Contemporary Engineering Sciences, Vol. 10, 2017, no. 17, 829 - 840
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ces.2017.7768

Opinion Mining using a Knowledge Extraction System from the Web

Taniana Rodríguez[1,2], Jose Aguilar[1,2,3] and Alexandra Gónzalez[3]

[1] Centro de Estudios en Microelectrónica y Sistemas Distribuidos (CEMISID), Facultad de Ingeniería, Universidad de Los Andes (ULA) Mérida-Venezuela

[2] Universidad Autónoma de Chile, Santiago, Chile

[3] Dpto. de Ciencias de la Computación y Electrónica Universidad Técnica Particular de Loja, Ecuador

Abstract

This paper designs and implements an opinion mining system based on the knowledge extraction from unstructured documents in HTML format. Basically, the system recovers information from the Internet about a subject, and transforms it in a structured and organized knowledge, described semantically in an ontology, using a Semantic Ontology. Therefore, the Semantic Ontology describes the opinion about this subject previously defined. This ontology is a semantic knowledge based on the extraction of entities and relationships, where entities are something about we can say something, and the relationships are the interactions between the entities. From this model of semantic knowledge, it is possible to infer the public opinion, and new knowledge, such as the terminological base around of the opinion.

Keywords: Opinion Mining, Knowledge Extraction, Semantic Ontology

1 Introduction

At present, there are a large number of unstructured natural-language documents, which have a large semantic content. These documents, for the most part, are generated and stored on the Internet. They describe different aspects, and

have a large influence in the determination of the public opinion. Particularly, to analyze its impact in the public opinion, it is necessary to deal it as unstructured information. These contents can be structured and organized semantically, in order to extract the knowledge in a given context. Particularly, we are interested in determining the influence in the context of public opinion.

In this work, we study how the information on the Internet defines a public opinion. For that, it is necessary to extract the knowledge included in the unstructured documents on the Internet. In this way, the idea is to generate semantic knowledge from unstructured documents about a subject, understanding as semantic knowledge the entities and relationships found in the texts. With this generated semantic knowledge, we are going to make inference processes in order to mine this information to determine the public opinion, and specifically, to generate the specific terminological basis of the opinion studied.

Our approach extracts the knowledge of the entities and relationships that are present in the unstructured documents, to be used in the semantic mining of a public opinion about a given subject. We design the semantic knowledge model as a semantic ontology. From this ontology is made the processes of reasoning, to infer and interpret the opinions about a given subject.

2 Our Extraction Systems for Opinion Mining

A general architecture of an Information Extraction Systems based on Ontologies, is proposed in [3]. The system proposed in this paper is based on these ideas. The processing of unstructured text is based on the next linguistic resources: lexicons, dictionaries, corpus and Onomasticon (see Fig. 1). The input is the text to be processed, and with the linguistic resources, the learning graph is generated. The scheme is the following: we start with a query about a subject, to know the public opinion about it. This query is interpreted by our opinion mining system, to recover a set of unstructured texts from the Internet with information on this subject. Our knowledge extraction process creates the learning graph, and in parallel, the pattern extraction process generates a set of patterns (pattern graph). The knowledge extraction process extracts entities and relationships, and the result is stored in the learning graph. In particular, the Knowledge Extraction process focuses on the extraction of entities and relationships found in the input texts, with the support of different linguistic resources [2, 11]. Specifically, the linguistic resources used are the lexicon, which is an ontology of Spanish language terms, and an Onomasticon, which is an ontology of proper names. Our system uses the CONLL-2002 corpus for Spanish [10] and WordReference.

The pattern graph describes the extraction of patterns using natural language processing techniques (using OpenNLP library). For that, it extracts the patterns of sentences, e.g. Noun Verbo_aux adjective Noun. Finally, the classification opinion process identifies the properties of an entity and determines the opinion. This process is divided into three steps: 1) Detect the properties of an entity in the given sentence. 2) Classify the opinion associated with these properties as positive, negative, or neutral. 3) Generate the opinion graph for the result analysis.

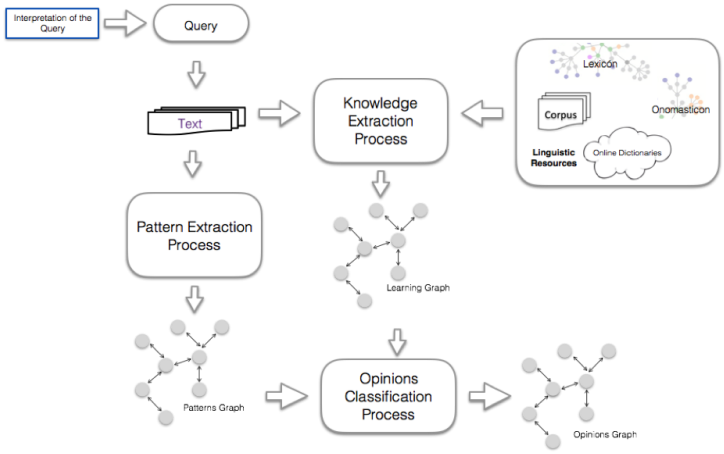

Fig. 1. General Scheme

Once the opinion graph is obtained, the analysis of the opinions can be carried out, in order to determine the public opinion about the subject, or to create a terminological base (instanced in the opinion graph).

A. Conceptual Model for the processing of the Text

The conceptual schema, with all the elements involved in the text treatment, is shown in Fig. 2. The interpretation of the text begins with the recognition of terms found in the text, which can be simple words (e.g., escuela, Estados Unidos) or compound words (e.g. "University of Houston", "Salt Lake"). Each simple word has a lemma (for example, University has the lemma Universit). In addition, both simple and compound words have a Category (it represents the type of the word, if it is a verb, a noun, if the noun is a common type (e.g. University) or own (of Houston), if the noun is an entity, etc. On the other hand, each one of these entities has a semantic classification (for example, people, places, organizations, etc.). Figure 2 shows the learning graph (in Spanish), which allows to describe ontologically (semantically) all the knowledge extracted from the text.

B. The Learning Graph

The Learning Graph is a semantic ontology, which allows to describe the knowledge contained in a collection of texts. Their instances will be that knowledge extracted from the texts. For this, the Learning Graph is based on the conceptual scheme shown in Fig. 2, and is defined as follows: a tuple (T, R) where, T is the set of all the concepts involved in the characterization of the terms found in the text, R is the set of their relationships.

To make this process of characterization of terms in a text, T contains the following classes (see Fig. 3): Term Graph: They represent all the nouns and verbs that exist in the text; Simple Word: is a subclass of the Term Graph, which represents all atomic words; Composite Word: is a subclass of the term Graph,

which represents words that are composed of several atomic words; Category: Represents the grammatical category of the term Graph, and is composed by Noun and Verb; Entity: represents the entities of place or location, organization, people and others; Events: represents verbs that execute an action between entities; Lexicon: in this case, the lexicon only contains nouns that are not proper names; Onomasticon: represents all the proper names found in the texts.

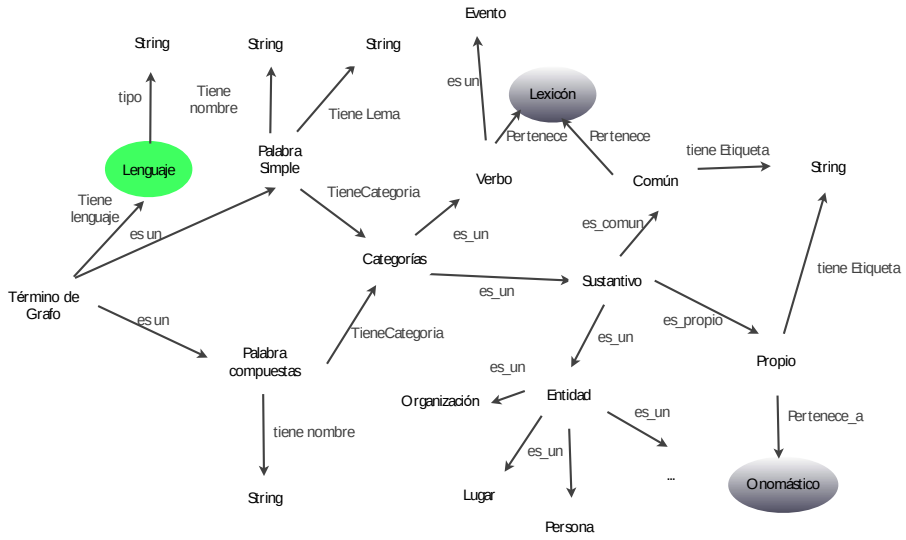

Fig. 2. Conceptual scheme for the text processing

On the other hand, the relationships between these concepts are (see Fig. 4): Has Gender: Indicates whether it is female, male or neutral; Has Number: Indicates whether the term of the tree is one or more than one. Specifically, in Spanish there are two numbers: singular and plural; Has semantics: represents the semantic classification of the term (if it is an organization, person, place, other); HasType: represents the type of the term (if it is a common name, proper name); esComponent word: represents whether the term is composed of several words; SimpleSword: represents whether the term is a single word; esUn: represents if the term is a verb, hasMode: In the case that the term of the graph is a verb, the mode refers to the attitude of the speaker. There are three verbal modes: Indicative (for example, I have arrived in the city), Subjective (for example, maybe it reaches the city), and Imperative (for example, come here).

Finally, the class term of the graph has a set of properties (see Fig. 5): hasEAGLESLabel: is a label that represents the morphological information of the term in the Learning Graph; hasFrequency: represents the number of times the term is repeated in the texts being processed. With this frequency, it is determined whether the term is relevant or not; hasName: represents the name of the term in the Learning Graph.

In this way, each instance of the Class Term of the Graph is described according to its morphological characteristics. For example, the term Academic is described as follows: Academy is a simple word (the relationship is Simple Word), Academy is

female (the Relationship HasGeneral), Academy is singular (the Relationship Has Number), Academy is a common name (the Relationship Has Type) and an organization (the Relationship Type), it is repeated in the documents twice, the name is Academic and has the eagles tag np00000 (see Fig. 6).

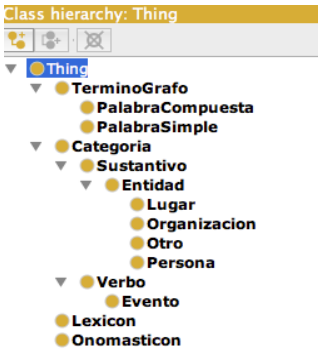

Fig. 3. The classes in the learning graph

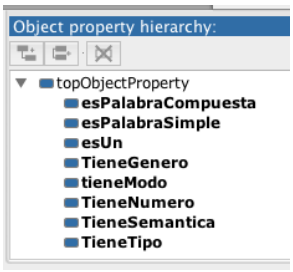

Fig. 4. The relationship in the learning graph

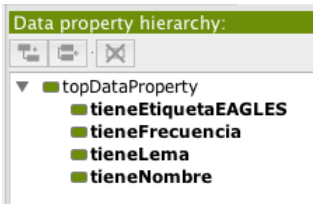

Fig. 5. The type of data by default in the learning graph

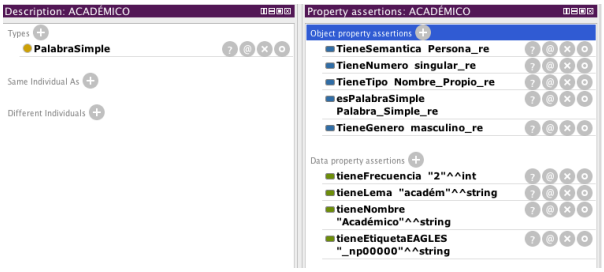

Fig. 6. Description of the class term in the graph, instanced with the term: Académico

C. Extraction of Entities and Relationships from unstructured text.

The process of extracting entities and relationships in unstructured texts uses the Learning Graph, and is divided into 3 steps, and one of them in 4 sub-processes:

1. Separation of the text in Sentences: this process consists of dividing the text into sentences.
2. For each sentence: a) Recognition of Proper Names: consists in the detection and classification of the terms of the text, such as names of people, organizations, places, numerical expressions, time, place etc.; In this process, the Onomasticon and Corpus are used as linguistic resources. b) Morphosyntactic Analysis: This analysis consists in labeling the words of the sentence with the morphological and syntactic information of them, that is, according to their grammatical category (verbs, nouns, adjectives, verbs, etc.). This process uses the linguistic resources lexicon, onomasticon and Corpus. c) Recognition of Entities: in this process are extracted the words labeled as proper names and common names in the sentence, in the morphosyntactic analysis. They will be the candidate entities

that will be used in the Learning Graph. In the recognition of simple entities we get the entity's lemma, which aims to normalize the entities (for example, the simple entities Universitario and Universitarios have the same lemma Universitario); d) Acknowledgments of Relations: in this process are extracted the words, the type verbs of the sentence, labeled in the morphosyntactic analysis. They are the candidate relationships that will be used by the Learning Graph and pattern graph.

3. Learning Graph Update: In this process all the entities and relationships candidate, found in the processed text, with their respective relationships and data types, are updated on the Learning Graph. Once the set of patterns of the morphosyntactic analysis is obtained, are selected the relevant patterns to be incorporated in the pattern graph.
4. Classification of opinion: This process is divided into three steps: a) Detect the aspects of an entity in the given sentence; b) Classify the opinion associated with this aspect, i.e. if it is positive, negative, neutral, c) Generate the opinion graph for later analysis of results

D. Update the learning graph

The previous step (analysis of each sentence) generates a collection of words in a specific domain: a) Candidate Entities, all nouns that are found in the texts processed; b) Candidate Relations, all the verbs that are in the processed texts.

In this step, the relevant entities and relationships are selected, which should be incorporated into the Learning Graph. They are defined as: a) Relevant Entities, all nouns that meet the relevance criteria; b) Relevant Relationships all verbs that meet the relevance criteria.

For this purpose, we propose the following measures, which classify the terms in the documents processed as relevant or not, inspired by the works in [5, 6, 7, 9]:

- Tf (Frequency term): is the frequency of a term j. It is a measure of relevance of a term in the Learning Graph. If the term appears many times in the texts, the weight of the term is high. Where, $tf_j = HasFrequency\_j$
- fi (reverse of the frequency): it is used to reflect the importance of the terms in the texts processed, prioritizing the accuracy and discriminatory power of the same. Thus, more important is a term, then it appears less in the processed texts of the collections. On the other hand, if a term appears in all processed texts, its accuracy and discriminatory power will be less. fi is calculated as

$$fi_j = \log\left(\frac{Number\ of\ texts \text{ processed}}{Number\ of\ texts\ with\ the\ term\ j}\right) \quad (1)$$

- wj (weight of term j) indicates that the greater is for a term j, it is because that term is more relevant, since it is very discriminant.

$$w_j = \begin{cases} tf_j * fi_j & for\ texts\ processed > 1 \\ fi_j & otherwise \end{cases} \quad (2)$$

Specifically, the criterion used to determine if an Entity or Relationship is relevant is: it will be relevant if its weight is greater than or equal to the average of the weights. For the evaluation of the Generated Learning Graph, the following criteria are used:

- Correct use of language, which consists in evaluating the quality of the Learning Graph in terms of the way it is written.
- Accuracy of the taxonomic structure. In this phase, the following is checked with the Protégé editor: identification of inconsistencies, completeness of concepts and existence of redundancy in classes, instances and relationships.

3. Our Extraction Systems for Opinion Mining

A. Characterization of the process of extraction of knowledge from the Internet

In this case study, we want to extract the relevant entities and relationships, which will allow to instantiate the learning graph (semantic ontology), from texts from the Internet.

Initially, we focus on the extraction of patterns, from the sentences. For example, if we have the following tweet "# TracaTrumpM4 You'll see how Donald Trump will be a good President". The pattern recognized in this sentence is Noun Verbo_aux adjective Noun. (Donald Trump) _Sustainable (will be a) _Verbo_aux buen_Adjetivo Presidente_Sustantivo

The classification process focuses on to identify the properties or characteristics of an entity to determine the opinion, in order to build the opinion graph. For example: "Donald Trump will be a good President", the associated opinion is positive.

Once the opinion graph is obtained, the analysis of the opinions can be carried out. Following the process shown in the previous section, the text is separated into sentences. Next, each sentence is analyzed, the proper names that appear in the text are recognized, then the morphosyntactic analysis of the words is performed, and the candidate entities and relationships are recognized. The analysis phase of the sentences uses the different sources of knowledge. Finally, we proceed to the last phase, which consists of determining the relevant entities and relationships.

In the last phase, which consists of determining the relevant entities and relationships, the metrics defined in the previous section are used. The relevant entities and relationships determined, when are used to instantiate the ontology that is derived from the learning graph, allow to make a semantic description of the analyzed document.

B. Construction of the terminological base

In this case study, we evaluate the possible uses of the semantic knowledge model (the semantic ontology) generated from unstructured documents. In particular, we will evaluate its ability to generate the terminological base of an opinion. For this, 12 documents of the Websites were processed, recovered from a

query about the last American elections and Trump. These documents are:

- http://www.lavanguardia.com/internacional/20161108/411675232210/elecciones-estados-unidos-2016-en-directo.html
- http://www.el-nacional.com/leopoldo_martinez_nucete/resultado-elecciones-Unidos-primera-lectura_0_956304582.html
- http://internacional.elpais.com/internacional/2016/11/08/actualidad/1478603130_588370.html
- http://www.20minutos.es/noticia/2883316/0/trump-presidente-eeuu-magnate-negocios-despide-hillary-clinton/
- http://www.20minutos.es/noticia/2880964/0/datos-curiosos-elecciones-presidenciales-estados-unidos/
- http://www.mundodeportivo.com/elotromundo/actualidad/20161108/411678490280/elecciones-eeuu-ultima-hora-directo-2016.html
- http://www.elconfidencial.com/mundo/2016-11-11/trump-evitar-ser-mal-presidente-elite-eeuu-voto-confianza_1288001/
- https://actualidad.rt.com/actualidad/214437-razones-trump-presidente-moor
- http://www.bbc.com/mundo/noticias-internacional-37918294

In addition, with these documents were calculated the metrics defined in the previous section, for which the following queries were made in SPARQL:

- To determine the total and average weights of the entities, the following query is performed:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT (COUNT( ?subject) AS ?Entidades_Candidatas )  (AVG(?object) AS ?Promedio_Peso)
WHERE { {?subject j.0:Peso ?object . ?subject j.0:TieneTipo  j.0:Nombre_Propio_re } UNION
{ ?subject j.0:Peso ?object . ?subject j.0:TieneTipo  j.0:Nombre_Comun_re } }
```

To determine the total and average of the weights of the relationships, a similar query is performed. The results are shown in table 1. In this table, we can see that there are 602 candidate entities, and the average weights of the entities is 4,89. The analysis for the case of the candidate relations is similar.

Table 1. Metrics to determine the relevant entities and Relationships.

	Total	**Average Weight**
Candidates Entities	602	4,89
Candidates Relationships	205	4,92

From these values, we can determine the relevant entities and relationships. Table 2 shows the relevant entities, based on the criterion: "all terms greater than or equal to the average of the weights", knowing that the average is 4.89 (see

Table 1). The specific query in SPARQL to obtain the relevant entities, with their respective weights, is:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?nombre ?peso WHERE { {?subject j.0:tieneFrecuencia ?object . ?subject j.0:TieneTipo j.0:Nombre_Propio_re . ?subject j.0:tieneNombre ?nombre . ?subject j.0:TieneTipo ?tipo . ?subject j.0:Peso ?peso } UNION { ?subject j.0:tieneFrecuencia ?object . ?subject j.0:TieneTipo j.0:Nombre_Comun_re . ?subject j.0:tieneNombre ?nombre . ?subject j.0:TieneTipo ?tipo . ?subject j.0:Peso ?peso } FILTER (?peso>4.89)}
```

With this query we find the relevant entities, which are defined as "Own Names" and "Common Names", to generate Table 2. The same process is followed, to determine the relevant relationships. The weight column is calculated using eq. 2.

Tabla 2. Preliminary Results

Entities	Weight	Relationships	Weight
Trump	183.88	SERA	37.63
Clinton	54,67	SER	29,11
Elecciones	35,83	PODRÁ	24,84
Votación	32,30	PRESENTAR	21,50
Republicano	32,30	DEBERÁ	14,33
Voto latino	24,85	ESTARÁ	12,43
…		SERÁN	12,42

Finally, the final Learning Graph that is generated is composed of the relevant entities and relationships (their instances). Now, from the Learning Graph there are many things that can be done, next we will describe some of them for this case study. Starting from the fact that the Learning Graph contains a set of basic axioms that allow to infer new knowledge, some of the potential applications using the inference engine on the Learning Graph is: Build a terminological base in a specialized domain.

The Learning Graph allows to extract a list of specialized terms of a domain from the corpus of documents studied. For this the weight criterion defined above is taken into account. For example, the following rules on the Learning Graph could be used:

- TieneTipo(?x, Nombre_Propio_re), tienePeso(?x, ?pes), greaterThan(?pes, 4,89 -> Onomasticon(?x),

- TieneTipo(?x, Nombre_Comun_re), tienePeso (?x, ?pes), greaterThan(?pes, 4,89) -> Lexicon(?x)
- Onomsticon((?x)or(Lexicon((?x)-> TerminosEspecializados

These rules characterize the relevant terms (proper names, common names, etc.) (weight 4.89) that characterize the original query. These rules, when are executed by the reasoner on the Learning Graph, would obtain the result of specialized terms about the opinion mining. This specialized terminology acquisition capability can be used in word processing, text mining tasks, the creation of specialized ontologies, etc.

4 Conclusions

In this paper, we present a system of knowledge extraction from unstructured texts, which is organized semantically in a Learning Graph. The Learning Graph is a semantic ontology that can be represented in RDF, OWL, etc., whose instances represent all the knowledge extracted from a collection of documents. From the Learning Graph, it is possible to use the system to generate new semantic knowledge, such as specialized terminological bases.

For the construction of the Learning Graph, we have used a metric that allows to classify the terms contained in the processed documents, as relevant or not. This metric, combines the frequency and the ability to discriminate, of the terms in the documents processed. This gives the Learning Graph a fundamental characteristic, being a semantic ontology of processed documents, formed only by the relevant information (individuals).

This information is used, in our case, to mine the opinion about a public subject. We can characterize the main aspects about a subject, in order to know the public perception about it.

The main difference in our work with previous researches is that we propose an opinion mining method that builds a knowledge graph, which can be used to analyze other aspects, for example, the vocabulary or taxonomy around the opinion. The classical approaches classify the source of information, identify the main information (twitter, website, etc.), which is used to infer information [12, 13, 14, 15, 16, 17]. We extend it with the knowledge graph, which has a great potential in different contexts, for example, for the analysis of textual citizens' contributions, the general attitudes-sentiments (positive, negative or neutral) of the public in the discussion, among other things.

References

[1] S. Nirenburg and V. Raskin, Ontological Semantics, MIT Press, Cambridge, Massachusetts, London, England, 2004.

[2] J. Aguilar, Sistema semántico para la búsqueda inteligente de información por contexto para la Web, *Revista Ciencia e Ingeniería*, **32** (2011), no. 3, 141-152.

[3] D. Wimalasuriya and D. Dou, Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches, *Journal of Information Science*, **36** (2010), no. 3, 306-323. https://doi.org/10.1177/0165551509360123

[4] A. Rodríguez and A. Cuevas, Método para la extracción de información estructurada desde texto, *Revista Cubana de Ciencias Informaticas*, **7** (2013), 1, 55-67.

[5] T. Rodríguez, E. Puerto, J. Aguilar, Dynamic Semantic Ontological Framework for Web Semantics, *Proceeding of the 9th WSEAS Intl. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS '10),* (2010), 91- 98.

[6] E. Puerto, J. Aguilar, T. Rodríguez, Automatic Learning of Ontologies for the Semantic Web: experiment lexical learning, Respuestas: Revista, **17** (2012), no. 2, 5-12.

[7] T. Rodríguez and J. Aguilar, Aprendizaje ontológico para el marco ontológico dinámico semántico, *DYNA*, **81** (2014), no. 187, 56-63.

[8] L. Ramshaw and R. Weischedel, Information extraction, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005. https://doi.org/10.1109/icassp.2005.1416467

[9] T. Rodríguez and J. Aguilar, Implementación del marco ontológico dinámico semántico, *Ingeniare. Revista Chilena de Ingeniería*, **25** (2017), no. 3, 430-448. https://doi.org/10.4067/s0718-33052017000300430

[10] E. Sang and T. Sang, Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, *Proceedings of the 6th Conference on Natural Language Learning CoNLL-2002*, (2002), 155-158. https://doi.org/10.3115/1118853.1118877

[11] J. Aguilar and J. Altamiranda, Conceptos sobre Minería Web, *Revista GTI*, **3** (2004), no. 7, 71-77.

[12] K. Guo, L. Shi, W. Ye, X. Li, A survey of Internet public opinion mining, *2014 IEEE International Conference on Progress in Informatics and Computing*, (2014), 173-179. https://doi.org/10.1109/pic.2014.6972319

[13] M. Maragoudakis, E. Loukis, Y. Charalabidis, A Review of Opinion Mining Methods for Analyzing Citizens' Contributions in Public Policy Debate, Chapter in *Electronic Participation*, Vol. 6847, Springer, 2011, 298-313. https://doi.org/10.1007/978-3-642-23333-3_26

[14] E. Hridoy, M. Ekram, M. Samiul, F. Ahmed, R. Rahman, Localized twitter opinion mining using sentiment analysis, *Decision Analytics*, **2** (2015), no. 8. https://doi.org/10.1186/s40165-015-0016-4

[15] D. Kim, J. Kim, Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power, *Advanced Science and Technology Letters*, **51** (2014), 224-228. https://doi.org/10.14257/astl.2014.51.51

[16] J. Aguilar, O. Téran, H. Sánchez, J. Gutiérrez de Mesa, J. Cordero, D. Chávez, Towards a Fuzzy Cognitive Map for Opinion Mining, *Procedia Computer Science*, **108** (2017), 2522-2526. https://doi.org/10.1016/j.procs.2017.05.287

[17] O. Téran, J. Aguilar, Social media and free knowledge: Case study: public opinion formation, Chapter in *Societal Benefits of Freely Accessible Technologies and Knowledge Resources*, Advances in Knowledge Acquisition, Transfer, and Management Series, IGI GLOBAL, 2015 156-190. https://doi.org/10.4018/978-1-4666-8336-5.ch007

Received: August 12, 2017; Published: October 8, 2017