Semi-automatic extraction and validation of concepts in ontology learning from texts in Spanish

Manuela Gómez-Suta
madegomez@utp.edu.co
Faculty of Business
Universidad Tecnológica de Pereira

Julián D. Echeverry-Correa
jde@utp.edu.co
Faculty of Engineering
Universidad Tecnológica de Pereira

José A. Soto-Mejía
jomejia@utp.edu.co
Faculty of Business
Universidad Tecnológica de Pereira

ABSTRACT

The construction of ontologies from texts in Spanish is a challenge since this language lacks conceptual databases to validate abstract ontology structures as concepts and relations between them. The preceding generates the necessity of using manual evaluation by human experts; carrying high expenses that limit the calibration of algorithm parameters and large-scale evaluations. This document presents a proposal to evaluate abstract ontology structures through the task of semantic clustering of documents, without the expensive necessity of using manual evaluation or conceptual databases. The proposal is not only affordable but also applicable to model data and domains that lack structured knowledge resources. The experiments lead to the extraction and validation of the ontology structures from texts in Spanish regarding the domain of the Colombian armed conflict.

CCS CONCEPTS

• **Information systems → Ontologies**; **Information extraction**; **Language models**.

KEYWORDS

Ontology learning, Spanish, evaluation, concepts.

ACM Reference Format:

Manuela Gómez-Suta, Julián D. Echeverry-Correa, and José A. Soto-Mejía. 2020. Semi-automatic extraction and validation of concepts in ontology learning from texts in Spanish. In *The 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020), June 30-July 3, 2020, Biarritz, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3405962.3405977

1 INTRODUCTION

The processes like automatic reasoning on the semantic web, data extraction on retrieval systems, and natural language tasks use ontologies [10]. It occurs since an ontology represents shares and reuses knowledge of a specific domain [31]. It is an instrument to capture semantic information through concepts and relations between them to provide a valid and meaningful representation of knowledge.

WIMS 2020, June 30-July 3, 2020, Biarritz, France

ACM ISBN 978-1-4503-7542-9/20/06.
https://doi.org/10.1145/3405962.3405977

Nonetheless, the ontology construction implies challenges on the acquisition and the updating of knowledge since these processes are usually manually executed, and it makes them more prone to have errors that demand qualified resources and time constraints [20]. The ontology learning field overcomes these challenges since it allows to build, extend, or adapt ontologies in a semi-automatic way through collecting and structuring knowledge on texts available on the web [31].

It is estimated that 66.1% of the 520,777,464 Spanish-speakers of the world are internet users, and this number has been increasing in the last ten years[1]. Consequently, the "computerization of the internet domains to Spanish is an undeniable truth" [30, p. 2058]. The earlier underscores the necessity of applying ontology learning, considering the particularities of the Spanish language [36].

Spanish document modeling has been developed until now by the adaptation of techniques designed for the treatment of English texts, which has traditionally been the focus of ontology learning studies. The adaptability of such techniques is relevant for two main reasons: 1) the linguistic differences between both languages; as in comparison with English, Spanish has a vast verbal inflection system, besides the number and gender markers, 2) the contribution to quantifying the different ontology learning techniques for the treatment of the Spanish documents.

It is a demanding task to adapt techniques initially designed to model texts in English because there is no record of any structured knowledge resource nor available training datasets in Spanish that facilitate the creation of abstract structures as concepts and relations between them [11, 14]. The preceding is the main reason by which the evaluation of techniques for the extraction of abstract ontology structures is commonly done by human experts or implementing knowledge sources like LAR-WordNet, YAGO3, and Wikicorpus in Spanish.

The manual evaluation suggests high costs since access to a significant number of experts is required to decrease the bias of the evaluation carried out by humans [11]. Furthermore, every variation on the algorithms involves the same high expenses of the first one, making parameter calibration and large-scale evaluations unfeasible [36].

On the other hand, the validation using general knowledge resources can produce inaccurate conclusions due to impossibility to make an exact comparison between extracted abstract ontology structures and the ones in the referential source [16]. It arises as a result of the difference in analyzed domains and the granularity levels between the vocabulary corpus and the labels of the conceptual database.

[1] Data extracted on 11/02/2020 from https://www.internetworldstats.com/stats7.htm

Evaluating ontology learning techniques using Spanish texts is more difficult when analyzing specific domains as the Colombian armed conflict, which was treat in this work. The importance and complexity of this domain are widely known, but it lacks computational efforts to generate structured knowledge resources. Currently, the thesaurus made by the National Center for History Memory of Colombia [13] is the only instrument that captures semantic information of this domain. Nonetheless, this thesaurus does not have formal language, and as previously stated, using it to extract and validate abstract ontology structures might produce an inaccurate evaluation.

This paper presents the semi-automatic transformation of Spanish texts to ontology structures as terms, concepts, and relations between them. The main contribution of this paper is the proposal to validate abstract ontology structures without manual evaluation or general knowledge resources. Consequently, the proposal is affordable in terms of time and qualified resources. Furthermore, it applies to data and domains that lack structured knowledge resources. The presented experiments involve Spanish texts regarding the domain of the Colombian armed conflict.

The rest of this paper is organized as follows: section 2 characterizes the ontology learning studies from texts in Spanish, section 3 describes the adapted techniques to extract and validate ontology structures from documents in Spanish, section 4 states the experimental framework and results, and section 5 indicates the conclusions and future works.

2 RELATED WORK

Ontology learning presents the construction of structures as terms, concepts, and relations to build an ontology. The terms are textual representations of concepts; in other words, they are lexical units that are explicitly shown on the corpus. A concept is a mental symbol in which representation and interpretation arrive in the form of a group of related terms [36]. Therefore, the knowledge contained in a concept arises from the context provided by their clustered lexical units. The relations could be classified into taxonomic and non-taxonomic, and allow modeling the interactions among the concepts.

Different techniques are applied for the construction of these ontology structures. Clark et al. [10] expose these techniques under three paradigms. The first, based on data, employs statistical techniques to extract terms and concepts. This paradigm uses co-occurrence analysis to build taxonomic relations. The second paradigm is knowledge-driven, as a result, it utilizes training datasets and structured knowledge resources to obtain semantic information of the concepts and non-taxonomic relations. Also, the second paradigm integrates human expert knowledge to acquire higher precision in the learned structures. The third paradigm is a mixture of the first and second paradigm.

The first paradigm, on ontology learning from texts in Spanish, applies the *bag-of-words* hypothesis. In this way, the paradigm weights the terms in the documents and filters out those considered irrelevant when comparing their presence against a *gold standard* usually generated by humans. The hierarchical clustering establishes term groups identified as concepts [17]. Furthermore, logistic regression allows determining which terms are significantly similar to the concepts of a reference ontology [14]. The above mentioned statistical techniques do not extract concepts and relations between them simultaneously. For this reason, the relations emerge from the co-occurrence analysis.

Term validation is an easy and executable task using manually constructed *gold standards*. This situation may be the reason by which some researchers [16, 30] have considered that the extracted terms are equivalent to concepts, even when it is inconsistent with the definition of concepts as a group of terms. Besides, the lack of annotated corpus with ontology knowledge (concepts and relations between them) and the high costs that implicate the construction of this resource might be the causes to use terms as concepts.

[14] had evaluated the similarity between the ontology structures extracted and the ones present in a general ontology to validate concepts. The assessment outcomes depend on the content collection obtained from the general knowledge resource. Accordingly, this evaluation weakness is its dependence on the comparative measures (such as *precision* and *recall*) that are unable to detect phenomena like synonymy and polysemy [34] that affect the granularity levels of the vocabulary extracted and the labels of the ontology. Consequently, this evaluation does not set an exact comparison between the ontology and the analyzed texts [16, 17].

In this sense, the first paradigm encompasses techniques that are not costly and applicable to data from different languages and domains. Nevertheless, these techniques do not regard the nature of concepts and types of relations. As a consequence, it will produce light ontologies that are not interpretable by machines. Therefore, this reduces the usefulness of the learning models in contexts such as the semantic web. The ontology evaluation with the *gold standard* is feasible and straightforward to assess extracted terms; nonetheless, the training datasets with ontology knowledge are expensive to build, and they are not available for specific domains and languages. Additionally, this evaluation employs metrics that do not quantify the granularity levels between the vocabulary of the texts and the general knowledge resources.

In the second paradigm, the concept extraction arises using structured knowledge resources [2], even when the recovered conceptual structures may be absent in the analyzed texts. Additionally, it is usual to employ conceptual databases and multilingual resources, which show the syntactic features that describe a concept [4]. On the other hand, some studies [1] construct relations with vocabulary patterns and external resources that link the language forms to the cognitive patterns of taxonomic and non-taxonomic relations [30]. The ontology evaluation usually arises with expert judgment [1].

By the second paradigm, the extracted structures have details about the types of relations; it results in ontologies that can be formalized and used in automatic reasoning tasks. Nonetheless, the associated techniques are not simple to apply in domains or languages that lack of conceptual databases, for example, Ochoa et al. [30] report the use of ADESSE, which is a database of verbs for the Spanish language that offers semantic information; nevertheless, the authors do not make this resource available. The utilization of lexical patterns and static knowledge resources restrict the findings to the options previously learned; accordingly, if it has not been considered a form, it may lose information presented in the texts. Due to the weakness mentioned, this paradigm emphasizes the

Table 1: Ontology learning studies from texts in Spanish

Study	Initial requirement	Learning approach	Degree of automation	Ontology structures built
[17]	GO	S	A	C
[14]	GO	S	A	C
[30]	CD - GS	S - P	SA	T - TR - NR
[16]	GO - HE	P	SA	T - TR - NR
[2]	HE	P	SA	T - C - TR
[1]	HE	P	SA	C - TR - NR
[4]	HE	P	SA	C - TR - NR
Our proposal	GS	S	SA	T - C - TR

manual evaluation by human experts, which makes the process expensive, and it does not allow the scalability of the techniques and results [36].

The third paradigm is a conjunction between the techniques of the first and second paradigm. Some proposals suggest the concepts extraction through knowledge resources and the relations formation as a result of the co-occurrence analysis. Other studies recognize concepts as terms recovered by their frequency of occurrence, and the lexical patterns facilitate building relations [16]. Some works [17] recommend homogenizing the representation of the documents by replacing the terms with the corresponding label of WordNet into Spanish to cluster hierarchical the tags, consequently forming concepts. According to Alemán et al. [2], human experts can recognize concepts, and then, semi-automatically identify similar terms regarding the syntax and semantics of the recovered conceptual structures. The drawbacks in ontology evaluation of the first two paradigms are presented in the third one since it is recurrent the manual evaluation and the comparison with general ontologies [16].

The previous paragraphs allow to point out that the techniques of the first paradigm are proper to model texts for languages and domains that lack linguistic resources. These techniques do not employ external resources to recognize the nature of the concepts and types of relations; in consequence, this produces light ontologies. Nevertheless, the research and results of the field have generated a growing awareness of the complexities of modeling the knowledge in ontology structures. This situation has raised the question about the feasibility of automatically building a formal ontology or, on the contrary, the need to address more pragmatic objectives by focusing on the automatic construction and extension of light ontologies [36].

This paper supports the practical approach of the ontology learning task because it sought to construct an ontology minimizing the human effort and giving scalability to the suggested techniques. Therefore, it does not require to utilize general resources or expert knowledge. This study employs the techniques of the first paradigm to build an ontology from texts in Spanish regarding the domain of the Colombian armed conflict. To prevent the evaluation drawbacks of the paradigm used, this paper proposes a task-based evaluation that does not require *gold standard* or conceptual databases. In this sense, the authors do not assume that the terms are equivalent to concepts, and do not validate the ontology structures through an inaccurate comparison with the tags of a general ontology. The main contribution of this research is the extraction and validation of abstract ontology structures from texts, without the obligation to utilize general knowledge resources or manual evaluation by human experts.

Table 1 comparatively presents studies associated with the three paradigms, and our proposal. The categorization criteria is an adaptation of [31]. We employ it because this classification has been utilized previously (in [2, 30]) to examine ontology learning studies. The comparative framework analyzes:

(1) The requirement of general ontology (GO), *gold standard* (GS), conceptual database (CD) or knowledge from human experts (HE).
(2) The learning approach to extract knowledge could be statistical (S) or pattern (P) based.
(3) The degree of automation according to the construction and validation of ontology structures. For example, if the study reports using *gold standard* then the process is semi-automatic (SA), otherwise it is automatic (A).
(4) The ontology structures such as terms (T), concepts (C), taxonomic relations (TR) or non-taxonomic relations (NR) built that imply the formalization reached.

Consequently, this paper is remarkably different from previous ones because it proposes the learning and validation of abstract ontology structures without requiring conceptual references or expert judgment. In this sense, the methodology could be scalable to other domains and languages. The proposal is semi-automatic since the term validation task employed human-made resources, nevertheless, these data are reusable in later assessments.

3 METHODOLOGY

Figure 1 summarizes the two stages for ontology learning. The first one is the vocabulary building through the preprocessing of texts, the index-terms selection based on statistical weighting, and the evaluation based on *gold standard*. The second one is the extraction and validation of concepts and relations supported by the proposed semantic clustering task.

3.1 Vocabulary building

This paper considered semantic related words and collocations. The first one came from the thesaurus of the National Center for History Memory of Colombia [13]. Named entity recognition facilitated the collocations identification, meaning that the authors assessed the

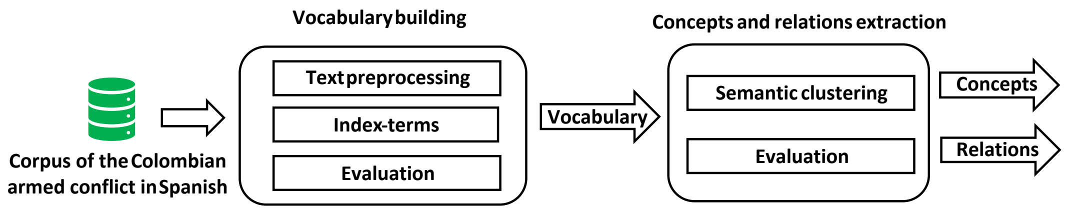

Figure 1: Methodology

significance of the entities composed for more than one term, using a log-likelihood test.

This work automatically identified the *part-of-speech* (POS) tags of the terms and filtered out the words whose syntactic information was different from adjectives, nouns, and verbs. These tags were selected because it is usual to associate them with informative and useful terms that form concepts [30]. Then, the texts suffered the lemmatization process and normalization of the lowercase. Additionally, the authors removed stopwords, punctuation, numeric characters, and special markers. Terms with semantic related words were standardized, leaving the more informative description; for example, the term *International Human Rights Law (Derecho Internacional Humanitario)* replaced *IHRL (DIH)*. Finally, this study contemplated unigrams and collocations for the tokenization process.

This research implemented and compared the index-terms selection based on four statistical weighting proposals to build a vocabulary of relevant terms that described the corpus. The chosen schemes come from ontology learning studies [16, 30]. Below, we briefly introduced each of them:

TF-IDF (Term frequency-inverse document frequency): It quantifies the relative frequency of the terms in a document compared to the inverse proportion of that word in the corpus. This paper employed the TF normalized by the length of each document.

TF – Entropy: It expresses the relative frequency of the terms in a document and the number of times the term appears in each analyzed text. This study approximated to zero the logarithm of the term frequency in the text and the corpus, to avoid math uncertainties when a word did not appear in a document.

Ochoa et al. [30] work: The study calculates TF to all terms, and it differences the global weighting schemes. For unigrams, there is a quantification of IDF, and for collocations exist measurement of the NC-value and C-Value.

The NC-Value finds the adjectives, verbs, and nouns that compose the neighborhood of a candidate term, considering the 10 word history. It uses this syntactic information to calculate a weighting factor that assigns a high value to collocations surrounded by words with the tags of interest. The C-Value (Eq. 1) where a is the candidate term, $|a|$ is the length of the candidate term, *f(a)* is the frequency of the candidate term in the corpus, T_a is the set of terms that contain a, $P(T_a)$ is the frequency in the corpus of the longest word that contains the candidate term, and $\sum_{b \in T_a}$ is the frequency of occurrence of the candidate term as a sub-term of any candidate term b. C-Value assigns a high value to terms that appear frequently in the corpus and sparsely inside of other terms.

$$C - Value = \begin{cases} \log_2 |a| * f(a) & \text{if } T_a \in \emptyset \\ \log_2 |a| * \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & \text{if } T_a \notin \emptyset \end{cases} \quad (1)$$

Modification Ochoa et al. [30] work: The authors of this paper proposed to modify [30] work. The alteration is quantifying the unigrams through the entropy scheme. This scheme seeks to obtain an accurate vocabulary.

For the selection of the index-terms, this paper utilized the evaluation based on a *gold standard* for two reasons. On the one hand, the results are reproducible and comparable examining the same corpus [20]. On the other hand, this assessment regards metrics like *precision*, *recall*, and *F-measure* that characterize the functionality of the learned ontology through the terms.

The drawback of employing the *gold standard* was the cost associated with its development. Nonetheless, the investment only came about once because later assessments were fully automated [11]. For the building of the reference list, this study regarded relevant terms that explicitly appeared in the corpus, but not the domain concepts that might be inferable of the texts. For example, the *gold standard* has the terms *crime gang (banda crimen)*, *organized crime gang (banda crimen organizado)*, and *criminal gang (banda criminal)* even when the concept *organized crime (crimen organizado)*[2] can represent them given their syntactic similarity in Spanish language.

3.2 Concepts and relations extraction

This research is away from studies like [16, 30] that utilize the statistical weighting to extract concepts. Besides, it captures taxonomic relations simultaneously to the concept construction. The relations are of subsumption since they emerge from the co-occurrence between the terms.

This paper proposes the validation of concepts and relations through a task-based evaluation. Researches like [6] point out that this evaluation is not ideal for comparing techniques of ontology structures extraction as the task and the algorithm used may influence the results. On the other hand, Dellschaft and Staab [11] presented two standards for filtering off the task intervention to obtain valid conclusions about the compared techniques. The standards establish an assessment that: 1) must allow the techniques characterization; through the utilization of different measurement that will enable to weigh the strengths and weaknesses of the procedures, and 2) must ensure low-costs

[2]The concept is in the thesaurus of the National Center for History Memory of Colombia

for frequent and large-scale evaluation of different experimental settings.

This research proposes the semantic clustering of documents that means crowding the texts in topics or concepts that are extracted from the corpus [3]. This type of clustering has appeared to respond to two shortcomings of the ordinary clustering algorithms supported in the vector space model. The first is the construction of clusters with low levels of accuracy caused by the high dimensionality of the data [27]. The second is to regard the term frequency as the only classification feature, ignoring the inherent semantic relations that link documents to each other [3]. In this sense, the semantic clustering aims to group the texts in topics or concepts by considering the relations implied in the corpus.

The proposal task accomplishes the two standards indicated by [11] since this study utilized metrics to quantify the results of automatic clustering accordingly to the density, overlap, and similarity with the reference data besides, it automatically evaluates the coherence of the identified concepts. Therefore, the proposal facilitates frequent and large-scale evaluation.

To the best of our knowledge, this article is the first to proposed the semantic clustering of documents as a task to evaluate abstract ontology structures. We briefly present the task proposal and the metrics used during the evaluation.

3.2.1 Extraction of ontology structures by semantic clustering of documents.

There are two approaches to execute the semantic clustering of documents. The generative models, being the most used the Latent Dirichlet Allocation model (LDA) [7], and algorithms for community detection [15] with special emphasis on modularity-based methods. This paper utilizes both approaches.

LDA model is widely used for information retrieval and has been classified as a technique to automatic clustering as it finds topics of a corpus and mapped distributions of these to each document, additionally, the model determines term distributions on the topics [7]. Each topic is a concept because it is a term group with a high probability of belonging to this. Besides, the texts associated with each topic are representations of this concept [8]. The LDA model builds concepts using only the corpus. Consequently, this model is from the first ontology learning paradigm explained in section 2.

This study utilized *Topic Coherence* (TC) for the LDA calibration. It is a measure for the semantic interpretability of each discovered topic [21]. Therefore, the LDA model builds a semantic and interpreted decomposition of the clustered texts [8]. This paper employed the metric of [29] and presented in Eq. 2.

$$TC\left(t; V^{(t)}\right) = \frac{2}{N(N-1)} \sum_{m=2}^{N} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)} \quad (2)$$

Where $V^{(t)} = (v_1^{(t)}, ..., v_N^{(t)})$ is the list of the N top terms, this means the most probable terms within the topic *t*. *D(v)* is the frequency of term v in the corpus, and $D(v, v')$ is the number of texts that contain both v and v'.

The community detection algorithms utilize the co-occurrence matrix to establish a word associated network. These algorithms cluster nodes (terms) if they share common properties or have similar roles within the graph [15]. In this way, concepts that describe the content of the corpus are established, then, the documents semantically associated are clustered in each established community or concept [27].

This paper employed a directed word association network to exploit the semantic information that arises from the ordered appearance of terms. Figure 2 presents an example of the network where the weight of the arc between the term t_i andt_j, being $i \neq j$, exists if the co-occurrence between words was significant under a log-likelihood test with 95% of confidence.

This paper employed the Directed Louvain algorithm [12], which displays a hierarchical structure of communities using the notion of directed modularity from Leicht and Newman [23]. The use of Directed Louvain in asymmetric networks has shown to build accurate communities as it discusses the imbalance in the degrees of input and output, which is a feature of the word association networks. Nonetheless, the algorithm has a shortcoming since the results may vary depending analysis order of nodes during community detection. In consequence, this research employed *consensus clustering*; in specific, the proposal of [22] since it exhibits a consensus matrix where d_{ij} indicates the number of partitions in which vertex i and j are assigned to the same cluster. The consensus matrix only encompasses the significant values through a log-likelihood test with 95% confidence.

After community detection, the followed step was the semantic clustering of documents in each community. This study employed the *Normalized Pointwise Mutual Information* (NPMI) to measure the association degree between documents and concepts. The strategy used is similar to that reported in [27]. Eq. 3 presents the NPMI between the document d_k and the concept C_l.

$$NPMI\left(d_k, C_l\right) = \frac{\log\left(p\left(d_k\right) p\left(C_l\right)\right)}{\log\left(p\left(d_k, C_l\right)\right)} \quad (3)$$

Where $p(d_k, C_l)$ is the probability of the document d_k, and the concept C_l. It comes from the cosine similarity as Eq. 4 presents it.

$$p\left(d_k, C_l\right) = \frac{\sum_{i=1}^{n} w\left(i, d_k\right) w\left(i, C_l\right)}{\sqrt{\sum_{i=1}^{n} w^2\left(i, d_k\right) \sum_{i=1}^{n} w^2\left(i, C_l\right)}} \quad (4)$$

Where $w(i, d_k)$ is the weight of term i in document d_k by using the information of the *term–document matrix* with TF weighting. $w(i, C_l)$ is the weight of the term i in concept C_l. It proceeds from the *Weighted LeaderRank* (WLR) algorithm [28]. It is based on random walking to establish the importance of the node (term) in the network (concept), regarding the number and weight of its links. WLR is very similar to the PageRank algorithm; nevertheless, WLR is adaptive and non-parametric. During each iteration, the WLR algorithm establishes the probability that the random walker visits the next node regarding the vertices previously examined. Therefore, WLR outperforms PageRank in the robustness to manage noisy data leading to higher accuracy in weighing the importance of the nodes [28]. This research utilized the WLR algorithm presented in [37].

In Eq. 5, $p(d_k)$ is the probability of the document d_k.

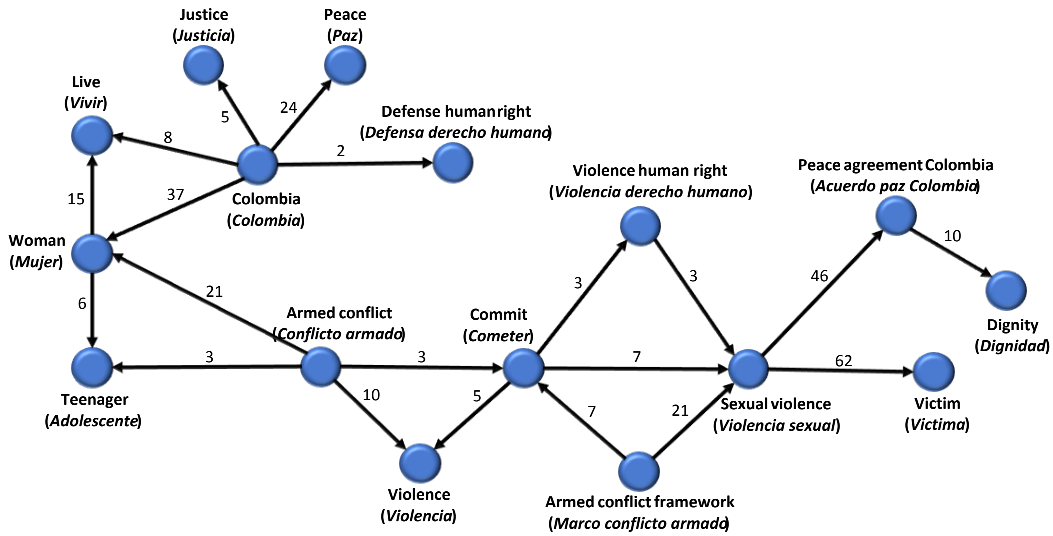

Figure 2: Directed weighted word association network

$$p\left(d_k\right) = \frac{\sum_{i=1}^{n} w\left(i, d_k\right)}{\sum_{j=1}^{m}\sum_{i=1}^{n} w\left(i, d_j\right)} \tag{5}$$

Additionally, $p(C_l)$ is the probability of the concept C_l as presented in Eq. 6.

$$p\left(C_l\right) = \frac{\sum_{i=1}^{n} w\left(i, C_l\right)}{\sum_{j=1}^{m}\sum_{i=1}^{n} w\left(i, C_j\right)} \tag{6}$$

NPMI established the association between the document d_k and the concept C_l in a range of [-1,1], being -1 that these objects never co-occurre, 0 that they are independent and 1 that they always co-occurre [5]. This study set all negative associations as zero and did not change the other values. This strategy consented to generate an association degree matrix between documents and concepts. In consequence, the research detected a *possibilistic partition* of the documents based on the concepts. It is essential to underline that the association degree between the document d_k and the concept C_l is a possibility; therefore, it should never be interpreted as a probability.

3.2.2 *Metrics to evaluate extracted structures.*

This paper regarded metrics: 1) to evaluate the performance of the concepts in the semantic clustering task and 2) to automatically quantify the coherence of conceptual structures. We employed indexes of the soft partition since a *possibilistic partition* generated.

Dunn's index: It is the most recommended density measure and the easiest to calculate. It promotes to identify separate and densely distributed clusters. This study regarded only the values greater than 0.6 of the probability matrix for the LDA model and the association degree matrix for Directed Louvain. This metric gives a close value to zero when a document has the same association degree with the concepts where it intends to cluster.

Overlap measure: It evaluates the overlapping of the groups regarding the information of the probability matrix and the association degree matrix. This research utilized [26] proposal. It established a threshold H of 0.4 to determine whether a document is in the overlapping area of two clusters or not. The index has a high value when the document clusters are overlapping.

Adjusted frand index: It is a measure of the fuzzy rand index family. It compares the similarity between the texts clustered automatically and a manual reference of such groups. The advantages and drawbacks of employing a manual reference are similar to those of *gold standard* and they are described in section 3.1. This research applies [18] proposal because it accomplishes four practical criteria: i) reaches its maximum value when two equivalent responses are compared, ii) detects the best solution given a set of responses, iii) shows progressively better evaluations for superior solutions, and iv) it is corrected for randomness, which ensures that the result is not produced by the randomness fluctuations inherent to the measurement. This metric computes all pairwise information between texts clustered through the membership matrix U where u_{ij} is the association degree of document i to cluster j. [18] report to calculate matrices $J = UU^T$ and $S = U(1_k - I)U^T$ where 1_k is a matrix with 1 in each entry. Therefore, J and S are square matrices with the number of rows equal to the analyzed texts. Particularly, J_{ij} indicates the probability of documents d_i and d_j belonging to the same cluster according to U, in the other hand, S_{ij} gives the opposite information. We utilize the algorithm exposed in [18] to calculate the fuzzy rand index and to make the correction for randomness. This metric gets a value of one when automatic and manual clustering are the equal.

This paper utilized the proposal of [29] to automatically quantify the coherence of the concepts. This metric has been treated in section 3.2.1 and allows to calculate the TC concerning the corpus. This research used WLR in order to identify the N most relevant terms within each extracted concept with Directed Louvain algorithm.

4 EXPERIMENTS AND RESULTS

The authors of this paper make available[3] the analyzed corpus, the list of semantic related words, the stopwords, the gold standard, the built vocabulary, the manual clustering reference, as well as the identified and validated concepts. All the schemes and approaches described in section 3 were implemented in Python 3.7. The the experiments were run on a machine with AMD Ryzen 5 @ 2.2 GHz/4 core processor and 8 GB of RAM memory.

This section presents the information following the same organization as the section 3.

4.1 Vocabulary building

The corpus was 240 texts in Spanish available on the internet and retrieved manually through the equation "armed conflict" (*"conflicto armado"*) and "Colombia." The corpus had 8,461 sentences and 12,955 tokens before preprocessing.

To vocabulary construction, the authors employed the packages of named entity recognition, tokenization, and lemmatization from Freeling 4.0[4]. This work examined the occurrence of the entities with more than one word through a log-likelihood test with 95% confidence, intending to identify the significant collocations. The stopwords used were the compilation of the available lists for the treatment of the Spanish language in the applications Orange3, NLTK 3.4.5, and Google code project. Accordingly, the created vocabulary had 10,182 terms considering collocations and unigrams.

This study manually constructed the *gold standard* utilized for index-terms selection. The reference list encompasses 5,113 terms, of which 2,806 are unigrams and the other collocations.

The index-terms selection integrated the filtration method [35], to identify the vocabulary that efficiently represented the corpus. This method comprises:

(1) Starting with a list of terms, we sorted it in ascending order concerning the value of the weighting schemes designated in section 3.1.
(2) We removed the term in the last position to eliminate the least relevant word, given a weighting scheme. This step produced a new term list.
(3) We computed the *F-measure* to the new listing.
(4) We repeated the second and third steps until there were no terms to represent the corpus.
(5) We compared the *F-measure* results for all iterations and weighting. So, we choose the index-terms with higher similarity against the *gold standard*.

Table 2 presents the features listings that showed the best *F-measure* for each weighting scheme. In Table 2, the last two columns show the number of collocations and unigrams recovered that appear in the *gold standard*. All schemes reported in Table 2 employed the same local weighting, meaning that this study utilized the TF and different global weights.

Ochoa et al. [30] work can recover a vocabulary with a higher number of relevant collocations since these metric places them in upper positions in comparison with the index produced under the other schemes. Thus, this metric has the best *recall* level.

TF-Entropy retrieved unigrams more likely to the ones of the reference list. Therefore the entropy extracts unigrams more relevant than the IDF weighting because it simultaneously regards the term frequency in the corpus and each text. The TF-Entropy scheme has very similar *recall* to [30] work because it recovered a high number of unigrams that genuinely belonged to the *gold standard*.

The modification to the Ochoa et al. [30] work has the best behavior when examining the *F-measure*. This result is a consequence of the *precision* achieved when recovering a smaller index-term constituted by relevant words.

Nevertheless, the *F-measure* of each scheme is low in comparison with investigations [16, 30] that reported findings above 0.7 with the use of a similar methodology for extracting terms from Spanish texts. The research cited filtered out terms that did not follow lexical patterns drawn from specialized resources. This paper did not regard pattern utilization since no knowledge database includes syntactic information from the analyzed domain.

Besides, the Table 2 results may be effect of the human-made *gold standard*. Therefore, it is possible that the extracted index-terms capture data that have been not conceived by the person who constructed the reference list [31].

Given the arguments exposed, the *F-measure* value of the index-term acquired through the modification to the [30] work is acceptable, even though this measure is lower than the reported in other investigations. In consequence, the input for the concepts and relations extraction was the vocabulary built through the modification to Ochoa et al. [30] work.

4.2 Concepts and relations extraction

This study segmented the corpus following the classical division 70-30 to perform the semantic clustering task. The training dataset was composed of 6,431 sentences, and the evaluation dataset had 1,799 sentences. It is necessary to clarify that the same training dataset was used to extract the ontology structures using the LDA model and Directed Louvain algorithm.

To LDA calibration, it was not necessary to fit the model to previously unseen data since this research employed the TC proposal by [29]. The TC application regarded the range of the 15 to 45 top terms within each topic. Fifty-one topics got the highest TC value. The prior to document topic distribution was 13, and the prior of topic word distribution was 0.07. It allowed obtaining a TC with a maximum mean of -1.83±0.056 for N=15 and a minimum mean of -1.448±0.029 for N=60 with 95% confidence. This study clustered the evaluation dataset with the constructed model.

For the Directed Louvain[5] algorithm, the input was a directed weighted word association network. The graph had 6,810 nodes and 26,896 arcs. The Directed Louvain gave partitions with average modularity of 0.4025±0.000373 at 95% confidence and four hierarchical levels.

This research constructed the concepts following a similar methodology to that reported in [33]. Nevertheless, these works differ in two aspects. First, we generated a directed network to capture the semantic information associated with the order of the terms, unlike [33] reported the employment of an undirected

[3]https://github.com/madegomez/Ontology-learning-spanish
[4]http://nlp.lsi.upc.edu/freeling/
[5]https://github.com/nicolasdugue/DirectedLouvain

Table 2: Term listing features

Scheme	Vocabulary size	F − measure	Precision	Recall	Number of recovered and relevant collocations	Number of recovered and relevant unigrams
TF-IDF	7,935	0.651	0.535	0.829	1,777	2,469
TF-Entropy	7,997	0.652	0.534	0.835	1,785	2,488
Ochoa et al. [30] work	8,153	0.661	0.538	0.857	1,903	2,484
Modification Ochoa et al. [30] work	6,810	0.676	0.592	0.787	1,731	2,300

Table 3: Semantic clustering results

Algorithm	Dunn's index	Overlap measure	Adjusted frand index
LDA	0.446	114.216	0.992
	(0.127)	(2.789)	(0.095)
Directed Louvain 1	0.014	195.494	1.0
	(0.0001)	(5.5e−5)	(1.49e−7)
Directed Louvain 2	0.154	195.615	1.0
	(6.33e−17)	(5.93e−14)	(4.44e−17)
Directed Louvain 3	0.354	195.259	1.0
	(0.0001)	(0.0002)	(5.9e−17)
Directed Louvain 4	0.316	195.368	1.0
	(0.001)	(0.003)	(3.5e−17)

network. Second, the validation of the constructed concepts. In [33] evaluated the concept quality through the modularity and the comparison against a *gold standard*. Modularity-guided assessment favors concepts built using algorithms focused on optimizing modularity. Therefore, the evaluation is biased. For example, if this work had utilized this framework, we would have prejudiced the concepts constructed by LDA compared to those of Directed Louvain. Additionally, the evaluation based on *gold standard* could not be applied in this paper because this conceptual reference does not exist in the investigated domain, and we avoided its construction that implied high costs.

Following the semantic clustering task to validate the constructed concepts, Table 3 presents the evaluation results; each interval has 95% confidence. This study set up a manual reference of clustering documents to measure the adjusted frand index. The reference has a significant number of overlapping clusters. Figure 3 shows the TC mean and the error bar of the experimental settings when the top terms vary from 15 to 60. For Directed Louvain information, the level of the hierarchy is shown with the number after the name, with one being the first level. The Figure 3 has a windows zoom to point out the significant difference between the TC of Directed Louvain 3 and Directed Louvain 4.

Directed Louvain 1 built clusters with a degree of association very similar to each conceptual structure. The addressed algorithm and Directed Louvain 2 had the highest overlap rates. It warns that these scenarios have concepts whose semantic information is not differential and produce that the texts cannot be segregated. Consequently, these algorithms do not produce appropriate concepts and relations to describe the corpus.

LDA constructed dense clusters with the lowest overlap index. Additionally, the partitions are the least similar to the manual reference. The concepts of LDA have poor coherence based on the co-occurrence of their terms. Notably, the TC of LDA is significantly lower than the coherence of Directed Louvain 3, Directed Louvain 4, and Directed Louvain 2. This scenario is consistent with the results reported in [25], where short text topics are automatically identified by experimenting with the LDA model and a Louvain algorithm.

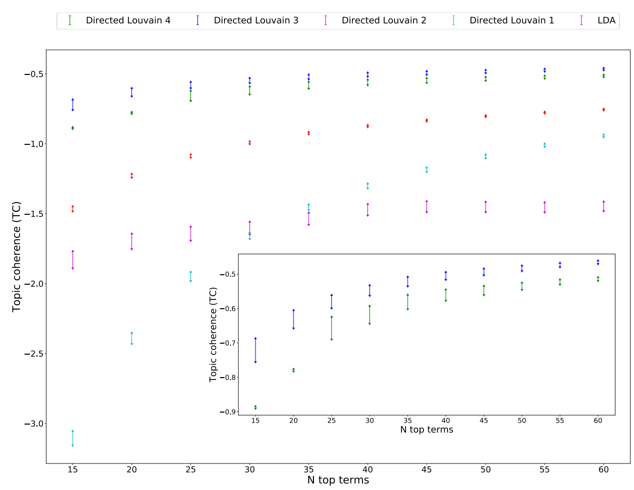

Figure 3: *Topic coherence* with top terms between 15 to 60

The Directed Louvain 3 and Directed Louvain 4 algorithm generate dense clusters with acceptable overlap rates, making the grouped data remarkably like the manual reference. The scenarios have the highest levels of TC, namely their concepts are semantic meaning. The performance of Directed Louvain 3 concepts is significantly better than the results obtained with Directed Louvain 4. This situation is notable when analyzing the density index and TC in the zoom window of the Figure 3.

The findings of Directed Louvain are like those reported in [24], where texts are semantically clustered using a modularity optimization algorithm, producing documents group with low overlap and very similar to the manual reference. Besides, in [32], the employ of a Louvain algorithm for the concept construction helped the visual narration of scientific articles.

Nonetheless, the presented results are different from those reported in [9, 19]. These researches utilized partitioning clustering algorithms (following the classification of [15]) to detect communities, in this way, they constructed groups of dense documents and very similar to the manual reference. Partitional clustering is characterized by bringing the terms (from the vocabulary) to a metric space where a measure of distance is determined to build communities. The named studies can be questioned since they use the terms frequency in the corpus for the extraction of concepts, disregarding the semantic relations that may exist between words [3]; we utilized the arcs between the terms to capture the semantic information. Additionally, partition clustering is not appropriate because it may be sensitive to the distance measure used during community detection [15].

This paper utilized a semantic clustering methodology very similar to the presented in [27]. Our work classifies the terms (nodes) within each concept (community) using the WLR metric. Therefore, this research regarded the node relevance given the number and weight of its neighbors [28]. On the contrary, in [27] the node degree is the only measure to determine a relevant term, producing conclusions such as that a node is not influential even though it is linked only to one distinguished vertex.

These differences may be the reason why Liu et al. [27] reported that the *clique percolation method* constructed appropriate concepts for semantic clustering of texts. In addition, in [27] validated the concepts quality against a manual reference list. This study did not carry out the evaluation based on manual reference because, as already mentioned, this resource does not exist for the studied domain, and its construction implied high costs that we wanted to avoid in order to propose an appropriate economic evaluation for domains without conceptual datasets.

5 CONCLUSIONS AND FUTURE WORK

This work exposes an ontology learning method from texts in Spanish regarding the domain of the Colombian armed conflict. To the best of our knowledge, this work is the first to present a computational proposal for the semi-automatic extraction and evaluation of concepts and relations for this domain. The findings may be the foundation to extend an ontology in this line.

Following the experimental conditions, the vocabulary built using the modification of Ochoa et al. [30] work, has the highest F-measure value since it recovers a higher number of collocations and unigrams that truly belong to the *gold standard*. The concepts detected by the Directed Louvain algorithm at the third hierarchical level generated dense clusters with an overlap index of 195.258 ± 0.0002, additionally, the generated partition is very similar to the manual reference. Besides, the TC result indicates that the extracted concepts are semantically relevant. Therefore, the Directed Louvain algorithm at the third hierarchical level detects appropriate concepts and relations from Spanish texts regarding the Colombian armed conflict.

It is appropriate to report that the vocabulary building using the modification of Ochoa et al. [30] work, as well as the extraction and validation of abstract ontological structures following the Directed Louvain algorithm at the third hierarchical level takes 7058 seconds, moreover it consumes 1203.64583 MiB using the equipment and data specified at the beginning of the section 4.

This study proposes the evaluation of abstract ontology structures through the semantic clustering task. To the best of our knowledge, we are the first ones to evaluate concepts and relations extracted from texts utilizing this task. The proposal evaluation, unlike the existing ones, allows validating ontological structures, without the costly need to use manual validation or conceptual databases. Therefore, this evaluation is not only inexpensive but also applicable to model data and domains that lack structured knowledge resources.

This research proposes to automatically quantify the results of the texts clustering through density, overlap, and similarity metrics with reference data, and automatically measure the coherence of the identified concepts. Therefore, this study is committed to low-cost and large-scale evaluation. The proposed approach satisfies the standards indicated in [11]. In consequence, the proposal filters out the task intervention to obtain valid conclusions about the ontology learning techniques compared in this work.

The metrics, presented in section 3.2.2, allowed a robust characterization of the techniques to extract abstract ontology structures since this paper regarded the coherence of the concepts, as well as the density, overlap, and similarity of the semantic clustering of the texts. Therefore, it was found that LDA generates dense groups of texts with low levels of overlap, unlike the Louvain algorithms. On the other hand, the LDA concepts have a poor semantic interpretation among the regarded scenarios, since their most probable terms do not co-occurre in the corpus.

In future works, experimental scenarios with other local and global weighting schemes (section 3.1) will be regarded to analyze the influence that the local measure applied had on the results of the Table 2. Future research may utilize lexical patterns if these are built through unsupervised tools or applying a semi-supervised approach. Moreover, evaluating the consistency and eliminating the redundancies of the constructed abstract ontology structures will be considered.

Following the reported in [33], we will explore different strategies to build the network that serves as an input for community detection. We will examine community detection algorithms that build concepts analyzing clustering patterns and overlapping communities. In future studies, we will apply the evaluation proposal in a widely investigated corpus and conceptual database. It will strengthen the exposed results in this paper.

ACKNOWLEDGMENTS

This research is supported by Universidad Tecnológica de Pereira under the project "Propuesta de enriquecimiento ontológico a partir de datos textuales para el idioma español en el dominio del conflicto armado colombiano". Project code: E7-20-1. Additionally, M. Gómez-Suta's work is funded by Colciencias under the program "Jóvenes Investigadores e Innovadores - Convenio 321-2019".

REFERENCES

[1] C. Aguilar, O. Acosta, G. Sierra, S. Juárez, and T. Infante. 2016. Extraction of definitional contexts from biomedical corpora [Extracción de contextos definitorios en el área de biomedicina]. *Procesamiento de Lenguaje Natural* 57 (2016), 167–170.
[2] Y. Alemán, M.J. Somodevilla, and D. Vilariño. 2019. Similarity metrics analysis for principal concepts detection in ontology creation. *Journal of Intelligent and Fuzzy Systems* 36, 5 (2019), 4753–4764. https://doi.org/10.3233/JIFS-179024
[3] I. Ali and A. Melton. 2018. Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory. *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018* 2018-January (2018), 243–247. https://doi.org/10.1109/ICSC.2018.00042
[4] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief, and Y.F. Hassan. 2019. CLOE: a cross-lingual ontology enrichment using multi-agent architecture. *Enterprise*

Information Systems 13, 7-8 (2019), 1002–1022. https://doi.org/10.1080/17517575.2019.1592232

[5] H. Amoualian, W. Lu, E. Gaussier, G. Balikas, M.-R. Amini, and M. Clausel. 2017. Topical coherence in LDA-based models through induced segmentation. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2017), 1799–1809. https://doi.org/10.18653/v1/P17-1165

[6] N. Aussenac-Gilles and D. Sörgel. 2005. Text analysis for ontology and terminology engineering. *Applied Ontology* 1 (2005), 35–46.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 4-5 (2003), 993–1022.

[8] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D.M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference* (2009), 288–296.

[9] S.-Y. Chen, C.-N. Chang, Y.-H. Nien, and H.-R. Ke. 2012. Concept extraction and clustering for search result organization and virtual community construction. *Computer Science and Information Systems* 9, 1 (2012), 323–355. https://doi.org/10.2298/CSIS101124020C

[10] M. Clark, Y. Kim, U. Kruschwitz, D. Song, D. Albakour, S. Dignum, U.C. Beresi, M. Fasli, and A. De Roeck. 2012. Automatically structuring domain knowledge from text: An overview of current research. *Information Processing and Management* 48, 3 (2012), 552–568. https://doi.org/10.1016/j.ipm.2011.07.002

[11] K. Dellschaft and S. Staab. 2008. Strategies for the evaluation of ontology learning. *Frontiers in Artificial Intelligence and Applications* 167, 1 (2008), 253–272.

[12] N. Dugué and A. Perez. 2015. *Directed Louvain : maximizing modularity in directed networks*. Research Report. Université d'Orléans. https://hal.archives-ouvertes.fr/hal-01231784

[13] L Espinosa. 2018. *Tesauro con enfoque diferencial sobre graves violaciones a los DDHH e infracciones al DIH con ocasión del conflicto armado colombiano*. Technical Report. Centro Nacional de Memoria Historica. 1–226 pages.

[14] J. Farreres, K. Gibert, H. Rodríguez, and C. Pluempitiwiriyawej. 2010. Inference of lexical ontologies. The LeOnI methodology. *Artificial Intelligence* 174, 1 (2010), 1–19. https://doi.org/10.1016/j.artint.2009.09.004

[15] S. Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3-5 (2010), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

[16] S.N. Galicia-Haro and A. Gelbukh. 2014. Extraction of semantic relations from opinion reviews in Spanish. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8856 (2014), 175–190.

[17] K. Gutiérrez-Batista, J.R. Campaña, M.-A. Vila, and M.J. Martin-Bautista. 2018. An ontology-based framework for automatic topic detection in multilingual environments. *International Journal of Intelligent Systems* 33, 7 (2018), 1459–1475. https://doi.org/10.1002/int.21986

[18] D. Horta and R.J.G.B. Campello. 2015. Comparing hard and overlapping clusterings. *Journal of Machine Learning Research* 16 (2015), 2949–2997.

[19] C. Jia, M.B. Carson, X. Wang, and J. Yu. 2018. Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition* 76 (2018), 691–703. https://doi.org/10.1016/j.patcog.2017.09.045

[20] A. Konys. 2019. Knowledge repository of ontology learning tools from text. *Procedia Computer Science* 159 (2019), 1614–1628. https://doi.org/10.1016/j.procs.2019.09.332

[21] D. Korenčić, S. Ristov, and J. Šnajder. 2018. Document-based topic coherence measures for news media text. *Expert Systems with Applications* 114 (2018), 357–373. https://doi.org/10.1016/j.eswa.2018.07.063

[22] A. Lancichinetti and S. Fortunato. 2012. Consensus clustering in complex networks. *Scientific Reports* 2 (2012). https://doi.org/10.1038/srep00336

[23] E.A. Leicht and M.E.J. Newman. 2008. Community structure in directed networks. *Physical Review Letters* 100, 11 (2008). https://doi.org/10.1103/PhysRevLett.100.118703

[24] X. Li, J. Chen, and O. Zaiane. 2013. Text document topical recursive clustering and automatic labeling of a hierarchy of document clusters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7819 LNAI, PART 2 (2013), 197–208. https://doi.org/10.1007/978-3-642-37456-2_17

[25] K.H. Lim, S. Karunasekera, and A. Harwood. 2017. ClusTop: A clustering-based topic modelling algorithm for twitter using word networks. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017* 2018-January (2017), 2009–2018. https://doi.org/10.1109/BigData.2017.8258147

[26] P.-L. Lin, P.-W. Huang, and C.-Y. Li. 2016. A validity index method for clusters with different degrees of dispersion and overlap. *Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016* (2016), 222–229. https://doi.org/10.1109/ICACI.2016.7449829

[27] Y. Liu, N. Borhan, X. Luo, H. Zhang, and X. He. 2013. Association link network based core events discovery on the web. *Proceedings - 16th IEEE International Conference on Computational Science and Engineering, CSE 2013* (2013), 553–560. https://doi.org/10.1109/CSE.2013.88

[28] L. Lü, Y.-C. Zhang, C.H. Yeung, and T. Zhou. 2011. Leaders in social networks, the delicious case. *PLoS ONE* 6, 6 (2011). https://doi.org/10.1371/journal.pone.0021202

[29] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (2011), 262–272.

[30] J.L. Ochoa, R. Valencia-García, A. Perez-Soltero, and M. Barceló-Valenzuela. 2013. A semantic role labelling-based framework for learning ontologies from Spanish documents. *Expert Systems with Applications* 40, 6 (2013), 2058–2068. https://doi.org/10.1016/j.eswa.2012.10.017

[31] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos. 2011. Ontology population and enrichment: State of the art. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6050 (2011), 134–166. https://doi.org/10.1007/978-3-642-20795-2_6

[32] Q. Ping and C. Chen. 2018. LitStoryTeller+: an interactive system for multi-level scientific paper visual storytelling with a supportive text mining toolbox. *Scientometrics* 116, 3 (2018), 1887–1944. https://doi.org/10.1007/s11192-018-2803-x

[33] J. Qiu, Y. Chai, Z. Tian, X. Du, and M. Guizani. 2020. Automatic Concept Extraction Based on Semantic Graphs from Big Data in Smart City. *IEEE Transactions on Computational Social Systems* 7, 1 (2020), 225–233. https://doi.org/10.1109/TCSS.2019.2946181

[34] H. Sfar, A.H. Chaibi, A. Bouzeghoub, and H.B. Ghezala. 2016. Gold standard based evaluation of ontology learning techniques. *Proceedings of the ACM Symposium on Applied Computing* 04-08-April-2016 (2016), 339–346. https://doi.org/10.1145/2851613.2851843

[35] C. Silva and B. Ribeiro. 2010. *Inductive Inference for Large Scale Text Classification*. Vol. 255. Springer-Verlag Berlin Heidelberg. 155 pages. https://doi.org/10.1007/978-3-642-04533-2

[36] W. Wong, W. Liu, and M. Bennamoun. 2012. Ontology learning from text: A look back and into the future. *Comput. Surveys* 44, 4 (2012). https://doi.org/10.1145/2333112.2333115

[37] J. Xuan, H. Jiang, Z. Ren, and W. Zou. 2012. Developer prioritization in bug repositories. *Proceedings - International Conference on Software Engineering* (2012), 25–35. https://doi.org/10.1109/ICSE.2012.6227209