Conceptual Representation Using WordNet for Text Categorization

H. Nezreg, H. Lehbab, and H. Belbachir

***Abstract*—The explosion of textual documents on the Web (journals, social networks, blogs) generated a need to treat this mass of data to extract knowledge. We are interested in this paper in a particular treatment which is text categorization. We propose a conceptual representation of texts by using Wordnet for documents categorization. This representation is based on terms disambiguation by using Wordnet concepts. Disambiguated concepts are extracted from representative terms of a document, and three representations (terms, concepts, terms+ concepts) are applied with three training algorithms: SVM, Decision trees, KNN for the categorization. Experiments were applied on two corpora: 11 categories of reuters-21578 articles and 7 categories of 20 newsgroup discussion documents. The use of (terms+ concepts) gave better results for the three training algorithms and especially for the decision trees.**

***Index Terms*—Text mining, categorization, classification, wordnet.**

I. Introduction

Actually, the Web contains billions of textual documents, and there are different ways to share text using social networks and blogs. These texts need a treatment to improve different services proposed to users. One of these treatments is text categorization, which consist of assigning one ore several categories to a text according to its content [1]. Text categorization process operates on two fundamental steps, texts representation and classification as shown in the Fig. 1.

In our work, the text categorization consists of assigning a value of the set **{**0, 1**}**, to each pair where $D=\{d_1,d_2,...,d_i\}$ is the set of documents, and $C=\{c_1,c_2,...,c_j\}$ is the set of categories.

Nevertheless, it is necessary to take into account the text representation problem. Methods of texts representation are extremely important and determine the success or failure of any method of classification or categorization. Therefore, we have to choose a method that determines at best the document content, to be able to classify it in an adequate category. Bag of Words and N-grams are texts representation methods that lack of semantic and can reduce the precision of the classifiers. Using a concept based representation may increase text representation semantic, which leads to a better interpretation and thus a better rate of classification. This representation also makes possible to define the exact sense of an ambiguous word, by using the constructed hierarchy from the concepts and the various relations between them, referring to an external resource of knowledge such as Wordnet.

We propose in this paper to use the disambiguation method of [2] by bringing some improvements for texts representation with an aim of categorization. The major problem of text categorization is the semantic extraction from the text, knowing that the membership of a document to a category is closely related to the meaning of the text, in addition the nature of texts influences significantly the difficulty of the classification task for example: direct style of newspaper articles, varied vocabulary of literary coprus and a characteristic vocabulary of scientific texts, which makes classification task more difficult [3].

This paper contains five Sections, in Section I, we present a state of the art of the use of Wordnet for textual documents classification, and Section II describes the proposed approach for text representation. The third Section describes the training algorithms used for the categorization. Finally Section IV includes the evaluation and results part followed by the conclusion and possible future works in Section V.

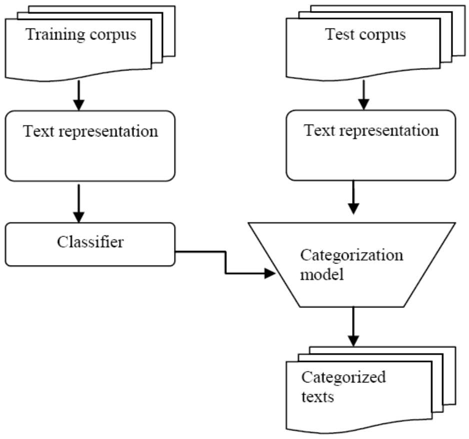

Fig. 1. General diagram of automatic text categorization.

II. State of the Art

Many studies have recently been applied to enrich the textual representation for different applications including text classification, In [4], WordNet has been used to improve the document classification by improving Rocchio algorithm. Their method was supervised and required manual annotation of term vectors. The method of Khan [5] is based on the notion of ontology region and semantic distance between concepts to attach words to concepts. In [6], WordNet has been used for documents clustering. They used Wordnet Synsets to enhance the documents representation, but without word sense disambiguation, results showed no improvement

Manuscript received November 2, 2013; revised January 23, 2014.

The authors are with the “LSSD” Laboratory and the Computer Science Department, Faculty of Mathematics and Computer Science of the University of Sciences and Technology of Oran Mohammed Boudiaf, Algeria (e-mail: nezreg.houaria@univ-usto.dz, lehbabhayet@yahoo.fr, h_belbach@yahoo.fr).

DOI: 10.7763/IJCCE.2014.V3.286

with the only use of synsets. In [2] a more general model of Information Retrieval is proposed based on concepts, which represent documents and queries as sub-trees of concepts extracted from ontology. Documents representations and the queries are not only sets of concepts appearing in their content, but they are also complemented by intermediate concepts, results showed that the conceptual approach improves the results in replies of the IRS (Information retrieval system). In [7] The proposed method extract generic concepts of WordNet for all terms in the text, and then combines them with document terms in different ways to form a new representing vector for each document. This approach was tested using two methods of similarity (chi 2, cosine distance) and gave satisfactory results. In [8] an approach for documents clustering extracts key terms ofrom all documents and the initial representation of all documents is enhanced by the use of hypernyms of WordNet to exploit semantic relations between terms.

III. Proposed Approach for Text Conceptual Representation

Wordnet is a lexical database developped by linguists of the Cognitive Science Laboratory of Princeton University [9] in the aim of indexing, classifying and relating in various ways the lexical and semantic content of the English language.

Wordnet groups English words (nouns, verbs, adjectives, adverbs) into sets of synonyms called “Synset”, each one expresses a distinct concept (see Table I). Synsets are related between them by lexical and semantic conceptual relations. The resulting network of related words and concepts can be navigated with the browser.

Each synset is linked to another synset with conceptual relationships. The relationship most often found on WordNet between synsets is hypernymy, hyponymy or an IS-A relationship. This relationship connects the synsets that are more general to those more specific. All nominal hierarchies finally back to the root node (entity).

TABLE I: Lexical Categories and Their Corresponding Semantic Relationships.

Semantic relation	Lexical category	Examples
Synonymy	N, V, adj, adv	Horse-knight/ remember-reward/ happy-euphoric/ rapidly-speedly
Antonymy	Adj, adv	Wet-dry/ powerfull-powerless
IS - A	N	Car-motor vehicule
component-Composed	N	Oxygen-air/ car-air bag
Troponymy	V	March-walk
sequential position	V	Divorce-marry

The synonym sets are associated by semantic relations: hyponymy-hypernymy (is-a), antonymy (relation between sets of words which have opposite meaning) etc…

We use Wordnet for enhancing text representation with disambiguated concepts. We present the method used for the conceptual representation of textual documents, based on the disambiguation method of [1].

A. Steps of Documents Representation in the Proposed Approach

Documents representation goes by the following steps:

1) Removing special characters and punctuation for each document.
2) Annotation of extracted terms according to their grammatical category using the tool “TreeTagger” [10].
3) Extraction of multi-word concepts with a maximum size of 5 (including five (05) maximum words) that represent inputs in WordNet (Fig. 2).
   In the example of Fig. 2 it retains the multi word "u_s_ house" and continues treatment from "Agriculture Committee approved Proposals to" and so on.

5 terms:	**u_s_ house agriculture committee approve proposal** → no synset
4 terms:	**u_s_ house agriculture committee approve**→ no synset
3 terms:	**u_s_ house agriculture committee** → no synset
2 terms:	**u_s_ house agriculture**→ no synset
1 term:	**u_s_ house** → synset in wordnet,

Fig. 2. Example of multi-word concepts extraction.

4) Terms weight: The results of this step are simple and multi-word concepts labeled with TreeTagger [10], which will be used to enrich the representing terms vectors of each document. When concepts are extracted from the document using WordNet, selected concepts are weighted according to a variant TF.IDF noted CF.IDF [1]:

$$cf.idf\left(c\right) = count\left(c\right) + \sum_{sc \in sub\_concepts(c)} \frac{length\left(sc\right)}{length\left(c\right)} \times count\left(sc\right) \quad (1)$$

Where **Length (c)** represents the number of words in the concept **c** and **sub_concept (c)** all concepts derived from **c**.

5) Selecting important concepts: To select the important concepts we defined a threshold ***th*=*2***, which is not very high for a sufficient number of concepts that reflect document semantic.

With that we are looking for strong links between terms senses of the text, by reducing the number of relationships and respecting the lexical categories comparing to [1] without reducing the disambiguation semantic.

We use the following relations:

- Definition: Is the definition of a concept,
- Hyperonymy (Hypernyms): hyperonyms Class contain fathers concepts for a generalization relationship.
- Hyponymy (Hyponyms): Is the inverse relationship of hyperonymy.
- Entailments: the class of inferences of the verb, which may be involved from the verb.
- Outcomes: the verb results class.
- Attributes: the class of concepts where the adjective is an attribute.

- Relative adjectives (related) in the concept.
- Topics related to the concept.
- Pertainyms: the class of adjectives which the adverb is derived.

TABLE II: RELATIONSHIPS ACCORDING TO THEIR GRAMMATICAL CATEGORY

Grammatical category	Relationships
Nouns	Hypernyms , hyponyms
Verbs	Hypernyms, entailment, outcomes, hyponyms(troponyms)
Adjectives	Attribut, related, similar, topics
Adverbs	Pertainyms ,topics

We use labels obtained by TreeTagger to define grammatical categories of concepts terms, then, for each concept term we extract the different relationships that exist in its grammatical category according to Table II.

The concept choice of an ambiguous term is determined by calculating the similarity between the important concepts two by two, by summing the intersection of the results of WordNet relationships applied to these concepts as follows:

Given a set of the ontology relations $R=\{R_{1,}R_2...,R_n\}$, and two concepts C_k et C_l , assigned to them two senses J1 and J2 : $S_{j_1}{}^k$ and $S_{j_2}{}^l$. The semantic similarity between $S_{j_1}{}^k$ and $S_{j_2}{}^l$ noted $P_l^k\left(S_{j_1}{}^k, S_{j_2}{}^l\right)$ is defined as follows [1]:

$$P_l^k\left(S_{j_1}{}^k, S_{j_2}{}^l\right) = \sum_{(i,j)\in\{1,\dots,n\}} R_i\left(S_{j_1}{}^k\right)\cap R_j\left(S_{j_2}{}^l\right) \qquad (2)$$

This similarity is the intersection of the number of words in common between the informations returned by the relations R_i when they are applied to concept-sense S_{j1}^k and S_{j2}^l.

A score C_score is calculated for each concept-sense, which is equal to the sum of all similarity measures with other concepts-sense except those who are in the same set of senses of the concept-sense[1]:

$$C_{score}\left(S_k{}^i\right) = \sum_{l\in[1..m],\, l\neq i,\, j\in[1..n]} P_{i,l}\left(S_k{}^i, S_j{}^l\right) \qquad (3)$$

For a concept it represents the score of the Kth sense, where m is the number of concepts of D_t , the concept-sense that maximize C_score is chosen as the best concept-sense that represent at best the sense of the concept[1]:

$$Best_{score}\left(C_i\right) = Max_{k=1..n}\, C_{score}\left(S_k{}^i\right) \qquad (4)$$

The selected concept-sense can disambiguate the concept C_i.

The document is represented by their disambiguated keys concepts. After disambiguation, we obtain the set of concepts-sense disambiguated with their weight in the document (CF-IDF calculated previously) (see Table III).

TABLE III: EXAMPLE OF SELECTED CONCEPTS AFTER DISAMBIGUATION

Concept	CF-IDF
a vehicle carry many passenger use for public transport	3.912023005428146
available for purchase	4.605170185988092
a university in Philadelphia Pennsylvania	4.605170185988092
a ticket good for a ride on a bus	4.605170185988092

IV. TEXTS CATEGORIZATION

Features and their frequencies are extracted from documents, and then weighted vectors $d_i=\{w_{i1},w_{i2},...,w_{ij}\}$ are obtained for representing each document, this in three cases which are represented as follows:

Case 1: Representation based on concepts: We gather all the concepts to build a concepts dictionary. For each document where the concept appears, it will be represented by its weight (w_i = CF-IDF), otherwise its weight is zero.

Case 2: Representation based on terms: We represent each document by terms, their weights are calculated by the formula TF-IDF s follows:

$$W_i = tf.idf\left(T,d\right) = N\times\log\left(T/T'\right) \qquad (5)$$

where *N*: the number of occurrences of a term t in the text, T: total number of texts in the corpus, T ': the number of texts in which the term t appears at least once.

Case 3: Representation based on the terms and concepts: To avoid losing the contribution of the concepts in the text representation, we set a threshold for selected terms defined as follows:

We collect all terms that appear in the documents of each category. If a term belongs to more than S categories, it will be removed from the representation (**S = 4**, a term that belongs to more than 4 categories may not represent at best the semantic of the document).

Each document is represented by terms and concepts weights that appeared in it (TF-IDF for terms, CF-IDF for concepts).

For text categorization we used the support vector machines, decision trees and k-nearest neighbors. These learning algorithms take as input feature vectors mentioned above. We consider the weight of each term as a feature of the document.

V. EVALUATION AND RESULTS

We evaluated our work on two corpora: Reuters-21578 and 20 Newsgroups, each document of both corpus have been treated for stopwords removing and stemming using the TreeTagger [10].

For the Reuters corpus we took 11 categories "Acquisition, balance of payment, crude, dollar, housing, interest, industrial production, jobs, reserves, retail, and trade" with an average of 30 documents for each category with a total of 307 documents for learning. For the test 86 documents were used

(10 documents per category). And the second corpus, we used 7 of the 20 categories that compose the corpus “Composant.system.IBM, Composant.system.MAC, for sale, science.electronics, science.medical". We used the formula of precision for evaluating the classification:

$$\Pr ecison = \frac{number\ of\ documents\ categorized\ correctly}{number\ of\ all\ documents} \quad (6)$$

Table IV summarizes the results obtained with each representation (Term, Concepts, Concepts+Terms) using the following learning algorithms (SVM, Decision trees and K-nearest neighbors).

The only use of concepts was not an improvement for text categorization (a precision from 31% to 54%) compared to the representation based only on terms (the precison from 34% to 69%). However, experiments on the 20 newsgroups corpus show that the concepts have made an improvement (from 22% to 44%) compared to the representation based on terms (15% -32%); this can be explained by the style of the vocabulary in this two different corpus. The enrichment of the representation based on terms with concepts gave a good contribution to the categorization and for both corpus. We note that the precision reaches its maximum with the use of decision trees (Reuters: 74.41% 20Newgroups: 55.71%).

TABLE IV: COMPARISON OF THE THREE REPRESENTATIONS (WORDS, CONCEPTS, TERMS AND CONCEPTS) FOR TEXT CATEGORIZATION

	SVM		**Arbre de décision**		**Kppv**	
	Reuters	**20news Groups**	**Reuters**	**20news Groups**	**Reuters**	**20news Groups**
Termes	34,88%	27,14%	69,76%	32,86%	59,30%	15,71%
Concepts	31,39%	25,71%	54,65%	44,29%	40,69%	22,86%
Termes +concepts	39,53%	21,42%	**74,41%**	**55,71%**	46,51 %	21,42%

VI. CONCLUSION AND FUTURE WORKS

In this paper we have presented a conceptual representation approach based on [1] using a WordNet concepts disambiguation ontology in order to improve the process of text categorization. Decision trees give better results than SVM and k-NN for their semantic aspect in classification. For the 20 newsgroups corpus we see an improvement of 22.85% and 4.65% for the Reuters-21578 corpus, the use of other classification methods such as association rules that could gave good results by giving more semantic to the learning phase. It remains improving these results for a better precision, while maintaining the conceptual representation and test this approach on Web documents that are longer and therefore more semantically rich.

REFERENCES

[1] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[2] M. Baziz, M. Boughanem, and N. A. Gilles, “A conceptual indexing approach based on document content representation,” in *Proc. Fifth International Conference on Conceptions of Libraries and Information Science*, Glasgow, UK, 2005. pp. 171-186.

[3] S. Réhel, “Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés, ” 2005.

[4] M. D. B. Rodriguez, J. M. G. Hidalgo, and B. D. Agudo, “Using WordNet to complement training information in ttext categorization,” *Computation and Language*, Sep. 1997.

[5] L. R. Khan. “Ontology-based information selection,” PhD dissertation, Faculty of the Graduate School, University of Southern California. August, 2000.

[6] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” in *Proc. The 12th International Conference on World Wide Web*, 2003, pp. 519--528.

[7] Z. Elberrichi, A. Rahmoun, and M. A. Bentaalah, “Using WordNet for text categorization,” *The International Arab Journal of Information Technology*, vol. 5, no. 1, January, 2008.

[8] C. L. Chen, F. S. C. Tseng, and T. Liang “An integration of WordNet and fuzzy association rule mining for multi-label document clusters,” *Data & Knowledge Engineering*, vol. 69, no. 11, pp. 1208-1226, 2010.

[9] G. Miller, “Nouns in WordNet: a lexical inheritance system,” *International Journal of Lexicography*, vol. 3, no. 4, 1990.

[10] G. Schmid. (1994). Tree tagger–a language independent part-of-speech tagger. Manuscript. [Online]. Available: http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html

H. Nezreg was born in Oran, Algeria, on December 13, 1987, she is a PhD student in computer science at the Faculty of Mathematics and Computer Science of the University of Sciences and Technology of Oran “USTO-MB” and belongs to the Laboratory of Signals, Systems and Data “LSSD”, she get her master degree in 2010 on information systems engineering and now her researches are focused on information retrieval, and especially text mining.

H. Lehbab was born in Oran, Algeria, on February 1, 1990, she is a PhD student at the University of Sciences and Technology of Oran “USTO-MB” and belongs to the: Laboratory of Signals, Systems and Data “LSSD”. Now her researches are focused on combinatorial optimization and multi agent systems.

H. Belbachir is a professor in computer science at the Faculty of Mathematics and Computer Science of the University of Sciences and Technology of Oran “USTO-MB” she is a member of the “LSSD” Laboratory”