Neurocomputing 179 (2016) 176–185

Contents lists available at ScienceDirect

Neurocomputing

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

An empirical comparison of latent sematic models for applications in industry

Constanza Contreras-Piña, Sebastián A. Ríos

Business Intelligence Research Center, Industrial Engineering Department, University of Chile, Av. Beauchef 851, Santiago, Chile

ARTICLE INFO

Article history:
Received 3 September 2015
Received in revised form
13 November 2015
Accepted 14 November 2015
Communicated by Grana Manuel
Available online 15 December 2015

Keywords:
Latent semantics
Text mining
Quality measures
Evaluation assessment

ABSTRACT

In recent years, topic models have been gaining popularity to perform classification of text from several web sources (from social networks to digital media). However, after working for many years in the web text mining area we have notice that assessing the quality of topics discovered is still an open problem, quite hard to solve. In this paper, we evaluated four latent semantic models using two metrics: coherence and interpretability which are the most used. We show how these pure mathematical metrics fall short to asses topics quality. Experiments were performed over a dataset of 21,863 text reclamation.

1. Introduction

Over the past 15 years, powerful methods to analyze large collection of text has been extensively studied. These models growth in popularity attracting commercial interest for potential applications. Particularly, Topic Models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help to develop new ways to search, browse and summarize large archives of texts.[1]

They have proved to be useful matching human concepts attracting widespread interest as many applications showed. Moreover, many authors had improved their limitations creating new models such as Correlated Topic Model [1], MedLDA [2] or adapting them to other structures like LDA-G [3] or Spatial Latent Dirichlet Allocation (SLDA) [4].

Nevertheless these works, nowadays it is still difficult to assess the usefulness of learned topics and how many of them should be learned.

Hence, there is still no clear and overall understanding of the properties of each model and how they could be suitable for a given application. The evaluation of models and topics is done by human judgment or using intrinsic statistical measures, such as Perplexity.

Researchers have created new measures oriented to automate and generalize the human judgment in the evaluation of topics. These measures focus on evaluate the Coherence [5,6], Similarity [7] and Interpretability [8,6] of topics. Besides these contributions, each researcher applied his own metric to measure the performance of proposed models. Sato proposed the Pitman–Yor Topic Model (PYTM) [9] and showed that outperforms Latent Dirichlet Allocation (LDA) [10] in terms of perplexity. Arora et al. [11] encouraged the use of Non-Negative Matrix Factorization (NMF) [12] model over other topics based on the interchangeability property. But [13] demonstrated that NMF is less effective in constructing semantic spaces than Latent Semantic Analysis (LSA) [14], using two tests and cosine similarity.

Additionally, another researcher applied two automatic semantic evaluations to three distinct topic model (LDA, NMF, LSA), describing the strengths and weakness of each model [15]. This work was more in line with our desire. However, their work is focused on the coherence and word similarity over models and number of topics. It is not considered the interpretability of them, which is important to create new applications in industry.

We want to know how many useful topics are learned in each model and how each model performs in dimensions like coherence and interpretability. We expect to create a match between human interpretability and mathematical meaning and we want to match these properties with the extraction of useful topics in each model.

We explore four latent semantic models (Latent Dirichlet Allocation, Latent Semantic Analysis, Non-Negative Matrix Factorization and Pitman–Yor Topic Model) comparing them in terms of coherence and interpretability.

Three of these models represent different approaches to model topics (LDA, NMF and LSA). PYTM presents a particular and

E-mail addresses: ccontreras@ceine.cl (C. Contreras-Piña), srios@dii.uchile.cl (S.A. Ríos).

[1] http://www.cs.princeton.edu/blei/topicmodeling.html

http://dx.doi.org/10.1016/j.neucom.2015.11.080
0925-2312/© 2015 Elsevier B.V. All rights reserved.

apparently desirable property which leads us to include it in this analysis.

We measured their capability to match with human judgment and concepts for a given application.

This paper present our experiments and results evaluating the coherence and interpretability of each model and its characterization. In the next section the core of each topic model applied is explained. Section 3 presents the measures used to assess coherence and interpretability. Then, it is described the corpus used in this research in Section 4. Section 5 presents experiments done with the data and results of them are showed in Section 6. Finally at Section 7 key points of results are discussed.

2. Topic models

2.1. Latent Semantic Analysis (LSA)

In this work [14], authors modeled the relationship between terms and documents in a document collection. Transforming text in a term-document matrix, they extracted a latent semantic space using Singular Value Decomposition (SVD) reducing the matrix dimensionality. In other words, LSA constructs a semantic space where terms and documents related are placed near one another.

Given a term-document matrix, X, for example a $t \times d$ matrix of t terms and d documents, X can be decomposed into the product of three other matrices:

$$X = TSD \tag{1}$$

This is called Singular Value Decomposition (SVD). T and D are matrices of singular vectors and S is the matrix of singular values. Eingenvectors are used to compute similarities between terms, documents and term-document as explained in [21–23]. If m is the rank of X, the dimension of T and D are $t \times m$ and $m \times d$ respectively.

SVD gets an optimal approximate fit using smaller matrices. Thus, if S is ordered by the size of its singular values, the first k largest may be kept and the remaining smaller ones set to zero. Then, the product of the three matrices results in an approximate matrix of X with k rank

$$X \approx \hat{X} = T'S'D' \tag{2}$$

Therefore, T' and D' (reduced matrices) are dimension $t \times k$ and $k \times d$. Rows of singular vectors are points that represent documents and terms in a k dimensional space.

2.2. Latent Dirichlet Allocation (LDA)

Blei et al. proposed a new model called Latent Dirichlet Allocation (LDA) [10] arguing that it is not clear why one should adopt the LSA methodology instead of Bayesian methods.

LDA is a three level hierarchical Bayesian model, in which each document of the collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities [16]. Finally, the topic probabilities provide an explicit representation of a document. The representation of the model is shown in Fig. 1.[2]

Given the smoothing parameters β and α and a joint distribution of a topic mixture θ, the idea is to determine the probability distribution to generate – from a set of topics $\mathcal{K}$ – a message composed of a set of N words w (where $\mathbf{w} = (w_1, \ldots, w_N)$ is the vocabulary),

$$p(\theta, z, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{3}$$

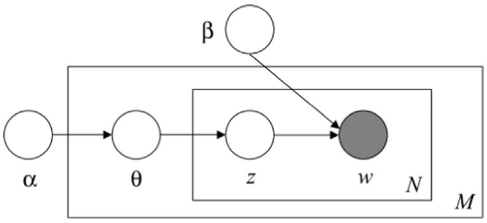

Fig. 1. Representation of Latent Dirichlet allocation.

where $p(z_n \mid \theta)$ can be represented by the random variable θ_i; such a topic z_n is presented in document i ($z_n^i = 1$). A final expression can be deduced by integrating Eq. (3) over the random variable θ and summing over topics $z \in \mathcal{K}$.

Defining:

- A document D as a sequence of N words, $D=(w_1,w_2,\ldots,w_N)$, where w_n is the nth word in the sequence.
- A corpus C as a collection of M documents denoted by $C=\{D_1, D_2,\ldots,D_M\}$.

The generative process of LDA to calculate Eq. (3) for each document D in a corpus C is:

1. Choose $N \sim$ Poisson(ξ).
2. Choose $\theta \sim$ Dir(α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ)
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Where N draws as a Poisson distribution with parameter ξ and θ draws as a Dirichlet distribution with parameter α.

2.3. Pitman–Yor Topic Model (PYTM)

The Pitman–Yor Topic Model (PYTM) proposed by Sato [9] changes the LDA model by using the Pitman–Yor process to generate the prior. Due to PY prior, the PYTM captures the power-law phenomenon of a word distribution.

The called PY process is a distribution over distributions over a probability space. It uses three parameters: a concentration parameter γ, a discount parameter d ($0 \le d \le 1$) and a base distribution G_0 that is understood as a mean of draws from PY process. The discount parameter controls the power-law property and generalizes the Dirichlet process when d is 0.

The representation of PYTM is shown in Fig. 2.[3]

The PY document model has its perspective based on the Chinese Restaurant Process. In this process n customers sit down in a Chinese Restaurant with an infinite number of tables. The first customer always sits at the first table.[4] The next customers can sit at an occupied table or at a new (unoccupied) table. The probability of sitting at any table t is [17]:

$$\begin{cases} \dfrac{|b_t|}{n+1} & \text{Probability of sitting at an occupied table} \\ \dfrac{1}{n+1} & \text{Probability of sitting at a new table} \end{cases} \tag{4}$$

where $|b_t|$ is the size of customers in t table. This process defines an exchangeable distribution on customer partitions. Moreover, it defines a prior over the number of tables and the parameters associated with each tables.

[2] Source: ammai2012.blogspot.com

[3] Source: original PYTM publication [9].

[4] videolectures.net/icml05_jordan_dpcrp/

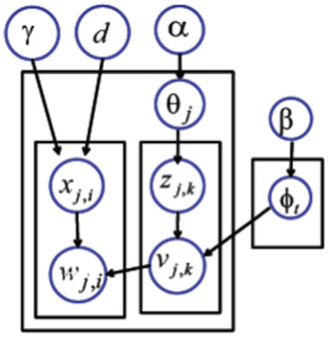

Fig. 2. Graphical representation of Pitman–Yor topic model.

For the PY topic model the CRP representation is composed of four elements: a customer, a table, a dish and a restaurant. The customer is represented by a word in a document. The table is represented by a latent variable. The dish is represented by a word type. The restaurant is represented by a document. Thus, PYTM constructs a multiple topics distribution over each document with this representation as a prior.

As it is observed in CRP, if n increases, the number of occupied tables increases, giving a long tale to the distribution. In other words, it is useful when the trend has many frequency-1 words. This is how CRP in this context captures the power-law distribution over words.

Sato adapted this process modifying the two probabilities as shown in the following equation:

$$\begin{cases} \text{The } k\text{th occupied table with probability} & \dfrac{N^c_{j,k} - d}{\gamma + N^c_{j,\cdot}} \\ \text{A new unoccupied table with probability} & \dfrac{\gamma + dK_j}{\gamma + N^c_{j,\cdot}} \end{cases} \quad (5)$$

$N^c_{j,k}$ is the number of customers sitting at kth table, $N^c_{j\ ,\cdot} = \sum_t N^c_{j,k}$ indicates the document length N_j and K_j denotes the total number of tables in restaurant j.

With this prior, the generative process of LDA was modified, creating the PYTM. Hence, the generative process for PYTM is:

```
1. Draw $\phi_t \sim \mathrm{Dir}(\phi|\beta)(t = 1, \ldots, T)$
2. for all document $j(= 1, \ldots, M)$ do:
3.    Draw $\theta_j \sim \mathrm{Dir}(\theta|\alpha)$
4.    for all word $i(= 1, \ldots, N_j)$ do:
5. Sit at the $k$th occupied table in proportion to $N^c_{j,k} - d$
6. Sit at a new unoccupied table in proportion to $\gamma + dK_j$, draw a
   topic $z_{j,k^{new}} \sim Multi(z|\theta_j)$ and draw a word type $\nu^{new} \sim p(w|z_{j,k^{new}}, \phi)$ at the new table
7.     end for
8. end for
```

Where $\phi_t (t = 0, 1, \ldots, T)$ is the word distribution for each document and θ_j is the topic distribution for each document. PYTM generates topics as much as tables are created, while the number of topics in LDA is equal to the number of words.

2.4. Non-Negative Matrix Factorization (NMF)

Non-negative Matrix Factorization [12] proposes an optimization problem over SVD extracting 2 matrices instead of 3.

Given a term-document matrix X of dimensions $n \times m$, it can be decomposed in the following equation:

$$X = GF + E \quad (6)$$

where G is an unknown factor matrix (scores) of dimensions $n \times p$ and F is an unknown factor matrix (loadings) of dimensions $p \times m$. E is the matrix of residuals with dimensions $n \times m$.

To approximate X matrix, E has to be minimized so it is necessary to estimate residuals, i.e. estimating standard deviation of each element in E. This problem has a weighted least squares sense: G and F are determinate so that the Frobenius norm of E divided (element by element) by standard deviation is minimized.

Defining σ as the standard deviation, p as the selected rank and $\| B \|_F$ as the Frobenius norm of any matrix B, NMF model is described as:

$$\{G, F\} = \arg\min \| X - GF \|_F \quad (7)$$

$$\{G, F\} = \arg\min \| E \|_F \quad (8)$$

$$\{G, F\} = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{E_{ij}^2}{\sigma_{ij}^2} \quad (9)$$

3. Measures

3.1. Topic coherence

David Mimno et al. created an automatic evaluation metric for identifying semantic coherence in topic models [5]. They argued that the presence of poor quality topics reduces user confidence in the utility of statistical topic models. Thus, authors proposed a new topic coherence score that corresponds well with human coherence judgments. This makes it possible to identify specific semantic problems such as semantic coherence without human evaluations or external references.

Based on the idea that "words belonging to a single concept will co-occur", they defined *topic coherence* as:

$$C(t, V^t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v^t_m, v^t_l) + \epsilon}{D(v^t_l)} \quad (10)$$

where $D(v)$ is the document frequency of word type v (the number of documents with at least one token of type v), $D(v, v')$ is the co-document frequency of word types v and v' (the number of documents containing one or more tokens of type v and at least one token of type v') and $V^t = (v^t_1, v^t_2, \ldots, v^t_M)$ is a list of M most probable words in topic t. In the original work, ϵ is set as 1.

Authors demonstrated that "standard topic models do not fully utilize available co-occurrence information, and that held-out reference corpus is therefore not required for purposes of topic evaluation".

The result of this metric is a vector with negative numbers, in which topics with numbers closer to zero indicate higher coherence. To validate the results of this metric, authors also compare it with the work of Chang et al. [8] (*Word Intrusion*) presented in next section.

3.2. Word intrusion

Chang et al. [8] created the *word intrusion* task which involves finding an "intruder" between a bag of words. In their words, this task tries to evaluate if a topic has human-identifiable semantic coherence. The goal was to measure the success of interpreting topic models across number of topics and modeling assumptions. The task is constructed as follows:

1. Select a random topic from the model.
2. Select 5 of the most probable words from that topic.
3. Randomly select the *intruder*: a word with low probability in the current topic but high probability in any other topic.
4. Shuffle the 6 words selected before. Present them to subject (a person).

5. The subject has to choose the word which is out of place or does not belong with the others.

With S as the number of subjects, $i_{k,s}^{m}$ is the intruder selected by the subject s on the bag of words related to topic k from model m and ${w_k}^m$ is the intruding word among the other words, *word intrusion* is calculated as the number of subjects agreeing with the model as is shown in the following equation:

$$MP_k^m = \sum_s \mathbb{1}(i_{k,s}^m = w_k^m)/S \tag{11}$$

where $\mathbb{1}$ is the indicator function which is 1 if the intruder word selected by the subject is the intruded word and zero otherwise. Comparing ${MP_k}^m$ with the model's estimate of the likelihood of the intruding word, authors found that higher probabilities and higher predictive likelihood did not have higher interpretability.

4. About the Corpus

For the purpose of our research we have gathered data from Chilean government (through its agency called SERNAC).[5] The data base of consumer's complaints from 2012 were facilitated to us.

SERNAC receives complaints from consumers against companies (retailers, financial or telecoms) and mediates between both parties. Since we are exploring models capability to extract useful information for a given application, we used this data set as a real applied case where topics extracted could lead further insights about the complaints or a new categorization of them. All complaints related to department store's credit cards (21,863 complaints) were selected.

Complaints are received by 2 channels: in office or by an on-line form. When complaints are presented in office, they are written in third person by an executive. In the case of online complaints, consumers fill an on-line form in which they describe in their own words their claims.

Generally, people that write complaints fail to construct a cohesive story. They use legal words in wrong contexts or have many misspellings and typos.

First we selected which complaints are useful. We found some complaints with no explanation because the explanation is in an “attached letter”, which is not in our database. Then, we pass all texts to lower case and remove non-alpha/numeric characters. Additionally, we created a dictionary of common misspelled words. For example, the word “celular” (cell phone in English) is misspelled (“CELUALR”) in some complaints. We identified most common cases and corrected them. Most of them are related to companies, i.e. the names of many companies were misspelled.

There are many words that together have a special meaning in complaints context like “Letra Chica” (abusive clause in English) or the name of some companies such as “Banco de Chile” (Bank of Chile). Thus, “bank” and “Chile” by themselves have a meaning but together mean something different (the name of a bank). Therefore, we identified most of this words calling them *Key Entities*. We use the first letter of the second word in upper case to recognize them (now “Banco de Chile” is “bancoChile”). For this function we created a list of keywords that have to be replaced.

5. Experiments

We extracted 25 topics with each model. The number of topics was chosen observing that extracting less than 15 topics generate topics too aggregated. Also, generating more than 50 topics turns tedious the analysis of all of them and many of them were too fine-grained or with no sense. In ours and executives appreciation of topics, 25 is a good number of topics to extract. Thereby we fixed the number of topics in 25.

Table 1
Number of useful and half useful topics in each model.

Number of topics	LDA	PYTM	LSA	NMF
Useful	4	4	4	5
Half useful	7	9	5	6
Topic size range (%)	5.91–2.37	7.55–1.74	21.23–0.6	13.2–0.96

Each topic extracted was labeled and interpreted by executives. They also labeled topics as useful if it is, half useful if it could be helpful and no value. We considered a numeric value of 1, 0.5 and 0 respectively.

Every model gives us a range of negative numbers but in different scales when we calculated Topic Coherence. As there is not a clear cutoff to compare between models, we therefore used it to compare the semantic coherence of a topic with the other metrics proposed and then, characterize the model.

To calculate Word Intrusion we polled 100 people by an on-line questionnaire (see Appendix B). Since in our first trial the questionnaire was too long and people got confused because the words were repeated in many questions, we decided not to survey all topics (4 models with 25 topics each = 100 topics). We surveyed 10 topics from each model.

6. Results

Generally speaking, in the four models were found useful topics in similar proportions. Additionally, some topics extracted are similar or common over models. Table 1 shows how many topics are useful and half useful in each model. Topic Size refers to the range between the topic with most complaints related and the topic with fewer complaints related.

Comparing topics of LDA and PYTM, about a third part of them are useless. PYTM has a bigger difference in topics size than LDA. But, regarding executives observations, the useful topics found in PYTM are easier to interpret than LDA.

Frequency models are quite different from Bayesians in results, especially LSA. The size difference between topics in LSA is the biggest of the 4 models. Reading top words, the topics were not easy to understand. It was completely necessary to read complaints related to the topic. NMF has the same problem, except for useful topics.

6.1. Coherence

The authors of Topic Coherence worked with experts that classified topics as being good or bad. A topic is good if the words of the topic belong to a single concept. Then, the authors calculated their metric and also applied the word intrusion task with the experts. They realized that good topics have high accuracy at identifying the intruder and bad topics have uniform accuracy. But topic coherence was successful identifying good topics (good topics presented higher topic coherence than bad topics).

Therefore, their contribution was a metric capable of identifying a class of topics with low quality that can not be detected with word intrusion tests. Based on that, topics with higher coherence should be “good” topics. Therefore, they should have a high accuracy in word intrusion task, because their words represent a single concept.

In Table 2 topics from each model with high topic coherence and their value in word intrusion task are presented. Position in coherence is the position of the topic ordered by coherence, being the first position the topic with highest coherence in the model.

[5] Servicio Nacional del Consumidor, http://www.sernac.cl

Table 2
Topics from the 4 models with high topic coherence value and their word intrusion value.

Model	Topic	Position in coherence	Word intrusion value
PYTM	Topic 23	1st	0.25
PYTM	Topic 3	4th	0.45
LDA	Topic 18	4th	0.38
LSA	Topic 24	2nd	0.34
LSA	Topic 5	4th	0.73
NMF	Topic 22	2nd	0.48
NMF	Topic 16	3rd	0.50

Table 3
Topics with low topic coherence value and their usefulness.

Model	Topic	Position in coherence	Usefulness
PYTM	Topic 14	25th	0.5
PYTM	Topic 0	24th	0.5
LDA	Topic 4	25th	0.5
LDA	Topic 1	24th	1
LSA	Topic 0	25th	0
LSA	Topic 8	24th	1
NMF	Topic 23	25th	0.5
NMF	Topic 3	24th	0

As it is observed, topics with high coherence do not have high value in word intrusion. All of them, except one, have a value below 0.5, which means that less than half people polled could identify the intruder.

Then, bad topics (topics with lowest coherence in their models) were checked comparing them with our usefulness scale. Results are in Table 3.

Apparently, bad topics are not bad topics for some models (LDA and LSA), because some of them are useful. Hence, in our opinion, this metric may not be helpful enough identifying good topics.

We did a new experiment based on those results. In the original work, the word intrusion task was done with experts. We polled common people, then we decided to repeat the test with our experts (two SERNAC employees). Results are in Table 4 for good topics.

SERNAC word intrusion value has the same trend as our previous value. Even the single “good topic” with high word intrusion value has a 0 for SERNAC executives.

Since we could not obtain the same results of the original work of topic coherence, a different approach was taken. We observed the coherence in useful topics from each model. If useful topics are coherent, then it could be easier for experts to identify value in topics. In other words, when experts are analyzing topics, it should be easier to classify them between useful or not if the model is capable of generating coherent-useful topics.

The average of coherence in each model was calculated and we counted how many useful topics are over the average. We counted how many useful topics are top 4 in coherence too (because most model extracted 4 useful topics). Table 5 shows the results.

The only model that has all its valuable topics in the most coherent list of topics is PYTM. But none of its topics is top 4 in coherence. NMF instead has 2 of 3 valuable topics in top 4.

6.2. Interpretability

Chang et al. consider Perplexity and held-out likelihood as not useful metrics to explore common goals in topic models. Those metrics are useful to evaluate how predictive a model is. Therefore, the goal of authors of word intrusion is to measure the success of interpreting topic models over a number of topics and model assumptions.

Table 4
Topics with high topic coherence, their word intrusion value and SERNAC word intrusion value (SERNAC WI).

Model	Topic	Position in coherence	Word intrusion	SERNAC WI
PYTM	Topic 23	1st	0.25	0
PYTM	Topic 3	4th	0.45	0.5
LDA	Topic 18	4th	0.38	0
LSA	Topic 24	2nd	0.34	0
LSA	Topic 5	4th	0.73	0
NMF	Topic 22	2nd	0.48	0
NMF	Topic 16	3rd	0.50	0

Table 5
Number of useful topics in each model over the average of coherence and top 4 in coherence.

Number of topics	LDA	PYTM	LSA	NMF
Useful	4	4	4	5
Over average	2/13 (15%)	4/15 (26%)	1/14 (7%)	3/10 (30%)
Top 4	1/4 (25%)	0/4 (0%)	1/4 (25%)	2/4 (50%)

They found that traditional metrics are negative correlated with topic quality. Also, extracting high number of topics produces more fine-grained topics but less useful to humans.

Moreover, it is known that the use of words fits a Zipf-law distribution. According to Ferrer and Solé [18] people make the least effort when they have to think of words to express what they want. Hence, people tend to use the most ambiguous words in their language. It is a receptor task to interpret words and extract the semantic context.

Considering this, it is possible to obtain low values in word intrusion task, because the test exposes people to the top probable words which means the most ambiguous words in the topic. Therefore, we believe that models with high values in word intrusion task are capable of defining semantic context mixing ambiguous words with specific words (co-occurrence).

We selected the useful topics in each model that were polled. Then, we counted how many were over 0.7 in the word intrusion task and over 0.5. Table 6 shows the results. The percentage was calculated counting the number of polled topics over 0.7 (or 0.5 respectively).

Because of bias (we polled 10 of 25 topics from each model, randomly chosen), we can not say that a model finds more interpretable topics than other (counting topics). But we can analyze the relation between useful topics and their interpretability in each model. It is observed that the only model that has all its useful-polled topics over 0.5 is PYTM. Over 0.7 LSA and LDA have one useful and highly interpretable topic.

6.3. Similarity

Regarding to useful topics, it may be reasonable to suppose that useful topics extracted in one model are not present in other models. Therefore, we want to analyze if based on topics we can find complementary models, instead of choosing just one for this application. We expect to find common topics along models, topics represented in other topics in other model and topics that just were extracted in one model.

To do this analysis we applied cosine similarity and intersected topics along models using complaints related.

LDA and PYTM have more similar topics between them. The same it is observed between LSA and NMF. Since our research is focus on how useful topics are to humans, we decided to analyze deeper the useful topics. Fig. 3 shows the intersection of useful topics between models.

Table 6
Number of useful topics over 0.7 and 0.5 in word intrusion task.

Number of topics	LDA	PYTM	LSA	NMF
Useful	4	4	4	5
Useful polled	3	3	4	3
Over 0.7	1/1 (100%)	0/1 (0%)	1/1 (100%)	0/2 (0%)
Over 0.5	2/3 (66%)	3/8 (38%)	1/3 (33%)	1/5 (20%)

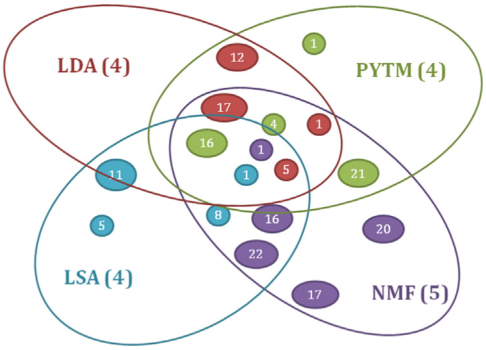

Fig. 3. Intersection of useful topics between models.

LDA has its 4 useful topics present in the other models. In fact, 3 of them are present in all the other models. Topic 1 of PYTM refers to a specific problem in a specific company. It has some similarity with topic 14 in LDA but the problem is completely different (the company involved is the same).

The topic 5 in LSA model refers to many known and serious problems against retailers. We found the same in topic 17 in NMF. While topic 20 in NMF is the opposite. Topic 20 refers to many serious problems in a supermarket chain. None of them were extracted in other models as clear as they are. Table 7 presents these topics described above.

6.3.1. *Common topics*

In this section, useful topics that were extracted in more than one model are analyzed. First of all, most useful topics are present in all the models. The differences are in how they were described (top words) and interpreted by experts. For example, topic 12 in LDA is present in PYTM. But in LDA this topic is useful and in PYTM it was classified as half useful.

Topic 5 in LDA, topic 16 in PYTM, topic 1 in LSA and topic 1 in NMF are the same. The only difference is in how they were interpreted. Table 8 shows the description of those 4 topics. In particular, topic 5 (useful), 14 (no value) and 17 (useful) from LDA are present in topic 16 of PYTM.

Topic 1 in NMF, topic 17 in LDA and topic 4 in PYTM refers to the same theme too. Because of the way they are described and interpreted make them different from Topic 1 in LSA. See Table 9.

Topic 1 in LDA is present in PYTM and NMF but the topics in those models are not valuable. The topic in PYTM (half useful) refers to 3 sub-problems in the theme and in NMF there was not a clear problem. Instead, in LDA the topic is against a specific credit card.

There are many similar topics between LSA and NMF. Specifically, topics 1, 16 and 22 in NMF are better represented than in LSA. Topic 8 in LSA refers to many problems in a specific company. The problems are present in topics in other models but not strongly related to this company. They are focus more on the problem itself.

Topic 11 in LSA is close to topic 1 in LDA and other topics related to unilateral renegotiations but it has no specificity about the process or company related.

Finally, topic 21 in PYTM was extracted in NMF too. But in NMF it has no specificity in the interpretation. This topic refers to an event that happened in 2011 in the country (unilateral renegotiations done by La Polar, a retailer).

Table 7
Topics extracted in their respective models that are not present in other models.

Topic	Model	Description
Topic 1	PYTM	Payments made with CMR visa or to Falabella (retailer). Consumer asks for a refund because the product was returned or the credit card did not work but the charge appears anyway
Topic 5	LSA	People drowned in debts. High debts and they can not afford them. Minimum payments are not clear. Some of them are reported in the Commercial Bulletin although the client paid
Topic 17	NMF	Unilateral Renegotiation or excessive increases in debts
Topic 20	NMF	Complaints against Presto (credit card) for keeping improper charges or charges already paid. Also it is because the company does not refund the interests of the purchase. They charge commissions in credit cards that have not been used and report people to DICOM for debts in blocked credit cards

Table 8
Common topics over models.

Topic	Model	Description
Topic 1	LSA	Credit card problems. Fraud, improper charges, cards unreasonably blocked, among others
Topic 1	NMF	Credit cards blocked, closed or never used that have insurance charges, collection commissions or fraud
Topic 5	LDA	Improper charges for fraud/clonation/stolen credit cards. Many of them were blocked and used anyway
Topic 16	PYTM	Completely unknown charges in Falabella (a retailer) credit cards or to Falabella clients

Table 9
Common topics over models NMF, LDA and PYTM.

Topic	Model	Description
Topic 17	LDA	Consumer could not use an old credit card that had not been uses for long time
Topic 1	NMF	Credit cards blocked, closed or never used that have insurance charges, collection commissions or fraud
Topic 4	PYTM	Charges in closed, blocked or unused credit cards

6.4. *Optimal number of topics*

As we explained before, we extracted 25 topics to facilitate SERNAC experts analysis. Since results shows that measures did not perform as expected for this application, we calculated how many topics should be learned and we repeated experiments on them. We expected to avoid the bias of extracting too (dis)aggregate topics. We extracted from a range of 10 to 150 topics in LDA and PYTM. Then, Perplexity was computed. Fig. 4 shows Perplexity curve for LDA and PYTM.

Fig. 4 shows that PYTM has lower perplexity than LDA. Also, it is observed the curve's elbow is near 35 topics and over 40 topics does not change considerably. Therefore, we choose 35 as an approach of the optimal number of topics to extract.

Results on 35 topics does not differ from 25 topics results.

LDA, NMF and PYTM has very similar proportion of useful and half useful topics. The main difference is in the distribution of topics size. There is a big difference between the biggest and the smallest topic in NMF while it is similar between LDA and PYTM. LSA differs from other models in the three aspect presented in Table 10. It has a big topic with 88.7% of complaints related and the

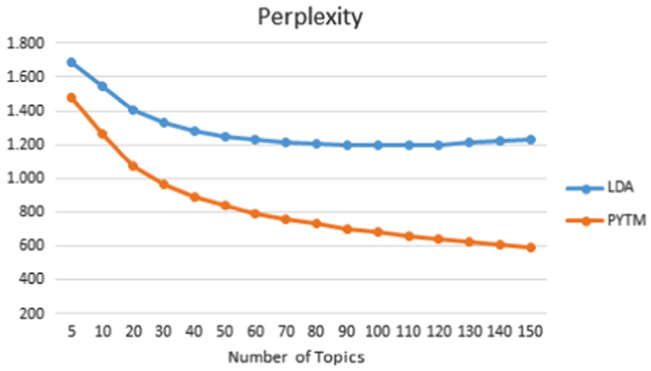

Fig. 4. Perplexity over number of topics for LDA and PYTM.

Table 10
Number of useful and half useful topics in each model.

Number of topics	LDA	PYTM	LSA	NMF
Useful	6	8	1	7
Half useful	18	15	20	18
Topic size range (%)	6.82–2.77	6.38–1.82	88.7–1.22	13.38–1.61

second in the list only has 5.67%. Therefore, we confirm that LSA on large documents set has poor precision.[6]

Regarding to useful topics, half of them are above size average. Chang et al. proved that fine-grained topics are less useful to humans. In our results, just one topic was near to bottom of size. Despite, we do not observe a relation between the size and the usefulness of the topic.

Observing coherence, for LDA and PYTM, results are similar. But for LSA and NMF there is a big difference in Top 4 coherence topics. For NMF, results are consistent with experiments in [15]. Lower coherence topics are learned while number of topics increase. But this does not affect the percentage of useful topics extracted (20% in both cases) as is shown in Table 11.

Finally, for word intrusion task, we use PMI score to identify the intruder [6]. Results agree with 25-tops results. PYTM is the model with more topics over 0.7 in this metric. But it is observed that interpreting level is lower for all models regarding 25-tops results.

7. Discussion and conclusions

Previous investigations were carried out to develop new automated measures to assess coherence, interpretability and similarity of topics generated from latent semantic techniques. However, topic models have not been evaluated under these measures. Besides, experiments conducted did not relate these metrics with topic models implementation for industry applications.

To solve this gap, we performed many experiments testing the measures mentioned above to provide an overall understanding of models and their suitable properties for real-case applications.

Table 12 summarizes the analysis of this work. Appendix C shows the values of metrics in each model.

Results confirm previous appreciations about models supporting their use in industry. However, we differ from the quality expected of the metrics tested. As far as concerned, actual metrics are not enough to analyze the quality of models or their use in end-user applications. It was observed that there is not a straightforward relation between highly interpretable, coherent and useful as previous works strongly supported. We showed that coherent topics are not the most useful or highly interpretable topics (opposite to what is stated on the literature). Topics with high word intrusion value are not the most coherent or useful as well.

Table 11
Number of useful topics in each model over the average of coherence and top 4 in coherence.

Number of topics	LDA	PYTM	LSA	NMF
Useful	6 (17%)	8 (23%)	1 (3%)	7 (20%)
Over mean	2/21 (10%)	6/25 (24%)	1/29 (3%)	6/24 (25%)
Top 4	1/4 (25%)	0/4 (0%)	0/4 (0%)	0/4 (0%)

Table 12
Summary of metrics: usefulness, topic coherence and word intrusion.

Number of topics	LDA	PYTM	LSA	NMF
Useful	4	4	4	5
Useful polled	3	3	4	3
Over average (coherence)	2	4	1	3
Over 0.5 (word intrusion)	2	3	1	1

In contrast, our work provides insights into these models application and usefulness of evaluation measures. Exploring interpretation, coherence and similarity of topics extracted; we proved that each property contributes to discriminate the usefulness of the topic. In other words, coherent or highly interpretable topics are easier to tag as useful or not.

We identified the strengths and weakness of all models in each dimension analyzed. NMF and PYTM show more useful-coherent topics than LSA and LDA. Also, both models show useful topics with high word intrusion values. LSA obtained more topics over the mean of coherence than other models but their usefulness decreases. In general, topics learned by Bayesian models are easier to understand and interpret than LSA and NMF. Regardless, topics extracted with NMF are more understandable and close to LDA and PYTM, than LSA. We believe that NMF could be as good as a Bayesian models (like LDA) if its interpretability is enhanced. We identified this as a good research line for future research.

Analyzing the similarities between useful topics we believe that intersecting results from a Bayesian model with a frequency model should be reasonable to obtain useful and specific topics. Besides this, if we focus on an end-user application, our study lead us to implement PYTM. We observed that the addition of the power-law distribution in this specific context benefits the extraction of more coherent, interpretable, and in consequence, more identifiable useful-topics, as PYTM results shows.

Summarizing, we evaluated the human usefulness of four topics models for a given application. We explored the interpretation, coherence and similarity of topics extracted adding some key views in topic models application. This case of study has proved to be promising in recommending approaches for modeling topics and evaluating the usefulness of them. Our work is a starting point and should be replicated with many large collection of text of real applications. Specifically, we encourage to research in increasing the interpretability of topics extracted for them. One idea is increasing the number of specific words in the description of topics (top words) and mix them in some proportion with ambiguous words to obtain a label for the topic. Regarding this, it should be possible to create an indicator to predict how many significant topics are possible to extract considering the model used and the number of topics.

Finally, libraries and implementations used to apply each model are described in Appendix A.

[6] http://staff.scem.uws.edu.au/~lapark/lt.pdf

Acknowledgments

The authors would like to acknowledge the support of the Servicio Nacional del Consumidor SERNAC. We also would like to acknowledge the support of the Chilean Millennium Institute of Complex Engineering Systems (ICM: P05-004-F FIN. ICM-FIC) and of the Business Intelligence Research Center (CEINE.

Appendix A. Models implementations

A.1. LDA model implementation

We use JGibbLDA, a Java implementation of Latent Dirichlet Allocation (LDA) using Gibbs sampling for parameter estimation and inference [19,20].[7] The default value of α in this implementation is $50/k$, with k being the number of topics. In our case, α is 2. β instead is 0.1 by default.

The input for this implementation is a file in which the first line contains the total number of documents. Then each line is one document represented by its words separated by the blank character. The parameters used for this implementation were:

- $\alpha=2$
- $\beta=0.1$
- number of topics: $k=25$
- number of iterations$=2000$
- number of most likely words for each topic$=30$

A.2. PYTM model implementation

Victor Chahuneau has repositories in GitHub with his implementations of many models. One of them is "vpyp",[8] an implementation of LDA with a Pitman–Yor prior instead of a Dirichlet prior. The code is in python.

The input is a file in which each line is one document represented by its words separated by the blank character (similar to LDA implementation). The parameters used were the number of topics (25), the number of iterations (2000) and the number of top words (30).

A.3. LSA model implementation

Scikit-learn is an open source package for machine learning in python. It has simple and efficient tools for data mining and data analysis.[9] It has a function that we used to apply LSA model. We were able to process a data set with 21,800 complaints and other with 38,500 complaints. Topics make sense. However we have to reduce the number of unique words – we do not consider words that appear less than 4 times in the data set - we were satisfied with this implementation.

This implementation receives the path of a folder in which each file contained is a document. Every document is represented by words separated by blank character. The parameters used are the number of topics (25) and the number of top words (30).

A.4. NMF model implementation

Scikit-learn has a function for NMF model. Based on the tutorial[10] of DARIAH-DE[11] we applied this code in a data set of 38,500 complaints and another one of 21,800 complaints. We also considered just words with more than 3 appearances in the data set.

As LSA model implementation, the code uses as an input the path of a folder with all complaints. Each complaint is in one file in the folder. The parameters fixed are the number of topics – 25 – and the number of top words – 30.

Appendix B. Questionnaires

First we created a pre-test. We constructed a survey consisting in 10 topics of LDA model, with its top 5 words and a fixed intruder for each topic. We selected the intruder as in the original task but it will be the same in all the surveys. We polled 10 people. Then we calculated the word intrusion. Our appreciation of this pre-test is that it is necessary to poll a significant amount of people. We decided to poll 100 people. Also, as the original metric state, the intruder has to be random and vary among subjects.

Hence, we created a second pre-test consisting in 4 forms. Each form has 25 questions (one for each topic of LDA experiment). We randomly mixed the top 5 words with an intruder. We selected 4 intruders for each topic. The first form has 25 questions using the first intruder. The second form has 25 questions using the second intruder, and so on. These forms were made in *Google docs*. Then, we decided to spread the 4 forms by Facebook and e-mail. The idea is to obtain 25 answers from different people of each form. We will have 100 answers of each topic with 4 different intruders.

It took us 3 days to complete the 100 answers. The feedback was that the questionnaire was too long and people got confused because the words were repeated in every question. Additionally we observed that answers differ in every questionnaire. For some topics every form has similar results but for other topics results highly differ. Therefore, it is useful to use many intruders and the questionnaires have to be shorter. We decided not to survey all topics of every model.

Considering concerns described above, we designed the final experiment. We selected 10 topics randomly for NMF, LSA and PYTM model. We designed a form with 3 pages (one model in each page). Each page has 5 questions related to 5 topics. Each question has 5 top words of its related topic and one intruder. We created 2 phases: phase 1 has first 5 topics of each model and phase 2 the other topics.

Then we selected 4 intruders for each topic as before. Thus, for phase 1 we created 4 forms. Form 1 has one specific intruder. Form 2 has another intruder and so on. For phase 2 we did the same. In total, there were 8 forms. Each form has 15 questions.

The instruction was to choose just one form of each phase. Then, each person is answering 30 questions related to 10 topics per model. We distributed the forms via Facebook and e-mail. The reason for using social media to distribute the forms was to avoid greedy bias. If we pay someone to respond to the questionnaire, the person can be willing to respond randomly just to finish quickly and obtain the payment.

Appendix C. Metrics for each model

We summarized the metrics for each model in Tables C1, C3, C2 and C4 for LDA, PYTM, LSA and NMF model respectively.

C.1. LDA topics

See Table C1.

C.2. LSA topics

See Table C2.

[7] Source: http://jgibblda.sourceforge.net/

[8] https://github.com/vchahun/vpyp

[9] Source: scikit-learn.org/stable

[10] https://de.dariah.eu/tatom/topic_model_python.html

[11] https://de.dariah.eu/

Table C1
Metrics calculated in topics obtained with LDA model.

Topic	Value	Size[a] (%)	Word intrusion	Topic coherence[b]
Topic 0	0.5	3.27	0.14	−3894.95
Topic 1	1	5.66	0.64	−3939.88
Topic 5	1	5.18	0.37	−3639.32
Topic 12	1	4.72	0.88	−3845.34
Topic 14	0	4.66	0.49	−3709.64
Topic 18	0	2.37	0.38	−3601.52
Topic 19	0	5.74	0.46	−3718.81
Topic 20	0.5	4.31	0.24	−3672.23
Topic 22	0	5.63	0.42	−3830.65
Topic 23	0.5	5.06	0.62	−3653.32

[a] The range of topic size is (5.91%, 2.37%).
[b] The range of topic coherence value is (−4002.25, −3537.98).

Table C2
Metrics calculated in topics obtained with LSA model.

Topic	Value	Size[a] (%)	Word intrusion	Topic coherence[b]
Topic 0	0	5.16	0.5	−3417.22
Topic 1	1	5.28	0.47	−3130.70
Topic 4	0	2.93	0.64	−3167.26
Topic 5	1	3.54	0.73	−3028.71
Topic 8	1	6.42	0.31	−3202.59
Topic 9	0.5	7.55	0.44	−3038.67
Topic 11	1	3.05	0.24	−3140.23
Topic 12	0	3.02	0.24	−3162.53
Topic 22	0	4.01	0.13	−3041.20
Topic 24	0.5	2.77	0.34	−2983.12

[a] The range of topic size is (21.23%, 0.6%).
[b] The range of topic coherence value is (−3417.22, −2964.58).

Table C3
Metrics calculated in topics obtained with PYTM model.

Topic	Value	Size[a] (%)	Word intrusion	Topic coherence[b]
Topic 0	0.5	5.16	0.68	−4332.97
Topic 3	0	5.28	0.45	−3909.74
Topic 4	1	2.93	0.69	−3994.97
Topic 10	0.5	3.54	0.55	−4312.99
Topic 13	0.5	6.42	0.72	−4188.13
Topic 16	1	7.55	0.53	−4008.03
Topic 17	0.5	3.05	0.62	−4199.07
Topic 21	1	3.02	0.66	−3984.36
Topic 22	0.5	4.01	0.60	−4298.51
Topic 23	0	2.77	0.25	−3620.94

[a] The range of topic size is (7.55%, 1.74%).
[b] The range of topic coherence value is (−4383.29, −3620.94).

Table C4
Metrics calculated in topics obtained with NMF model.

Topic	Value	Size[a] (%)	Word intrusion	Topic coherence[b]
Topic 0	0	10.23	0.49	−2951.15
Topic 2	0	5.96	0.72	−2811.39
Topic 6	0	7.21	0.68	−2911.85
Topic 13	0	2.86	0.18	−2887.87
Topic 16	1	2.55	0.50	−2707.46
Topic 17	1	2.43	0.30	−2861.43
Topic 18	0.5	1.91	0.82	−2878.71
Topic 21	0.5	1.49	0.28	−2857.11
Topic 22	1	1.56	0.48	−2692.79
Topic 23	0.5	1.83	0.57	−3002.69

[a] The range of topic size is (13.20% −0.96%).
[b] The range of topic coherence value is (−3002.69, −2641.01).

C.3. PYTM topics

See Table C3.

C.4. NMF topics

See Table C4.

References

[1] D.M. Blei, J.D. Lafferty, Correlated topic models, in: Proceedings of the 23rd International Conference on Machine Learning, MIT Press, ACM, New York, USA, 2006, pp. 113–120.

[2] J. Zhu, A. Ahmed, E.P. Xing, Medlda: maximum margin supervised topic models, J. Mach. Learn. Res. 13 (2012) 2237–2278.

[3] K. Henderson, T. Eliassi-Rad, Applying latent Dirichlet allocation to group discovery in large graphs, in: SAC '09: Proceedings of the 2009 ACM Symposium on Applied Computing, ACM, New York, USA, 2009, pp. 1456–1461.

[4] X. Wang, E. Grimson, Spatial latent Dirichlet allocation, in: NIPS, 2007.

[5] D.M. Mimno, H.M. Wallach, E.M. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: EMNLP, ACL, 2011, pp. 262–272.

[6] D. Newman, Y. Noh, E. Talley, S. Karimi, T. Baldwin, Evaluating topic models for digital libraries, in: The ACM/IEEE Joint Conference on Digital Libraries (JCDL2010), ACM, Gold Coast, Australia, 2010.

[7] V. Rus, N. Niraula, R. Banjade, Similarity measures based on latent Dirichlet allocation, in: Computational Linguistics and Intelligent Text Processing, Springer, Samos, Greece, 2013, pp. 459–470.

[8] J. Chang, J.L. Boyd-Graber, S. Gerrish, C. Wang, D.M. Blei, Reading tea leaves: how humans interpret topic models, in: Neural Information Processing Systems, vol. 22, 2009, pp. 288–296.

[9] I. Sato, H. Nakagawa, Topic models with power-law using Pitman–Yor process, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, USA, 2010, pp. 673–682.

[10] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[11] S. Arora, R. Ge, A. Moitra, Learning topic models—going beyond svd, in: FOCS, IEEE Computer Society, New Brunswick, New Jersey, USA, 2012, pp. 1–10.

[12] P. Paatero, U. Tapper, Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (1994) 111–126.

[13] A. Utsumi, Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: comparison to latent semantic analysis, in: SMC, IEEE, Istanbul, Turkey, 2010, pp. 2893–2900.

[14] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.

[15] K. Stevens, W.P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, in: J. Tsujii, J. Henderson, M. Pasca (Eds.), EMNLP-CoNLL, ACL, 2012, pp. 952–961.

[16] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.

[17] D.J. Aldous, Exchangeability and related topics, in: École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Mathematics, vol. 1117, Springer, Berlin, 1985, pp. 1–198.

[18] R.F.i. Cancho, R.V. Solé, Least effort and the origins of scaling in human language, Proc. Natl. Acad. Sci. USA 100 (3) (2003) 788–791.

[19] X.-H. Phan, L.-M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: WWW '08: Proceeding of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 91–100.

[20] I. Bíró, J. Szabó, A.A. Benczúr, Latent Dirichlet allocation in web spam filtering, in: C. Castillo, K. Chellapilla, D. Fetterly (Eds.), AIRWeb, ACM International Conference Proceeding Series, 2008, pp. 29–32.

[21] S. Ríos, F. Aguilera, F. Bustos, T. Omitola, N. Shadbolt, Leveraging social network analysis with topic models and the Semantic Web extended, Web Intell. Agent Syst. 11 (4) (2013) 303–314.

[22] G. L'Huillier, A. Hevia, R. Weber, S. Ríos, Latent semantic analysis and keyword extraction for phishing classification, in: Intelligence and Security Informatics (ISI), IEEE 2010, Vancouver, Canada, pp. 129–131.

[23] T. Omitola, Rios, S. Ríos, J. Breslin, Social Semantic Web Mining, Morgan & Claypool, San Rafael, California, USA, 2015, p. 154.

Constanza Contreras-Piña received her B.E. on Industrial Engineering (2011) and her P.E. on Industrial Engineering (2014) from the University of Chile. Since 2013, she is with the Business Intelligence Research Center (CEINE) as a junior researcher, at the Industrial Engineering Department. Her research interests include Latent Semantics, Natural Language Processing and Machine Learning.

Sebastián A. Ríos received the B.E on Industrial Engineering on 2001, the B.E on Computer Science, the P.E. on Industrial Engineering on 2003 from the University of Chile, Chile; and the Ph.D. on Knowledge Engineering from the University of Tokyo, Japan. He was lecturer at the University of Chile since 2002 and became an assistant professor on 2008 in the Industrial Engineering Department of the University of Chile. He is the Founder and Director of the Business Intelligence (BI) Research Center (CEINE) at the University of Chile since 2012, a collaborative applied research effort. His research interests include data mining algorithms in big dataset and its applications to different industry domains (medicine, marketing, operations, etc.); he is also interested in generative topic models for text mining in social networks and knowledge representation using semantic web technologies.