UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Visión Electrónica
Más que un estado sólido

http: revistas.udistrital.edu.co/ojs/index.php/visele/index

VISIÓN ELECTRÓNICA

A RESEARCH VISION

Bayesian methods for classification inappropriate web pages

Métodos bayesianos para la clasificación páginas web inapropiadas

Jorge E. Rodríguez R.[1], ***Jenny P. Ortiz P.***[2]

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Enviado: 25/09/2016
Recibido: 19/04/2017
Aceptado: 09/08/2017

Keywords:
Bayesian methods
Data Preprocessing
Machine Learning
Web Content Filtering
Web Mining

Palabras clave:
Métodos bayesianos
Preprocesamiento de datos
Aprendizaje automático
Filtrado de contenido web
Minería web

ABSTRACT

The incursion of the Internet has created new forms of information and communication, but it can also carry great dangers, when its use is related to inappropriate content, such as, access to harmful contents and the rise of new kinds of crimes. In this situation, automatic filtering systems identify improper Internet content. This paper describes the use of an algorithm, to automatically filter out inappropriate Web pages. To accomplish this (automatic filtering) task implementation method TAN (Tree Augmented Naive Bayes) is plasma. Data mining algorithms and computational learning for the extraction process, representation and classification of web pages are implemented.

RESUMEN

La incursión de Internet ha creado nuevas formas de información y comunicación, pero también puede conllevar grandes peligros cuando su uso está relacionado con contenido inapropiado, como el acceso a contenidos dañinos y el surgimiento de nuevos tipos de crímenes. En esta situación, los sistemas de filtrado automático identifican contenido inapropiado de Internet. Este documento describe el uso de un algoritmo para filtrar automáticamente las páginas web inapropiadas. Para lograr este método de implementación de tareas (filtrado automático) TAN (Tree Augmented Naive Bayes) es plasma. Se implementan algoritmos de minería de datos y aprendizaje computacional para el proceso de extracción, representación y clasificación de páginas web.

[1] Master of Systems Engineering. Systems Engineer. Teaching Universidad Francisco José de Caldas. Colombia, E-mail: Je_rodrod@hotmail.com, jerodriguez@udistrital.edu.co.

[2] BSc. in Telematics Engineering and Technology in Data Systematization Universidad Francisco José de Caldas. Current position: DANE Colombia. E-mail: jportizp@dane.gov.co, jportizp@correo.udistrital.edu.co.

Citar este artículo como: J. E. Rodríguez R., J. P. Ortiz P. "Bayesian methods for classification inappropriate web pages". Visión Electrónica, algo más que un estado sólido, Vol. 11, No. 2, 179-189, july-december 2017. https://doi.org/10.14483/22484728.13135

1. Introduction

Web content has become in one of the largest sources of data and information for studies, research and development of applications, however, the difficulty to control the Internet also takes significant risks for its use. The information transmitted on the Internet must have legal contents, but there are great difficulties when a page offers different links levels, where the user is sent to some other sites, where he may find information not consistent with his understanding as it happens specifically with the infant population [1].

That is why, there is now a growing need for filtering out harmful and inappropriate Internet contents, given the amount of information and diversity of content that it handles [2]. Integration of Information Technology and Communications (ICT), due to the fact that it has contributed to the progress of communications, access to information and mobility, among others; but also due to its substantially distributed and difficult to control nature has created hazards such as access to harmful content and the emergence of new criminal types such as pornography [3].

Occasionally, information filtering systems are totally effective, because there is too much filtering (content blocked by mistake) or in some other cases a faulty filtering (content issues not locked) [4]. One cause of these problems is that filtering systems do not consider all the existing multimedia information on the Web page (text, images, audio, or video).

Given these difficulties, machine learning techniques, including use agents to allow filtering of information through their evolution over time and their learning environment, optimizing the results gotten [5].

In [6, 7], different algorithms have been used for filtering improper Web pages, such as neighbor- based methods, support vector machines, radial basis functions, etc.

This paper describes the use of Bayesian methods that are proposed in order to compare the results with some existing developments, and thus concludes the feasibility or otherwise of Bayesian methods. The paper is structured following a methodology and procedure of filtering pages as well: security incidents undue pages, filtering systems, data collection and preparation of selected algorithm, test and analysis results, and conclusions of the investigation.

2. Problem

One of the main risks of the Internet is the amount of content that could be described as unfit for users who are freely through the network. Under the label of harmful content, they could be grouped pages with pornographic material with a high degree of violence, promoter of terrorism, racial discrimination and xenophobia, collective suicide, anorexic methods, pages that establish loving contacts through the network to under age, etc. [8, 9].

Illegal contents and harmful conducts online are a constant concern for lawmakers, industry and end-users, particularly parents and educators [9].

Colombian law has already regulated the content circulating in the media such as television, creating slots in accordance with the viewer's age in; checking the contents of the broadcasted programs on national channels through the National Television Commission. The contents of the press and radio programs are also subject to control by the state, but against the regulation of content on the Internet has some provisions designed to protect minors from virtual pornography only [10]. Today, in Colombia telecommunications providing companies, offer products that block violent content, such as ETB (Bogota's Telecommunications Company) which offers the contents guardian, where parents can manage pages that can be accessed by their children. However, every day new Web pages are created with improper content, which involves entering each of these URLs manually, that proves to be a tedious task [7].

Likewise, LAN administrators must make a daily check of creating Web pages and undue contents, enter them manually in the squid (web proxy cache server) so that these are blocked by proxy servers; In undertaking this work sometimes lacks sufficient time, which favors the omission of some of which should be restricted Web sites.

3. Filtering Systems

Filters are programs that prevent users from accessing harmful content on the Web are some techniques used for selecting and filtering information are the tools for access control and monitoring system for the user and the ISP (Internet Service Provider), with use of semantic analysis for blocking keywords leveraging artificial intelligence techniques, access control according to access profiles and schedules, classification and content filtering, among others [11].

Filtering Web systems are classified as knowledge-based with profiles that are explicitly defined by and for the user, based on the individual's behavior with profiles that are defined upon the system according to user behavior and collaborative filtering with profiles are defined by the system to like-minded people using data mining techniques [12].

In spite of several companies have developed Web filtering systems, some of the deficiencies found in the models, is a difficult categorization of sites allowing more or less information than necessary and some language differences that hinder the filtrate was blocked. Some R & D (research and development) that have recently worked are POESIA and TEFILA (filtering Open source for a Safer Internet Environment Access) is an open system for filtering inappropriate content in school settings code. TEFILA guides the filtering in the workplace [13].

For the identification of risky content, filtering systems implementing content classification methods that identify patterns in the information evaluated. Besides, these methods allow content classification, categorizing Web pages by learning embodied in a set of preselected and preprocessed data, called training set. This dataset of training methods of content rating learn the characteristics of the classes to evaluate new documents as belonging or not belonging to the classes considered in training [14].

3.1. *Bayesian Methods*

One of the simplest ways is to classify contents by using Bayesian methods; which are based on subjective interpretation of probability and its central point Bayes Theorem [15].

Among the applications of probability theory, it applies to state the Bayes theorem as an expression of conditional probability that demonstrates the benefits obtained from estimates based on intrinsic knowledge. Bayesian methods specify a probability model containing any previous knowledge about a research setting, thus the probability model is conditioned to adjust the assumptions [15].

The probabilistic Thomas Bayes theorem is useful when we know the outcome of an experiment, but we do not know any of the intermediate steps in which you are interested in. [16].

The Bayes theorem is given by the following statement: Let B1, B2, ..., Bn a complete system of events with $p(Bi) > 0$ for all i. If A is any verified event:

$$P(Bk|A) = \frac{P(A|Bk) * P(Bk)}{\sum_{i=1}^{n} P(A|Bi) * P(Bi)} \tag{1}$$

Let:
$P(Bi)$: A priori probability
$P(Bi|A)$: A posteriori probability
$P(A|Bi)$: verosimilarities

$$p(spam|words) = \frac{p(words|spam)p(spam)}{p(words)} \tag{2}$$

The Bayesian filter needs a database containing words and some other criteria (IP addresses, hosts...), to calculate the probability that a specific email is spam, drawn from a sample of spam and another valid email. Each word is assigned a probability value based on the frequency of occurrence of that word in a spam facing the same frequency of occurrence in a valid email.

These assignments are made through a process of analyzing the email, divided by the probability of finding those words in any e-mail [17].

By having the database, the filter may act. When a new mail is received, the analysis breaks the text into words and select the most relevant, which will process the Bayesian filter by calculating the probability that we have received mail is spam or not. If the probability exceeds a set threshold, it will be considered spam.

4. Data Collection And Preparation

Web content filtering corresponds to the process that restricts or allows access to an HTML (Hypertext Markup Language), based on some kind of analysis done on this.

4.1. *Removing information*

The techniques of extracting information in the first instance refer to the method of collecting the documents to be evaluated in this case Web pages, and secondly the form of extracting information from these documents.

The dataset was performed manually, considering the deficiencies and disadvantages presented by other methods described below. First, a dataset of pornographic and non- pornographic sites available on the Web use was found, if there are sets already collected pages, as it can be seen in [18], where they use a dataset of pornographic and not Web pages, that are part of POESIA (Public Open-Source Environment for a Safer Internet Access, http://www.poesia-filter.org/) project, which is an open source software for filtering content funded by the European Union EU. However, it was not possible to access to this data set because it is not available on the official website. Another option for this step are the crawlers, which are programs that scan the Web automatically and its basic operation is that from one or a set of URLs extracted and added links to later visit and perform different tasks.

You can use these tools to automatically download pages to your hard drive, providing a base assembly the URLs; so you can visit and download.

However, the difficulty with this approach is that it cannot be assured that if provided a pornographic URL, all visited pages will be of this type; which requires validation of each collected document [18].

The manual collection of pages needs more work and time, but it ensures that, all the dataset is for the domain being addressed and ruled noisy information (e.g. pages made entirely in flash). 2500 pages were collected, (see Table 1).

Table 1: Technical Data Sheet of the Survey.

KIND	AMOUNT	PERCENTAGE
Pornographic	750	30 %
Non-pornographic	1750	70 %
Total	2500	100 %

For extracting information from Web pages, it is necessary to use some tools called parsers.

4.2. Representation of Web pages

Within this process, a representation of the pages is created, which is the dataset of characteristics that best describe the content, so, the classification model has high levels of accuracy and better performance. The representation should include the types of pages being addressed and the characteristics of the content they present. In general, the representation and analysis of hypertext can be classified into two types based on the content and hyperlinks based on [19], which are directly related to the Content and Web Mining and Web Mining Structure.

Representation based on the content: the Web Content Mining extracts information from Web pages. The content corresponds to the collection of facts used in pages to provide information to users, that is, transforming from Web data to Web knowledge. For example, it includes text, images, audio, video or structured records. The HTML content included on the target page provides useful information. The URL itself DOM tags such as title, subtitled pages and embedded metadata, such as keywords, language, etc. these help to describe the contents of a Web page [6].

Hyperlinks based representation: the Web is a vast collection of documents linked together by links or references. The language of communication used by each document is based on hypertext, embedded in HTML code. This language describes the way they should be shown a web page in a browser.

In general, the Web can be seen as a directed graph, where nodes are Web pages and hyperlinks are represented by URLs. The importance of this structure or topology, is reflected in the tasks performed by Web search engines to determine the ranking and relevancy of each page. This task often develops the reference to other documents page in terms of links or according to weight or participation of words within the document.

This approach has taken great interest in recent years and has been the subject of many studies [20, 21] and [22]. In this approach, the hyperlinks on a page and its structure or topology are studied in order to extract information for a better representation of the target page; daughter information pages, parent pages, information pages and sisters as the anchor text hyperlinks are used for this purpose. Moreover, as the Page Rank algorithms [23] and HITS [19] are used to determine the relevance and authority of a Web page based on the structure of these link representations. Considering that, you can decide to create a dictionary of terms to compare information extracted from each page to provide more effective representation as used in [8] and [24, 25]. The selected ones within the page attributes are:

4.3. Based on the content Features:

- Title (p): number of terms found in the dictionary and present on the title page p.
- Keywords (p): number of terms found in the dictionary and present in the keyword metatag on page p.
- Description (p): number of terms found in the dictionary and present in the page description metatag p.
- Totalwords (p): total number of words on the page p. H1 (p): number of subtitles (tag) of the page containing p dictionary terms.

- TotalImages (p): total number of images on the page p.
- AltImages (p): number of page images containing p dictionary terms in the "alt" attributes or "title" attributes.

4.4. *Characteristics based on hyperlinks:*

- TotalLinks (p): total number of links on the page p.
- DictionaryLinks (p): number of links on the page containing p in your anchor text dictionary terms.
- Titlelink (p): average (number of terms found in the dictionary and present in the title of the daughter pages (q) 5 daughters' pages (q) p random page.
- Keywordlinks (p): average (number of terms found in the dictionary and present in the keywords (metatags) daughters pages (q) 5 daughters pages (q) p random page.

5. Selected Algorithm

Data mining receive a classifier training data, in which each entry is marked with a label or class of a finite set. The classifier is trained using these dataset, and once trained, you are provided with unlabeled entries to be allocated by [26]. This same procedure can be followed with hypertext documents, in which a classifier learns from previously tagged documents and this is the ability to assign the label to new unclassified documents. The techniques used for classification of hypertext can be either used in Data Mining as hypertext documents can be represented in a way that fit the specific requirements of each

5.1. *Description of the method TAN*

Get a Bayesian network from data is a learning process, which is divided into two aspects [27]:

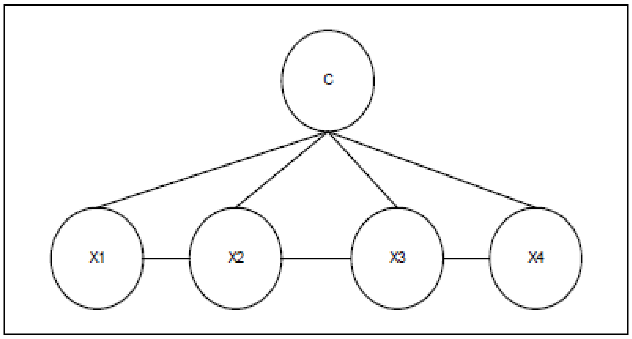
1 Parametric study: given a structure, obtain the prior probabilities and conditional required.

2 Structural Learning: Getting the Bayesian network structure, that is, relationships of dependence and independence between the variables involved.

Structural learning techniques depend on the type of network model such as: trees, trees poly and multiconnected networks. Another alternative is to combine subjective expert knowledge of learning. For it is part of the structure given by an expert, which is validated and improved by using statistical data.

For this type of structure, a tree structure is started by the predictor variables, later to connect the variable class with each of the predictor variables. Figure 1 illustrates an example of naïve Bayes augmented tree structure [17].

Figure 1: Example of naïve Bayes augmented tree structure - TAN.



Source: [8].

In [28–30] show an algorithm called Augmented Tree Network (TAN) which is basic in an adaptation algorithm Liu Mr. Chow. In the algorithm takes into account the amount of mutual information class conditional variable, rather than the amount of mutual information in which the algorithm is Chow-Liu based on. The amount of mutual information between discrete variables X and Y conditional on the C variable is defined as:

$$I(X,Y|C) = \sum_{i=1}\sum_{j=1}\sum_{r=1} P(xi,yi,cr) = \frac{logP(xi,yj,cr)}{P(xi|cr)P(yi|cr)} \quad (3)$$

As it can be seen from the pseudocode in Table 2, is built in five steps. In the first step the amounts of mutual information for each pair of variables (Xi, Xj) conditioned by the variable C. Then it must build a complete undirected graph with n nodes, one for each of the predictor variables are calculated, in which the weight of each edge is given by the amount of mutual information between the two variables together by the class subject to the variable edge. Kruskal's algorithm of the weights obtained in the previous step to build the maximum weight spanning tree as follows:

1 Assign the two heavier edges of the tree you want to build.

2 Examine the next edge of greater weight, and add it to the tree unless it forms a cycle, if it is discarded and the next heavier edge is examined

3 Repeat step 2 until the edges are selected

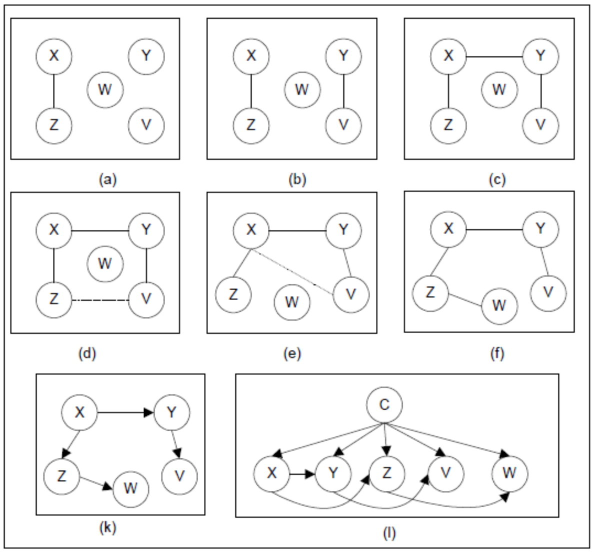
The theoretical properties of this TAN algorithm construction are similar to those of Chow-Liu algorithm. That is, if the data has been generated by a structure Augmented Tree Network, TAN algorithm is asymptotically correct, in the sense that if the sample is large enough case, recover the structure that generated the case file. An example of application of the algorithm is shown in Figure 2.

Table 2: Pseudocode algorithm TAN.

Step 1. Calculate $I(Xi, Xj\|c)$ with $i > j$, $i, j = 1, ..., n$
Step 2. Build a complete and undirected graph whose nodes correspond to the predictor variables. Assign each edge connecting variables and weight given by $X_1, \dots X_n$. Assign each edge connecting variables X_i y X_a weight given $I(X_i X_j\|C)$
Step 3. From the graph above and following the complete Kruskal's algorithm we build a maximum spanning tree.
Step 4. Transform the resulting tree undirected one run, choosing a variable as root, then address it to the remaining edges.
Step 5. Build a model TAN adding a node labeled and then an arc from each predictor variable

Source: [31].

Figure 2: Illustration of TAN algorithm with five predictor variables X, Y, Z, V and W is assumed that the order of the amounts of conditional mutual information was



Source: own.

$$I(X,Z|C) > I(Y,V|C) > I(X,Y|C) > I(Z,v|C) > I(X,V|C) > I(Z,W|C) > I(X,Y|C) > I(X,W|C) > I(Y,Z|C) > I(Y,W|C) > (V,W|C) \tag{4}$$

Figures (a) to (f) relate to the application Krustall's algorithm The subfigure (g) corresponds to step 4 of the algorithm TAN and finally figure (h) TAN step 5 is performed. The qualification model is obtained:

$$P(C|x,y,z,v,w) \propto p(c)p(x|c)p(y|x,c)p(z|x,c)\,p(v|y,c)p(w|z,c) \tag{5}$$

5.2. *Why TAN method?*

Given the good performance offered by the Naïve Bayes classifier (NB), despite the strong assumption made, regarding the independence of the attributes given the class, one wonders if the results will not be better after removing this assumption [30]. Bayesian networks are an alternative for data mining, which has several advantages as follow:

- Allow easy learn about relationships of dependence and causality.
- Allow combine knowledge with data.
- Avoid over-fitting the data.
- They can handle incomplete databases.

Given that the Naïve Bayes classifier is the simplest Bayesian network that can build oriented classification, TAN is an extension of Naïve Bayes classifier, with which it is intended to keep the computational simplicity of Naïve Bayes classifier but trying to improve the success rate during classification, this algorithm has two distinct advantages: a low computational complexity and, secondly, ensuring that the network structure obtained is the maximum likelihood of the set of all possible structures TAN [31].

5.3. *Algorithm for classification of features based on entropy [31].*

The similarity measure S of two numerical samples is

$$S_{ij} = e^{\alpha D_{ij}} \tag{6}$$

Where Dij is the distance between samples Xi and Xj

$$D_{ij} = \sum_{k=1}^{n} \frac{[Xik - Xjk]^2}{(maxk - mink)} \quad (7)$$

And α is a parameter mathematically expressed as:

$$\alpha = -(ln0{,}5)/D \quad (8)$$

D is the average distance between samples in the data set, but in practical applications approaches.

The entropy measure is given by:

$$E = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (Sij{*}logSij{+}(1{-}Sij){*}log(1{-}Sij) \quad (9)$$

Where N is the number of records.

The steps of the algorithm are defined as follows (Table 3):

Table 3: Pseudo selection algorithm for entropy.

Step 1. Start with the initial set of characteristics F
Step 2. For each function $f \in F$: Remove a function f from F and get a subset F_f. find the difference between the entropy and the entropy for all Ff. whether to compare the differences : $(E_F - E_{F-F1})(E_F - E_{F-F2})(E_F - E_{F-F3})$
Step 3. Let F_k be the characteristic such that the difference between the entropy F and entropy for F is f_k minimal
Step 4. Updating the feature set : $F = F - \{f_k\}$ Where - is a set difference operation. For our example, if the difference $(E_F - E_{F-F1})$ is minimal, then the reduced feature set is $\{F2, f3\}$ return to the bottom of the sorted list.
Step 5. Repeat steps(2)-(4) until there is only one F function

Source: own.

6. Implementation and Testing

6.1. *Attributes Selection [31]*

Within the data set 70 % of the data in order to eliminate noise and work with the most relevant information is selected, this selection is made through the selection algorithm from the entropy where expert assembly data to determine where to stop, then the most important attributes relate (see Table 4):

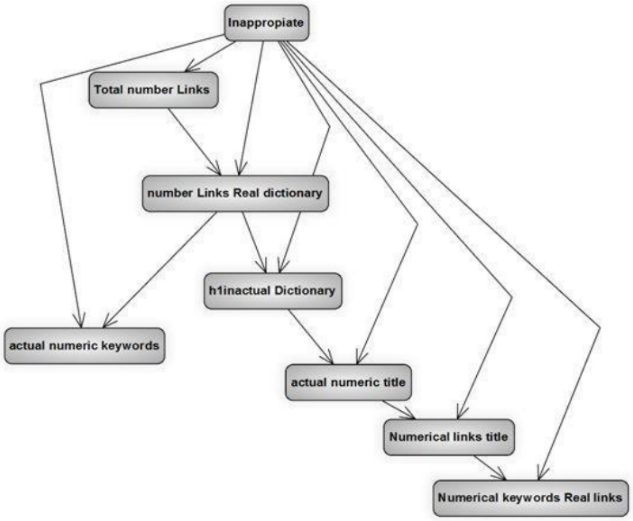
Table 4: Pseudo selection algorithm for entropy.

RELEVANT ATTRIBUTES
@attribute actual numeric title
@attribute actual numeric keywords
@attribute Total number Links Real
@attribute number Links Real dictionary
@attribute h1inactual Dictionary
@attribute Numerical links title
@attribute Numerical keywords Real links

Source: own.

6.2. *Bayesian Network*

Figure 3: Shows the Bayesian network, with dependence and independence relations between variables involved.



Source: own.

Testing each of the algorithms and the initial basic configuration is established: 2500 of pages obtained in the collection phase, the dataset for training and testing phase split. (See Table 5).

Table 5: Pseudo selection algorithm for entropy.

Type	**Training**	**Test**	**Total**
Pornographic	525	225	750 (30 %)
Non-pornographic	1225	525	1750 (70 %)
Total	1750(70 %)	750(30 %)	2500(100 %)

Source: own.

To compare the implemented algorithms; Decision Trees, Naive Bayes, KNN were chosen, since reviewing the literature, these are often used and selected.

When testing in each of the algorithms the overall percentage of precision and classical assessment measures used in classification processes such as precision, coverage and F- measure (harmonic mean of precision and coverage) was obtained.

Precision refers to the fraction of individuals that have been classified as of the relevant class and, in fact are of that class [32]. Therefore,

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

TP: true positive rate (cases correctly classified
FP: false positive rate (cases falsely classified as a particular class).

And then there is the recall (sensitivity), it refers to the fraction of examples of the kind of the whole that is classified correctly,

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Recall: the proportion of cases classified as a given class divided by the actual total of that class (equivalent to the TP rate). And the combination of these two is a measure -F

$$f - measure = \frac{2 * precision * Recall}{precision + Recall} \tag{12}$$

For more information about the assessment methods training / testing and cross-validation and evaluation measures (Table 6), see [33, 34] and [35].

Table 6: Precision and general error of classification algorithms Hypertext.

Algorithm	**Evaluation method** **Cross validation**	**Percentage accuracy**		**Error rate**	
Decision tree	(n=90)	96.2 %	Precision Average 96.09 %	3.8 %	Error Average 3.91 %
	(n=80)	96.04 %		3.96 %	
	(n=70)	96.12 %		3.88 %	
	(n=60)	96.12 %		3.88 %	
	(n=50)	95.96 %		4.04 %	
	(n=40)	96.08 %		3.92 %	
Naive Bayes	(n=90)	95.12 %	Precision Average 95.13 %	4.88 %	Error Average 4.87 %
	(n=80)	95.16 %		4.84 %	
	(n=70)	95.04 %		4.96 %	
	(n=60)	95.16 %		4.84 %	
	(n=50)	95.16 %		4.84 %	
	(n=40)	95.16 %		4.84 %	
KNN	(n=90)	93.72 %	Precision Average 93.68 %	6.28 %	Error Average 6.32 %
	(n=80)	93.64 %		6.36 %	
	(n=70)	93.72 %		6.28 %	
	(n=60)	93.68 %		6.32 %	
	(n=50)	93.64 %		6.36 %	
	(n=40)	93.68 %		6.32 %	
TAN	(n=90)	96.28 %	Precision Average 96.41 %	3.72 %	Error Average 3.59 %
	(n=80)	96.44 %		3.56 %	
	(n=70)	96.52 %		3.48 %	
	(n=60)	96.4 %		3.6 %	
	(n=50)	96.4 %		3.6 %	
	(n=40)	96.44 %		3.56 %	

Source: own.

6.3. *Analysis of Results*

The results of each of the tests are evaluated and classifiers are chosen for using in implementing default filter (see Table 7).

Whereas decision tree and TAN and considering that an ideal model would be one that does not leak or reject the 100 % non-pornographic content and also filter the content higher percentage if it is; we can say that the greater coverage and greater accuracy non- pornographic Web pages in a better model will be achieved. Therefore, and analyzing these measures, technology with higher levels of accuracy in each of these is TAN.

Table 7: Evaluation measures with the cross validation method.

Algorithm	**Pornographic**		**Non- pornographic**	
	Recall	**F-Measure**	**Recall**	**F-Measure**
Decision Tree	0.909	0.933	0.983	0.972
Naive Bayes	0.886	0.916	0.978	0.965
KNN	0.887	0.893	0.958	0.955
TAN	0.925	0.939	0.980	0.974

Source: own.

7. Conclusions

Today, Web content filtering features with tools and methods that require lots of effort and management by users or network administrators, and its efficiency is based on the constant updating of the information used for filtering. With constant and accelerated growth of the World Wide Web, the operation and design of these tools do not provide the full power required.

In the observation of the obtained results, it is seen that the effectiveness of the method is so increasing in all cases, surpassing other classification methods because the algorithm is contemplated that rather than assuming all independent variables, given the class, support between attributes.

Design and develop a software prototype will allow to analyze the information and decide to restrict or not a page online using incremental learning techniques that allow the self-adapt in order to respond and update their software mechanisms decision.

This work can be used as a starting point for the extension and applicability, for example by filtering pages from other domains and in different languages, filtering the different multimedia contents of web pages, building Web directories or search engine optimization information, so it is possible to perform a performance comparison of the filters in terms of efficiency.

8. Future Work

Studying and testing with incremental learning algorithms which reviewed the concept learned to receive new examples of efficient considering that the learning is mainly characterized incremental learning be able to incorporate the information to provide new experiences (that were not previously available in the data set) to the model is leading [36] and able to make it, evolves to represent increasingly more complex concepts, see [37, 38].

References

[1] M. Villarreal, “Regulación de contenidos en Internet. Estudio cualitativo, Colombia y derecho comparado”, *Revista Estudios Socio-Jurídicos*, vol 10, no. 2, pp. 254-281, december 2008.

[2] Netcraf, “Web Server Survey”, june 12 th 2009, [Online]. Available: http://news.netcraft.com/archives/2009/06/17/june_2009_web_server_survey.html

[3] M. A. Hernández, P. López, “Contenido nocivo en la red. ¿Qué Hacer? “, Universidad de Murcia, june 12 th 2009, [Online]. Available: http://www.congresointernetenelaula.es/virtual/archivosexperiencias/200806041531262008_COM_Contenido_nocivo.doc

[4] RED USI, “Preguntas Frecuentes - RED USI”, june 12 th 2009, [Online]. Available: http://www.usi.org.uy/es/preguntasfrecuentes/index.html#faqs-5

[5] A. García, “La Regulación De Los Contenidos Audiovisuales En Internet”, Comisión del Mercado de las Telecomunicaciones, june 12 th 2009, [Online]. Available: http://serbal.pntic.mec.es/~cmunoz11/casti.pdf

[6] J. E. Rodríguez, H. A. Barrera, S. P. Bautista, “Software para el filtrado de páginas web pornográficas basado en el clasificador KNN – UDWEBPORN”, *Revista Avances en Sistemas e Informática*, vol. 8, no. 1, pp. 43-49, march 2011.

[7] J. E. Rodríguez, A. P. Herrera, M. L. Rojas, “Sistema de bloqueo automático para páginas web que incitan

a la violencia a través de un algoritmo híbrido de aprendizaje computacional". *Revista Vínculos*, vol 10, no 2, july 2013.

[8] A. L. Rotta, "La protección de los niños y niñas en internet – Los sistemas de filtrado", I Congreso internacional sobre ética en los contenidos de los medios de comunicación en internet, october 01 th 2001. [Online]. Available: http://www.ugr.es/~sevimeco/congreso.html

[9] I. M. Solano, M. A. Hernández, "La seguridad de los menores en Internet", Universidad de Murcia, april 19 th 2005. [Online]. Available: http://ticemur.f-integra.org/vticemur/documentos/mesa5/C2.pdf

[10] A. García, "La Regulación De Los Contenidos Audiovisuales En Internet", Comisión del Mercado de las Telecomunicaciones, june 12 th 2009, [Online]. Available: http://serbal.pntic.mec.es/~cmunoz11/casti.pdf

[11] M. Heins, C. Cho, A. Feldman, "Internet filters a public policy report", Brennan Center for Justice, june 12 th 2009, [Online]. Available: http://www.fepproject.org/policyreports/filters2.pdf

[12] L. M. Quiroga, "Sistemas de filtrado: Un puente tecnológico entre oferta y demanda de información en línea al servicio de la toma de decisiones", june 12 th 2009, [Online]. Available: http://www.cepal.org/dds/noticias/paginas/2/14632/ppt_LMQuiroga_Hawaii.ppt

[13] J. M. Gómez, E. Puertas, F. Carrero, M. de Buenaga, "Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet", june 12 th 2009, [Online]. Available: http://www.esp.uem.es/jmgomez/papers/sepln03.pdf

[14] A. I. Oviedo, C. A. Manco, J. E. Guerra, "Sistema Multiagente para el filtrado de pornografía mediante la evaluación del contenido multimedial de las páginas Web". *Revista en Telecomunicaciones e Informática*, vol. 3 no. 5 pp.55 -73, june 2013.

[15] O. Mesa, M. Rivera, J. Romero, "Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión", Universidad del Rosario. La simulación al servicio de la academia, Edición 2, 2011, pp. 1-3.

[16] A. Gascón, M. de la Puente, "Clasificación Jerárquica de contenidos Web", june 12 th 2009, [Online]. Available: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/30.pdf

[17] E. Fernández, "Análisis De Clasificadores Bayesianos", june 12 th 2009, [Online]. Available: http://materias.fi.uba.ar/7550/clasificadores-bayesianos.pdf

[18] J. Gomez, F. Carrero, E. Puertas,"Named Entity Recognition for Web Content Filtering", Natural Language Processing and Information Systems, pp. 286-297, 2005.

[19] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, vol. 46, pp. 604-632, 1999.with anchor extraction and links Analysis, https://doi.org/10.1145/324133.324140

[20] W. Cohen, "Improving A Page Classifier with Anchor Extraction and Link Analysis", *Advances in Neural Information Processing Systems*, vol. 15, pp. 1481-1488, 2002.

[21] E. Glover, E. K. Tsioutsiouliklis, S. Lawrence, D. Pennock, G. Flake, "Using Web Structure for Classifying and Describing Web Pages", Proceedings of the eleventh international conference on World Wide Web, june 12 th 2009, [Online]. Available: //dpennock.com/papers/glover-www-2002-using-Web-structure.pdf

[22] A. Prakash, K. Kumar, "Web Page Classification based on Document Structure", International Institute of Information Technology, june 12 th 2009, [Online]. Available: http://www.iiit.net/students/stud_pdfs/kranthi1.pdf

[23] B. Sergey, P. Lawrence, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Stanford University Computer Networks and ISDN Systems 30, pp. 107- 117, 1998, https://doi.org/10.1016/S0169-7552(98)00110-X

[24] S. Guy, L. Jian, M. Ying, L. Sheng, "Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model". *Journal of Zhejiang University Science*, vol. 5, no.9, pp. 1106-13, 2004, https://doi.org/10.1631/jzus.2004.1106

[25] M. Chau, H. Chen, "A machine learning approach to web filtering using content and structured analysis, Decision Support Systems", *Decision Support Systems*, vol. 44, issue 2, pp. 482-494, 2008, https://doi.org/10.1016/j.dss.2007.06.002

[26] S. Chakrabarti, "Mining the Web: Discovery Knowledge from Hypertext Data", USA: Morgan Kaufmann, pp. 125-173, 2003, https://doi.org/10.1016/B978-155860754-5/50006-9

[27] E. Morales, "Redes Bayesianas en Minería de Datos", september 21 th 2016. [Online]. Available: http://dns1.mor.itesm.mx/~emorales/Cursos/KDD03/node44.html

[28] N. Friedman, D. Geiger, M. Goldizmitdt, "Bayesian networks classifiers", Machine Learning, 1997, https://doi.org/10.1023/A:1007465528199

[29] T. Mitchell, "Machine Learning", USA: McGraw-Hill, 1997, pp. 230-247.

[30] J. Hernández, M. Ramírez, C. Ferri, "Introducción a la Minería de Datos", España: Prentice Hall. 2004, pp. 97-125.

[31] M.Kantardzic. DATA MINING. "Entropy measure for features ranking Algorithm" 2 nd Edition. August 2011 p.p 29-30.

[32] J. Botía, "Herramientas de Minería de datos: WEKA (Waikato Environment for Knowledge Analysis)". june 12 th 2009, [Online]. Available: http://webs.um.es/juanbot/miwiki/lib/exe/fetch.php?id=tiia&cache=cache&media=practica_tiia2.pdf

[33] W. Lan, E. Frank, "Data Mining, Practical Machine Learning Tools and Techniques", USA: Morgan Kaufmann, 2005, pp. 143-184.

[34] D. Larose, "Discovering Knowledge in Data", USA: Wiley Interscience, 2004, pp. 90- 106, https://doi.org/10.1002/0471687545

[35] W. K. Chen, "Linear Networks and Systems". Belmont: Wadsworth, 1993, pp. 123– 135.

[36] J. Schlimmer, D. Fisher, "A Case Study of Incremental Concept Induction". Proc. 5 th National Conf. on Artificial Intelligence, 1986, pp. 495–501.

[37] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann. 1993.

[38] R. Morales, G. Ramos, "Algoritmo multiclasificador con aprendizaje incremental que manipula cambios de conceptos", Universidad de Granada, june 12 th 2009, [Online]. Available http://digibug.ugr.es/bitstream/10481/35217/1468964.pdf