A Method of Topic Detection for Great Volume of Data

Flora Amato, Francesco Gargiulo, Antonino Mazzeo and Carlo Sansone
Dipartimento di Ingegneria Elettrica e delle Teconolgie dell'Informazione (DIETI) University of Naples Federico II,
Via Claudio 21, Naples, Italy.

Keywords: Topic Detection, Clustering, $tf-idf$, Feature Reduction.

Abstract: Topics extraction has become increasingly important due to its effectiveness in many tasks, including information filtering, information retrieval and organization of document collections in digital libraries. The Topic Detection consists to find the most significant topics within a document corpus. In this paper we explore the adoption of a methodology of feature reduction to underline the most significant topics within a *document corpus*. We used an approach based on a clustering algorithm (X-means) over the $tf-idf$ matrix calculated starting from the corpus, by which we describe the frequency of terms, represented by the columns, that occur in each document, represented by a row. To extract the topics, we build n binary problems, where n is the numbers of clusters produced by an unsupervised clustering approach and we operate a *supervised* feature selection over them considering the top features as the topic descriptors. We will show the results obtained on two different corpora. Both collections are expressed in Italian: the first collection consists of documents of the University of Naples Federico II, the second one consists in a collection of medical records.

1 INTRODUCTION

Topic extraction from text has become increasingly important due to its effectiveness in many tasks, including information retrieval, information filtering and organization of document collections in digital libraries.

In this paper we explore the adoption of a methodology of feature extraction and reduction to underline the most significant topic within a *corpus* of documents.

We used an approach based on a clustering algorithm (X-means) over the TF-IDF matrix calculated starting from the corpus.

In the proposed approach, each cluster represents a topic. We characterize each of them through a set of *words* representing the documents within a cluster.

In literature a standard method for obtaining this kind of result is the usage of a non-supervised feature reduction method, such as the Principal Component Analysis (PCA). However the high dimensionality of the feature vectors associated to each document make this method not feasible.

To overcome this problem we build n binary problems, where n is the numbers of clusters produced by the X-means. Then we operate a *supervised* feature selection over the obtained clusters, considering the cluster membership as the class and the selected top features (*words*) as the researched topic.

We exploit two different *ground truth* made by domain expert in order to evaluate the most significant feature that separates the class of interest from all the rest.

The paper is organized as follows. Section 2 describes the most recent related work for topic detection. Section 3 outlines the proposed methodology. In Section 4, we present the performed experiments, showing dataset used for experimental validation and the obtained results. Eventually, in Section 5 we discuss conclusions and possible future directions for our research.

2 RELATED WORKS

Nowadays we have available a large amount of information on web but we would access to this information in the shortest time and with the highest accuracy. To help people to integrate and organize scattered information several machine learning approaches have been proposed for the topic detection and tracking technology. Moreover Topic detection and tracking (TDT) is to study the problem of how to organize information which is based on event in natural language information flow and Topic Detection is a sub-

Amato F., Gargiulo F., Mazzeo A. and Sansone C..
A Method of Topic Detection for Great Volume of Data.
DOI: 10.5220/0005145504340439
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (KomIS-2014), pages 434-439
ISBN: 978-989-758-035-2

task of TDT. In the Topic Detection (TD) task the set of most prominent topics have been found in a collection of documents. Furthermore, in other words, TD is based on the identification of the stories in several continuous news streams which concern new or previously unidentified events. Sometimes unidentified events have been retrieved in an accumulated collection (“retrospective detection”) while in the “online detection” the events could be labelled when they have been flagged as new events in real time. Give the TD is a problem to assign labels to unlabelled data grouping together a subset of news reports with similar contents, most unsupervised learning methods, proposed in literature as (Wartena and Brussee, 2008), (Jia Zhang et al., 2011), exploit the text clustering algorithms to solve this problem.

Most common approaches, as (Wartena and Brussee, 2008), given list of topics the problem of identifying and characterizing a topic is a main part of the task. For this reason a training set or other forms of external knowledge cannot be exploited and the own information contained in the collection can be used to solve the Topic Detection problem. Moreover the method, proposed in (Wartena and Brussee, 2008), is a two-step approach: in the former a list of the most informative keywords have been extracted; the latter consists in the identification of the clusters of keywords for which a center has been defined as the representation of a topic. The authors of (Wartena and Brussee, 2008) considered topic detection without any prior knowledge of category structure or possible categories. Keywords are extracted and clustered based on different similarity measures using the induced k-bisecting clustering algorithm. They considered distance measures between words, based on their statistical distribution over a corpus of documents, in order to find a measure that yields good clustering results.

In (Bolelli and Ertekin, 2009), a generative model based on Latent Dirichlet Allocation (LDA) is proposed that integrates the temporal ordering of the documents into the generative process in an iterative fashion called Segmented Author-Topic Model (S-ATM). The document collection has been split into time segments where the discovered topics in each segment has been propagated to influence the topic discovery in the subsequent time segments. The document-topic and topic-word distributions learned by LDA describe the best topics for each document and the most descriptive words for each topic. An extension of LDA is the author-topic model (ATM). In ATM, a document is represented as a product of the mixture of topics of its authors, where each word is generated by the activation of one of the topics of the document author, but the temporal ordering is discarded. S-ATM is based on the (ATM) and extends it to integrate the temporal characteristics of the document collection into the generative process. Besides S-ATM learns author-topic and topic-word distributions for scientific publications integrating the temporal order of the documents into the generative process.

The goals in (Seo and Sycara, 2004) are: i) the system should be able to group the incoming data into a cluster of items of similar content; ii) it should report the contents of the cluster in summarized human-readable form; iii) it should be able to track events of interest in order to take advantage of developments. The proposed method has been motivated by constructive and competitive learning from neural network research. In the construction phase, it tries to find the optimal number of clusters by adding a new cluster when the intrinsic difference between the presented instance and the existing clusters is detected. Then each cluster moves toward the optimal cluster center according to the learning rate by adjusting its weight vector.

In (Song et al., 2012), a text clustering algorithm C-KMC is introduced which combined Canopy and modified k-means clustering applied to topic detection. This text clustering algorithm is based on two steps: in the former, namely C-process, has been applied Canopy clustering that split all sample points roughly into some overlapping subsets using inaccurate similarity measure method; in the latter, called K-process, has been employed a modified K-means that take X-means algorithm to generate rough clusters from the canopies which share common instance. In this algorithm, Canopies are an intermediate result which can reduce the computing cost of the second step and make it much easier to be used, although Canopy is not a completed cluster or topic.

The authors of (Zhang and Li, 2011) used vector space model (VSM) to represent topics, and then they used K-means algorithm to do a topic detection experiment. They studied how the corpus size and K-means affect this kind of topic detection performance, and then they used TDT evaluation method to assess results. The experiments proved that optimal topic detection performance based on large-scale corpus enhances by 38.38% more than topic detection based on small-scale corpus.

3 METHODOLOGY

The proposed methodology, aiming to extract and reduce the features characterizing the most significant topic within a corpus of documents, is depicted in

Fig. 1.

The considered features are computed on the basis of the *Term Frequency-Inverse Document Frequency* ($tf-idf$) matrix from the corpus.

Process implementing the methodology is composed by several steps. The first step is the evaluation of a $tf-idf$ matrix (Manning and Schütze, 1999), consisting of an $m \times n$ matrix where m is the number of documents in the collection and n is the number of *tokens* we are considering. The tokens could be Lemmas, Terms, Synonymous or more complex Lexical Structures (Amato et al., 2013b),(Amato et al., 2010),(Amato et al., 2013a).

Each row represents a document and each column represents the $tf-idf$ value calculated for each document's token.

The $tf-idf$ value is defined as follows:

$$tf-idf = tf \times idf \tag{1}$$

where tf is the number of the token occurrence in the document and idf is a measure of whether the occurrence is common or rare across all collection items.

Each rows of $tf-idf$ matrix contains a vector of real numbers which represents the projection of each document in an n-dimensional feature space, for seek of simplicity in the Figure 1(a) we represent the documents as projected in a bi-dimensional space.

To overcome the high-dimensionality feature space (more than 60.000 tokens) that makes the usage of an unsupervised feature selection method, such as *Principal Component Analysis* (PCA), not feasible, we adopted a novel solution based on a preliminary documents clustering. In the Figure 1(b) is represented the clustering process.

After that we built n binary problems where n is the number of clusters obtained. Each problem is created using a *one vs all* strategy such as for each cluster i, all the documents that belong to the cluster i were labelled as *True* otherwise all the other documents were labelled as *False*, see Figure 1(c).

In order to obtain the topics characterizing each cluster, corresponding to a binary problem, we used a supervised feature selection method on each of them, using the previous assigned labels indicating the group membership as class on which evaluate each feature. The feature selection gave in output a set of ranked features that represent the tokens that more discriminate the cluster under analysis from all the rest. These tokens figure out the searched topic.

4 RESULTS

In order to evaluate the effectiveness of this approach, we used the Weka library (Holmes et al., 1994), which is a collection of data mining algorithms.

We selected for the clustering process the X-means algorithm (Dan Pelleg, 2000). The X-means is an evolution of the K-means, the main difference between them is on the choice of the optimal number of clusters to use, while X-means set automatically this number, the K-means use a manual parameter (**K**).

More in details we configure the X-means with: 4 as *Max Number of Iteration* and 15 as *Max Number of Clusters*.

The proposed approach has been applied on two different corpora. Both collections are expressed in Italian: the first one, called **UNINA**, consists of documents of the University of Naples Federico II, the second one, called **Medical Records**, consists in a collection of medical records.

We compare the performances obtained on these two corpora in terms of *precision*, *recall* and *cluster coverage*.

$$precision = \frac{N_{out} - N_{false}}{N_{out}} \tag{2}$$

$$recall = \frac{N_{out} - N_{false}}{N_{ins}} \tag{3}$$

$$coverage = \frac{N_{out}}{N_{tot}} \tag{4}$$

where N_{out} is the number of instances in result clusters, N_{false} is the number of mistake instances in result clusters, N_{ins} is the sample size of the category under test and N_{tot} is the total number of instances.

4.1 Corpora Description

The corpus **UNINA** was originally collected from the web site of the University of Naples Federico II (Unina) and labelled by a domain expert. It consists of a total of 469 documents, divided into four categories statistically characterized in the Table 1.

The corpus **Medical Records** consists of about 5.000 medical diagnoses coming from various health

Table 1: Categories distribution for corpus **UNINA**.

Class	Number of documents
Amministrativo	17
Bandi	111
Didattica	224
Relazioni Attivitá	117
TOT.	469

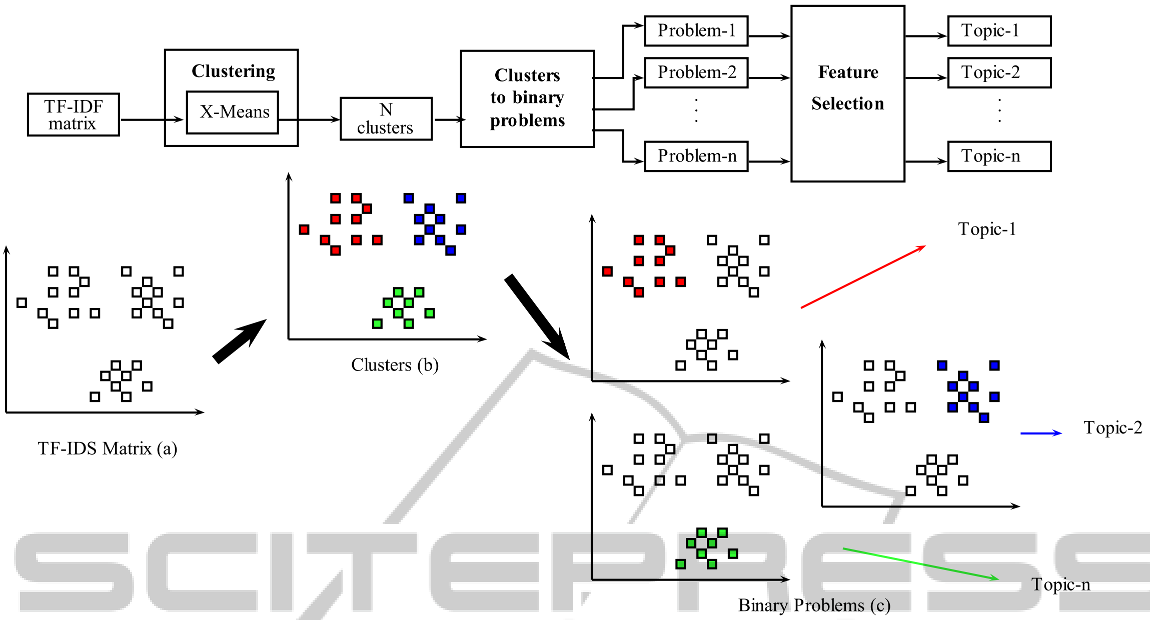

Figure 1: Main Schema.

Table 2: Categories distribution for corpus **Medical records**.

Class	Number of documents
Consulenze	96
Doppler	593
Ecoc	702
Ecografia	1171
Endoscopia	365
Intervento	346
Radiologia	1726
TOT.	4999

care organization in Campania (Italy). Since each fragment is a document produced from Medical Center, containing a description of a patient health care. For privacy issues, these documents are opportunely anonymized. The medical records are divided into seven categories detailed within the Table 2.

4.2 Experimental Results

The results of the two case studies are reported in Table 3 and Table 4. The first consideration is that not all the classes are characterized with a cluster, this fact is directly correlated to the number of documents belong to each of them.

For example, the absence of the class *Amministrativo* within the Corpus *UNINA* is due to the few documents that belongs to that class, see Table 1. For the same motivation it is possible to justify the absence of the classes *Consulenze*, *Intervento* and *Endoscopia* in the corpus *Medical Records*, see Table 2.

Table 3: Clusters assignments for Unina.

Cluster ID	Assigned category	Precision	Recall	Coverage
0	Didattica	100.0%	4.0%	1.9%
1	Didattica	100.0%	3.1%	1.5%
2	Didattica	100.0%	12.5%	6%
3	Bandi	100.0%	36.0%	8.5%
4	Relazioni Attivitá	99.0%	81.2%	20.5%
5	Didattica	41.1%	13.4%	15.6%
6	Didattica	67.5%	62.1%	43.9%
7	Didattica	100.0%	2.2%	1.1%
8	Didattica	100.0%	1.8%	0.9%
9	Didattica	100.0%	0.4%	0.2%
Average Parameters		76.4%	50.0%	100.0%

Table 4: Clusters assignments for medical records.

Cluster ID	Assigned category	Precision	Recall	Coverage
0	Ecografia	91.6%	59.8%	15.3%
1	Ecoc	99.7%	99.4%	14.0%
2	Ecografia	61.0%	35.1%	13.5%
3	Doppler	100.0%	67.3%	8.0%
4	Radiologia	41.0%	35.0%	29.4%
5	Radiologia	94.5%	12.9%	4.7%
6	Radiologia	100.0%	24.5%	8.5%
7	Radiologia	100.0%	1.0%	0.4%
8	Radiologia	100.0%	0.8%	0.3%
9	Radiologia	100.0%	17.4%	6.0%
Average Parameters		75.9%	47.22%	100.0%

Another important results is that the cluster with the highest coverage is composed in large part by noise. A possible motivation could be that the biggest cluster contains a collection of documents that are

Table 5: Topic detection for dataset *Unina*.

Cluster	Rel. Att.	Amm.	Bandi	Didat- tica	Coverage	Top features
6	3%		29%	67%	43.9%	emanato, norme, seguito, conto, u.s.r
4	99%			1%	20.5%	legittimata, odierna, giudizi, alfabetico, riunione
5	21%	23%	15%	41%	15.6%	gara, foro, cauzione, possedute, addebito
3			100%		8.5%	incombenza, destituzione, rimborsabile, risiedere, assunzione
2				100%	6%	facilitarne, aggrega, prefiggono, consecutivo, pianifica
0				100%	1.9%	regolamentazioni, oligopolio, monopolio, microeconomia, macroeconomia
1				100%	1.5%	ebbe, opzione, roffredo, lictera, spagnole
7				100%	1.1%	pneumoconiosi, 2607, extraepatiche, ards, propriocettiva
8				100%	0.9%	ril, discip, coordinatoredel, periimplantari, dental
9				100%	0.2%	anafilotossine, passivazione, overjet, overbite, otori

Table 6: Topic detection for dataset *Medical records*.

Cluster	Radio- logia	Eco- grafia	Ecoc	Consu- lenze	Endo- scopia	Doppler	Inter- vento	Coverage	Top features
4	41.0%	4.1%	0.3%	2.6%	15.3%	13.2%	23.5%	29.4%	norma, alterazioni, esame, eseguito, volume
0	8.4%	91.6%						15.3%	fegato, pancreas, milza, colecisti, reni
1			99.7%	0.3%				14.0%	sezioni, normali, aortica, sinistre, deficit
2	12.0%	61.0%		6.2%	20.8%			13.5%	tiroide, lobo, prevalente, nodulari, componente
6	100.0%							8.5%	compatibile, cardiaca, lesioni, aortica.immagine, focolaio
3						100.0%		8.0%	intimale, vasale, ispessimento, assi, succlavio
9	100.0%							6.0%	carattere, periferica, attivit , ilare, distribuzione
5	94.5%			5.5%				4.7%	ingrandita, stasi, cardiaca, piccolo, costo
7	100.0%							0.4%	prominenza, attivit .regolare, arco, periferica.marginale,
8	100.0%							0.3%	aerea, bozzatura, tracheale, emidiaframma, scoliotica

very close to each other, even if they belong to different categories. A possible solution is to not consider this cluster of documents or to use approaches to clean this data (Gargiulo and Sansone, 2010) or to use multi classification schemas (Gargiulo et al., 2013).

Within the Table 5 and Table 6 we represented the *top features* evaluated for each cluster with the percentage of documents belongs to each original category.

5 CONCLUSION

Topic extraction from documents is a challenging problem within the data mining field. The main motivation is due to its effectiveness in many tasks such as: information retrieval, information filtering and organization of documents collection in digital library.

In this paper we presented a methodology to implement an unsupervised topic detection for high dimensional datasets.

To this aim we used a preliminary clustering approach over the $tf-idf$ matrix computed starting from the corpora and we built *n* binary problems, one for each cluster obtained; we considered a supervised features selection over such problems to select the most important *features* and consequently the associated *topics*.

We showed the effectiveness of this approach on two different corpora, *UNINA* and *Medical Records*, obtaining interesting results.

As feature work we are planned to evaluate a set of distance measures to automatically figure out the *degree of belonging* between the *selected features* set and the most interesting topics set.

ACKNOWLEDGEMENTS

This work has been partially supported by the “SMART HEALTH CLUSTER OSDH SMART FSE STAYWELL” project, funded by MIUR (Ministero dellIstruzione, dellUniversità e della Ricerca).

REFERENCES

Amato, F., Casola, V., Mazzeo, A., and Romano, S. (2010). A semantic based methodology to classify and protect sensitive data in medical records. In *Information Assurance and Security (IAS), 2010 Sixth International Conference on*, pages 240–246. IEEE.

Amato, F., Casola, V., Mazzocca, N., and Romano, S. (2013a). A semantic approach for fine-grain access control of e-health documents. *Logic Journal of IGPL*, 21(4):692–701.

Amato, F., Gargiulo, F., Mazzeo, A., Romano, S., and Sansone, C. (2013b). Combining syntactic and semantic vector space models in the health domain by using a clustering ensemble. In *HEALTHINF*, pages 382–385.

Bolelli, L. and Ertekin, Giles, C. (2009). Topic and trend detection in text collections using latent dirichlet allocation. In Boughanem, M., Berrut, C., Mothe, J., and Soule-Dupuy, C., editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 776–780. Springer Berlin Heidelberg.

Dan Pelleg, A. M. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco. Morgan Kaufmann.

Gargiulo, F., Mazzariello, C., and Sansone, C. (2013). Multiple classifier systems: Theory, applications and tools. In Bianchini, M., Maggini, M., and Jain, L. C., editors, *H. on Neural Information Processing*, volume 49 of *Intelligent Systems Reference Library*, pages 335–378. Springer.

Gargiulo, F. and Sansone, C. (2010). SOCIAL: Self-organizing classifier ensemble for adversarial learning. In Gayar, N. E., Kittler, J., and Roli, F., editors, *MCS*, volume 5997 of *Lecture Notes in Computer Science*, pages 84–93. Springer.

Holmes, Donkin, A., and Witten, I. H. (1994). Weka: a machine learning workbench. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361.

Jia Zhang, I., Madduri, R., Tan, W., Deichl, K., Alexander, J., and Foster, I. (2011). Toward semantics empowered biomedical web services. In *Web Services (ICWS), 2011 IEEE International Conference on*, pages 371–378.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Seo, Y.-W. and Sycara, K. (2004). Text clustering for topic detection.

Song, Y., Du, J., and Hou, L. (2012). A topic detection approach based on multi-level clustering. In *Control Conference (CCC), 2012 31st Chinese*, pages 3834–3838. IEEE.

Wartena, C. and Brussee, R. (2008). Topic detection by clustering keywords. In *Database and Expert Systems Application, 2008. DEXA’08. 19th International Workshop on*, pages 54–58. IEEE.

Zhang, D. and Li, S. (2011). Topic detection based on k-means. In *Electronics, Communications and Control (ICECC), 2011 International Conference on*, pages 2983–2985.