual information.

Discussion and conclusion

At the technical level, we proposed a quality metric to assess the ontology entropy with information gain processed by connectivity matrix. Ontology can be considered as a graph or a network, the entropy rate of which is a measure of the complexity of the graph. Graph connectivity is generally used in graph entropy calculations; connectivity depending directly on the single point connectivity and equal path weight is considered as the weakest measure of network connectivity. Compared with existing measures of network connectivity, we defined a meaningful measure for connectivity that considers both connectivity path of the whole network and dynamic weight between different nodes. The experimental results proved that the EAPB could effectively evaluate ontologies, especially ontologies in the medical informatics field.

At the application level, the EAPB provides possible evaluation indicators for ontology engineers. We applied this new approach to evaluate an ontology that we developed as well as several well-known infectious disease-relevant ontologies including IDO, IDODEN, DO. The knowledge representation and information assessment do not simply "look up" structure or text but rather offer a reproducible "process" to assess the ontology by considering the number of node connections, the relationship between nodes, and the textual information included in the path.

Several points require further investigation in the next phase of research. First, the SST errors will be considered as one of the ontology quality evaluation metrics. SST errors can be used to relieve the problem of scalability because convergence significantly degrades as the network size increases [25]. Second, we will attempt to unify DeepWalk, LINE, and node2vec as a matrix factorization [26]. These network embedding algorithms are considered structure-only methods. However, their unification can address the problem of the relationship between the word-context matrix and the network. Third, we will explore ontology structure evaluation through normalized graph Laplacian, which is closer to the graph theory of ontology. Moreover, besides the exploration of asserted knowledge in the ontology, we will exploit the inferred knowledge from the asserted descriptions and compare their result differences to further improve our model.

Methodology

This study proposes a metric for ontology quality that utilizes information divergence by considering both text-based and structure-based embeddings. Specifically, we first learn structure-based embeddings via a similar fashion with Node2vec, and learn the text-based embeddings via a CNN model with mutual attention [27], meanwhile, we conduct the optimization jointly to encode both path and text information into same representation space. Then we concatenate structure-based and text-based embeddings as vertices embeddings. Finally, the connectivity matrix of entropy computation is adjusted using the information gain obtained by the vertices embeddings.

Structure-based embedding

The graph structure information is encoded by maximizing the log-likelihood of all directed edges. The structure-based energy function is given by:

$$L_s(e) = \ logp(v^s|u^s) \tag{1}$$

As with Node2vec, we calculate conditional probability of vertices v generated by vertices u as:

$$p(v^s|u^s) = \frac{exp(u^s \cdot v^s)}{\sum_{k \in V} exp(u^s \cdot k^s)} \tag{2}$$

Text-based embedding

Given the word sequence of a vertex, we adopt CNN to capture the text information included in the ontology. The illustration of text-based embedding is presented in Fig. 5.

- **Input representation:** We use distributed word representation and transform sentence $S = \{w_1, w_2, \ldots, w_n\}$ into corresponding word embedding sequence $\mathbf{W} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_n}\}$ as input of CNN, where $\mathbf{w_i} \in \mathbb{R}^d$, d is dimension of the word embeddings.
- **Convolution:** For connected edge $e_{u,v}$ with vertices u and v, we perform convolution operation over a sliding window to extract local features of their textual embeddings $\mathbf{W_u}$ and $\mathbf{W_v}$ where $\mathbf{W_u} \in \mathbb{R}^{d * m}$, $\mathbf{W_v} \in \mathbb{R}^{d * n}$, m and n represents the lengths of $\mathbf{W_u}$ and $\mathbf{W_v}$ respectively.

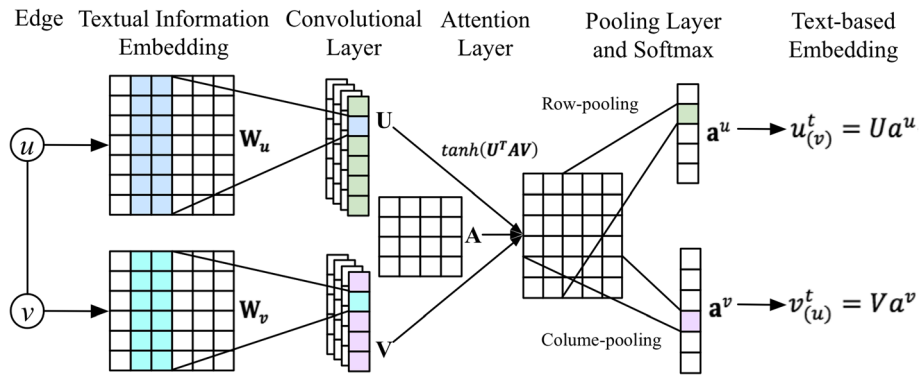

Fig. 5 An illustration of text-based embedding

- **Attentive Pooling:** To encode the vertices interactive information facing different neighbors, we apply mutual attention into the pooling layer. To be specific, based on the output of convolution layer **U** and **V** respectively for vertice u and v, we introduce an attentive matrix $A \in \mathbb{R}^{d*d}$ and calculate the correlation matrix $C \in \mathbb{R}^{m*n}$, which represents the pair-wise correlation score between U and V, as follow:

$$C = \tanh\left(\boldsymbol{U}^T \boldsymbol{A} \boldsymbol{V}\right) \tag{3}$$

Intuitively, the attentive matrix is used to assign different weights according to the specific role each vertex plays when interacting with other vertices. Then we conduct mean pooling operations along the rows and columns of C to generate the row-pooling and column-pooling respectively:

$$\begin{aligned} h_i^u &= mean\left(C_{i,1}, \ldots, C_{i,n}\right) \\ h_i^v &= mean\left(C_{1,i}, \ldots, C_{m,i}\right) \end{aligned} \tag{4}$$

where h_i^u and h_i^v indicate importance score for word i when interacting with vertice v and u respectively. Next we obtain attention vectors $\mathbf{a}^u$ and $\mathbf{a}^v$ from $\mathbf{h}^u = [h_1^u, \ldots, h_m^u]\hat{}T$ and $\mathbf{h}^v = [h_1^v, \ldots, h_n^v]\hat{}T$ by employing softmax function.

Finally, the text-based embeddings of u and v are calculated as:

$$\begin{aligned} u_{(v)}^t &= Ua^u \\ v_{(u)}^t &= Va^v \end{aligned} \tag{5}$$

Optimization

We learn the text-based and structure-based representations by maximizing their energy function jointly as:

$$\mathcal{L} = \sum_{e \in E} L_t(e) + L_s(e) \tag{6}$$

where E indicates all edges of learned ontology, $L_s(e)$ is structure-based energy function in Eq. (1). As for text-based energy function $L_t(e)$, we aim to map two types of vertex embeddings into the same representation space and define it as:

$$L_t(e) = \alpha \cdot L_{tt}(e) + \beta \cdot L_{ts}(e) + \gamma \cdot L_{st}(e), \tag{7}$$

$$\begin{aligned} L_{tt}(e) &= \log p(v^t|u^t), \\ L_{ts}(e) &= \log p(v^t|u^s), \\ L_t L_{st}(e) &= \log p(v^s|u^t) \end{aligned} \tag{8}$$

where α, β and γ are used to control the weights of $L_{ts}(e)$, $L_{st}(e)$ and $L_{tt}(e)$, the former two indicates mutual generation based on text and structure, meanwhile, we expect them can contain their own characteristics by $L_{tt}(e)$. All of them are adopted similar softmax computation as structure-based as Eq. (1).

Weighted connectivity matrix

Entropy can be calculated via the network probability distribution function. The connectivity matrix $Cn*n$ is dedicated to calculating node probabilities, where n stands for the number of nodes in the ontology. The connectivity matrix includes the connection information between u and v, where 0 demonstrates no connection between u and v and 1 indicates a connection. The probabilities P_k of each concept k can be processed by:

$$P_k = \frac{\sum_{u=1}^{n} C_{uk}}{\sum_{v=1}^{n} \sum_{u=1}^{n} C_{uv}} \tag{9}$$

Given the structure-based and text-based embedding concatenation results, the relevancy between nodes u and v can be evaluated using a cosine similarity measure. When nodes u and v are not directly connected, the selected path is their shortest path. To avoid multiple

calculations of the same path's weight, the relevancy is divided by the shortest path, the result of which is considered to be the information gain and represented as O_{uv}. We multiply O_{uv} by the connectivity matrix C_{uv} of the entropy computation. As a result, the probabilities P_k of each concept k can be given by:

$$P'_k = \frac{\sum_{k \neq u} \sum_{u=1}^{n} C_{uk} * O_{uk}}{\sum_{v=1}^{n} \sum_{u=1}^{n} C_{uv} * O_{uv}} \quad (11)$$

The path-based text-aware calculation formula for determining entropy using diverse node weights can be presented as:

$$S' = -\sum_{k=1}^{n} P'_k \log_2 P'_k \quad (12)$$

Endnotes

[1]http://infectiousdiseaseontology.org/page/Main_Page

[2]https://code.google.com/archive/p/dengue-fever-ontology/

[3]doi: https://doi.org/10.1093/nar/gku1011

[4]www.obofoundry.org/ontology/ncbitaxon.html

[5]human-phenotype-ontology.github.io/

[6]https://www.drugbank.ca/

[7]http://www.hopkinsmedicine.org/amp/guidelines/antibiotic_guidelines.pdf

[8]https://www.nlm.nih.gov/

Abbreviations

DO: Disease ontology; EAPB: Entropy-Aware Path-Based metric for ontology quality; IDDAP: Ontology-driven clinical decision support system for infectious disease diagnosis and antibiotic prescription; IDO: Infectious disease ontology; IDODEN: Infectious disease ontology for dengue; NLM: The United States national library of medicine

Funding

This work was financially supported by the National Natural Science Foundation of China (No.61602013), and the Shenzhen Fundamental Research Projects (Grant No. JCYJ20150301543307 11 (Key Project) and JCYJ20170818091546869).

Availability of data and materials

We will release the source code and IDDAP ontology of this work after publication (https://github.com/AnonymousResearcher1/ontologyEvaluate). Other datasets generated and/or analyzed during the current study are available in the Disease Ontology (DO) repository (http://www.obofoundry.org/ontology/doid.html) [19], Infectious Disease Ontology (IDO) repository (http://purl.obolibrary.org/obo/ido.owl) [20], and Dengue ontology (IDODEN) repository (http://purl.obolibrary.org/obo/ido.owl) [21].

Authors' contributions

YS carried out the application of mathematical techniques and the assessment of system operation. DC realized the development methodology and the creation of models. BT conducted an investigation process, implemented algorithms and programming. MY analyzed and counted ontology information. KL was responsible for the management and coordination responsibility for the research activity planning and execution. All authors read and approved the final manuscript.

Authors' information

Ying Shen is now an Assistant Research Professor in School of Electronics and Computer Engineering (SECE) at Peking University. She received her Ph.D. degree from the University of Paris Ouest Nanterre La Défense (France), specialized in Medical & Biomedical Information Science. She received her Erasmus Mundus Master degree in Natural Language Processing from the University of Franche-Comté (France) and University of Wolverhampton (England). Her research interest is mainly focused in the area of Medical Informatics, Natural Language Processing and Machine Learning.

Daoyuan Chen received the BS degree in computer science from University of Electronic Science and Technology of China, in 2016. He is working toward the MS degree in computer science at Peking University. His research interest is mainly focused in the area of deep learning and knowledge graph.

Buzhou Tang is now an Associate Professor in School of Computer Science and Technology at Harbin Institute of Technology. He received his Ph.D. degree and master degree from the Harbin Institute of Technology (China), specialized in Natural Language Processing. He received his bachelor degree in Computer Science from the Jilin University (China). His research fields include Artificial Intelligence, Machine Learning, Data Mining, Natural Language Processing and Biomedical Informatics.

Min Yang is currently an Assistant Research Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. She received her Ph.D. degree from the University of Hong Kong in February 2017. Prior to that, she received her B.S. degree from Sichuan University in 2012. Her current research interests include machine learning, deep learning and natural language processing.

Kai Lei received the Ph.D. in C.S. from Peking University, China, in 2015, M.Sc in C.S. from Columbia University in 1999 and B.Sc in C.S. from Peking University in 1998. He had worked for companies including IBM T.J Waston Research Center, Citigroup, Oracle, Google from 1999 to 2004. He currently is an associate professor in the School of Electonic and Computer Engineering (SECE), Peking University, Shenzhen, and participates in the CENI project supported by National Development and Reform Commission since 2016. His research interests include, knowledge graph, big data technologies and named data networking.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Shenzhen Key Lab for Information Centric Networking & Blockchain Technology (ICNLAB), School of Electronics and Computer Engineering, Peking University Shenzhen Graduate School, 518055 Shenzhen, People's Republic of China. [2]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), 518055 Shenzhen, People's Republic of China. [3]SIAT, Chinese Academy of Sciences, 518055 Shenzhen, People's Republic of China.

Received: 17 April 2018 Accepted: 30 July 2018
Published online: 10 August 2018

References

1. Gruber T, Liu L, Özsu MT. Encyclopedia of database systems: New York: Springer; 2009.
2. Lee CS, Kao YF, Kuo YH, Wang MH. Automated ontology construction for unstructured text documents. Data Knowl Eng. 2007;60(3):547–66.
3. Hempelmann CF, Sakoglu U, Gurupur VP, Jampana S. An entropy-based evaluation method for knowledge bases of medical information systems. Expert Syst Appl. 2016;46:262–73.
4. Valdez AC, Dehmer M, Holzinger A. Application of graph entropy for knowledge discovery and data mining in bibliometric data. Math Foundations Appl Graph Entropy. 2016;6:174.
5. Gurupur VP, Sakoglu U, Jain GP, Tanik UJ. Semantic requirements sharing approach to develop software systems using concept maps and

information entropy: a personal health information system example[J]. Adv Eng Softw. 2014;70:25–35.
6. Calmet J, Anusch D. From entropy to ontology. Na, 2004.
7. Doran P, Tamma V, Palmisano I, Payne TR, Iannone L. Evaluating ontology modules using an entropy inspired metric. In: Proceedings of 2008 Web intelligence and intelligent agent technology, vol. 1: Washington, DC: IEEE Computer Society. 2008;918–22.
8. Shen Y, Yuan K, Chen D, Colloc J, Yang M, Li Y, Lei K. An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. Artif Intell Med. 2018;86:20–32.
9. Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S. Quality assessment for linked data: a survey. Semantic Web. 2016;7(1):63–93.
10. Färber M, Bartscherer F, Menne C, Rettinger A. Linked data quality of DBpedia, freebase, OpenCyc, Wikidata, and YAGO. Semantic Web. 2016:1–53.
11. Gangemi A, Catenacci C, Ciaramita M, Lehmann J. Modelling ontology evaluation and validation. In Proceedings of European Semantic Web Conference. Berlin: Springer. 2006;140–54.
12. Duque-Ramos A, Fernández-Breis J, Stevens R, Aussenac-Gilles N. OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. J Res Pract Inf Technol. 2011;43(2):159.
13. Tsatsaronis G, Macari N, Torge S, Dietze H, Schroeder M. A maximum-entropy approach for accurate document annotation in the biomedical domain. J Biomed Semantics. BioMed Central. 2012;3(1):2.
14. Yang C, Liu Z, Zhao D, Sun M, Chang EY. Network representation learning with rich text information. In: IJCAI; 2015. p. 2111–7.
15. Zhang Y, Tao C, Jiang G, Nair AA, Su J, Chute CG, Liu H. Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. J Biomed Semantics. 2014;5(1):33.
16. Tu C, Zhang W, Liu Z, Sun M. Max-margin DeepWalk: discriminative learning of network representation. In: IJCAI; 2016. p. 3889–95.
17. Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, Paterson DL. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. Clin Microbiol Infect. 2012;18(3):268–81.
18. Yamagata Y, Kozaki K, Imai T, Ohe K, Mizoguchi R. An ontological modeling approach for abnormal states and its application in the medical domain. J Biomed Semantics. 2014;5(1):23.
19. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Kibbe WA. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2011;40(D1):940–6. https://doi.org/10.1093/nar/gkr972
20. Cowell LG, Smith B. Infectious disease ontology. In: Infectious disease informatics: New York: Springer; 2010. p. 373–95.
21. Mitraka E, Topalis P, Dialynas E, Dritsou V, Louis C. IDODEN: an ontology for dengue. In: ICBO; 2012. https://doi.org/10.1371/journal.pntd.0003479.
22. Topalis P, Mitraka E, Dritsou V, Dialynas E, Louis C. IDOMAL: the malaria ontology revisited. J Biomed Semantics. 2013;4(1):16.
23. Plummer MD. Some covering concepts in graphs. J Comb Theory. 1970;8(1):91–8.
24. Sayood K. Introduction to data compression: Massachusetts: Morgan Kaufmann; 2017.
25. Kitano H. Designing neural networks using genetic algorithms with graph generation system. Complex Syst. 1990;4(4):461–76.
26. Chen Y, Sun P, Fu X, Xu T. Improving prediction accuracy of matrix factorization based network coordinate systems. In: Proceedings of 19th International Conference on Computer Communications and Networks (ICCCN); 2010. p. 1–8.
27. dos Santos CN, Tan M, Xiang B, Zhou B. Attentive pooling networks. In CoRR 2016: 2(3), 4.