Statistical Methods for Performance Evaluation of WEB Document Classification

Daniel Volovici[1], Macarie Breazu[2], Gabriel Dacian Curea[3], Daniel Ionel Morariu[4]

"Lucina Blaga" University of Sibiu, 10, Victoriei Blv., 550024, Sibiu, Romania

[1]daniel.volovici@ulbsibiu.ro; [2]macarie.breazu@ulbsibiu.ro; [3]adi.mitea@ulbsibiu.ro; [4]daniel.morariu@ulbsibiu.ro

Abstract: The principal aim of this paper is to make a review of main statistical methods for classifying documents that could be easily adapted in the context of Web document retrieval. After presenting the most popular methods of classification we will also define the most accurate indicators for assessment of classifiers performance. Thus we will refer to the *recall*, *precision*, *fscore*, *sensitivity* and *specificity*. We will also describe how these indicators can be calculated in the context of Web documents.

Keywords: Information retrieval, Classification, Naïve Bayes, Evaluation metrics.

1. Introduction

Access to information is an increasingly frequent topic discussed both at national and international level. Today we can not talk about some traditional information skills, where each country can have access to virtual planetary database.

The main reason why people require information on another medium than the traditional one concerns the need of some specialised information. For example, in the field of science, it takes a long time to update information. Some specialized, cutting-edge information can be found only on this media, because the lengthy book publication process required for a traditional format makes printed books obsolete.

Development of online services must be the main concern in librarian world. In "WWW Library Directory" magazine [15] are identified over 30 types of services involving using of internet and reference services, databases and indexes sites, search guides, information services for trade and industry, banks of images, and so.

In December, 1999 the European Commission launched an initiative entitled "eEurope: An Information Society for All", [16] initiative which proposed ambitious targets, namely to provide the benefits of information society to all Europeans. The initiative focuses on ten areas of priority, from education to transportation, from health to disability issues. The idea behind this initiative was to build a strategy to modernize the European economy, hoping that it will become "the most competitive and dynamic knowledge-based economy in the world" [4]. In the same idea was started also a project for Romania [1].

Recently there was a new generation of Web technologies designed under the concept of Semantic Web project launched by Tim Berners-Lee [2]. The semantic Web seeks to access the data with heterogeneous semantics and obtain some useful knowledge from data through various services offered in the Web space. Semantic Web claims to improve communication between peoples using different technologies, extending the interoperability of databases and providing new mechanisms for agent-based data computation in which the people and the machines will work online and make possible a new level of interaction between scientific communities [5] [12].

2. Analyzing Text Data and Information Retrieval

Information retrieval (IR) is a field developed in parallel with database systems. Information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. A typical information retrieval problem is to locate relevant documents based on user input, such as keywords or example documents. Usually information retrieval systems include on-line library catalogue systems and on-line document management systems. Since information retrieval and database systems each handle different kinds of data, there are some

database system problems that are usually not present in information retrieval systems such as concurrency control, recovery, transaction and management. There are also some common information retrieval problems that are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords and the notion of relevance.

Basic measure for text retrieval

There are some indicators for measure efficiency of information retrieval algorithms. May [Relevant] be the set of documents relevant to a query and [Retrieved] be the set of documents retrieved. The set of documents that are both relevant and retrieved is denoted by $[\text{Relevant}] \cap [\text{Retrieved}]$. There are tow basic measures for assessing the quality of text retrieval:

1. ↳ *Precision*: is the percentage of retrieved documents that are in fact relevant to a query. It is defined as follows:

$$precision = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|} \qquad (1)$$

2. ↳ *Recall*: is the percentage of documents that are relevant to the query and were in fact retrieved:

$$recall = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \qquad (2)$$

Precision ranges from 1 (all retrieved documents are relevant) to 0 (none of relevant document is retrieved). *Recall* range from 1 (all relevant documents are retrieved) to 0 (none of retrieved document is relevant). In fact *precision* represents a quantitative measure of the information retrieval system while *recall* represents a qualitative measure of this system.

Keyword-based and similarity-based retrieval

Most information retrieval systems support *keyword-based* and *similarity-based* retrieval. In keyword-based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A user provides a keyword or an expression formed out of a set of keywords, such as "car and repair shop". A good information retrieval system needs to consider synonyms when answering such query. This is a simple model that can encounter two difficulties: (1) the *synonyms* problem, keywords may not appear in the document, even though the document is closely related to the keywords; (2) the *polysemy* problem: the same keyword may mean different things in different contexts.

The information retrieval system based on similarity finds similar documents based on a set of common keywords. The output for this system is based on the degree of relevance measured by using keywords closeness and the relative frequency of the keywords. In some cases it is difficult to give a precise measure of the relevance between keyword sets. In modern information retrieval systems, keywords for document representation are automatically extracted from the document. This system often associates a stop-list with the set of documents. A stop-list is a set of words that are deemed "irrelevant" and can vary when the document set varies. Another problem that appears is *stemming*. A group of different words may share the same word stem. A text retrieval system needs to identify groups of words where the words in a group are small syntactic variants of one another, and collect only the common word stem per group.

Let's consider a set of d documents and a set of t terms for modelling information retrieval. We can model each of the documents as a vector v in the t dimensional space $\mathbb{R}^t$. The i^{th} coordinate of $v()$ is a number that measures the association of the i^{th} term with respect to the given document: it is generally defined as 0 if the document does not contain the term, and nonzero otherwise. The element from a $v()$ vector, v_i can indicate the frequency of the term in the document and there are a lot of methods to define frequency of the terms. Similar documents are expected to have similar relative term frequency, and we can measure the similarity among a set of documents or between a document and a query. There are many metrics for measuring the document similarity. The used one is the Euclidean distance but the most used is cosine similarity defined as:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \qquad (3)$$

where $v_1 \cdot v_2$ the standard dot products defined as $\sum_{i=1}^{t} v_{1i} v_{2i}$ and ||v1|| is defined as $\|v_1\| = \sqrt{v_1 \cdot v_1}$.

The similarity ranges from 1 (perfectly similar) by 0 (orthogonal) to -1 (dissimilar). Great values of similarity represent a small angle between vectors and therefore the vectors (the documents) are similar.

3. Statistical Methods for Classification

It is known (without scientific proof but with statistical proof) that the classification performance depends on the area of the data that need to be classified. This empirical observation justifies the need to introduce new algorithms for classification and to see their performance in different contexts.

In generally, the complex applications of digitized there are used the following documents classification techniques: the technique of naive Bayesian classifier, TF-IDF technique, Latent Semantics Indexing technique, Support Vector Machine (SVM) technique, the technique of Artificial Neural Network (ANN), the technique of nearest value k (k-nearest neighbour KNN), Concept Mining technique. The algebraic algorithms have been less used, or not used at all, perhaps because of the lack of effective implementation.

The algebraic methods for documents classification support the design and implementation of the adaptive applications and the systems advice and recommendation by completing:

- Analysis of data on the Web;
- Analysis user logins;
- Link analysis scoring string in web browsing (click-stream sites);
- Create models of users with specific interests.

In the Web Mining area and related areas the algebraic tools and especially heterogeneous hierarchies of algebraic structures are suitable to create a framework of operator space for web document classification. The worktable framework in the web classification stage is provided by modelling using HAS hierarchy. This model provides cooperation and collaborative work of the classifiers in applications of Web space [9].

You can create a framework for working effectively in the task of classification for Web Mining. Thus, the running of several classifiers in a collaborative framework provides final results which become federal tools, adaptive and recommendation for Web applications. Modelling the hierarchy of heterogeneous algebraic structures ensure the creation of this working framework by providing accessibility to a system of classifiers that can be the basis for designing and implementing any adaptive Web system (Adaptive Web System - AWS). The classification accuracy in this new created working framework ensures the possibility to create new classification based on the existing hybrids.

Evaluation metrics in classification

Most evaluation metrics in the classification process is designed to achieve uniformity of classes induced by a certain characteristic from a set of samples. Other metrics are designed to realize the differencing power in the context of feature selection as a method to combat the problem of interaction characteristics [7].

Definition 1 (*Metric based on purity* [3]). An evaluation metric based on purity M quantifies a quality of partitions induced by a feature X_k on a lot of training samples T.

Metrics based on purity define M by measuring the amount of class uniformity obtained by decomposition of T into subsets of samples {Tm} induced by X_k. Be $\vec{P}$ the vector of class probabilities estimated from the full set T, be $\vec{P}_m$ the vector of class probabilities estimated from T_m and be I a measure of impurity of a class probabilities vector. M is defined as:

$$M(X_k) = I(\vec{P}_m) - \frac{|T_m|}{|T|} \sum_m I(\vec{P}_m) \qquad (4)$$

Different varieties of M can be obtained by altering the function of impurity I. For example, to gain information (Information

Gain [14]), impurity is defined in terms of entropy as:

$$I_{entropy}(\vec{P}) = -\sum_i p_i \log_2 p_i \quad (5)$$

Another example is the Gini index (Gini Index [13])

$$I_{gini}(\vec{P}) = -\sum_i p_i^2 \quad (6)$$

Previous equations cover most traditional metrics, but there are two major limitations:

- First is the tendency of features with more permitted values. Induction of several subsets of samples result in increased likelihood of finding common subsets of classes, but the cost of processing. To solve this problem several solutions have been proposed [14];
- Second is the inability to detect the relevance of a characteristic when its contribution is hidden target concept by combination with other features. This problem is known as feature interaction [8].

Another category of metrics are based on the discrimination power of each feature, i.e. on the ability of a characteristic to separate the samples into different classes.

Definition 2 (*Metric based on discrimination* [13]) Let $\vec{X}_i$ and $\vec{X}_j$ two samples very close in relation to measure of the distance D. To the characteristic (feature) X_k is assigned a certain power of discrimination if it has different values when the class values $\vec{X}_i$ and class $\vec{X}_j$ are different.

An example of discrimination is when $x_k^i = x_k^j$ where $C(\vec{X}_i) \neq C(\vec{X}_j)$. Most often this condition is true for pairs of similar samples, high quality feature of X_k.

Two examples of discrimination based metrics are most used in the algorithms Contextual Merit and RELIEF [6]. The distance between samples is defined as follows:

$$D(\vec{X}_i, \vec{X}_j) = \sum_{k=1}^{n} d(x_k^i, x_k^j) \quad (7)$$

For nominal features $d(x_k^i, x_k^j)$ is defined at:

$$d(x_k^i, x_k^j) = \begin{cases} 1 & if\ x_k^i \neq x_k^j \\ 0 & f\ x_k^i = x_k^j \end{cases} \quad (8)$$

For numerical features $d(x_k^i, x_k^j)$ is defined as:

$$d(x_k^i, x_k^j) = \frac{\left| x_k^i - x_k^j \right|}{TH(x_k^i, x_k^j)} \quad (9)$$

where, TH is defined as the normalization factor, for example, $MAX(X_k) - MIN(X_k)$ (the difference between maxim and minim value observed for feature X_k from T).

Other metrics are obtained by varying the update function. The Relief algorithm, for example, gives the result for the q_k metric as:

$$q_k = \begin{cases} q_k + d(x_k^i, x_k^j) & if\ C(\vec{X}_i) \neq C(\vec{X}_j) \\ q_k - d(x_k^i, x_k^j) & f\ C(\vec{X}_i) = C(\vec{X}_j) \end{cases} \quad (10)$$

The Relief algorithm updates q_k when values of two characteristics of neighboring sample differ; the result increases if their classes' values differ and decreases if they are the same. The Contextual Merit Algorithm updates q_k when both the values of characteristics differ and the values of classes differ, it is used the following update function:

$$q_k = q_k + \frac{d(x_k^i, x_k^j)}{D(\vec{X}_i, \vec{X}_j)} \quad if\ C(\vec{X}_i) \neq C(\vec{X}_j) \quad (11)$$

Statistical classification

Statistical classification is a statistical procedure whereby individual elements are placed in groups based on quantity criteria or based on some features (properties) using a training set with previously classified items.

The problem can be formalized as: Given the set of training:

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \quad (12)$$

to produce a classifier

$$h : \chi \to \gamma \quad (13)$$

that maps an object $x \in \chi$ on his label classification $y \in \gamma$.

For example if the problem is to filtrate the spam e-mails then x_i means an e-mail and y is either "Spam" or "Non-Spam".

The probabilistic model of Naïve Bayes classifier

A **Naïve Bayes** classifier is a simple probabilistic classifier that applies the Bayes theorem with strong conditions of independence.

The probabilistic model of any classifier is actually a conditional model:

$$p(C|F_1,...,F_n) \tag{14}$$

for a dependent class variable C, with a small number of results, respectively *class*, subject of characteristic variables F_1 ,.., F_n.

If the number n of features is large or when a characteristic has a large area (may take a large number of values), the model can not be built practically. Therefore, using Bayes's theorem, the model can be reformulate as follows:

$$p(C|F_1,...,F_n) = \frac{p(C)\,p(F_1,...,F_n|C)}{p(F_1,...,F_n)} \tag{15}$$

It is noted that the denominator is independent of C and it is know the values for the features F_i, so what really interests us is the numerator of the fraction which is a composed probability model (joint):

$$p(C,F_1,...,F_n) \tag{16}$$

Or, expressed otherwise:

$$\begin{aligned} &p(C,F_1,...,F_n)\\ &= p(C)\,p(F_1,...,F_n|C)\\ &= p(C)\,p(F_1|C)\,p(F_2,...,F_n|C,F_1)\\ &= p(C)\,p(F_1|C)\;p(F_2|C,F_1)\\ &\quad p(F_3,...,F_n|C,F_1,F_2)\\ &= p(C)p(F_1|C)\;p(F_2|C,F_1)\\ &\quad p(F_3|C,F_1,F_2)\\ &\quad p(F_4,...,F_n|C,F_1,F_2,F_3) \end{aligned} \tag{17}$$

etc.

It is assumed that each feature F_i is **conditionally independent** of every other feature F_j for all i ≠ j, so:

$$p(F_i|C,F_j) = p(F_i|C) \tag{18}$$

and the composed probabilistic model becomes:

$$\begin{aligned} p(C,F_1,...F_n) &= p(C)p(F_1|C)p(F_2|C)p(F_3|C)..\\ &= p(C)\prod_{i=1}^{n} p(F_i|C) \end{aligned} \tag{19}$$

with independence condition, conditional distribution over the class variable C is:

$$p(C|F_1,...,F_n) = \frac{1}{Z}\,p(C)\prod_{i=1}^{n} p(F_i|C) \tag{20}$$

where Z is a scaling factor depending only by F_1, ..., F_n, which are constant if the values of characteristic variables are known.

Such models have a *class prior p(C)* and independent probability distributions $p(F_i|C)$. If there are k classes and a model for $p(F_i)$ can be expressed in terms of r parameters, then the Naive Bayesian model has (k - 1) + $n \cdot r \cdot k$ parameters. In practice, the most common models have $k = 2$ (binary classification) and $r = 1$ (characteristics are Bernoulli variables) the total number of parameters of the naive Bayesian model is $2n + 1$, where n is the number of binary features used for prediction.

All parameters of model, priory classes and probability distributions for characteristics can be approximated with relative frequencies from the training data set (maximum likelihood estimators for the probability). If features are not discrete, they must be divided into discrete parts, unsupervised or supervised, using the training set.

The Naive Bayesian classifier combines Bayesian probability model with a rule of decision. The most commonly used rule is that which is to take the case most likely - the rule of maximum an apriory or MAP decision. Classifier will be given by the following function:

$$\begin{aligned} &\text{classify}(f_1,...,f_n)\\ &= \text{argmax}_c P(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c) \end{aligned} \tag{21}$$

With the MAP rule it will reach a correct classification if the correct class is more likely than all the others.

Although independence restrictions are hard to follow, the naive Bayesian classifier has

certain properties that are very useful in practice (eg, separation by class conditional distributions of characteristics). In addition, the classifier does not require training large data sets to estimate parameters (mean and variation of variables) as independent variables assumption is only necessary the change of variables for each class (not the entire covariance matrix).

Bayesian Networks

Bayesian network or **belief network** is a **probabilistic graphic model** that represents a set of variables and probabilistic dependences between them.

For example, a Bayesian network can be used to calculate the probability that a patient is suffering from a disease, once the presence or absence of symptoms, assuming known probability of dependency relations between symptoms and disease.

From formal point of view the networks are **directed acyclic graphs**, which have variable nodes (parameter measured latent variable, hypothesis, etc.) and arcs represent conditional dependencies between variables. Nodes are not restricted to representing random variables. Bayesian networks that model sequences of variables are called dynamic Bayesian networks.

If there is an arc from node A to another node B, A is called a *parent's B* and B is a *child* of a node A. The set of parents for a node is denoted by X_i. The joint distribution of node values can be written as the product of local distributions of each node and its parents:

$$P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i|\text{parents}(X_i)) \quad (22)$$

If the node X_i has no parents, it is said that the local distribution of probability is *unconditional*, otherwise been *conditional*. If a value of a node is an observed value it is said that the node is an *evidence node*.

Conditional independence is represented by the property of d-separation graph: nodes X and Y are d-separated in the graph, given some evidence nodes, if and only if the variables X and Y are independent given the evidence variables. Set of other nodes which may depend directly the node X is given by Markov's characteristics of X.

Because the Bayesian networks are complete models for variables and their relations can be used to answer to probabilistic queries about theme. For example, you can update the knowledge dataset base relative to the status of a subset of variables when other variables are observed (evidence variables), through an inference process.

4. Evaluating the Classification Performance for Web Documents

The classifier performance can be measured or estimated in various ways. The used method depends on the type of classifier and data classification. Quality of classification can be evaluated using a confusion matrix. For example, matrix with numerical elements of samples identified as correct or incorrect for each class. Table 1 is a confusion matrix for binary classification [11].

Table 1. Confusion matrix for binary classification

		Prediction class	
		Class=Yes	Class=No
Observed class	Class=Yes	***tp***	***fn***
	Class=No	***fp***	***tn***

Confusion matrix of Table 1 contains the following items: ***tp*** = true positive (number of true positive cases), ***fn*** = false negative (number of false negative cases), ***fp*** = false positive (number of false positive cases) and ***tn*** = true negative (number of true negative cases).

Retrieving the relevant documents, or a positive class, is the most important task in web classification process, so the focus is on classification ***tp***. The importance of retrieval of positive examples is reflected by the choice of performance metrics for text classification: *accuracy* - precision, *revocation*, *Fscore* and *BreakEvenPoint*:

$$Precision=\frac{tp}{tp+fp} \quad (23)$$

$$Recall=\frac{tp}{tp+tn} \quad (24)$$

$$Fscore=\frac{(\beta^2+1)tp}{(\beta^2+1)tp+\beta^2 fn+fp} \quad (25)$$

$$BreakEvenPoint = \frac{tp}{tp + fp} = \frac{tp}{tp + fn} \quad (26)$$

The first three metrics evaluate the performance of classifiers by calculating the ratio of positive samples correctly classified and samples labelled as positive (Precision), positive samples of data (Recall), or total positive samples labelled with the data (Fscore). The BreakEvenPoint metric estimates essentially when the disagreement between data and algorithm for labelling samples as positive ($fp = fn$) is balanced. All these measures fail to consider number of true negative cases ***tn*** in their formulas, so do not take into account the correct classification of negative samples.

The problems of retrieving a positive class, the discrimination between classes, balancing between classes are possible retrieval tasks whose importance depends on the problem arising in the classification of documents. So far, there is not a consensus choice of measures used for performance evaluation of classifiers for Web documents.

In the classification process from Web space some measures to performances evaluation are used, such as the following 3 formulas:

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}, \quad (27)$$

this is used for example in [9] and other works, or *Recall*, *Fscore*, with the following correspondence:

$$Sensitivity = \frac{tp}{tp + fn} = Recall \quad (28)$$

and

$$Specificity = \frac{tn}{fp + tn} \quad (29)$$

presented in [10]. With the use of different measures, it is important to know how the performance, produced by these measures, is changing.

5. Conclusions

The present review of the most important indicators for assessment of classifiers performance emphasis the relevance of few classical indicators in the context of WEB documents retrieval. The recall, precision, sensitivity and specificity defined in classical text document retrieval works well also in the WEB context. For future research it is our intention to develop a methodology for non-text data retrieval for example searching and retrieving of images based on search words.

Acknowledgements

This work was partially supported by the Romanian National Council of Academic Research (CNCSIS) through the grant CNCSIS no. 12133/2008-2011.

REFERENCES

1. BANCIU, D., **e-Romania- A Citizens' Gateway towards Public Information**, Journal of Studies In Informatics and Control, Vol. 18, No. 3, 2009
2. BERNERS-LEE, T., J. HENDLER, O. LASSILA, **The Semantic Web**, Scientific American, May 2001, Vol. 284, No. 5, 2001, pp.34–43.
3. GOLLER, G., J. LONING, T. WILL, W. WOLFF, **Automatic Document Classification: A thorough Evaluation of Various Methods**, Internationalen Symposiums für Informationswissenschaft, Darmstadt, Nov. 2000, pp. 145-162.
4. HAND, D., H. MANNILA, P. SMYTH, **Principles of Data Mining**, MIT Press, Cambridge, MA., 2001, ISBN 0-262-08290-X,.
5. HENDLER, J., **Science and the Semantic web**, Science, 299, January 2003.
6. HONG, S. J., **Use of Contextual Information for Feature Ranking and Discretization**, in proceedings of IEEE Transactions on Knowledge and Data Engineering – 1997, available at www.research.ibm.com/dar/papers/pdf/ tkde-cm_with_cover.pdf, (acc. Feb. 2010)
7. HUANG, J., C. X. LING, **Constructing New and Better Evaluation Measures for Machine Learning**, available at http://www.ijcai.org/papers07/Papers/IJCAI07-138.pdf, accessed in august. 2007
8. KONONENKO, I., **On Biases in Estimating Multi-Valued Attributes**,

International Joint Conference on Artificial Intelligence, 1995, pp. 1034-1040.

9. POP, I., **Strategies for the Classification Cost Calculus**, International Journal of Computers, Communications & Control, Volume: II (2007), No:4, ISSN 1841 – 9844, accepted august 2007
10. SOKOLOVA, M., **Learning from Communication Data: Language in Electronic Business Negotiations** Ph.D. dissertation, 2006, available at www.etud.iro.umontreal.ca/~sokolovm/ThesisSokolova.pdf, accessed august 2007
11. SOKOLOVA, M., G. LAPALME, **Performance Measures in Classification of Human Communications**, a study available at http://www-etud.iro.umontreal.ca/~sokolovm/PMF.pdf, accessed 2007.
12. OPREAN, C., C. KIFOR, B. BARBAT, D. BANCIU, **E-Maieutics in Post-industrial Engineering Education**, Journal of Studies in Informatics and Control, Vol. 19, No. 1, 2010.
13. VILALTA, R., M. BRODIE, D. OBLINGER I. RISH, **A Unified Framework for Evaluation Metrics in Classification Using Decision Trees**, Machine Learning: EMCL 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September, 2001, Proceedings, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, Vol. 2167/2001, pp. 503-511.
14. WITTEN, I. H., E. FRANK, **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**, Academic Press, Morgan Kaufmann Publishers, ISBN: 1-55860-552-5, 1999.
15. WWW Library Directory - http://travelinlibrarian.info/libdir/ - accessed in January 2010
16. eEurope: An Information Society For All - http://www.w3.org/WAI/References/eEurope - accessed in January 2010