Combining Syntactic and Semantic Vector Space Models in the Health Domain by using a Clustering Ensemble

Flora Amato, Francesco Gargiulo, Antonino Mazzeo, Sara Romano and Carlo Sansone
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, University of Naples Federico II, Naples, Italy

Keywords: Semantic Processing, Clustering Ensemble.

Abstract: The adoption of services for automatic information management is one of the most interesting open problems in various professional and social fields. We focus on the health domain characterized by the production of huge amount of documents, in which the adoption of innovative systems for information management can significantly improve the tasks performed by the actors involved and the quality of the health services offered. In this work we propose a methodology for automatic documents categorization based on the adoption of unsupervised learning techniques. We extracted both semantic and syntactic features in order to define the vector space models and proposed the use of a clustering ensemble in order to increase the discriminative power of our approach. Results on real medical records, digitalized by means of a state-of-the-art OCR technique, demonstrated the effectiveness of the proposed approach.

1 INTRODUCTION

Nowadays the adoption of services for automatic information management is one of the most interesting open problems in various professional and social fields. In the medical domain, the adoption of computers and web technologies lead to the so-called e-Health. The e-Health aim is to enhance the way of interaction between healthcare actors (as doctors, nurses and patient) by means of the use of innovative data management technologies and exploiting information and communication technologies.

In the health domain, the information availability coming from different sources can improve the health services quality. For this purpose, the medical data should be conveniently organized. In this work we propose a methodology for automatic document categorization based on a clustering ensemble. Clustering ensemble (or clustering aggregation) is an alternative approach that combines different clustering results in order to improve the quality of the clustering solution. In general, a clustering ensemble method is composed by two steps: generation and consensus. The generation step consists on the production of the set of clusterings obtained with different clustering algorithms or the same algorithm with different parameter initialization. The consensus step represents the main challenge in the clustering ensemble algorithm. In literature there are several works that address the problem of document categorization by means of clustering ensemble techniques (Fodeh et al., 2009) (Gonzàlez and Turmo, 2008) (Domeniconi and Al-Razgan, 2009). In (Vega-Pons and Ruiz-Shulcloper, 2011) is presented a good analysis of the existing techniques of clustering ensemble method. Our approach combines results coming from the X-means clustering algorithm executed on three different vector space models which include both a syntactic and a semantic content representation of a document. The corpus on which we based our work is composed by real medical records belonging to Italian hospitals (Boccignone et al., 2008), digitalized by means of a state-of-the-art- OCR. We address the document categorization problem dealing with many aspects as the use of noisy data and the Italian language for which natural language processing tools and thesaurus are not available as the for the English language. Our experimental results suggest that the proposed methodology improves the clustering quality when applied to a dataset affected by noise. A quite similar approach has been proposed in (Fodeh et al., 2009), where the authors propose an ensemble clustering algorithm combining the statistic information of the data with the sense information from WordNet. Their method, however, is not directly applicable in our case since we have no available semantic resources (WordNet) of the medical domain in Italian. Moreover, while they use two type of features

Amato F., Gargiulo F., Mazzeo A., Romano S. and Sansone C..
Combining Syntactic and Semantic Vector Space Models in the Health Domain by using a Clustering Ensemble.
DOI: 10.5220/0004250403820385
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2013), pages 382-385
ISBN: 978-989-8565-37-2

extracted from the documents, we propose to exploit three different vector spaces.

The reminder of the paper is structured as follows: in Section 2 we explain how we built the syntactic and semantic vector space models; in Section 3 we report the proposed clustering ensemble methodology reporting some experimental results in Section 4. Finally in Section 5 some conclusion and future work are drawn.

2 A METHODOLOGY FOR BUILDING SYNTACTIC AND SEMANTIC VECTOR SPACE MODELS

In this work we adopt three vector space models that include syntactic and semantic aspects based respectively on frequencies of terms, lemmas and concepts. In order to represent our document collection in the vector space models, we extracted a set of terms from the documents corpus, and therefore, the set of synonyms corresponding to them. For this aim we adopted a semantic methodology proposed in (Amato et al., 2011) for the automatic extraction of concepts of interest.

The implemented set of procedures aiming at extracting terms, the corresponding lemmas and the associated concepts from the input documents are described in the following.

Extracting Terms (I Criterium). Starting from the input documents, by using *Text Tokenization* procedures, text is arranged into tokens, sequences of characters delimited by separators. Applying *Text Normalization* procedures,variations of the same lexical expression are reported in a unique way.

Tokenization and Normalization procedures perform a first grouping of the extracted text, introducing a partitioning scheme that establishes an equivalence class on terms. At this point we built the doc-features matrix, having a column for each term in the terms list, which contains the evaluation, for each document, of the TF-IDF value for every terms in the list. The TF-IDF values are computed taking into account both the number of occurrences of each term for every documents and the terms distribution in the whole document corpus. This matrix is considered as input for the clustering algorithm according to the I Criterium.

Extracting Lemmas (II Criterium). In order to obtain the lemmas starting from the list of relevant text, we applied the procedures of *Part-Of-Speech* (POS) *Tagging* and *Lemmatization*. These procedures aim at enriching the text with syntactical aspects, aiming at performing a second type of grouping of the words, on the basis of reduction of terms in a basic form, independently from the conjugations or declinations in which they appear. *Part-Of-Speech (POS) Tagging* consists in the assignment of a grammatical category to each lexical unit, in order to distinguish the content words representing noun, verb, adjective and adverb from the functional words, made of articles, prepositions and conjunctions, denoting not useful information.

Text Lemmatization is performed in order to reduce all the inflected forms to the respective lemma, or citation form. Lemmatization introduces a second partitioning scheme on the set of extracted terms, establishing a new equivalence class on it.

We built a doc-features matrix, having a column for each lemma in the list, which contains, for each document, the TF-IDF value of each lemma comparing in it. This value is computed considering the sum of the number of occurrences of each term that can be taken back to the same lemma appearing in the document. The lemma based doc-features matrix is considered as input for the clustering algorithm according to the II Criterium.

Extracting Concepts (III Criterium). In order to identify concepts, not all words are equally useful: some of them are semantically more relevant than others, and among these words there are lexical items weighting more than others. In order to “weight” the importance of a term in a document, we recurred to TF-IDF index.

Having the list of relevant terms, concepts are detected by relevant token sets that are semantically equivalent (synonyms, arranged in sets named synset). In order to determine the synonym relation among terms, we exploit external resources (Moscato et al., 2009) like thesaurus, codifying the relationship of synonymy among terms.

The number of occurrence of a concept in a document is given by the sum of the number of occurrences of all terms in its synonyms list that appear in the document. We built the concept based doc-features matrix, containing, for each document, the TF-IDF of every concepts comparing in it. The TF-IDF values of such matrix is then evaluated on the basis of the sum of the number of occurrences of each terms that is synonymous of the input terms, i.e. that is included in the synonyms list. The concept based doc-features matrix is considered as input for the clustering algorithm according to the III Criterium.

Once we have prepared the vector space models, we used these data for the clustering algorithms.

3 CLUSTERING ENSEMBLE

In this work we propose to combine a set of clusters for exploiting the different information levels provided by both syntactic and semantic features. As base cluster, we chose the X-means algorithm (Pelleg and Moore, 2000) that can be considered as an evolution of the standard K-means approach.

The proposed general methodology is depicted in Fig. 1.

The clustering ensemble method we used (Bagui, 2005) is shown in Fig. 2. It is composed by the following steps:

1. We considered the initial document matrix A for each criteria *Terms*, *Lemmas* and *Concepts*. A is a $n \times m$ matrix where n is the number of documents and m depends on the criteria selected and it could represent the number of *terms*, *lemmas* or *concepts*.
2. We generated $C_k, k = 1, 2, \ldots, L$ partitions of A by using the X-means algorithm. Each partition have a random number of clusters depending on the initial seed chosen.
3. We defined a co-association matrix for each partition: $M^k = m^k_{i,j}$, of dimension $n \times n$, where n is the number of documents and $k = 1, 2 \ldots, L$. The elements of the matrices M^k are calculated as:
$$m^k_{i,j} = \begin{cases} 1 & a_i = a_j \text{ (i.e. in the same cluster)}, \\ 0 & \text{otherwise.} \end{cases}$$
4. We defined a co-association final matrix obtained as $M = 1/L * \sum_{k=1}^{L} M^k$.
5. We selected a threshold σ that maximizes the adopted performance indexes (Kuncheva, 2004), and then we used an inverse function, denoted as *Clustering Evaluation* in the Fig. 2, in order to obtain the final documents partition C from M and σ.
6. We compared all the obtained results by using the Rand Index, the Normal Mutual Information (NMI) index (Kuncheva, 2004) and the number of generated clusters.

4 EXPERIMENTAL RESULTS

For the experimental campaign we used a corpus composed by real medical records belonging to four different hospital departments that was digitalized by means of a state-of-the-art OCR. As described in section 2 we built three different classes of vector space models considering respectively *terms*, *lemmas* and *concepts*, that are the doc-matrices (A). We used these vector space models to generate, for each one of them, $L = 12$ different instances of the X-means clustering algorithm. The dataset used to validate the adopted strategy is made up of 143 documents, which are scans of medical records, obtained from different Italian hospitals.

Table 1: Evaluation of the quality of the cluster solutions; the best case is highlighted in bold.

	Rand index	NMI	Number of Clusters
Average values and standard deviations over the 12 partitions			
I Criterium (Terms)	0.5523 ± 0.0035	0.1946 ± 0.0055	2.83 ± 0.88
II Criterium (Lemmas)	0.5042 ± 0.0058	0.2507 ± 0.0039	2.58 ± 0.45
III Criterium (Concepts)	0.4925 ± 0.0044	0.2344 ± 0.0026	2.50 ± 0.45
Combined Criteria			
Terms + Lemmas	0.7164	0.3234	5
Lemmas + Concepts	0.7313	0.3674	7
Terms + Concepts	**0.7324**	**0.3787**	7
Terms + Lemmas + Concepts	0.7286	0.3491	7

Since the dataset used is composed by digitalized medical records, the data used for testing present some noise. The noisy terms are discarded by means of the preprocessing phase of the semantic methodology and so the documents are represented by the terms correctly recognized by the OCR procedure. It implies that the document representation in the vector space considers a subset of the original terms that occurs in the medical records.

These records were previously organized on the basis of department membership as follows: *Cardiology*: 41 documents; *Intensive Case*: 40 documents; *General Surgery*: 40 documents; *Oncology*: 22 documents. On the three doc-feature matrix A we have evaluated the proposed approach, obtaining the results reported in the Table 1.

Although the use of the III Criterium allows us to make documents' partition by topic, it introduces noise, making the partitions generated worse than the ones obtained by using only the I or II Criterium. On the other hand, the usage of this information combined with the I or the II Criterium allow us to always obtain a better partition. In particular, the best results are obtained by combining features coming from the

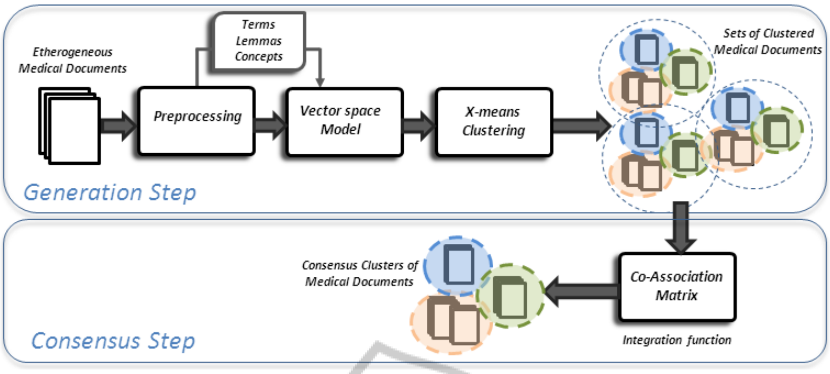

Figure 1: Document classification model.

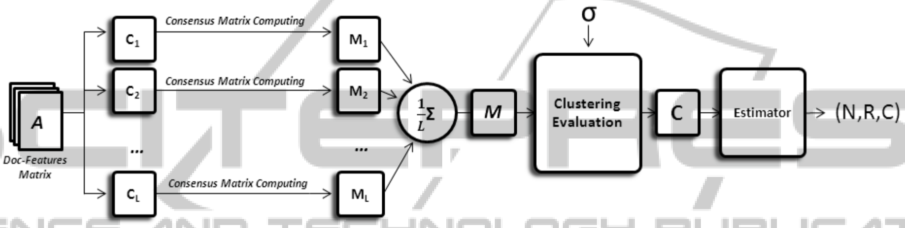

Figure 2: Generation and evaluation of the proposed clustering solution.

I Criterium and the III Criterium.

5 CONCLUSIONS AND FUTURE WORK

In this work we proposed a methodology for automatic document categorization based on a clustering ensemble technique. We combined results of different clustering algorithms executed on three different vectors space models which include both syntactic (Lemmas and Terms) and semantic (Concepts) content representation. Experiments were performed on a corpus of real medical records written in Italian. The results showed that although the use of concepts allows us to make documents' partition by topic, it introduces noise, making the generated partitions worse than the ones obtained by using only Lemmas or Terms. On the other hand, the usage of semantic information combined with the syntactical ones allowed us to improve the obtained results.

Future work will address the investigation of integrating results of different clustering algorithms as well as different document representations.

REFERENCES

Amato, F., Casola, V., Mazzocca, N., and Romano, S. (2011). A semantic-based document processing framework: a security perspective. In *Proceedings of CISIS 2011*, pages 197–202. IEEE Computer Society.

Bagui, S. (2005). Combining pattern classifiers: methods and algorithms. *Technometrics*, 47(4):517–518.

Boccignone, G., Chianese, A., Moscato, V., and Picariello, A. (2008). Context-sensitive queries for image retrieval in digital libraries. *JIIS*, 31(1):53–84.

Domeniconi, C. and Al-Razgan, M. (2009). Weighted cluster ensembles: Methods and analysis. *ACM Trans. Knowl. Discov. Data*, 2(4):17:1–17:40.

Fodeh, S. J., Punch, W. F., and Tan, P.-N. (2009). Combining statistics and semantics via ensemble model for document clustering. In *Proceedings of SAC*, pages 1446–1450, New York, NY, USA. ACM.

Gonzàlez, E. and Turmo, J. (2008). Comparing non-parametric ensemble methods for document clustering. In *Proceedings of NLDB*, pages 245–256, Berlin, Heidelberg. Springer-Verlag.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley.

Moscato, F., Di Martino, B., Venticinque, S., and Martone, A. (2009). Overfa: a collaborative framework for the semantic annotation of documents and websites. *IJWGS*, 5(1):30–45.

Pelleg, D. and Moore, A. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of ICML*, pages 727–734. Morgan Kaufmann.

Vega-Pons, S. and Ruiz-Shulclopher, J. (2011). A survey of clustering ensemble algorithms. *IJPRAI*, 25(3):337–372.