Modeling Students through Analysis of Social Networks Topics

María Emilia Charnelli[1], Laura Lanzarini[2], Javier Díaz[1]

[1]LINTI - Laboratory of Research in New Information Technologies
[2]III LIDI - Institute of Research in Computer Science LIDI
Computer Science School, National University of La Plata

mcharnelli@linti.unlp.edu.ar, laural@lidi.info.unlp.edu.ar,
jdiaz@unlp.edu.ar

Abstract. Educational Data Mining gathers the multiple methods that allow new and useful information extraction from great volumes of data coming from the educational context. The goal of this article is to obtain a model of the students of the Computer Science School of the UNLP from their participation in Facebook. The work describes the process of extraction of latent topics in posts made in public groups related to the School, and the modeling of the students from the topics discovered. Additionally, it includes the preprocessing done to the collected data, which constitutes a fundamental stage since it strongly conditions the performance of the models to be obtained. Finally, obtained results are presented together with conclusions and future lines of work.

Keywords: Educational Data Mining, Learning Analytics, Topic Modeling, User Modeling, Recommender Systems.

1 Introduction

In the last few years, educational institutions have embarked on their own exploration of big data sets to increase retention rates and provide students with a customized and higher quality experience. Applying data mining techniques in education has allowed to characterize the actors involved in teaching and learning processes. Generally speaking, it is very difficult to exploit available information to create models that describe students objectively. In particular, extracting information from the web constitutes a significant challenge due to the unstructured nature of the data it contains. Social networks make up an environment external to the educational institution, however containing valuable information regarding the interests of the students. Current research focuses on methods for extracting implicit information on student behaviors from social platforms in order to obtain dynamic models that are capable of easily adapting to changes in information and contributing to decision-making processes in education.

Analyzing latent topics has emerged as one of the most efficient methods to classify, group and retrieve textual data, such as those found in social network

posts. Many latent topic modeling methods have been developed and studied extensively, such as PLSA [1] and LDA [2]. In these models, documents are modeled as mixtures of topics, where a topic is a probability distribution of all the possible words in the documents. Statistical techniques are used to learn the topics and the mixing coefficients for each document. Conventional topic models reveal latent topics by discovering word co-occurrence patterns in all documents [3] [4]. Applying traditional topic modeling techniques on short texts such as tweets, Facebook status updates and instant messaging may not yield optimal results since they lack rich contexts. The main reason is that topic modeling implicitly captures word co-occurrence patterns by document in order to discover topics, therefore there is a severe data dispersion in shorter documents. More specifically, word occurrence in short texts plays a less discriminatory role in comparison to longer documents where the model has a number of words that is sufficient to know how they are related [5]. Discovering topics in short texts is crucial for a wide range of topic analyzing tasks, such as characterizing content [6] [7], modeling user interest profiles [8][9], and detecting latent or emerging topics [10]. BTM (Biterm Topic Model) [11] is an effective way to learn latent topics in short texts. BTM extracts underlying topics in a set of documents and a global distribution of each topic in each of them, through an analysis of the generation of biterms. Likewise, BTM has extensions to treat data flows and an incremental scheme for updating the model and giving more importance to the latest data collected.

This paper presents a method for obtaining a model of the students of the Computer Science School of the National University of La Plata (UNLP) through an analysis of the posts they make in public student groups in the Facebook social network.

This work is organized as follows: the second section describes the preprocessing effected on the collected data, the third section shows the extraction and modeling of latent topics, the fourth section shows the construction of student models through discovered topics, and the fifth section shows results obtained. Finally, the sixth section presents conclusions and future work.

2 Data preparation

The information used in this work comes from Facebook posts in groups created by Computer Science School students. The data were collected through the Facebook Graph API and involve over 3000 posts and 1500 students that write, comment, share and "like" posts. The Graph API is the main way of consulting and collecting data in the Facebook platform. Posts, group information and public user information were gathered. The following information was obtained from each of the posts: id of the post, creation date, author, likes, comments, shares and unstructured text content. When operating with textual information, it is necessary to use Text Mining techniques, in order to represent each post in a vector of terms. This was achieved through a process comprising many stages. The first stage consisted of the application of a stopwords filter, which

filters the words that match any indicated stopword. Stopwords were filtered in Spanish and English, using words from the social networks context such as smileys, greetings, etc; words from the context of the group such as university, school, informatics, etc; words that refer to web page addresses; words that are between symbols, etc. Likewise, students post code fragments that constitute solutions to exercises included in courses. In general, the most exercises are from the first year and written in Pascal. Therefore, each sentence or signature of the language used was reduced to a single word in order to better identify them. Following, each word in the text was reduced to its root applying the Snowball [12] stemming algorithm. The importance of this process lies in that it eliminates syntactic variations related to gender, number and tense. The algorithm was applied for both the English and Spanish languages. Once the roots of each of the words were obtained, the frequency of occurrence was calculated for each of them, and words that appeared more than once were chosen.

3 Extraction of latent topics

BTM was used for extracting topics in Facebook groups, which is an unsupervised learning technique that discovers topics characterizing a set of brief documents. In this context, each post and comment is considered a document.

Let a set of N_D documents be called corpus and let W be the set of all the words in the corpus, a topic is defined as a probability distribution of W. Therefore, a topic may be characterized by its T most likely words. Given a number K of topics, the goal of BTM is to obtain the K distributions of each of the words.

A "biterm" is a pair of words that co-occur without a set order in a short document. In this case, two different words in a document constitute a biterm. Given a corpus of N_D documents and a W vocabulary of unique words, it is assumed to contain N_B $\boldsymbol{B} = \{b_i\}_{i=1}^{N_B}$ biterms with $b_i = (w_{i,1} \in W, w_{i,2} \in W)$, and K topics expressed of W. Let $z \in [1, K]$ be a variable to indicate a topic. The $P(z)$ probability that a document in the corpus is of a z topic is defined as a $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^{K}$ K-dimensional multinomial distribution with $\theta_k = P(z = k)$ and $\sum_{k=1}^{K} \theta_k = 1$. The $P(w|z)$ distribution of words per topic can be represented as a $\Phi \in R^{K \times W}$ matrix where the ϕ_k k-th row is a W-dimensional multinomial distribution with $\phi_{k,w} = P(w|z = k)$ entry and $\sum_{w=1}^{W} \boldsymbol{\phi}_{k,w} = 1$. Given the α and β parameters, the main assumption of the model is that each of the documents of the corpus was generated thus:

1. A $\boldsymbol{\theta} \sim Dirichlet(\alpha)$ topic distribution is chosen for all the corpus
2. For each $k \in [1, K]$ topic
   - A $\phi_k \sim Dirichlet(\beta)$ distribution of words is extracted for the topic
3. For each $b_i \in \boldsymbol{B}$ biterm
   - A $z_i \sim Multimonial(\boldsymbol{\theta})$ topic assignment is extracted

– Two $w_{i,1}, w_{i,2} \sim Multimonial(\phi_{z_i})$ words are extracted

Taking into accont the generation mechanism assumed by BTM, likelihood for all the corpus can be obtained given parameters α and β from the probability of each of the biterms:

$$P(\boldsymbol{B}|\alpha,\beta) = \prod_{i=1}^{N_B} \int\int \sum_{k=1}^{K} P(w_{i,1}, w_{i,2}, z_i = k|\boldsymbol{\theta}, \boldsymbol{\Phi}) d\boldsymbol{\theta} d\boldsymbol{\Phi} \quad (1)$$

$$= \prod_{i=1}^{N_B} \int\int \sum_{k=1}^{K} \theta_k \phi_{k,w_{i,1}} \phi_{k,w_{i,2}} d\boldsymbol{\theta} d\boldsymbol{\Phi} \quad (2)$$

Obtaining exactly the $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ parameters that maximize the likelihood of equation 2 is an untreatable problem. Following the proposals in [13], parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ can be approximated using Gibbs sampling [14].

3.1 Co-occurrence Matrix

In order to evaluate the quality of the topics obtained, the coherence metric proposed by Mimno et al. [15] is used. Given a z and its T most likely words $V^{(z)} = (v_1^{(z)}, ..., v_T^{(z)})$ where $v_i^{(z)} \in W$ for $i = 1...T$, the coherence score is defined as:

$$C(z; V^{(z)}) = \sum_{t=2}^{T} \sum_{l=1}^{t} \log \frac{D(v_t{}^{(z)}, v_l{}^{(z)}) + 1}{D(v_l{}^{(z)})}$$

where $D(v)$ is the frequency of word v in all documents, $D(v, v')$ is the number of documents where words v and v' co-occur. The Coherence metric is based on the idea that words that belong to one concept will tend to co-occur in the same documents. This is empirically demonstrable since the coherence score is highly correlated with human criteria. In order to evaluate the general quality of a set of topics, the $\frac{1}{k}\sum_k C(z_k; V^{(z_k)})$ average of the coherence metric is calculated for each of the topics obtained. These results allow us to determine the amount of topics that best represent the entire corpus.

4 Modeling the Students

Once the K topics representing Facebook group posts collected were obtained, the students were modeled as vectors in a K-dimensional space. Each position of the vector represents the level of participation of the student in each of the topics. The level of participation is associated with possible actions performed by a user on the contents. Possible actions are creating content, commenting or liking a post. Given n users, $X \in R^{n,K}$ is defined as the matrix that contains in its rows the representation of each user in the new characteristics space. Let $A_{l,m} = \{a_1, a_2, ..., a_t\}$ be the set of actions of user l on posts classified in topic m. Then the l, m-th component of matrix X is defined as:

$$X[l,m] = \max\left(\sum_{i=1}^{t} w(a_i), 1\right), \text{ with } l = 1...n \text{ and } m = 1...K$$

where $w(a_i)$ is the weight associated with action a_i. $w(a_i)$ is defined as:

$$w(a_i) = \begin{cases} \frac{1}{20} & \text{if } a_i \text{ is a new post} \\ \frac{1}{40} & \text{if } a_i \text{ is a comment} \\ \frac{1}{50} & \text{if } a_i \text{ is a “like”} \end{cases}$$

Cosine similarity is used in order to evaluate the similarity between two users in the new space of characteristics. Given $v_1, v_2 \in R^K$, the similarity function is defined as:

$$d(v_1, v_2) = \frac{v_1 v_2}{||v_1 v_2||} = cos(\theta)$$

5 Results

The model obtained with BTM was evaluated in the set of Facebook group posts. For each number of topics between 2 and 45, the average of the obtained coherence was calculated sampling the test and training set randomly in 1000 iterations. Figure 1 shows the average of the coherence of the topics model according to the number of topics extracted. What is interesting is the number of topics in which there is a breaking point in the growth of the average coherence function. In this case, the optimal value is between 10 and 12 latent topics. Once the optimal number of topics was obtained, each of the Facebook group posts was classified into a topic according to the latent topics model that was obtained.

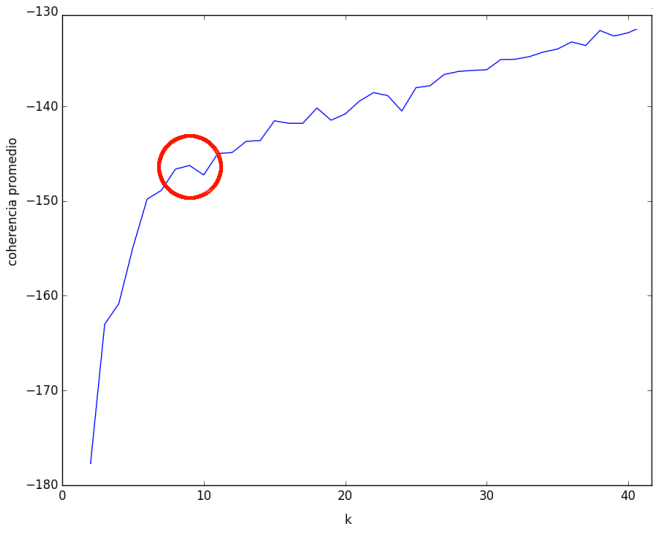

Fig. 1. Average coherence for different Ks

Table 1 shows topics obtained with K=10. For each of the topics, the six most important words are shown, i.e., those with the most likelihood of belonging to the topic. For example, topic 1 is about scholarships, topic 9 is about the first year programming subject where Pascal is used, and topic 10 contains posts on student accommodation search and offer.

Topic	Most important words of the topic				
1	enrollment	students	scholarships	university	national
2	support	workshops	courses	student	matters
3	community	sharing	tools	hacking	security
4	classroom	midterm	final	date	adp
5	courses	enrollment	information	page	systems
6	work	experience	knowledge	development	java
7	file	commands	linux	ubuntu	text
8	tutoring	classes	study	question	mathematics
9	pascal	close	file	enter	reset
10	students	double	simple	rooms	foreigners

Table 1. Model of the topics obtained with BTM

Afterwards, the actions performed by the users on each of the contents classified in the 10 topics were obtained which allowed for the generation of the users model described in section 4. Unsupervised learning techniques were applied to determine the underlying structure of student groups. In order to obtain the optimal number of user clusters automatically, the Silhouette [16] index was used, finding the value that optimized the criterion. Figure 2 shows the result of the evaluation of the model, generated by a hierarchical clustering algorithm with the average linkage criterion. It can be observed that the optimal value is implicitly imposed by the amount of topics obtained with BTM. In this context, it is interesting to model with greater detail the interests of the users. The following criterion-optimizing value is found in $k = 42$.

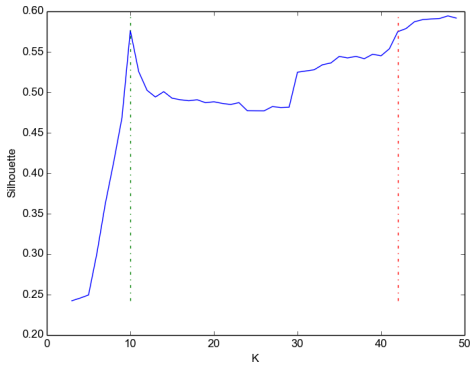

Fig. 2. Silhouette index on hierarchical clustering with average linkage

Figure 3 shows the dendrogram obtained by means of a hierarchical algorithm using the average linkage criterion with the cosine similarity metric, showing the obtained groupings and the correspondence with the cluster validity criterion.

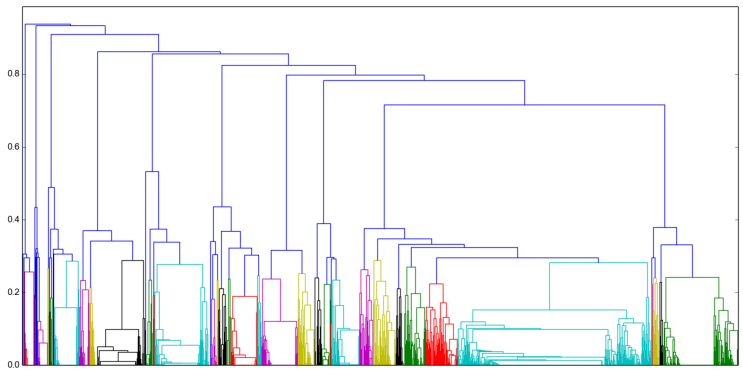

Fig. 3. Dendrogram obtained using average linkage and cosine distance

6 Conclusions and Future Work

This article presents a method for modeling the students of the Computer Science School of the UNLP through the detection of latent topics in their posts in public Facebook groups. This allows characterizing the students from a different context, knowing their topics of interest and how they relate to one another.

The methodology and the validation metrics used to obtain the user model presents satisfactory preliminary results. Topic coherence allows for automatic detection of the optimal number of topics and the user model presents compact groupings that characterize the behavior of the students in Facebook groups.

One of the future work points contemplates the implementation of an incremental model that allows updating the latent topics and therefore, the user modeling, from new posts in social networks.

The results of this work join those in [17] [18], which identifies features better characterizing students regarding their academic level through personal and academic information provided by the student management system. Results obtained will allow for the creation of an initial recommender system for the educational environment.

References

1. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1999) 50–57
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3** (2003) 993–1022
3. Boyd-Graber, J.L., Blei, D.M.: Syntactic topic models. In: Advances in neural information processing systems. (2009) 185–192
4. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2006) 424–433
5. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, ACM (2010) 80–88
6. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European Conference on Information Retrieval, Springer (2011) 338–349
7. Guo, J., Xu, G., Cheng, X., Li, H.: Named entity recognition in query. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM (2009) 267–274
8. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 261–270
9. Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. International Conference on Weblogs and Social Media **5** (2010) 130–137
10. Lin, C.X., Zhao, B., Mei, Q., Han, J.: Pet: a statistical model for popular events tracking in social communities. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2010) 929–938
11. Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering **26** (2014) 2928–2941
12. Gupta, V., Lehal, G.S.: A survey of common stemming techniques and existing stemmers for indian languages. Journal of Emerging Technologies in Web Intelligence **5** (2013) 157–161
13. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101** (2004) 5228–5235

14. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (1984) 721–741
15. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 262–272
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20** (1987) 53–65
17. Lanzarini, L., Charnelli, M.E., Baldino, G., Diaz, J.: Seleccion de atributos representativos del avance academico de los alumnos universitarios usando tecnicas de visualizacion: Un caso de estudio. Revista TE&ET (2015) 42–50
18. Lanzarini, L., Charnelli, M.E., Diaz, J.: Academic performance of university students and its relation with employment. In: Computing Conference CLEI, 2015 Latin American. (2015) 1–6