Conceptualization Effects on MEDLINE Documents Classification Using Rocchio Method

Shereen Albitar
Aix Marseille University LSIS UMR CNRS 7296
Domaine Universitaire de St Jérôme
13397 Marseille Cedex 20, France
shereen.albitar@lsis.org

Sebastien Fournier
Aix Marseille University LSIS UMR CNRS 7296
Domaine Universitaire de St Jérôme
13397 Marseille Cedex 20, France
sebastien.fournier@lsis.org

Bernard Espinasse
Aix Marseille University LSIS UMR CNRS 7296
Domaine Universitaire de St Jérôme
13397 Marseille Cedex 20, France
bernard.espinasse@lsis.org

***Abstract*- The aim of this paper is to propose a supervised text classification method for the biomedical domain using semantic resources. We choose the traditional text classification method, Rocchio, for its scalability and extendibility with semantic knowledge. This paper proposes to integrate semantic aspects into Rocchio through a conceptualization task. This conceptualization is realized by mapping terms that are extracted from text to their corresponding concepts in the UMLS® Metathesaurus® in order to take meaning into consideration during text classification. The proposed classifier is tested on the Ohsumed text corpus, which is composed of abstracts of biomedical articles retrieved from the MEDLINE® database. The effects of Conceptualization on Rocchio's performance are discussed according to different standard similarity measures and to a variety of conceptualization strategies.**

Keywords-Text Classification, Semantic classification, Information retrieval, Rocchio, Similarity measures, conceptualization.

I. INTRODUCTION

Nowadays and due to the explosive increase in published information, existing search engines seem to be unable to respond effectively to users' requests. This is often related to the traditional keyword-based indexing techniques that neglect search context [1]. Aiming at more effective and less time expensive search, it seems adequate to involve classification techniques in order to consider the contents of answers provided by search engines, applying thorough filtering and ranking.

Text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc.

Generally, text classification methods use syntactical and statistical models for text document representation. This applies to the most popular text classification methods such as: Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and K Nearest Neighbor (KNN). These models suffer the lack of sense in resulting representations, ignoring all semantics that reside in the original text that can help in text classification.

Vector-based (binary or TF/IDF) representation used by preceding methods permits semantic integration or "Conceptualization" that enriches document representation model using a certain background knowledge base [2, 3]. In addition, both KNN and Rocchio enable involving knowledge bases in decision making through semantic similarity functions [4].

In this work, we mainly focus on improving text classification in the biomedical domain using domain specific knowledge bases, particularly UMLS®. Many works have already tried to improve biomedical text representation for better classification using UMLS® [5, 6] or MeSH® [7, 8]. To the best of our knowledge, few works investigated in details the gain of integrating these semantic resources in classifying biomedical text.

Making a comparative study on the different traditional classification methods is out of the scope of this paper, for detailed comparisons please refer to [9]. We consider Rocchio an adequate baseline text classifier for its scalability and its extendibility with semantic resources at both levels: text representation and similarity calculation for decision making. Most of other traditional classification methods, such as SVM and NB, allow the integration of semantics essentially in text representation. Nevertheless, deploying semantic similarity functions allows a full exploitation of semantic resources (concept properties, relations between concepts, etc.) during decision making.

Next section proposes appropriate semantic solutions in order to improve text representation and classification. In third section, we apply Rocchio method to conceptualized Ohsumed corpus [10] using UMLS® (Unified Medical Language System) Metathesaurus® and MetaMap® tool. The forth section presents some preliminary results using these conceptualized documents. Conceptualization effects on Rocchio's classification according to different standard similarity measures and conceptualization strategies are discussed. Finally, we conclude with an assessment of our work, followed by research perspectives.

II. TOWARDS A ROCCHIO-BASED SEMANTIC CLASSIFICATION

In spite of being considered the most popular text representation method, VSM representation suffers from certain limitations [11, 12] especially for processing composed words, synonyms, polysemy, etc. In order to overcome these limitations, semantic resources (like thesaurus & ontologies) can be used to replace term-based

representation by concept-based one through "conceptualization". As an example of semantic resources that might be used for conceptualization: WordNet, Wikipedia and other domain specific resources usually called domain ontologies such as UMLS® thesaurus in the medical domain.

In general, vector conceptualization is realized in two steps: *(i)* search for corresponding concepts related to vector's terms and then *(ii)* the integration of these concepts in the vector producing the final conceptualized vector. Three different strategies have been proposed for conceptualization [2]: *(i) Adding Concepts*: Where the original vector is extended and corresponding concepts are added. *(ii) Partial Conceptualization*: Where terms are substituted by corresponding concepts. Terms having no related concepts are kept in the vector. *(iii) Complete Conceptualization*: Similarly to Partial Conceptualization, terms are substituted by concepts whereas remaining terms are eliminated from the final vector. The integrated concepts are assigned scores derived from related terms' frequencies.

While searching concepts that correspond to a polysemic term in semantic resources, multiple matches are detected. This introduces some ambiguities in document representation. For example: the term "Book" signifies in English a book and also a reservation (Ticket, accommodations…). According to [2], three strategies for disambiguation deploying WordNet can be used: *(i) All*: accept all candidate concepts as matches for the considered term. *(ii) First*: Accept the most frequently used concept among candidates using document language statistics. *(iii) Context*: Accept the candidates having the most similar semantic context compared to the term's context in text.

III. APPLYING ROCCHIO TO CONCEPTUALIZED CORPUS

This section presents details concerning the integration of conceptualization process in the Rocchio classifier (figure 1). Two resources, a knowledge base and a text-to-concept mapping utility, are needed in order to transform a plain text into a conceptualized one. We chose to use the UMLS® Metathesaurus® and the MetaMap® tool for mapping Ohsumed text to UMLS® concepts.

After introducing these resources, we present the new classification process that integrates a conceptualization task using these resources. Finally we introduce the different conceptualization strategies used in our experimentations.

A. Resources

In this section we present UMLS® and MetaMap® that are both used in our system during the conceptualization process. Both were developed by the National Library of Medicine (NLM) aiming at facilitating the development of sophisticated medical information systems.

1) Unified Medical Language System® (UMLS®)

The Unified Medical Language System® (UMLS®) [13] was developed to model the language of biomedicine and health. UMLS' sources enhance the development of information systems in the biomedical domain. The UMLS® knowledge base consists of three main resources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus is a multilingual vocabulary database of biomedical concepts, their names, their attributes and the relations among them. This database organizes concepts of various source vocabularies (like MeSH®, SNOMED-CT®, etc.) according to their senses grouping common concepts together. Concepts and relations among them are assigned at least one type from the Semantic Network. Indeed, the Semantic Network provides a higher level of abstraction through concept and relation categorization in inter-related Types constituting a network of 133 semantic types and 54 relationships. The SPECIALIST lexicon contains a large variety of general terms as well as medical terms and words. Text conceptualization using UMLS® as a semantic resource, if compared to other more generic semantic resources such as WordNet, is more relevant in the medical domain. This allows the mapping between text and the related specific concepts in UMLS®.

2) MetaMap® tool

In addition to the UMLS® knowledge resources, many tools are developed and provided by NLM in order to facilitate deploying these sources by medical information system developers. In this work we are interested particularly in MetaMap® [14]. The major goal of MetaMap's developers was to improve biomedical text retrieval using UMLS® sources. Indeed, MetaMap® can discover the links between biomedical text and the knowledge in the Metathesaurus. This mapping is the result of a rigorous linguistic analysis of each phrase of the text: First the text is tokenized and phrase boundaries are identified, then part of-speech-tags are added. Second, the Specialist lexicon and the shallow parser are used to analyze these phrases syntactically. Finally, different candidates are identified in the Metathesaurus and then final mappings combining these candidates are evaluated resulting in confidence scores for each mapping. In cases where ambiguities are detected, MetaMap® keeps the mappings that are the most semantically similar to the surrounding text following the context strategy (section 2).

B. Classification Process

Figure 1 illustrates the classification process that is realized through two phases: Training Phase and Classification Phase. The *Conceptualization* task is introduced before the *Vectorization* task using both MetaMap® and UMLS®.

During the training phase, all documents of the training corpus are conceptualized using UMLS® by means of MetaMap®; the system transforms all training corpus into conceptualized text documents. This *Conceptualization task* is realized on text documents directly and not on their vectors in order to keep all information in text. This information is needed by MetaMap® for realizing a better text to concept mappings. Inspired by vector conceptualization strategies described in the previous section, text conceptualization can also be done using several strategies. Then, the *Vectorization task* converts these conceptualized documents into vectors according to the well known VSM. Finally the *Rocchio Centroïd Computing task* computes classes' centroïds using vectors resulting from conceptualized training corpus vectorization.

During the classification phase, the new document to classify is first conceptualized and then vectorized through preceding *Conceptualization* and *Vectorization* tasks. Then, through a *Similarity Computing* task, the resulting class is the one having the most similar centroïd according to one of the five similarity measures (Cosine, Jaccard, Kullback-Leibler, Levenshtein and Pearson).

C. Conceptualization step

During the conceptualization task, different strategies can be implemented as described in section 2 (adding concepts, partial conceptualization and complete conceptualization). According to MetaMap® text-to-concept matching results, two strategies can be chosen: *Best concept strategy* that takes the best concept among several candidate concepts matched to the text according to a matching score computed by MetaMap® [14], or *All concepts strategy* that keeps all candidates. Candidates resulting from matching have many properties. We chose to use the *concept name* or the *concept ID*. In fact, during the tokenization step, the *concept ID* is considered as a single token so it stays intact. *Concept names*, being sometimes compound words, are cut during tokenization when applied to a conceptualized text according to concept name strategy. In this work, conceptualization is done using 12 combinations of the previous strategies. Classification step is executed using the same five similarity measures that we used in the original system without conceptualization. Next section shows a selection of the most relevant results using F-measure.

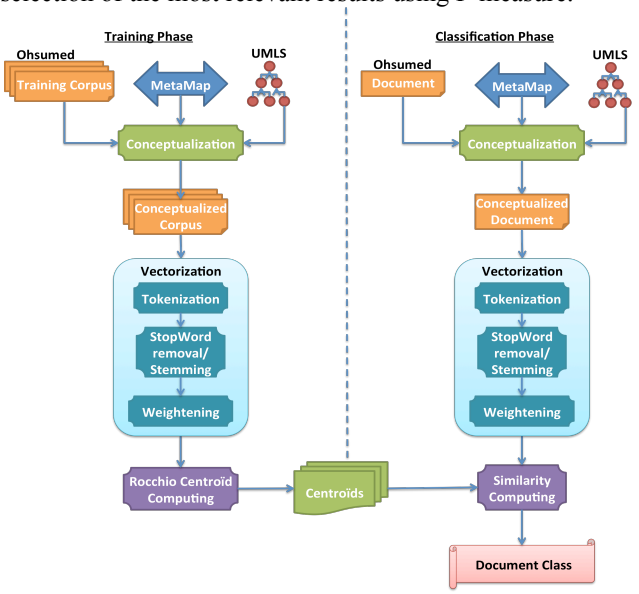

Figure 1. Conceptualization process: Rocchio applied to conceptualized corpus

IV. Experimentation and Results

These experimentations were realized in order to evaluate the effect of integrating MetaMap® into the classification process; Rocchio is tested on the Ohsumed before and after conceptualization. This integration enables mapping text to UMLS® concepts.

A. Results analysis

Table 1 shows the F1-measure obtained from applying Rocchio method to conceptualized Ohsumed corpus, for each of the five most frequent categories {C06, C23, C04, C14, C20}. These results are obtained first from using one of the five similarity measures that are also tested and compared on the original corpus without conceptualization. Then the classifier is tested with each similarity measure on the conceptualized corpus according to one of the twelve different conceptualization strategies. The two last columns present Micro and Macro averaged F-measure obtained for each similarity measure and conceptualized strategy. In Micro-averaging, F-measure is computed globally over all documents, whereas in Macro-averaging it is equal to the average of locally calculated F-measure for each class. As illustrated in the table, in most cases the original system outperforms the new system integrating conceptualization phase. In fact this applies to approximately 70% of results evaluated using the Micro-averaged F-measure. However, after a thorough look into the results, it seems clear that the system using the similarity measure of KullbackLeibler shows some amelioration. Indeed, results are improved in two thirds of the cases after conceptualization. Considering classes independently in the results, we can observe that conceptualization improves the outcome in about 70% of cases for the class "C23". The original system, showed the worst performance in treating this class. This difference is significant according to the McNemar [15] test on classification results of C23 documents before and after conceptualization ($\rho << 0.01$).

Concerning *conceptualization strategies*, one of the 12 conceptualization strategies tested in these experimentations seems to provide the maximal improvement. Indeed, *Best Concept* strategy outperforms others as it retrieves the best candidate concept provided by MetaMap® for the mapping result. *Best concept names* rather than *IDs* are used in text conceptualization. However this strategy does not present a significant improvement over the original system without conceptualization. Indeed, the gain is each time less than 1%. The highest increases are obtained when using conceptualization strategies taking into account all the candidates found by MetaMap® and not only the best. Thus, in some cases this increase exceeds 10%. Furthermore, largest increase in the Micro-averaged F-Measure values are attained using the strategy of adding Concepts applied to the classifier with KullbackLeibler similarity measure. Previous improvements seem to be significant according to McNemar [15] test having ($\rho << 0.01$).

Concerning *similarity measures*, the least improvement can be observed for the classifier using the similarity measure Levenshtein. In fact, among all Micro-averaged F-measure values for this classifier, only one surpasses its value when compared to system results using Levenshtein without conceptualization. As we noticed during experiments on the original corpus, Rocchio classifiers that use Cosine and Pearson have relatively similar behaviors. Indeed, both conceptualization strategies that add the names and the IDs of the best concepts seem to improve classification results for both similarity measures. Since the Micro-averaged F-measure's increases are in most cases less than 0.5%, this increase does not seem very significant.

Rocchio using Jaccard similarity measure outperforms the original system on conceptualized text using the best concept name according to all strategies.

Table 1. Results of applying Rocchio to conceptualized Ohsumed

System Configuration \ Category			C06		C23		C04		C14		C20		Macro		Micro	
Cosine	Original		55,61%		52,13%		69,67%		70,00%		57,33%		60,95%		61,14%	
	AddConcept	AllConcepts	51,94%	-3,68%	49,71%	-2,42%	67,10%	-2,57%	67,68%	-2,32%	53,20%	-4,12%	57,92%	-3,02%	58,58%	-2,56%
		AllConceptsID	53,68%	-1,93%	52,22%	+0,10%	68,87%	-0,80%	69,45%	-0,55%	57,79%	+0,46%	60,40%	-0,54%	60,85%	-0,29%
		BestConcept	55,74%	+0,12%	52,68%	+0,55%	69,71%	+0,03%	70,41%	+0,41%	57,07%	-0,26%	61,12%	+0,17%	61,38%	+0,24%
		BestConceptID	55,17%	-0,44%	53,35%	+1,22%	69,50%	-0,17%	69,71%	-0,29%	58,60%	+1,27%	61,27%	+0,32%	61,51%	+0,37%
	Partial	AllConcepts	50,23%	-5,38%	49,10%	-3,02%	66,26%	-3,41%	66,62%	-3,38%	52,78%	-4,55%	57,00%	-3,95%	57,75%	-3,39%
		AllConceptsID	52,89%	-2,72%	51,90%	-0,23%	68,44%	-1,23%	68,97%	-1,03%	57,70%	+0,37%	59,98%	-0,97%	60,47%	-0,67%
		BestConcept	54,36%	-1,26%	53,37%	+1,24%	69,29%	-0,38%	69,33%	-0,67%	55,90%	-1,43%	60,45%	-0,50%	60,98%	-0,16%
		BestConceptID	45,84%	-9,77%	53,38%	+1,26%	62,89%	-6,78%	61,94%	-8,06%	58,25%	+0,93%	56,46%	-4,49%	57,27%	-3,87%
	Complete	AllConcepts	50,03%	-5,58%	49,05%	-3,08%	66,29%	-3,38%	66,52%	-3,48%	52,92%	-4,40%	56,96%	-3,98%	57,71%	-3,43%
		AllConceptsID	52,68%	-2,93%	51,82%	-0,30%	68,44%	-1,23%	68,99%	-1,01%	57,63%	+0,31%	59,91%	-1,03%	60,42%	-0,72%
		BestConcept	54,38%	-1,23%	53,38%	+1,25%	69,33%	-0,34%	69,36%	-0,64%	55,99%	-1,33%	60,49%	-0,46%	61,01%	-0,13%
		BestConceptID	47,17%	-8,44%	53,05%	+0,92%	62,73%	-6,94%	62,18%	-7,82%	57,19%	-0,14%	56,46%	-4,48%	57,15%	-3,99%
Jaccard	Original		48,81%		7,82%		70,30%		67,99%		55,23%		50,03%		52,74%	
	AddConcept	AllConcepts	47,58%	-1,23%	10,99%	+3,18%	67,34%	-2,96%	64,68%	-3,31%	51,34%	-3,90%	48,39%	-1,64%	51,01%	-1,73%
		AllConceptsID	48,65%	-0,16%	11,30%	+3,48%	69,32%	-0,98%	67,32%	-0,67%	53,73%	-1,50%	50,06%	+0,03%	52,48%	-0,26%
		BestConcept	50,69%	+1,88%	8,40%	+0,59%	70,62%	+0,31%	67,52%	-0,47%	55,58%	+0,35%	50,56%	+0,53%	53,20%	+0,46%
		BestConceptID	48,26%	-0,55%	5,84%	-1,98%	69,99%	-0,31%	67,18%	-0,81%	56,63%	+1,39%	49,58%	-0,45%	52,34%	-0,40%
	Partial	AllConcepts	46,47%	-2,34%	12,33%	+4,52%	66,54%	-3,77%	63,74%	-4,25%	50,68%	-4,55%	47,95%	-2,08%	50,50%	-2,24%
		AllConceptsID	48,03%	-0,78%	12,33%	+4,51%	68,62%	-1,68%	67,04%	-0,95%	53,29%	-1,94%	49,86%	-0,17%	52,16%	-0,58%
		BestConcept	50,90%	+2,09%	8,98%	+1,16%	69,29%	-1,02%	66,86%	-1,13%	54,86%	-0,37%	50,18%	+0,15%	52,75%	+0,02%
		BestConceptID	40,30%	-8,52%	3,63%	-4,19%	60,62%	-9,69%	61,47%	-6,51%	56,54%	+1,30%	44,51%	-5,52%	46,48%	-6,26%
	Complete	AllConcepts	46,45%	-2,36%	12,25%	+4,44%	66,58%	-3,72%	63,72%	-4,27%	50,63%	-4,60%	47,92%	-2,11%	50,46%	-2,27%
		AllConceptsID	48,24%	-0,57%	12,56%	+4,75%	68,49%	-1,81%	66,93%	-1,06%	53,35%	-1,88%	49,92%	-0,11%	52,18%	-0,56%
		BestConcept	51,27%	+2,46%	8,98%	+1,17%	69,28%	-1,03%	66,65%	-1,34%	54,69%	-0,54%	50,17%	+0,14%	52,75%	+0,02%
		BestConceptID	41,23%	-7,58%	3,80%	-4,01%	60,79%	-9,51%	60,84%	-7,14%	56,73%	+1,50%	44,68%	-5,35%	46,70%	-6,03%
Kullback	Original		54,19%		18,97%		69,56%		68,92%		55,52%		53,43%		55,01%	
	AddConcept	AllConcepts	53,41%	-0,79%	27,91%	+8,94%	69,15%	-0,40%	69,65%	+0,73%	54,88%	-0,64%	55,00%	+1,57%	56,07%	+1,06%
		AllConceptsID	53,80%	-0,39%	25,86%	+6,89%	68,94%	-0,61%	69,13%	+0,21%	55,00%	-0,52%	54,55%	+1,12%	55,62%	+0,61%
		BestConcept	54,49%	+0,30%	20,63%	+1,67%	69,93%	+0,38%	68,79%	-0,13%	56,46%	+0,95%	54,06%	+0,63%	55,55%	+0,54%
		BestConceptID	53,12%	-1,07%	16,83%	-2,14%	69,52%	-0,04%	68,51%	-0,42%	55,86%	+0,35%	52,77%	-0,66%	54,61%	-0,40%
	Partial	AllConcepts	52,96%	-1,23%	29,50%	+10,54%	68,82%	-0,73%	69,13%	+0,21%	54,47%	-1,05%	54,98%	+1,55%	55,92%	+0,91%
		AllConceptsID	53,00%	-1,19%	26,68%	+7,72%	68,48%	-1,07%	68,84%	-0,08%	53,92%	-1,60%	54,19%	+0,75%	55,22%	+0,21%
		BestConcept	53,44%	-0,75%	19,02%	+0,05%	68,97%	-0,59%	68,11%	-0,81%	56,24%	+0,73%	53,16%	-0,27%	54,85%	-0,16%
		BestConceptID	45,92%	-8,27%	10,70%	-8,26%	64,12%	-5,44%	64,59%	-4,33%	54,90%	-0,62%	48,05%	-5,38%	50,14%	-4,87%
	Complete	AllConcepts	53,14%	-1,05%	29,64%	+10,67%	68,65%	-0,91%	69,03%	+0,10%	54,43%	-1,08%	54,98%	+1,55%	55,87%	+0,86%
		AllConceptsID	53,13%	-1,06%	27,30%	+8,34%	68,31%	-1,25%	68,89%	-0,03%	53,80%	-1,71%	54,29%	+0,86%	55,28%	+0,27%
		BestConcept	53,91%	-0,28%	20,75%	+1,78%	68,94%	-0,62%	68,21%	-0,71%	56,26%	+0,74%	53,61%	+0,18%	55,19%	+0,18%
		BestConceptID	46,32%	-7,87%	12,13%	-6,83%	64,25%	-5,31%	64,91%	-4,01%	54,39%	-1,12%	48,40%	-5,03%	50,46%	-4,55%
Levenshtein	Original		48,95%		4,71%		70,86%		66,55%		52,72%		48,76%		51,82%	
	AddConcept	AllConcepts	49,97%	+1,02%	7,14%	+2,43%	67,54%	-3,33%	65,86%	-0,69%	49,64%	-3,08%	48,03%	-0,73%	50,74%	+0,89%
		AllConceptsID	47,86%	-1,09%	6,70%	+1,99%	67,94%	-2,92%	66,39%	-0,16%	50,44%	-2,28%	47,87%	-0,89%	50,62%	-1,20%
		BestConcept	50,18%	+1,23%	4,88%	+0,18%	70,63%	-0,24%	65,51%	-1,04%	52,41%	-0,31%	48,72%	-0,04%	51,81%	-0,02%
		BestConceptID	48,12%	-0,83%	3,28%	-1,43%	69,98%	-0,88%	65,65%	-0,90%	54,23%	+1,52%	48,25%	-0,51%	51,31%	-0,51%
	Partial	AllConcepts	48,76%	-0,19%	8,31%	+3,61%	66,87%	-3,99%	65,61%	-0,94%	49,32%	-3,40%	47,78%	-0,98%	50,42%	-1,41%
		AllConceptsID	46,23%	-2,72%	6,78%	+2,07%	66,62%	-4,24%	65,93%	-0,62%	49,70%	-3,02%	47,05%	-1,71%	49,74%	-2,08%
		BestConcept	49,34%	+0,40%	5,04%	+0,34%	68,46%	-2,40%	64,88%	-1,67%	52,51%	-0,21%	48,05%	-0,71%	51,02%	-0,80%
		BestConceptID	39,64%	-9,30%	2,28%	-2,42%	62,03%	-8,83%	60,60%	-5,95%	54,82%	+2,11%	43,88%	-4,88%	46,13%	-5,70%
	Complete	AllConcepts	48,88%	-0,07%	8,73%	+4,02%	66,83%	-4,03%	65,63%	-0,92%	49,30%	-3,42%	47,87%	-0,88%	50,46%	-1,36%
		AllConceptsID	46,25%	-2,70%	6,86%	+2,15%	66,60%	-4,27%	65,89%	-0,66%	49,43%	-3,29%	47,00%	-1,75%	49,68%	-2,14%
		BestConcept	49,47%	+0,53%	5,13%	+0,42%	68,41%	-2,46%	64,59%	-1,96%	52,65%	-0,07%	48,05%	-0,71%	51,02%	-0,80%
		BestConceptID	40,85%	-8,10%	2,46%	-2,24%	62,27%	-8,59%	60,71%	-5,84%	54,37%	+1,65%	44,13%	-4,62%	46,51%	-5,31%
Pearson	Original		55,70%		52,60%		69,31%		69,63%		57,36%		60,92%		61,06%	
	AddConcept	AllConcepts	52,21%	-3,49%	49,93%	-2,67%	67,03%	-2,28%	67,56%	-2,07%	52,94%	-4,42%	57,93%	-2,99%	58,59%	-2,46%
		AllConceptsID	53,70%	-2,00%	52,77%	+0,17%	68,88%	-0,43%	69,00%	-0,64%	57,87%	+0,51%	60,44%	-0,48%	60,87%	-0,19%
		BestConcept	55,00%	-0,71%	52,94%	+0,34%	69,51%	+0,20%	69,92%	+0,29%	57,14%	-0,22%	60,90%	-0,02%	61,19%	+0,13%
		BestConceptID	55,26%	-0,45%	53,44%	+0,84%	69,20%	-0,11%	69,31%	-0,32%	58,46%	+1,09%	61,13%	+0,21%	61,32%	+0,26%
	Partial	AllConcepts	50,33%	-5,38%	49,07%	-3,54%	66,35%	-2,96%	66,47%	-3,16%	52,74%	-4,63%	56,99%	-3,93%	57,73%	-3,33%
		AllConceptsID	52,31%	-3,40%	52,35%	-0,25%	68,34%	-0,97%	68,59%	-1,04%	57,62%	+0,25%	59,84%	-1,08%	60,37%	-0,69%
		BestConcept	54,59%	-1,12%	53,53%	+0,93%	69,16%	-0,15%	68,92%	-0,72%	55,82%	-1,55%	60,40%	-0,52%	59,47%	-1,59%
		BestConceptID	45,97%	-9,73%	53,74%	+1,14%	62,51%	-6,80%	61,07%	-8,57%	57,51%	+0,15%	56,16%	-4,76%	57,01%	-4,05%
	Complete	AllConcepts	50,20%	-5,51%	49,04%	-3,56%	66,31%	-3,00%	66,57%	-3,06%	52,70%	-4,66%	56,96%	-3,96%	57,71%	-3,35%
		AllConceptsID	52,13%	-3,57%	52,39%	-0,21%	68,27%	-1,04%	68,62%	-1,02%	57,66%	+0,29%	59,81%	-1,11%	60,36%	-0,70%
		BestConcept	54,14%	-1,57%	53,75%	+1,15%	69,20%	-0,11%	68,91%	-0,72%	55,69%	-1,67%	60,34%	-0,59%	60,90%	-0,16%
		BestConceptID	54,14%	-1,57%	53,75%	+1,15%	69,20%	-0,11%	68,91%	-0,72%	55,69%	-1,67%	60,34%	-0,59%	60,90%	-0,16%

B. Discussion

According to the results presented in the preceding section, here we list some remarks. First of all, lowest results are observed when terms are replaced by IDs of their corresponding concepts in UMLS®. This performance degradation might be principally related to replacing all terms corresponding to a concept by its ID; only the IDs of

concepts can participate in vectorization. Terms that are shared among concept with different IDs are excluded from vectors even if they had an important influence on results.

Second, when the system performance has a good F1-measure value (i.e. exceeds 60%), no significant effect can be observed for the integration of the conceptualization task into the system. In fact, as the same similarity measures are used for both cases with/without conceptualization, results' improvements were limited.

Third, when the system performance using a specific similarity measure has a low F1-measure value, as it the case for the class "C23", introducing conceptualization can significantly improve this value with a maximum gain reaching (10%) in some cases. Indeed, the class "C23" is very large compared to others and so enriching class representation by semantics might result in a better identification of this class and also in better results.

Fourth, the best conceptualization strategy is *Add concept* adding the *Best concept* among mapped candidates into the text. In fact, best mappings retrieved by MetaMap® are added into text in order to enrich text representation with semantics avoiding any information loss.

Finally, even if the results are still preliminary, it seems useful to introduce semantic enrichments to the Rocchio method in order to ameliorate classification results. Nevertheless, the exploitation of semantic resources was limited in this work ignoring all relations (like Subsumption and Transversal relations) among concepts identified during the conceptualization task. Thus, it seems adequate to deploy these relations during classification through the introduction of new semantic similarity measures which can be coupled with traditional similarity measures.

V. CONCLUSION AND PERSPECTIVES

Due to the explosive growth of published data, many search engines demonstrate limited performance to meet the needs of users. This leads to a challenging need for effective filtering and ranking techniques. This paper concerns text classification in the biomedical domain, demonstrating the effects of involving semantic resources in text representation on classification effectiveness.

We choose Rocchio that demonstrates a good performance compared to its relatively minimal complexity in addition to its extendibility with semantics. Moreover, it can provide relevance feedback on classification results permitting better result understanding and potential improvements in classification. Moreover, some of its limitations could be overcome by means of semantic resources considering meaning in text classification. These extensions are promising.

Thus, in this research we have proposed to extend the text classification method, Rocchio, using semantic resources in order to improve its performance. Indeed, semantics can be integrated into Rocchio through conceptualization and also during decision making through different semantic similarity measures. Document conceptualization using knowledge bases helps to complete VSM approach with semantics. We realized some experiments using different conceptualization strategies and UMLS® on the Ohsumed corpus with standard similarity measures according to Rocchio's classification process. These experiments show in some cases considerable performance improvements. However we expect better improvements, through deploying semantic similarity measures that can be calculated by aid of semantic resources. These measures can be combined with standard similarity measures, already used in this work, permitting the development of effective semantic text classification methods. Moreover, we intend to realize comparative studies on the effects of conceptualization on other traditional classification algorithms such as SVM and NB.

VI. REFERENCES

[1] A. P. Asirvatham and K. K. Ravi, *Web page classification based on document structure*, 2001.

[2] A. Hotho, S. Staab, and G. Stumme, "Text clustering based on background knowledge," ed, 2003.

[3] E. Ferretti, M. Errecalde, and P. Rosso, "Does Semantic Information Help in the Text Categorization Task?," *Journal of Intelligent Systems,* vol. 17, pp. 91-107, 2008.

[4] A. Guisse, K. Khelif, and M. Collard, "PatClust : une plateforme pour la classification sémantique des brevets," in *IC 2009 Conference*, Hammamet, Tunisie, 2009.

[5] V. N. Garla and C. Brandt, "Ontology-guided feature engineering for clinical text classification," *J Biomed Inform,* in press.

[6] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," in *AMIA Annu Symp*, 2005, pp. 849-853.

[7] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measures on PubMed document clustering," in *12th international conference on Database systems for advanced applications*, Bangkok, Thailand, 2007, pp. 115-126.

[8] F. Camous, S. Blott, and A. F. Smeaton, "Ontology-based MEDLINE document classification," in *1st international conference on Bioinformatics research and development*, Berlin, Germany, 2007, pp. 439-452.

[9] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98,10th European Conference on Machine Learning*, Chemnitz, Germany, 1998, pp. 137–142.

[10] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: an interactive retrieval evaluation and new large test collection for research," in *17th annual international ACM SIGIR conference on Research and development in IR*, Dublin, Ireland, 1994, pp. 192-201.

[11] S. Bloehdorn and A. Hotho, "Boosting for text classification with semantic features," in *6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis*, Seattle, WA, 2006, pp. 149-166.

[12] P. Wang and C. Domeniconi, "Building semantic kernels for text classification using wikipedia," in *14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, Nevada, USA, 2008, pp. 713-721.

[13] *Unified Medical Language System (UMLS®)*. Available: http://www.nlm.nih.gov/research/umls/

[14] A. R. Aronson and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J Am Med Inform Assoc,* vol. 17, pp. 229-36, May-Jun 2010.

[15] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.,* vol. 10, pp. 1895-1923, 1998.