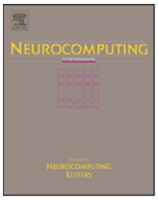
Neurocomputing 76 (2012) 125–133

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Topic model validation

Eduardo H. Ramirez [a,**], Ramon Brena [a], Davide Magatti [b], Fabio Stella [b,*]

[a] Tecnologico de Monterrey, Campus Monterrey, Monterrey, Mexico
[b] DISCo, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy

ARTICLE INFO

Available online 30 July 2011

Keywords:
Topic models
Soft clustering
Fowlkes–Mallows index
Monte Carlo

ABSTRACT

In this paper the problem of performing external validation of the semantic coherence of topic models is considered. The Fowlkes–Mallows index, a known clustering validation metric, is generalized for the case of overlapping partitions and multi-labeled collections, thus making it suitable for validating topic modeling algorithms. In addition, we propose new probabilistic metrics inspired by the concepts of recall and precision. The proposed metrics also have clear probabilistic interpretations and can be applied to validate and compare other soft and overlapping clustering algorithms. The approach is exemplified by using the Reuters-21578 multi-labeled collection to validate LDA models, then using Monte Carlo simulations to show the convergence to the correct results. Additional statistical evidence is provided to better understand the relation of the metrics presented.

1. Introduction

The continuously increasing amount of text available on the WEB, news wires, forums and chat lines, business company intranets, personal computers, e-mails and elsewhere is overwhelming [1]. Information is switching from *useful* to *troublesome*. Indeed, it is becoming more and more evident that while the amount of data increases exponentially our storing and processing capabilities do not grow with comparable speed [2]. This trend strongly limits the extraction of valuable knowledge from text and thus drastically reduces the competitive advantage we can gain. Search engines have exacerbated such a problem by dramatically increasing the amount of text available in a matter of a few key strokes.

Probabilistic topic modeling (PTM) methods comprise a new and promising family of unsupervised techniques to model text collections which go toward the direction suggested by Halevy et al. [3].

PTM is a particular form of document clustering and organization used to analyze the content of documents and the meaning of words with the aim to discover the *topics* mentioned in a document collection. Table 1 shows two topics, out of three hundreds, derived from the TASA corpus, a collection of over 37,000 text passages from educational materials [4].

A variety of PTM models [5–8] have been proposed, described and analyzed in the specialized literature. These models differ among them in terms of the assumptions they make concerning the data-generating process. However, they all share the same rationale, i.e. a document is a mixture of topics.

To describe how the PTM model works, let $P(z)$ be the probability distribution over topics z and $P(w|z)$ be the probability distribution over words w given topic z. The *topic-word distribution* $P(w|z)$ specifies the weight to thematically related words. A document is assumed to be formed as follows: the ith word w_i is generated by first extracting a sample from the *topic distribution* $P(z)$, then by choosing a word from $P(w|z)$. We let $P(z_i = j)$ be the probability that the jth topic was sampled for the ith word token while $P(w_i|z_i = j)$ is the probability of word w_i under topic j. Therefore, the PTM induces the following distribution over words within a document:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{1}$$

where T is the number of topics.

The main idea of PTM can be summarized as follows; a document is a linear combination of multiple topics (1). A topic is a probability distribution over a given vocabulary of words. Therefore, a document can be interpreted as a linear combination of probability distributions over a given vocabulary, where each probability distribution, i.e. topic, is associated with a specific argument, idea or theme.

Hoffmann [6,9] proposed the probabilistic Latent Semantic Indexing (pLSI) method which makes no assumptions about how the mixture weights in (1), i.e. $P(z_i = j)$, are generated. Blei et al. [7]

[*] Corresponding author.
[**] Principal corresponding author.
E-mail addresses: eduardo.ramirez@itesm.mx (E.H. Ramirez), ramon.brena@itesm.mx (R. Brena), magatti@disco.unimib.it (D. Magatti), stella@disco.unimib.it (F. Stella).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.04.032

Table 1
Illustration of two (out of 300) topics extracted from the TASA corpus (language and arts, social studies, health, sciences). The 10 most probable words (word) for each topic are listed together with the corresponding probability value (prob).

Word	Prob
Topic 247	
Drugs	0.069
Drug	0.060
Medicine	0.027
Effects	0.026
Body	0.023
Medicines	0.019
Pain	0.016
Person	0.016
Marijuana	0.014
Label	0.012
Topic 5	
Red	0.202
Blue	0.099
Green	0.096
Yellow	0.073
White	0.048
Color	0.048
Bright	0.030
Colors	0.029
Orange	0.027
Brown	0.027

improved the generalizability of this model to new documents. They introduced a Dirichlet prior with hyperparameter α on $P(z)$, thus creating the Latent Dirichlet Allocation (LDA) model. In 2004, Griffiths and Steyvers [4] introduced an extension of the original LDA model which associates a Dirichlet prior, with hyperparameter β, also to $P(w_i|z_i=j)$. The authors suggested that the hyperparameter to be interpreted as the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed. This choice can smooth the word distribution in every topic with the amount of smoothing determined by β. The authors showed that good choices for the hyperparameters α and β will depend on the number of topics T and vocabulary size W, and that accordingly to the results of their empirical investigation $\alpha=50/T$ and $\beta=0.01$ to work well with many different document collections.

PTM originated in the psychology community where semantic coherence of topics was evaluated by focusing on replicating human performances or judgments. This strategy was implemented by performing standardized tests, comparing sense distinctions, and matching intuitions about synonymy [10]. Topic model validation is an extremely important task which aims to check whether a PTM is semantically meaningful or not. Indeed, PTMs assume that the latent space, were topics are projected, is semantically meaningful and thus it can be exploited to validate the extracted topics, to summarize a given document corpus and to guide its contextual exploration. Furthermore, to ignore the analysis of the internal representation of topic models is in contrast with their rational and development. Several approaches have been proposed for topic model evaluation; measures based on held-out likelihood, sentiment detection or information retrieval as discussed in Chang et al. [10]. Wallach et al. [11] summarized several evaluation techniques based on tools from language modeling. These techniques measure the fit of the knowledge learned when applied to unseen documents. The main drawback of such methods is that they only measure the probability of observation while the internal representation of the topic models is ignored. An important exception is offered by Griffiths and Steyvers [8] where the authors have shown that different senses of a word are correlated with the number of topics that word appears in. However, Chang et al. [10] pointed out that this is still not a deep analysis of the structure of the latent space, as it does not examine the structure of the topics themselves. The problem of validating topic models without relying on the performance of a task was recently addressed by Chang et al. [10]. Their work, the first in that direction, shows by means of user studies that some problem exists when using supervised classification predictive metrics.

In this paper we are concerned with the analysis and validation of the semantic coherence of the results obtained through PTM and with the problem of their comparison with the results obtained through alternative models. The proposed approach consists of transforming each topic model into a “hard” overlapping partition through the discretization of the “soft” document-topic associations. Then, the validation is performed by exploiting novel probabilistic metrics, based on the interpretations of widely accepted concepts such as “precision” and “recall”. Also, we have generalized the Fowlkes–Mallows index [12,13], an existing cluster validation metric, in order to make it suitable to validate overlapping clusterings. The generalization of the FM index was performed on the basis of its underlying probabilistic interpretation and allows us to link the Fowlkes–Mallows index to the semantic coherence of the model rather than to the mere similarity between cluster partitions. Novel metrics inspired on the widely known precision and recall concepts are also presented. The proposed validation approach has the following advantages with respect to existing metrics:

- it offers an explicit probabilistic interpretation;
- it allows to validate overlapping partitions;
- it allows to validate incomplete partitions.

Harmonic mean of precision and recall can be computed to obtain a combined single-metric measurement of the quality of the partition. The paper also shows how the proposed metrics allow to perform a “drill-down” analysis into the individual clusters (topics) to make straightforward the determination of:

1. which are the best/worst clusters in the partition;
2. which topics are better recalled by any given cluster.

The rest of this paper is organized as follows. In Section 2 we establish the notation and the main objects involved in our analysis. In Section 3 we propose a dissection of the Fowlkes–Mallows index. Section 4 introduces and describes the proposed metrics to evaluate topic model quality. The results of the numerical experiments performed on the Reuters-21578 data set are described in Section 5. Finally, conclusions and research directions are reported in Section 6.

2. Notation and problem statement

Once the soft-clustering solution of a multi-labeled corpus is discretized to obtain the corresponding hard-clustering, the problem to be faced consists in correctly evaluating the quality of the resulting overlapping partitions. Let us first introduce the terminology and the notation used in the rest of the paper. Every “hard-clustering” problem applied to a multi labeled document corpus involves the following elements:

- a dataset $D=\{d_0,\dots,d_n\}$ consisting of n documents;
- a partition of D in K clusters: $U=\{u_1,\dots,u_K\}$;
- a partition of D in S classes: $C=\{c_1,\dots,c_S\}$.

Table 2
Contingency table.

		Classes				
		c_1	c_2	…	c_S	Σ
Cluster	u_1	n_{11}	n_{12}	…	n_{1S}	n_{1*}
	u_2	n_{21}	n_{22}	…	n_{2S}	n_{2*}
	…	…	…	…	…	…
	u_K	n_{K1}	n_{K2}	…	n_{KS}	n_{K*}
	Σ	n_{*1}	n_{*2}	…	n_{*S}	n_{**}

Most of the existing validation metrics [14] can be expressed in terms of a $|U| \times |C|$ contingency table (Table 2) where the content of each cell n_{ij} represents the number of documents belonging to cluster u_i and class c_j. In the special case where clusters do not overlap and the document corpus is uni-labeled, the following properties hold:

1. $\bigcup_1^K u_i = D$;
2. $u_i \cap u_j = \emptyset \ \forall i,j = 1,\ldots,K$ with $i \neq j$: there is no "overlap" in the cluster partition;
3. $c_i \cap c_j = \emptyset \ \forall i,j = 1,\ldots,S$ with $i \neq j$: there is no "overlap" in the class partition.

In this work we consider the case where the aforementioned properties cannot be assumed to hold, which happens when:

- the thresholding procedure used to move from soft to hard clustering results in some documents being assigned to no cluster;
- the thresholding procedure used to move from soft to hard clustering results in some documents being assigned to more than one cluster;
- the document corpus is multi-labeled, and thus in principle every document can be assigned to no, one or more classes.

3. Background

PTM can be thought of as soft-clustering methods that discover probabilistic associations of documents to latent topics. Therefore, to evaluate the quality of the knowledge extracted from PTM we can exploit metrics designed to compare heterogeneous soft-clustering methods in text collections. This strategy, based on external validation metrics, takes full advantage of multi-labeled corpora. To the best of our knowledge this is the first paper that proposes and develops such an approach. It is worthwhile to notice that most of the works on external clustering validation deal with the case of non overlapping partitions of uni-labeled corpora. A comprehensive review of the traditional metrics used to validate non-overlapping partitions can be found in [14,15]. A fuzzy clustering validation approach for complete solutions is presented in [16].

In this section we will focus on dissecting the Fowlkes–Mallows Index, as being a probabilistic metric which can be extended to work in the overlapping and incomplete case. Finally, it is important to clarify that clustering entropy, a probabilistic, external metric commonly referred in the literature is not considered in detail in our analysis despite its ability to validate overlapping clustering solutions. Indeed, a probability distribution over the class variable has to be constructed to compute the entropy of a cluster. However, such a probability distribution can be constructed only in the case where the classes do not overlap. In our opinion this is an undesirable property for a topic model validation corpora, such as Reuters-21578 where classes actually overlap.[1]

3.1. Dissecting the Fowlkes–Mallows index

Among the existing cluster validation metrics, a particular interesting one is the Fowlkes–Mallows index [12,13] (hereafter referred to as "FM"). Using the contingency table notation from Table 2, the FM index is defined as follows:

$$FM = \frac{\sum_{ij}\binom{n_{ij}}{2}}{\sqrt{\sum_i\binom{n_{i*}}{2}\sum_j\binom{n_{*j}}{2}}} \tag{2}$$

In order to analyze the FM index, the events associated with the experiment of randomly sampling two documents d_1 and d_2 without replacement from D are defined as follows:

- S_{u*}: d_1 and d_2 belong to the same cluster;
- S_{*c}: d_1 and d_2 belong to the same class;
- S_{uc}: d_1 and d_2 belong to the same cluster and class.

To denote the event of d_1 and d_2 belonging to class c_j we write S_{*c_j} whose probability is given by

$$P(S_{*c_j}) = \frac{\binom{n_{*j}}{2}}{\binom{n_{**}}{2}} = h(n_{**},n_{*j},2,2) \tag{3}$$

where $h(n_{**},n_{*j},2,2)$ represents the probability value, according to the hypergeometric distribution, to obtain two successes in a sampling without replacement of size 2 from a population of size n_{**} that contains n_{*j} successes. In a similar manner, we write S_{u_i*} to denote that d_1 and d_2 belong to cluster u_i, where the corresponding probability value is given by

$$P(S_{u_i*}) = \frac{\binom{n_{i*}}{2}}{\binom{n_{**}}{2}} = h(n_{**},n_{i*},2,2) \tag{4}$$

The probability of two documents belonging to the same class can be computed from expression (3) to be

$$P(S_{*c}) = \sum_j P(S_{*c_j}) = \frac{1}{\binom{n_{**}}{2}}\sum_j\binom{n_{*j}}{2} \tag{5}$$

while the probability of two documents belonging to the same cluster can be computed from expression (4) to be

$$P(S_{u*}) = \sum_i P(S_{u_i*}) = \frac{1}{\binom{n_{**}}{2}}\sum_i\binom{n_{i*}}{2} \tag{6}$$

Finally, the probability of two randomly sampled documents, without replacement, to belong to the same class and cluster is

$$P(S_{uc}) = \sum_{ij} P(S_{u_ic_j}) = \frac{1}{\binom{n_{**}}{2}}\sum_{ij}\binom{n_{ij}}{2} \tag{7}$$

[1] Certainly, a generalized version of the cluster entropy measure could be produced by mapping the original class labels into a new set of labels, where each distinct label combination is transformed into a new, composite class label. We consider, however that such an approach would be excessively "strict" with the evaluation of the resulting solutions in terms of semantic coherence.

Then, the conditional probability that two randomly sampled documents, without replacement, belong to the same class given they belong to the same cluster is

$$P(S_{*c}|S_{u*}) = \frac{P(S_{uc})}{P(S_{u*})} = \frac{\sum_{ij}\binom{n_{ij}}{2}}{\sum_i\binom{n_{i*}}{2}} \tag{8}$$

while the conditional probability that they belong to the same cluster given that they belong to the same class is

$$P(S_{u*}|S_{*c}) = \frac{P(S_{uc})}{P(S_{*c})} = \frac{\sum_{ij}\binom{n_{ij}}{2}}{\sum_j\binom{n_{*j}}{2}} \tag{9}$$

It is worthwhile to note that the FM index (2) can be obtained by computing the geometric mean of the conditional probability that the pair of sampled documents belongs to the same class given they belong to the same cluster ($P(S_{*c}|S_{u*})$), and the conditional probability that the pair of sampled documents belongs to the same cluster given they belong to the same class ($P(S_{u*}|S_{*c})$). Therefore, expressions (8) and (9) allow us to write the following:

$$FM = \sqrt{P(S_{*c}|S_{u*})P(S_{u*}|S_{*c})} \tag{10}$$

The previous formulations can also be expressed in terms of the hypergeometric distribution, defined as the probability of selecting exactly x successes in a sample of size m, obtained without replacement from a population of N objects from which k possess the characteristic of interest. Formally

$$h(N,k,m,x) = \frac{\binom{k}{x}\binom{N-k}{m-x}}{\binom{N}{m}} \tag{11}$$

In practice, when N is greater than 50 and $m/N \leq 0.10$, the hypergeometric distribution can be conveniently approximated by the binomial distribution [17]. So, the proposed formalization serves two goals. On the one hand, it is helpful for computational purposes as it allows the usage of cost efficient approximations; on the other hand, it is useful to better understand the properties of the considered metric.

For instance, by expressions (4) and (6), the probability of sampling two documents from the same cluster could be rewritten as follows: $P(S_{u*}) = \sum_i h(n_{**},n_{i*},2,2)$, while the probability of sampling two documents from the same class becomes $P(S_{*c}) = \sum_j h(n_{**},n_{*j},2,2)$. In a similar fashion, we have $P(S_{uc}) = \sum_{ij} h(n_{**},n_{ij},2,2)$. Thus, the conditional probabilities expressed above can be rewritten as

$$P(S_{*c}|S_{u*}) = \frac{\sum_{ij} h(n_{**},n_{ij},2,2)}{\sum_i h(n_{**},n_{i*},2,2)} \tag{12}$$

and

$$P(S_{u*}|S_{*c}) = \frac{\sum_{ij} h(n_{**},n_{ij},2,2)}{\sum_j h(n_{**},n_{*j},2,2)} \tag{13}$$

Finally, by geometrically averaging (12) and (13) the new expression for (2) is

$$FM = \frac{\sum_{ij} h(n_{**},n_{ij},2,2)}{\sqrt{\sum_i h(n_{**},n_{i*},2,2)\sum_j h(n_{**},n_{*j},2,2).}} \tag{14}$$

It is worthwhile to mention that Eq. (14) makes it easy to account the effects of overlapping clusters when computing the FM index.

3.2. Overlapping partitions

When validating using a multiply labeled corpus, such as Reuters-21578, the set of ground-truth classes result in overlapping partitions. In such a case the FM index cannot be computed by using Eq. (2) because the assumption of sampling without replacement does not hold. The main difficulty with overlapping when computing the FM index is due to the use of the contingency table notation, which hides the probability being computed that easily results in making the wrong assumptions that $n_{i*} = |u_i|$, $n_{*j} = |c_j|$ and $n_{**} = |D|$. The implications of such a wrong assumptions are shown through the following example.

Example 1

Consider a non-overlapping partition consisting of two clusters and two classes. Let $u_1 = \{d_1,d_2,d_3,d_4,d_5\}$ with $\{d_1,d_2,d_3\} \in c_1$ and $\{d_4,d_5\} \in c_2$ and let $u_2 = \{d_6,d_7,d_8,d_9,d_{10}\}$ with $\{d_6\} \in c_1$ and $\{d_7,d_8,d_9,d_{10}\} \in c_2$. The situation can be conveniently summarized through the following contingency table:

	c_1	c_2	$\sum$
u_1	3	2	$n_{1*} = 5$
u_2	1	4	$n_{2*} = 5$
	$n_{*1} = 4$	$n_{*2} = 6$	$n_{**} = 10$

According to (3) and (5) we can compute $P(S_{*c})$ as follows; $P(S_{*c}) = \sum_j P(S_{*c_j}) = \sum_j h(n_{**},n_{*j},2,2) = \binom{4}{2}/\binom{10}{2} + \binom{6}{2}/\binom{10}{2} = \frac{21}{45}$ to obtain the correct probability value.

The following class overlapping scenario, due to multi-labeled documents, is considered. Let c_3 be such that $\{d_1,d_4,d_8,d_9,d_{10}\} \in c_3$. The corresponding contingency table is

	c_1	c_2	c_3	$\sum$
u_1	3	2	2	$n_{1*} = 7$
u_2	1	4	3	$n_{2*} = 8$
	$n_{*1} = 4$	$n_{*2} = 6$	$n_{*3} = 5$	$n_{**} = 15$

Intuitively, we expect the intra-cluster overlap to increase the value of $P(S_{*c})$. However, Eq. (5) yields the incorrect result of 31/105, which is smaller than the correct one 21/45. This is due to the fact that the sampling without replacement assumption no longer holds. Indeed, there are not $\binom{15}{2} = 105$ ways to select two documents, as that would allow to sample the same document more than once. The correct number of ways to select two elements is still 45 and it is given by $\binom{|D|}{2} = \binom{10}{2}$. However, the events S_{*c_j} to sample two documents from the same class j are no longer independent. Therefore, they cannot be added as in (5). Basically, when class and/or cluster overlap exists, the contingency table bins do not represent mutually exclusive events. Thus, the value of $P(S_{*c})$ when classes overlap exists is given by

$$P(S_{*c}) = \sum_j h(|D|,|c_j|,2,2) - J(C) \tag{15}$$

where $J(C)$ is the probability that a selected pair of documents belongs to two classes simultaneously, defined by the expression

$$J(C) = \sum_j \sum_{j' > j} P(S_{*c_j} \cap S_{*c_{j'}})$$

or accordingly to the hypergeometric notation by the following expression:

$$J(C) = \sum_j \sum_{j' > j} h(|D|,|\{S_{*c_j} \cap S_{*c_{j'}}\}|,2,2)$$

However, the above formulas deal with the case where the classes overlap is restricted to pairs. The case where general classes overlap is concerned is more complex from both the theoretical and computational point of view and will be presented in a different work. Formula (15) is a re-expression of (5) under the general addition rule of probability for non-independent events.[2] For instance, if any of the pairs $\{(d_4,d_8),(d_4,d_9),(d_4,d_{10}),(d_8,d_9),(d_8,d_{10}),(d_9,d_{10})\}$ are sampled, then S_{*c_2} and S_{*c_3} are both true, and this results in a double count. The correct value of $P(S_{*c})$ is obtained by subtracting the probability of the classes intersection:

$$P(S_{*c}) = \left(\frac{\binom{4}{2}}{\binom{10}{2}} + \frac{\binom{6}{2}}{\binom{10}{2}} + \frac{\binom{5}{2}}{\binom{10}{2}} \right) - \frac{\binom{4}{2}}{\binom{10}{2}} = 0.55$$

3.3. *Incomplete partitions*

When hardening a soft-cluster solution we potentially obtain overlapping and incomplete partitions; thus, the validation metrics should be sensitive to some form of "recall." In the FM index computation the basic assumption would be that the column marginal totals correspond to the size of the classes, i.e. $n_{*j} = |c_j|$, and that the row marginal totals equal the size of the clusters. As shown before, such an assumption is false when an overlapping exists with the same applying to cases where the clusters are incomplete. Measuring incomplete partitions with the FM contingency table is wrong. Indeed, it incorrectly reduces the number of successes inside the population by using n_{i*} instead of $|u_i|$. Furthermore, the possibility of cluster overlapping has to be taken into account. Therefore, the correct probability of selecting two documents from the same cluster will be given by

$$P(S_{u*}) = \sum_i h(|D|,|u_i|,2,2) - J(U) \tag{16}$$

where $J(U)$ accounts for the probability of selecting a pair of documents belonging to two or more clusters, and it is given by adding up the probabilities of cluster intersections

$$J(U) = \sum_i \sum_{i' > i} P(S_{u_i *} \cap S_{u_{i'} *})$$

and by using the hypergeometric distribution

$$J(U) = \sum_i \sum_{i' > i} h(|D|,|\{S_{u_i *} \cap S_{u_{i'} *}\}|,2,2)$$

It is worthwhile to note that formula (16) is also valid in the case where clusters do not overlap. However, although FM can be corrected to take into account some of the effects of partitions' incompleteness and/or overlap, we consider that its interpretation is more biased toward measuring partition similarity, and thus we find it valuable to study new metrics that can serve better to estimate semantic coherence.

4. Proposed metrics

In this section, we introduce a version of the FM index adjusted for overlapping and incomplete clusters. Two overlapping "precision" metrics together with their probabilistic interpretations are given. Finally, we discuss the computation of a kind of cluster "recall" which can be used to achieve a single metric performance.

4.1. *Generalized Fowlkes–Mallows index*

As discussed in Section 3, if the FM index is expressed in terms of the contingency table, it cannot be used to validate overlapping or incomplete clusters. The reason is that while its additive terms come from the hypergeometric distribution, they would use an incorrect population size in the case where cluster overlapping is concerned. However, we have shown that when re-expressing the FM index in terms of the hypergeometric distribution and by correcting its formula in order to use the cluster size $|u_i|$ and the class size $|c_j|$, the probabilities $P(S_{*c})$ in (15) and $P(S_{u*})$ in (16) are correct under the assumption that the maximum overlap equals two. Therefore, the last step to obtain a generalized version of the FM index requires to generalize the computation of $P(S_{uc})$ in such a way that non-independent events $S_{u_i c_j}$ are correctly taken into account. This generalization requires to compute the probability of the intersection of elementary events. For the whole contingency table the sum of the probabilities of the intersection between "bins" will be denoted by

$$J(U,C) = \sum_{ij} \sum_{i'j'} P(S_{u_i c_j} \cap S_{u_{i'} c_{j'}})$$

where $i' > i$ and $j' > j$ and where by using the hypergeometric probabilities we obtain

$$J(U,C) = \sum_{ij} \sum_{i'j'} h(|D|,|\{S_{u_i c_j} \cap S_{u_{i'} c_{j'}}\}|,2,2) \tag{17}$$

Note that the computation of $J(U,C)$ requires the creation of an additional "overlap matrix" consisting of $(|U| \times |C|)^2$ elements. Finally, the generalized result for $P(S_{uc})$ is given by

$$P(S_{uc}) = \sum_{ij} h(|D|,n_{ij},2,2) - J(U,C) \tag{18}$$

Thus, the generalized version of the metric can be defined as the geometric average of

- the probability that two randomly sampled documents belong to the same class, given they belong to the same cluster, i.e.

$$P(S_{*c}|S_{u*}) = \frac{\sum_{ij} h(|D|,n_{ij},2,2) - J(U,C)}{\sum_i h(|D|,|u_i|,2,2) - J(U)} \tag{19}$$

- the probability that two randomly sampled documents belong to the same cluster, given they belong to the same class, i.e.

$$P(S_{u*}|S_{*c}) = \frac{\sum_{ij} h(|D|,n_{ij},2,2) - J(U,C)}{\sum_j h(|D|,|c_j|,2,2) - J(C)} \tag{20}$$

In conclusion, the generalized version of the FM index, referred to as GFM, is given by[3]

$$\frac{\sum_{ij} h(|D|,n_{ij},2,2) - J(U,C)}{\sqrt{[\sum_i h(|D|,|u_i|,2,2) - J(U)][\sum_j h(|D|,|c_j|,2,2) - J(C)]}} \tag{21}$$

4.2. *Partial class match precision*

This metric is inspired by the notion of precision utilized in the IR field. The partial class match precision (PCMP) measures the

[2] Which states that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

[3] We are aware that this formulation may not be accurate on extreme cases of *strongly* overlapped collections, however, extending the formulations to work on such cases is straightforward. Moreover, we will show experimentally that the hypothetical error, which is in fact an underestimation of actual probabilities is negligible in real-world corpora such as Reuters-21578.

probability of randomly selecting two documents from the same class taken from a randomly sampled cluster. In contrast to FM, where we are concerned with the random sampling of two documents d_1 and d_2 from the documents corpus, PCMP requires to first randomly sample a cluster and then to randomly sample two documents from the sampled cluster. In order to clearly differentiate both random events, we use $\tilde{S}_{c_*}$ to denote the event of selecting two documents belonging to the same class sampled from a given cluster. Formally, the PCMP metric is defined as follows:

$$P_{PM} = P(\tilde{S}_{*c}) = \sum_i P(\tilde{S}_{*c}|u_i)P(u_i) \tag{22}$$

where the prior probability of selecting the cluster u_i is given by $P(u_i) = n_{i*}/n_{**}$.

PCMP measures the probability of the event $\tilde{S}_{*c}$, i.e. to sample two documents from the same class, *after* having randomly selected a cluster. However, the computation of each individual $P(\tilde{S}_{*c}|u_i)$ also needs to be generalized in the case of class overlapping. Therefore, we need to add up the probability of selecting two documents from each class comprised within the cluster $P(\tilde{S}_{*c_j}|u_i)$ under the general rule of the addition for non-independent events, which implies discounting the probability of a success in two classes simultaneously. Thus, each individual $P(\tilde{S}_{*c}|u_i)$ would be given by

$$P(\tilde{S}_{*c}|u_i) = \sum_j P(\tilde{S}_{*c_j}|u_i) - J(u_i) \tag{23}$$

where $J(u_i)$, which represents the probability to sample two elements from two or more classes when selecting documents d_1 and d_2 which belong to cluster u_i, is given by

$$J(u_i) = \sum_j \sum_{j'>j} P(\{S_{u_i c_j} \cap S_{u_i c_{j'}}\}) \tag{24}$$

The previous equation represents the probability of selecting two elements from cluster u_i that simultaneously belong to two different classes. Thus, in order to obtain $J(u_i)$ we need to compute the individual probabilities of selecting two documents that simultaneously belong to every distinct pair of classes $(c_j, c_{j'})$ and then add them up to obtain the probability of selecting two documents that simultaneously belong to any pair of classes. The expression for the individual probabilities can also be represented using the formula of the hypergeometric distribution, where the parameter accounting for the number of successful outcomes is the number of elements in u_i that belong to both c_j and $c_{j'}$, that is, the "overlap" between c_j and $c_{j'}$:

$$J(u_i) = \sum_j \sum_{j'>j} h(|u_i|, |\{S_{u_i c_j} \cap S_{u_i c_{j'}}\}|, 2, 2) \tag{25}$$

This metric is designed to work well with multi-labeled documents corpus. The name *"Partial"* comes from the fact that in a multi-label setting the two randomly sampled elements d_1 and d_2 can be associated with many classes. As long as one of their classes matches we will consider the result to be semantically coherent, thus a success. We consider that this property of the metric is a valuable feature to focus on measuring semantic coherence rather than mere partition similarity. For instance, in contrast to similarity oriented metrics, more than one clustering solution can achieve the maximum evaluation in terms of the PCMP metric. In fact, we can think of two clustering solutions that will obtain a PCMP value of 1, where any pair of elements sampled from a given cluster will belong to the same class:

(a) Creating one cluster for every class, and assigning all the elements in c_i to u_i, so that $k = |C|$.
(b) Creating clusters of elements that share exactly the same class labels.

Finally, we should highlight that this metric can be efficiently approximated via a Monte Carlo simulation. Indeed, we will exploit this method to demonstrate the correctness of the metric.

4.3. Clustering recall

In the IR field the "recall" measure represents the probability that a relevant document is retrieved. Therefore, for the clustering scenarios under consideration, when the completeness of the partition cannot be assumed, it is critical to provide clear ways to measure the completeness of the clustering. Let N_c be the total number of class assignments, given by the sum of the sizes of every class:

$$N_c = \sum_j |c_j|$$

In overlapping and incomplete clustering we must not rely on the values of the contingency table to compute recall values, because they can account for duplicates. They also do not consider elements not assigned to any clusters.

4.3.1. Class recall

If we are interested in measuring which classes are better captured by the clustering it is straightforward to compute a class recall value. We define this "class recall" as the probability that a document d, randomly sampled from the class c_j, is included in any cluster:

$$R(c_j) = P([d \in \cup_i^k u_i]|c_j) = \frac{|\cap_i^k \{u_i \cap c_j\}|}{|c_j|} \tag{26}$$

In other words, Eq. (26) divides the number of documents, labeled with class c_j that were recalled by any cluster u_i, by the total number of documents labeled with class c_j.

4.3.2. Gross clustering recall

From the previous expression and recalling that the probability of selecting a class would be given by $P(c_j) = |c_j|/N_c$, it is possible to derive the following unconditional expression to measure the recall of the whole clustering:

$$R_U = P(d \in \cup_i^k u_i) = \sum_j P(d \in \cup_i^k u_i | c_j)P(c_j) \tag{27}$$

where the probability of selecting each class would be given by $|c_j|/N_c$. Therefore, (27) becomes

$$R_U = \frac{1}{N_c} \sum_j R(c_j)|c_j| \tag{28}$$

4.4. Single-metric performance

In retrieval and classification it is widely known that it is trivial to achieve high recall at the expense of precision and viceversa. Thus, traditionally they are averaged into a single metric, the F-Score.

The traditional F-Score is nothing but the harmonic mean between precision and recall. Almost any two probabilities can be averaged in this way. However, for the particular case of topic-model validation we are interested in balancing the best measurement for semantic coherence with the best measure for completeness, so our proposed metric is defined by the harmonic average of Eqs. (22) and (28) to obtain

$$F_o = \frac{2P_{PM}R_U}{P_{PM} + R_U} \tag{29}$$

Notice that the selection of (22) and (28) comes at the expense of not penalizing some clustering dissimilarities. Thus, if the ultimate performance criteria is the partition similarity, then

Table 3
Four out of the 90 topics extracted with the LDA algorithm running 1000 iterations of Gibbs sampling with parameters value $\beta = 0.01$ and $\alpha = 0.56$. The 10 most frequent words for each topic are listed together with their corresponding conditional probabilities. Furthermore, each topic is associated with its prior probability.

Topic 0	0.0097
Coffee	0.0620
Brazil	0.0397
Said	0.0372
Export	0.0348
Quotas	0.0261
Quota	0.0220
Producers	0.0183
ICO	0.0166
Brazilian	0.0154
International	0.0153
Topic 2	0.0081
Price	0.0623
Prices	0.0468
Oil	0.0336
Effective	0.0280
CTS	0.0223
Crude	0.0218
Increase	0.0211
Raised	0.0209
Barrel	0.0201
Raises	0.0186
Topic 49	0.0127
Rate	0.0945
Rates	0.0771
Interest	0.0551
PCT	0.0484
Cut	0.0354
Bank	0.0267
Market	0.0227
Money	0.0216
Prime	0.0204
Point	0.0163
Topic 56	0.0110
Wheat	0.0365
Agriculture	0.0340
US	0.0330
USDA	0.0322
Corn	0.0298
Grain	0.0285
Program	0.0278
Farm	0.0265
Said	0.0235
Farmers	0.0197

the GFM may be a best metric of choice. Both components of the F_o metric are micro-averaged so that every document has the same weight on the result. The micro-averaging effect is achieved by the marginalization step performed in (22) and (28) in order to work with unconditional probabilities.

5. Numerical experiments

Experimental evidence of the correctness of the theoretical formulations together with some insights on their characteristics are presented by using the Reuters-21578 corpus, ModApte split, including only documents labeled with topics.

Documents with less than 10 unique words were discarded while train and test documents were all included into the analyzed documents corpus. Therefore, a total of 10,468 documents were used which are labeled with 117 ground-truth classes.

5.1. Topic extraction

For demonstration purposes, the used algorithm was LDA ($\beta = 0.01$, $\alpha = 50/K$) running 1000 iterations of Gibbs sampling. The proposed measurement techniques require a discretization of the document-topic assignments. Thus, to show the effects of the discretization on the measurement we generated models using document-topic probability thresholds t equal to 0.05, 0.1, 0.2, 0.25 and a number of topics K equal to 10, 30, 50, 70, 90 and 117. Four topics extracted by LDA, in the case where $K{=}90$, are shown in Table 3. Each topic is associated with its prior probability $P(z_i)$ while each word is associated with its conditional probability $P(w_i|z_i)$.

5.2. Empirical approximation to the metrics

The correctness of the GFM and F_o formulations has been demonstrated by Monte Carlo simulation. Algorithm 1 for the estimation of the GFM metric can be summarized as follows: (i) randomly sample a pair of documents, (ii) check if they belong to the same class, (iii) check if they belong to the same cluster, (iv) check if they belong to the same class and cluster and *v)* compute estimates for $P(S_{uc})$, $P(S_{*c}|S_{u*})$, $P(S_{u*}|S_{*c})$ and *GFM* (lines 16–22).

Algorithm 1. Approximation to GFM

Require: $D = \{d_1, \ldots, d_N\}$, the input documents and *maxTrials*, the maximum number of trials.
Ensure: Γ, the empirical approximation to the Generalized Fowlkes–Mallows index GFM.

ClaSet(d) and *CluSet*(d) return respectively the set of classes and clusters associated with document *d*. *SampleDocsPair*(D) randomly samples a pair from the set of documents *D*.

```
1: $sameClassFreq \leftarrow 0$
2: $sameClustFreq \leftarrow 0$
3: $sameClassAndClustFreq \leftarrow 0$

4: for $trials = 1$ to $maxTrials$ do
5:   $sameClass \leftarrow False$
6:   $sameClust \leftarrow False$
7:   $d_x, d_y \leftarrow SampleDocsPair(D)$

8:   if $\{ClaSet(d_x) \cap ClaSet(d_y)\} \neq \emptyset$ then
9:     $sameClassFreq \leftarrow sameClassFreq + 1$
10:    $sameClass \leftarrow True$
11:  end if

12:  if $\{CluSet(d_x) \cap CluSet(d_y)\} \neq \emptyset$ then
13:    $sameClustFreq \leftarrow sameClustFreq + 1$
14:    $sameClust \leftarrow True$
15:  end if

16:  if $(sameClass \wedge sameClust)$ then
17:    $sameClassAndClustFreq \leftarrow sameClassAndClustFreq + 1$
18:  end if

19:  $P_{uc} \leftarrow sameClassAndClustFreq/trials$
20:  $P_{*c} \leftarrow sameClassFreq/trials$
21:  $P_{u*} \leftarrow sameClustFreq/trials$
22:  $\Gamma \leftarrow \sqrt{\frac{P_{uc}}{P_{u*}} \cdot \frac{P_{uc}}{P_{*c}}}$

23: end for

24: return $\Gamma$
```

Algorithm 2 estimates PCMP and F_o as follows: (i) sample a cluster, (ii) sample two documents from it and check if they belong to the same class, (iii) sample a class c_x, then sample a document d_z from c_x and check if it is included in the clustering solution (lines 15–19) and (iv) compute estimates for $P(\tilde{S}_{*c}|u_i)$, R_U and F_o (lines 20–22).

Algorithm 2. Approximation to F_o

Require: $U=\{u_1,\ldots,u_k\}$, the set of clusters, $C=\{c_1,\ldots,c_s\}$, the set of classes and *maxTrials*, the maximum number of trials.
Ensure: Φ_0, the empirical approximation to F_o.

SampleClass(C) randomly samples an element from the set of classes C. *SampleClust*(U) randomly samples an element from the set of clusters U. *SampleDocClass*(c) randomly samples a document associated with the class c. *SampleDocsClust*(u) randomly samples a pair of documents associated with the cluster u.

```
1:  $sameClassGivenClustFreq \leftarrow 0$
2:  $recDocsFreq \leftarrow 0$
3:  $RecalledDocs \leftarrow \emptyset$
4:  for all $u_j \in U$ do
5:      $RecalledDocs \leftarrow \{RecalledDocs \cup u_j\}$
6:  end for
7:  for $trials = 1$ to $maxTrials$ do
8:      $sameClass \leftarrow False$
9:      $sameClust \leftarrow False$
10:     $u_x \leftarrow SampleClust(U)$
11:     $d_x, d_y \leftarrow SampleDocsClust(u_x)$
12:     if $\{ClaSet(d_x) \cap ClaSet(d_y)\} \neq \emptyset$ then
13:         $sameClassGivenClustFreq \leftarrow sameClassGivenClustFreq + 1$
14:     end if
15:     $c_x \leftarrow SampleClass(C)$
16:     $d_z \leftarrow SampleDocClass(c_x)$
17:     if $(d_z \in RecalledDocs)$ then
18:         $recDocsFreq \leftarrow recDocsFreq + 1$
19:     end if
20:     $P_{PM} \leftarrow sameClassGivenClustFreq/trials$
21:     $R_U \leftarrow recDocsFreq/trials$
22:     $\Phi_0 \leftarrow \frac{2 \cdot P_{PM} \cdot R_U}{P_{PM} + R_U}$
23: end for
24: return $\Phi_0$
```

Results of an individual simulation for $K=90$, $t=0.2$, are shown in Fig. 1 where the convergence pattern of the empirical measurements to their correct values is depicted.

5.3. Relation of GFM and overlapping F_o

We are interested to measure how sensitive the F_o, PCMP and GFM metrics are with respect to the document-topic discretization probability threshold (t) and the number of topics (K). Furthermore we are interested to measure how the presented metrics correlate to each other. Therefore, we think it is important to report two statistical measures. First, in Table 4 we present the results of a cross-correlation tab between GFM and the components of F_o for the overall data set.

In Table 4 it is possible to observe a high correlation between GFM and F_o, although not high enough to make the metrics redundant. Recall is positively correlated with F_o and GFM while it is negatively correlated with Precision; this property is widely acknowledged in the retrieval field. A two-factor analysis of variance has been performed to evaluate the impact of the following factors; number of topics and probability threshold on the considered measurements.

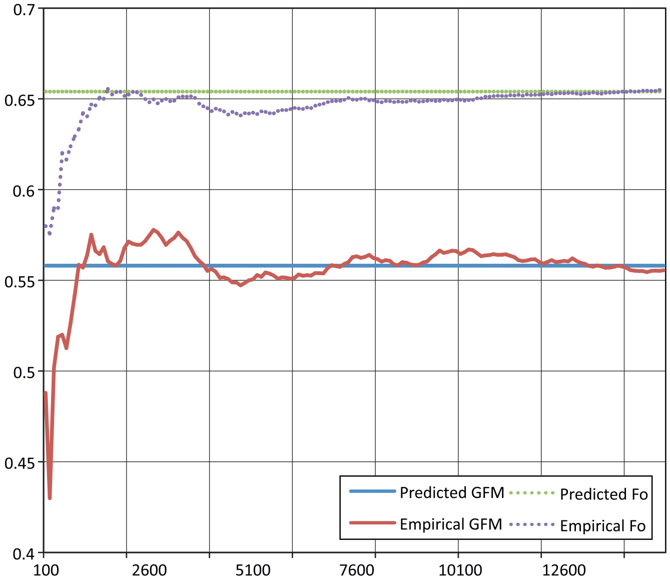


Fig. 1. Monte Carlo approximation of GFM and F_o.

Table 4
Correlation matrix for the following metrics; F_o, GFM, PCMP and Recall.

	F_o	GFM	PCMP	Recall
F_o	1			
GFM	0.57	1		
PCMP	0.91	0.33	1	
Recall	−0.04	0.38	− 0.41	1

Table 5
Two-factor ANOVA of the F_o and GFM metrics, statistics and p-values associated with the number of topics (K) and the probability threshold (t) factors.

Factor	F_o	GFM
Number of topics (K)	0.0523	0.5105
Probability threshold (t)	0.0004	0.1267

The results, summarized in Table 5 show that both factors, the probability threshold (t) and the number of topics (K), have a statistically significant effect on the F_o metric with confidence of above 94%, while this effect can only be moderately noticed on the GFM metric for the probability threshold factor with a confidence of about 88%. A potentially important consequence of the higher sensitivity of the F_o metric is that it becomes the natural candidate to perform model selection analysis.

6. Conclusions and future work

In this paper we have shown that it is possible to measure the semantic coherence of topic models by considering them to be special instances of soft-clustering algorithms and then using multi-labeled corpora as external validation input. In order to accomplish this goal, we have generalized existing metrics designed to evaluate non-overlapping partitions like the Fowlkes–Mallows Index. We have also proposed metrics with more straightforward probabilistic interpretations and of easier implementation. In both cases we have shown the correctness of the formulations by empirically approximating the predicted values through Monte Carlo simulation.

The usage of annotated collections to compute the validation metrics as proposed in this work, will result in additional methodological benefits for the research on probabilistic topic models. For instance, the approach will enable automated validation of models, therefore, producing a significant acceleration of the development–validation cycle. In addition, more robust and objective comparisons between independently developed models will be possible as objections related to the repeatability of experiments, such as judge selection or expertise, will be eliminated. All of the aforementioned benefits will be obtained while preserving the usage of valuable human input as the ultimate benchmark of model quality.

In future works we are interested in discussing how the different properties of a topic modeling algorithm like completeness, similarity between partitions or semantic coherence are stressed by the different metrics. Moreover, although this metric is already based on human input, it would be useful to more clearly visualize the predictive power of such probabilistic metrics on the performance of machine learning tasks like classification or retrieval.

References

[1] R. Feldman, J. Sanger, The Text Mining Handbook, Cambridge University Press, New York, 2007.

[2] The Economist, A special report on managing information: Data, data everywhere, The Economist ISSN 0013-0613, 〈http://www.economist.com/opinion/displaystory.cfm?story_id=15557443〉.

[3] A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, IEEE Intelligent Systems 24 (2) (2009) 8–12. ISSN 1541-1672, 〈http://doi.ieeecomputersociety.org/10.1109/MIS.2009.36〉.

[4] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of National Academy of Science United States of America 101 (Suppl 1) (2004) 5228–5235.

[5] T.L. Griffiths, M. Steyvers, A probabilistic approach to semantic representation, in: G.W., C. Schunn (Eds.), Proceedings of the Twenty-Fourth Annual Conference of Cognitive Science Society, 2002, pp. 381–386.

[6] T. Hofmann, Probabilistic latent semantic analysis, in: In the Proceedings of Uncertainty in Artificial Intelligence, UAI, 1999, pp. 289–296.

[7] D.M. Blei, N. Andrew, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[8] T.L. Griffiths, M. Steyvers, Probabilistic Topic Models, Erlbaum, 2007.

[9] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 42 (1-2) (2001) 177–196. ISSN 0885-6125, 〈http://portal.acm.org/citation.cfm?id=599631〉.

[10] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. Blei, Reading tea leaves: how humans interpret topic models, in: Neural Information Processing Systems (NIPS), 2009.

[11] H.M. Wallach, I. Murray, R. Salakhutdinov, D.M. Mimno, Evaluation methods for topic models, in: A.P. Danyluk, L. Bottou, M.L. Littman (Eds.), ICML, ACM International Conference Proceeding Series, vol. 382, ACM, 2009, p. 139. ISBN 978-1-60558-516-1, 〈http://dblp.uni-trier.de/db/conf/icml/icml2009.html#WallachMSM09〉.

[12] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, Journal of the American Statistical Association 78 (383) (1983) 553–569. ISSN 01621459, 〈http://www.jstor.org/stable/2288117〉.

[13] D.L. Wallace, A method for comparing two hierarchical clusterings: comment, Journal of the American Statistical Association 78 (383) (1983) 569–576. ISSN 01621459, 〈http://www.jstor.org/stable/2288118〉.

[14] J. Wu, H. Xiong, J. Chen, Adapting the right measures for K-means clustering, KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data MiningACM, New York, NY, USA, 2009, pp. 877–886 ISBN 978-1-60558-495-9, doi: http://doi.acm.org/10.1145/1557019.1557115.

[15] L. Denoeud, A. Guénoche, Comparison of distance indices between partitions, Studies in Classification, Data Analysis, and Knowledge Organization (2006) 21–28.

[16] A. Di Nuovo, V. Catania, On external measures for validation of fuzzy partitions, in: P. Melin, O. Castillo, L. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), Foundations of Fuzzy Logic and Soft Computing, Lecture Notes in Computer Science, vol. 4529, Springer, Berlin, Heidelberg, 2007, pp. 491–501.

[17] H.D. Brunk, J.E. Holstein, F. Williams, A comparison of binomial approximations to the hypergeometric distribution, American Statistician 22 (1) (1968) 24–26.

Eduardo H. Ramirez Rangel is a doctoral candidate on Intelligent Systems at the Tecnologico de Monterrrey, where he works on Large Scale Topic Modeling technologies. During the course of his studies he had interned with companies like Microsoft Research and Yahoo! Research working in areas related to search-spam, search quality measurement and personalized search. Before joining the Ph.D. program he was co-founder and software architect of Ensitech de Mexico, a startup company specializing on e-Commerce and Internet Marketing software. He has also performed consultancy on topics like XML standards, software engineering, website performance and information architecture. His research interests include unsupervised learning, data mining, cloud computing, Internet economics and contextual advertising.

Ramon F. Brena is full professor at the Tecnologico de Monterrey, Mexico, since 1990, where he is head of a research group in Distributed Knowledge and Multiagent Systems. Dr Brena is the head of the Master level graduate programs in Computer Science and Artificial Intelligence. Dr. Brena holds a PhD from the INPG, Grenoble, France, where he presented a doctoral Thesis related to Knowledge in Program Synthesis. His current research and publication areas include Intelligent Agents and Multiagent Systems, Ubiquitous computing and Ambient Intelligence, Formal Methods in Software Engineering, Knowledge representation and reasoning, Semantic Web, and Artificial Intelligence in general. He has been visiting professor at the U. of Texas at Dallas and the Université de Montréal. Dr Brena is member of the ACM, and is recognized as an established researcher by the official Mexican research agency, CONACyT (SNI level I).

Davide Magatti graduated in 2007 in Compuer Science at Università degli studi di Milano – Bicocca and he is now a C.S. PhD student. He works in the "Models and Algorithms for Data and Text Mining Laboratory" at Department of Informatics Systems and communications (DISCo) of Milano-Bicocca university. He will deliver his thesis on "Graphical Models for Text Mining: knowledge extraction and performance estimation" in 2010. His main research area is Text Mining and in particular document clustering and supervised models for information extraction. He is also interested in the interplays between Text Mining and document management systems. He mainly works with probabilistic graphical models for topic extraction and information extraction models. He is also interested in solutions that integrate semantic web with text mining and machine learning.

Fabio Stella graduated in Computer Sciences in February 1991 at the Università degli Studi di Milano. From 1991 to 1994 he worked as research assistant for the EEC IMPROD project on semiconductors failure diagnosis, analysis and quality improvement. In January 1994 he became Assistant Professor of Operations Research at the Università degli Studi di Milano. He received the Ph.D. in Computatioonal Mathematics and Operations Research in 1995. In 2001 he became Associate Professor of Operations Research at the Università degli Studi di Milano-Bicocca. He directs the Models and Algorithms for Data and Text Mining Laboratory and actively collaborates with many Italian SMEs in the area of document management, financial risk management, and analysis of clinical and microarray data. He has been advisor of several Ph.D. students in Computer Science at the Università degli Studi di Milano-Bicocca. His main research interests include; continuous time Bayesian networks, data mining, text mining with specific reference to topic models, and computational finance with specific reference to on-line algorithms for portfolio selection.