ISSN: 0067-2904

Modern Probabilistic Model: Filtering Massive Data in E-learning

Hachem Harouni Alaoui1*, Elkaber Hachem, Cherif Ziti
Mathematics & Computer Department, Faculty of Sciences, Moulay Ismail University, Meknes, Morocco

Abstract

So muchinformation keeps on being digitized and stored in several forms, web pages, scientific articles, books, etc. so the mission of discovering information has become more and more challenging. The requirement for new IT devices to retrieve and arrange these vastamounts of informationaregrowing step by step. Furthermore, platforms of e-learning are developing to meet the intended needsof students.

The aim of this article is to utilize machine learning to determine the appropriate actions that support the learning procedure and the Latent Dirichlet Allocation (LDA) so as to find the topics contained in the connections proposed in a learning session. Ourpurpose is also to introduce a course which moves toward the student's attempts and which reduces the unimportant recommendations (Which aren't proper to the need of the student grown-up) through the modeling algorithms of the subjects.

Keywords: E-learning platforms, IT tools, LDA, machine learning.

1. INTRODUCTION

NICTs (new information and communication technologies) have contributed in the development of new applications that are rich in both their mediums and functionalities. A digital document is enriched in the light of technological development, and can be composed from contents distributed on the Web. Multilingual, multimedia and ontological dimensions can be apprehended more easily through models and architectures, in order to build new digital libraries. Providing the user with new possibilities for building and accessing shared digital contents contribute in the construction of new systems that favor knowledge acquisition. This article is interested in creating a search engine that is capable of suggesting links selected by theme. These themes are adapted for each chapter of our course.

2. PROBLEMATIQUE

The information filtering systems take part in the allowance for the reception of documents considered interesting. Unlike search engines (Google, AltaVista, Yahoo, etc.), that requires the user to systematically formulate their need by using keywords. The result given back to the user is often filled with irrelevant documents. Thereby, the user is obliged to manually single out the relevant documents. This can be an annoying and frustrating task (Which causes the learner to drop out). Information filtering systems perpetuate this need for information and allow for the transportation of interesting documents over time.

This study uses information identified thanks to cluster analysis of actions in an online course, to support the course using combined information that come from processed documents, in order to update the profile of the learner. Course supplement will be presented as links that will be analyzed to discover the themes' contents with algorithms that list the subjects.

* Email: harouni.alaoui@gmail.com

The profile of the learner constantly translates the information needs of this latter, in order to minimize the irrelevant suggestions, with the help of log files that are automatically generated. These files contain registered actions that reflect precise learning behaviors of each individual learner.

3. INFORMATION FILTERING

The first information filtering systems emerged in the early 90s, experimentations [1] on methods of information filtering showed very promising results, [2] after that, and with many works being carried out, information filtering became a very active research method. Many applications have been developed, such as recommendation systems of books, CDs, or others by Amazon.com, of films with Movielens.

Filtering can be seen as the selection of relevant information about an inbound stream.

Our system makes a "prediction" when the interest is to present information to the user. This prediction relies on the profile of this user, to end-up in a decision-making process: "Recommend" or "not recommend"

4. CLUSTER

4.1. Cluster Analysis

In data sets that are too large for manual analysis, data mining techniques are ideal for automatically identifying and describing meaningful patterns despite the noise surrounding them [3].

This automatic extraction of implicit and interesting models from large and noisy data sets can lead to the discovery of new knowledge about how students solve their problems in order to identify interesting or unexpected learning models [4] And can be used to answer questions that could not be answered [5]. One of the most common educational data mining techniques is cluster analysis [6]. Cluster analysis is a technique of density estimation to identify models within users' actions that reflect differences in underlying attitudes, thought processes or behaviors [7] through general and sequential correlation analysis [3].

4.2. Radial Basis Function Network (RBFN) Classifier

Radial function networks (RBFs) are a special case of multi-layer networks and SVM networks. They are composed of three layers where each nodes of the hidden layer uses as a function of activation a kernel function such as "the Gaussian".

This function is centered at the point specified by the weight vector associated with the node. The position and width of the activation functions are learned from the learning data. More specifically, the input of each hidden neuron (radial unit) corresponds to the distance between the center of the unit (determined by the clustering K-Means algorithm) and the input vector. The output of the hidden neuron is then the application of the Gaussian function at this distance. The standard deviation of the Gaussian function can be calculated through some methods, e.g., the "K-nearest neighbor" method.

Each output knot implements a linear combination of these functions.

The algorithm "clustering K-Means" [8] Attempts to select a set of optimal points that are considered the centroids of the learning data groupings. This algorithm consists, in a first step, of assigning training data randomly to k sets and, in a second step, in calculating the centroid of each of the k sets. These two steps are repeated until the stop criterion is met. That is to say when no change is made in the assignment of the learning data.

RBF networks have several advantages, including the ability to model any nonlinear function using a single hidden layer. The linear transformation in the output layer can be optimized using the traditional linear modeling techniques. These networks are also faster than multi-layer networks in the learning. Nevertheless, they are slower in execution and consume more space than traditional multi-layered networks.

5. IDENTIFICATION OF SUGGESTED LINKS BY THEME

The LDA has an essential purpose of classification, it allows to associate a context to a document from the words contained in this document, which words taken individually could belong to different contexts [9].

Since there are many links proposed for each article by our search engine, we simply do not have the human power to examine and extract the theme of each individual article. To this end, trying to realize a probabilistic modeling of the subjects, to discover and annotate the proposed links with the thematic information of the subject suggested to the learner. The subject modeling algorithms are statistical methods that analyze the words of the texts to discover the themes that cross them, how these themes are related to each other and how they change.

We start off by highlighting the fundamental thoughts behind Latent Dirichlet Allocation (LDA), which is the most straightforward subject model [10]. The intent behind LDA is that documents showcase various subjects.

LDA is an element of probabilistic modeling. In generative probabilistic modeling, data are considered to emerge from a generative process involving hidden variables. This generative procedure is characterized by a joint probability distribution through the observed and hidden arbitrary variables. The analysis of the data is done by using this joint distribution to calculate the reliable circulation of the hidden variables, taking into account the perceived variables. This reliable circulation is also called posterior distribution.

LDA can be more "scientifically" portrayed with the annotation ahead:

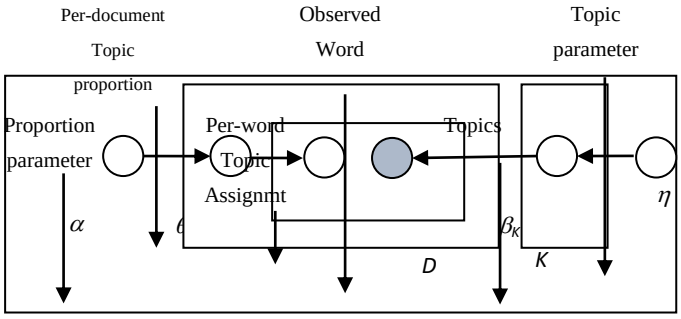

Figure 1-The distributions over words

With this annotation, the generative progression for LDA matches the following joint distribution of the concealed and perceived variables. [10]

$$\mathbf{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)} \quad (1)$$

Per-word topic assignment $(\mathbf{z_{d,n}})$

Per-document topic proportions $(\boldsymbol{\theta_d})$

Per-corpus topic distributions $(\boldsymbol{\beta_k})$

LDA was initiated to solve a problem with an earlier probabilistic model, probabilistic latent semantic analysis. (pLSI). [11]

We can calculate the Posterior computation for LDA (computing the dependable circulation of subject structure relying on the three documents. A process referred to, as mentioned above, the "posterior) by: [12]

$$\mathbf{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}} \quad (2)$$

6. SUMMARY OF THE PROCCESS OF LDA AND DATAMINING

Our content-based filtering system recommends documents similar to those that the user has already appreciated in the proposed links. This is calculated by approximating the interests of users (implicitly introduced through the monitoring of their behavior).

The filtering system suggests complements concerning the following axis:

1.0. Knowing the importance of this science.
1.1. The definition of this science its subdivisions.
1.2. Representation of each course section
1.3. General rules.
1.4. Some applications of this science.
1.5. The ability to derive these provisions.
1.6. Questions about the chapter.
…

During a learning session, every action taken by a learner will be automatically saved and stored in the form of a structured journal, written in a text file, delimited by tabs. Each entry in the log file contains general information about the type of action performed, specific information on the exact parameters of the action and relevant contextual information.

6.1. Presentation of how to fit a topic model

In our case we will take the three subjects represented in the following table:

Table 1-Subject Proposed by the platform

Topics	**Information system**	**Computing**	**Web development**
Key words	Information system	Computer	Web design
	Conception	Procedure	Page layout
	Analyze	Programming	Graphic design
	Architecture	language	Php
	Merise	algorithm	Perl
	Uml	Analyzing machine	Html
	Diagramme	Object-Oriented Programming	.net
	Use case	Class	Javascript
	Filter	Object	Css
	Data	Software	Dynamic web page
	….	Hardware	Static web page
		Processes	Interface
		…..	….

The following example shows how to extract the subject from a document

An information system (IS) is any sorted out system for the accumulation, association, stockpiling and correspondence of data. All the more particularly, it is the investigation of complementary networks that individuals and associations use to collect, filters, handles, makes and convey data.

A procedure is a depiction of a process. A basic process can be portrayed just by posting the steps. The rundown of steps is the procedure; the demonstration of tailing them is the process. A technique that can be taken after with no algorithm is called a mechanical procedure. An algorithm is a mechanical method that is ensured to eventually finish.

JavaScript and other scripting languages decide the way the HTML in the got page is parsed into the Document Object Model, or DOM, that represents the stacked website page.

Computing is any objective situated action requiring, profiting from, or making a numerical succession of steps known as an algorithm — e.g. through PCs. Computing incorporates planning, creating and building hardware and software systems; processing, organizing, and overseeing different sorts of data; doing logical research on and with PCs; making PC systems carry on brilliantly; and making and using communications and distraction media.

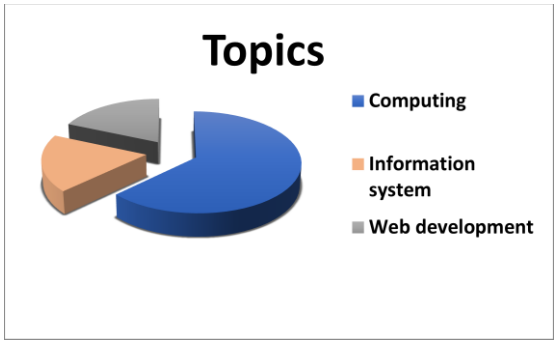

Figure 2-The distribution of different topics in the proposed document

We used the search engine, which attained us with a group of documents. We then saved the suggested documents in the text file specialized to each port of the course. However, the first document, "an information system (IS) is any sorted out system for the accumulation, association…" are words of a single document. And as it can be gathered, each line represents one document.

An information system (IS) is any sorted out system for the acc...
Thanks to the diverse nature of its higher education, Morocco is...
Since our collective information continues to be digitized and...
….

Figure 3-Representation of all proposed document

This can be turned into a sparse representation. This is the second drop in the (Figure 4). For instance, the first line contains 131 unique terms, and the term 683 happened only once. The term 547

happened once, and so on and so forth. It's a reading by the dispersed matrix of the number of words in each article.

```
131 683:1 547:1 432:1 347:1 425:5 235:1 234:6 12:1...
148 478:1 667:1 437:1 324:1 547:1 147:7 239:1 98:2...
231 671:1 321:1 539:3 421:1 342:1 88:4 116:1 543:1...
….
```

Figure 4-Representation of the word counts of each article

These results show the efficacy of the singular value decomposition of the document-term matrix.

```
docs <- read.documents("entree.dat")
k <- 3
alpha <- 1/3
eta <- 0.001
model <- lda.collapsed.gibbs.sampler(documents, k, vocab, 1000, alpha, eta)
```

Figure 5-Code of R by which subjects are found in the database with [13]:

documents A collection of documents in LDA format.
k An integer representing the number of topics in the model.
alpha The scalar value of the Dirichlet hyperparameter for topic proportions.
eta The scalar value of the Dirichlet hyperparamater for topic multinomials.
vocab A character vector specifying the vocabulary words associated with the word indices used in documents.

These five lines of R (figure 5), we can construct the Figure 4, in which subjects are found in the database. (We put words of arrest-words like "the", "but", "of", "and", and then take the time to highlight every word of the document). In this case, there are 50 articles related to science. Topic 3, as an example, is on web development. This process takes a few minutes on a computer.

6.2. Identification of data-driven suggestions

Results achieved with data mining will allow us to draw conclusions (positive or negative) on the level of course comprehension, and the use of links suggested as course complement (prioritizing links that yield a good course comprehension)

In addition, students' perception regarding the suggested links will be taken into consideration (I disagree, I agree).

7. CONCLUSION

In the light of recent scientific advancement, there is growing support for unsupervised machine learning. Flexible components for modeling, scalable algorithms for posterior inference, in addition to the boosted access to gigantic datasets, topic models assure to be a key element in summarizing and tackling the increasing amplification of digitized libraries of information. The handling of hefty archives can now be solved, and that is through probabilistic topic models, which are a set of algorithms that offer a statistical solution to the difficulty of handling these archives of documents.

Our work opens the door on a new perspective that takes into account in addition to the cognitive side, the motivation of the learner, which is essential in order to achieve a sense of satisfaction and well-being of the individual and to respond to psychological needs such as need of autonomy, skills and the need for a social relationship...

And to make the problem of obsolescence, our system must be reinforced by an iterative and incremental model to follow the perpetual change of the means of information.

REFERENCES

1. Foltz, P.W. and Dumais, S.T. **1992**. Personalized Information Delivery: An Analysis of Information Filtering Methods, *Communications of the ACM*, **35**(12): 51-60.
2. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. **1994**. GroupLens: An open architecture for collaborative filtering of netnews, Proc. of the 1994 Conference on Computer Supported Collaborative Work, Furuta, R. and Neuwirth, C., Eds. ACM Press, New York, 1994, p. 175-186.
3. Bonchi, F., Giannoti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., Renso, C. and Ruggieri, S. **2001**. Web log data warehouses and mining for intelligent web caching, *Data & Knowledge Engineering*, **39**: 165-189. (2001).
4. Siahpirani1, F. and SRoy, S. **2016.** A prior-based integrative framework for functional transcriptional regulatory network inference Nucleic Acids Research, 2016 doi: 10.1093/nar/gkw963
5. Wassan, J.T. **2014**. Discovering Big Data Modelling for Educational World, proc in International Educational Technology Conference, IETC 2014, 3-5 September 2014, Chicago, IL, USA.
6. Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaïane, O.R. **2009.** Clustering and sequential pattern mining of online collaborative learning data, (2009) *Knowledge and Data Engineering, IEEE Transactions* on, **21**(6): 759-772.
7. Kerr, D.S. **2014**. Into the Black Box: Using Data Mining of In-Game Actions to Draw Inferences from Educational Technology about Students' Math Knowledge, UCLA Electronic Theses and Dissertations, Ed.D., Education 0249UCLA. (2014)
8. E.W. Weisstein, E.W. **2005**. K-Means Clustering Algorithm. From MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html
9. https://www.scriptol.fr/seo/lda.php, last updated,2012.
10. Blei, D., Ng, A. and Jordan, M. **2003**. Latent Dirichlet allocation, *J. Mach. Learn. Res.* **3**(January 2003): 993–1022.
11. Asuncion, M., Welling, P., Smyth, and Y. **2009**. The On smoothing and inference for topic models. *In Uncertainty in Artificial Intelligence* (2009).
12. T. Hofmann. **1999.** Probabilistic latent semantic analysis. *In Uncertainty in Artificial Intelligence (UAI)* (1999).
13. Chang. J. **2015**. Collapsed Gibbs Sampling Methods for Topic Models, 2015