Why We Need an Interdisciplinary View in Information Retrieval

Daniel VOLOVICI
"Lucian Blaga" University of Sibiu, Romania
daniel.volovici@ulbsibiu.ro

Abstract: *Information Retrieval (IR) may be considered the scientific foundation of the librarians' activity. Information retrieval (Van Rijsbergen, 1979) is the science of searching for information relevant for a need specified by a user (Volovici & Volovici, 2013). The search could be inside a document, or for finding documents. It is possible to have searching using metadata that describes data and using databases of texts. In the recent years other fields of study develop methods to find relevant information in large collection of data, especially Data Mining (DM) (Witten et al., 2011). Some such methods are inspired from IR. But the process, relevant for both domains, of finding patterns in data (Crețulescu & Morariu, 2012) (not only in texts) is covered by the academic discipline present in all curricula of Computer Science, Machine Learning (ML) (Japkowicz & Shah, 2011).*

Keywords: Information Retrieval, Machine Learning, Artificial Intelligence, Statistics, Contingency Tables

Introduction

If we want to clarify all the agents that compete and try to solve problems in this area of activity it is necessary to write that ML could be considered (Alpaydin, 2016) as a branch of the field of **Artificial Intelligence (AI),** dedicated to programs (machines) that learn=improves the performance obtained performing a task. Some of the researchers in these fields consider that ML is nothing else that **Statistics** applied in AI.

For different subjects all these disciplines have solutions but, usually using different terminologies. For example, for measuring the quality of classification / clustering in IR are used *Precision* and *Recall*, *F-score,* and other measures related to the *Confusion Matrix.* The problem appear when the researcher want to make connections with others area of research. In other disciplines a similar matrix is used and other measures are more suggestive. For example, *matching matrix, association matrix* and all these matrices could be considered *contingencies tables.*

In different contexts and different times different measures become more important than others. In present time, probabilities of false positives and false negatives are more relevant under the pressure of **Epidemiology.**

More problems appear because the confusion between different disciplines when we intend to measure the quality of clustering items coming from m classes into n clusters, so using *nxm* contingency tables (association matrices,...)

Another subject of study that exists at the border where these fields intersect is *plagiarism detection.* In the present this detection is based on measures of similarities between documents. The measures are statistical and it is difficult to measure the similarity between documents written in different languages. And sometimes similarity shows only influences, not plagiarism.

Learning attributes of users could be an important area of research in the future related to *recommender system.*

https://doi.org/10.2478/9788395815065-001

In the present paper I am presenting all these links between fields as an overview of some subject of interest in research activities in the Department of Computer Science and Electrical Engineering from “Lucian Blaga” University of Sibiu.

Contingency Tables and their applications in different domains

For every proposed method for information retrieval we want to analyze how well it performs comparatively with other methods. The goal of all these methods is to obtain from a large collection of documents, those that are relevant for a need of a user. The method imply the searching in the collection based on a way of describing the need: a query, a list of key words, a text, or another list of items. The ideal result is to obtain all relevant documents and only relevant documents.

Usual performance measures in IR (Crețulescu & Morariu, 2012) are:

- *Precision,* the proportion of **retrieved** documents that are **relevant** for the search

$$\text{Precision} = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Retrieved\}}$$

- *Recall*, the proportion of **relevant** documents that are **retrieved** in this search

$$\text{Recall} = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Relevant\}}$$

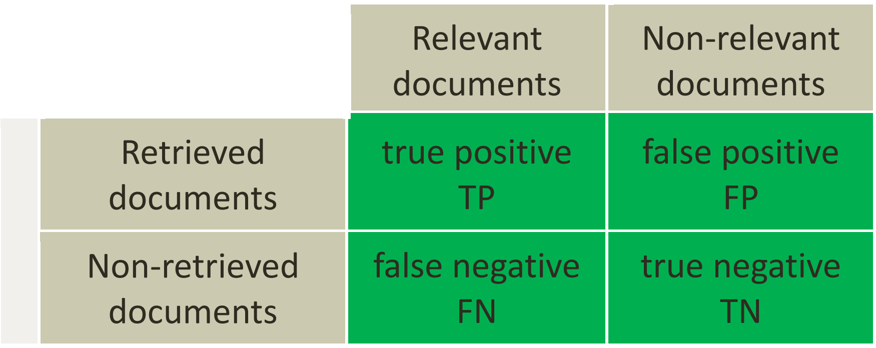
These measures could be visualized better in a Contingency Table, where:

TP=True Positive (retrieved documents for the search) = $\{Relevant\} \cap \{Retrieved\}$

FP=False Positive = $\{Nonrelevant\} \cap \{Retrieved\}$

TN=True Negative = $\{Nonrelevant\} \cap \{Nonretrieved\}$

FN=False Negative = $\{Relevant\} \cap \{Nonretrieved\}$



	Relevant documents	Non-relevant documents
Retrieved documents	true positive TP	false positive FP
Non-retrieved documents	false negative FN	true negative TN

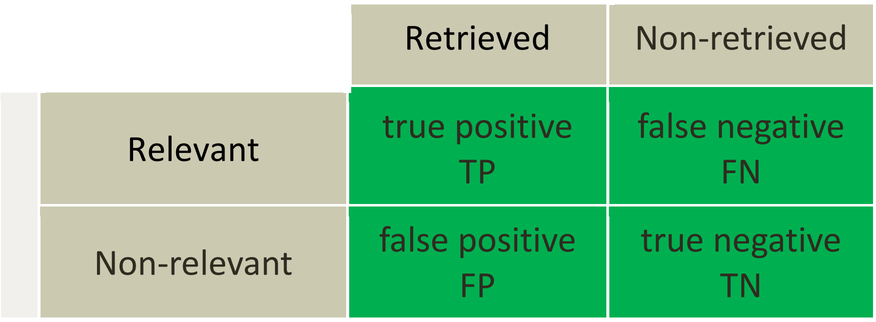
Figure 1: Results for the query search

In the literature (Morariu, 2008) it is preferred this representation with the results of the method on the rows of the table (2*x*2 matrix). But, for me, it is clearer to represent on the rows the ground truth so. I prefer to use the transposed matrix. On the columns we will have the results of one or more methods of searching, or of a combination of methods.

With this representation Precision and Recall will be calculated as:

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

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



	Retrieved	Non-retrieved
Relevant	true positive TP	false negative FN
Non-relevant	false positive FP	true negative TN

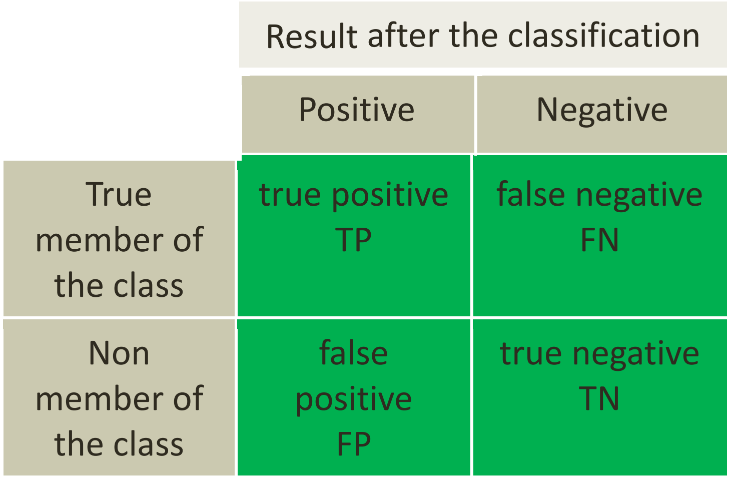
Figure 2: Results for the query search with the truth represented on rows

The same type of table it is used in other fields with different names. The most usual in Data mining and in Machine learning in the context of classification/clustering is **Confusion Matrix.** The name enhances the fact that methods used for classification or clustering in 2 classes (for IR, relevant and non-relevant) are not perfect because FP and FN are not equal to 0.

Beside the measures of Precision and Recall that emphesize the importance of TP, the result of a good classification must take into account the proportion of correct assigned items in both classes, so we must take into account also TN:

$$Accuracy = Success\ Rate = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ Rate = 1 - Success\ Rate = \frac{FP + FN}{TP + TN + FP + FN}$$



	Result after the classification	
	Positive	Negative
True member of the class	true positive TP	false negative FN
Non member of the class	false positive FP	true negative TN

Figure 3: Two class case with with classes *yes* and *no*

In different contexts and different times different measures become more important than others. In present time, probabilities of false positives and false negatives are more relevant under the pressure of **Epidemiology**. Because in this context TP represent people with the disease correctly identified by the test and TN, people in good health correctly identified as healthy, the focus is on *True Positive Rate*, *TPR*, and *True Negative Rate, TNR:*

$$TPR = \frac{TP}{TP + FN} = Sensitivity = Recall = 1 - FNR$$

$$TNR = \frac{TN}{TN + FP} = Specificity = Selectivity = 1 - FPR$$

	Positive	Negative
The disease is present	true positive TP	false negative FN
The disease is absent	false positive FP	true negative TN

Figure 4: Result of one test represented using a Confusion Matrix in Epidemiology

It is easy to observe that *Sensitivity* is the probability of a positive test of a person who has the disease and is the same as *Recall* from IR. *Sensitivity* measures the ability of the test to detect subjects that have the disease. *Specificity* in a diagnosis situation measures the ability to give a negative result for healthy subjects. If it has a value very close to 1.00 it is highly probable that the test will **select** only subjects that have the disease and this is the reason of naming it *Selectivity. Accuracy* is informative in this context too (Fletcher et al., 2014) because it combines in the same measure both aspects of the testing decisions: *Sensitivity* and *Selectivity.*
In the context of medical diagnosis and screening a similar measure as precision is not very useful. Instead, for bayesian analysis, together with *Sensitivity* and *Specificity,* it is very important to know the *Prevalence* of the disease in the population:

$$Prevalence = \frac{TP + FN}{TP + TN + FP + FN}$$

In the context of Hypothesis testing from the **Statistical Mathematics** false negatives examples are known as Type II errors and false positives examples are known as Type II errors. The probabilities of these errors represent the risks (Loftus & Loftus, 1987)

$$Probability\ of\ type\ I\ error = Risk\ of\ false\ positive = \alpha = \frac{FP}{FP + TN} = 1 - Sensitivity = 1 - \mathrm{Re}\,call$$

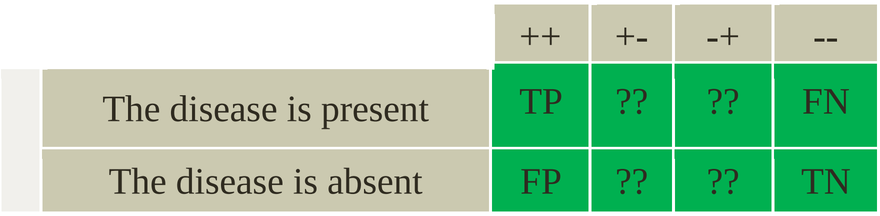
	Do not reject H_0 (positive test)	Reject H_0 (negative test)
Hypothesis H_0 is true	true positive TP	false negative FN Type II error
Hypothesis H_0 is false	False positive FP Type I error	true negative TN

Figure 5: Hypothesis-testing approach from *Statistics*

$$Probability\ of\ type\ II\ error = Risk\ of\ false\ negative = \beta = \frac{FP}{FP+TN} = 1 - Sensitivity = 1 - \mathrm{Re}\,cal$$

In designing a statistical test, the risk β is very important because it is used to determine the volume of the sample and the value is very important for $Power\ of\ the\ test = 1 - \beta$ witch measure the power of the test to correctly discriminate items for that H_0 is true.

Power of a statistical test is named *statistical sensitivity*. But according to some statisticians, the most informative measure is *Matthews correlation coefficient (MCC).*

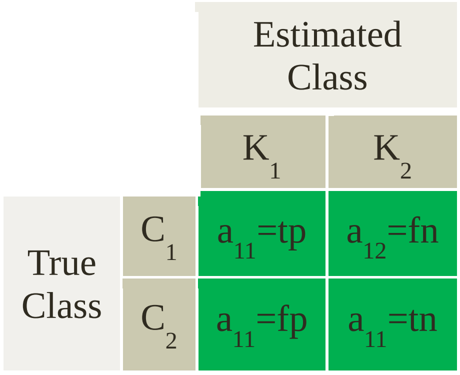


	++	+-	-+	--
The disease is present	TP	??	??	FN
The disease is absent	FP	??	??	TN

Figure 6: Results after 2 tests: A confusion matrix producing more *confusion*

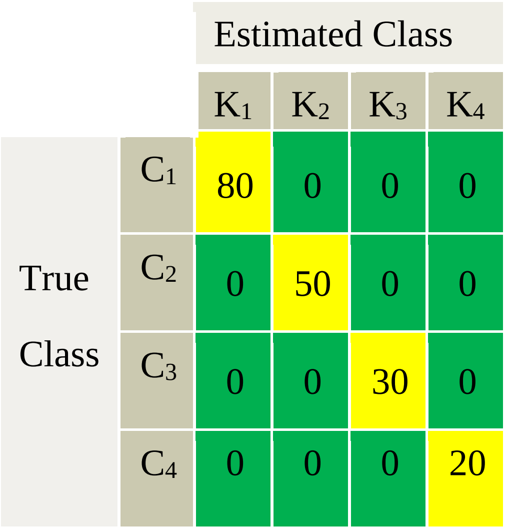
Sometimes when it is a need to eliminate some of the uncertainties related to a test is used the method to repeat tests, eventually of different type. In this case True Positive cases it is clear that are the situations when both test are positive and True Negatives, those with two negatives. FP and FN are evident on the table but, in general, for the other combination when at least one test is positive it is confusing how to interpret the results.

The solution is to consider again the methods used in Data Mining for analyzing the methods of grouping items from two classes in two estimated classes or clusters. The methods are based on the contingency matrix. This type of matrix could be extended to the case of n original classes mapped in other n estimated clusters.



		Estimated Class	
		K_1	K_2
True Class	C_1	a_{11}=tp	a_{12}=fn
	C_2	a_{11}=fp	a_{11}=tn

Figure 7: Contingency matrix for 2 classes



		Estimated Class			
		K_1	K_2	K_3	K_4
True Class	C_1	80	0	0	0
	C_2	0	50	0	0
	C_3	0	0	30	0
	C_4	0	0	0	20

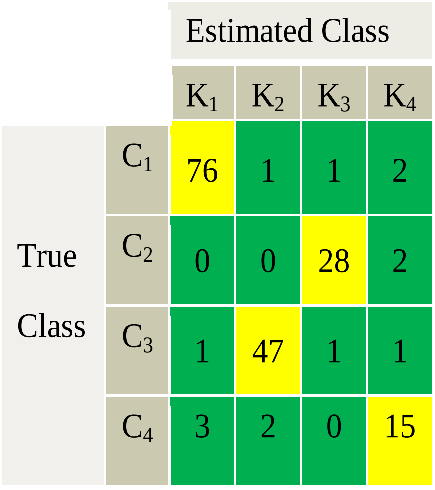
Figure 8: Example 1 for 4 classes (ideal)

Accuracy is a measure related to the association between the true sharing of examples in the two true classes on one side and the distribution of them in the two estimated clusters on the other side.

The term of *contingency matrix* is used in statistics for representing the frequency of a distribution of variables in form of a tableau. It was introduced by Karl Pearson and it is also named *cross tabulation*.

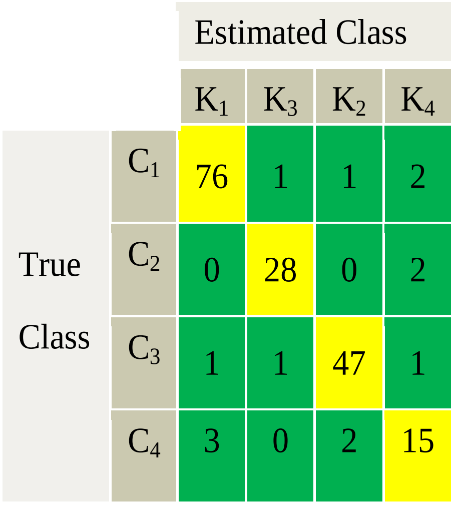
Transforming the Analysis of Contingency Tables in a Problem of Association

In the context of classification, the goal is to measure the degree of association between true (predefined) classes and those estimated. As a consequence, the variables are considered the membership to classes: true vs estimated. Being proved in time that Statistics offer very powerful tools for estimating the degree of association between 2 (or more) variables (Anderberg, 1973), (Fleiss, 1981), we suggest in (Volovici, 2016) transforming this problem of evaluation in a **problem of association.**
I will present the basic ideas of this method using some o the figures from that original paper (Volovici, 2016).
From that point on the *contingency table* will be considered an *association matrix* between true classes and estimated classes.
In the ideal situation, the classification method will assign all examples, each in one class, the correct one. There will be no mistakes, no false positive and no false negative. So for the example with 4 classes taken from (Volovici, 2016) presented in Figure 8, all positives examples are on the main diagonal of the matrix and all the elements not on that diagonal are equal with zero.
In a not so ideal situation it is possible to have a lot of non-zero elements spread all over the matrix, like in Figure 9. Because it is considered valid association that maximizing the sum of cells corresponding to the association that it was found we propose to maximize the sum of that cells and because we want to establish a clear link between the true classes and the estimated ones we will rearrange the matrix interchanging the columns so that the maximum values corresponding to best association to be present on the principal diagonal (Figure 10).



		Estimated Class			
		K_1	K_2	K_3	K_4
True Class	C_1	76	1	1	2
	C_2	0	0	28	2
	C_3	1	47	1	1
	C_4	3	2	0	15

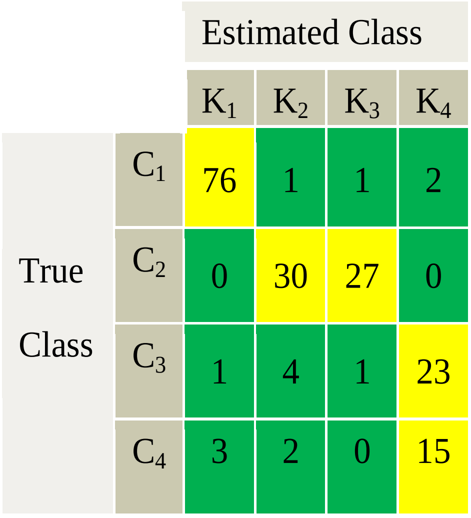
Figure 9: Example 2 for 4 classes (non ideal)



		Estimated Class			
		K_1	K_3	K_2	K_4
True Class	C_1	76	1	1	2
	C_2	0	28	0	2
	C_3	1	1	47	1
	C_4	3	0	2	15

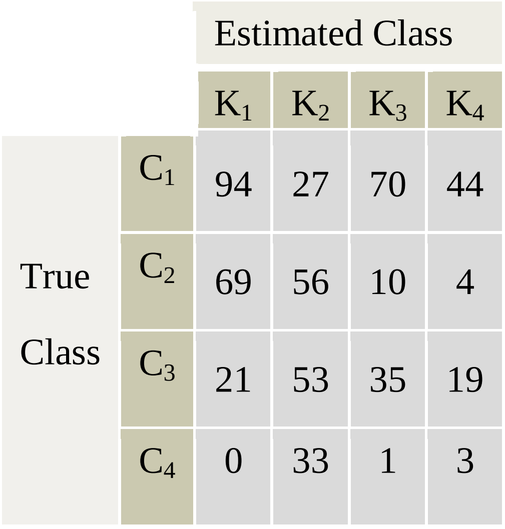
Figure 10: Example 2 for 4 classes with the main diagonal completed with true positives

In those situations we will try to maximize the sum of the values on the principal diagonal of the matrix (and the Accuracy). Sometimes it is possible to observe situation when the association matrix looks like that in Figure 11. There it is not possible to maximize the sum on the main diagonal and probably it will be better not to try to. This type of contingency table show that probably class C_2 is mapped in the reunion of two estimated classes K_2 and K_3. And, on the other side, the class K_4 brings together examples from two original classes: C_3 and C_4.



		Estimated Class			
		K_1	K_2	K_3	K_4
True Class	C_1	76	1	1	2
	C_2	0	30	27	0
	C_3	1	4	1	23
	C_4	3	2	0	15

Figure 11: Example 3 for 4 classes (unusual)

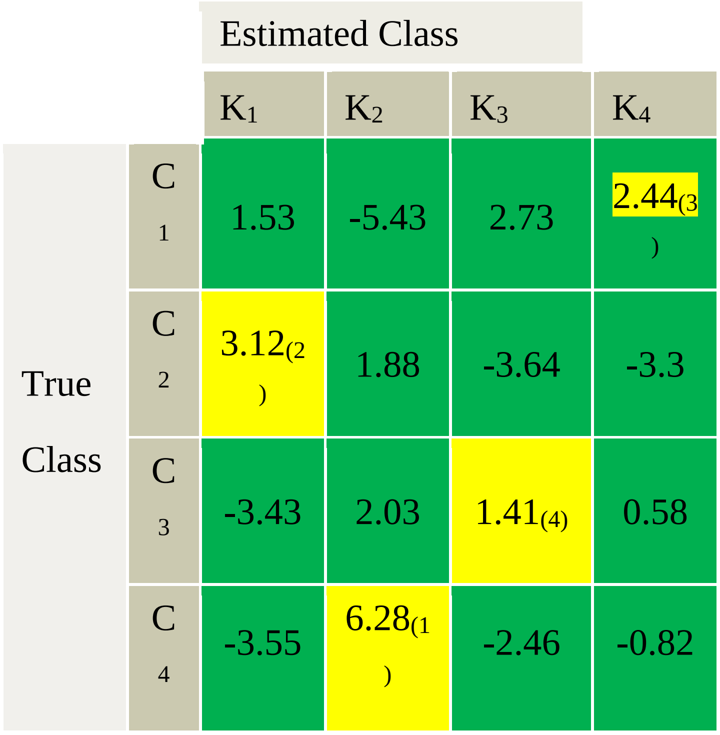


		Estimated Class			
		K_1	K_2	K_3	K_4
True Class	C_1	94	27	70	44
	C_2	69	56	10	4
	C_3	21	53	35	19
	C_4	0	33	1	3

Figure 12: Example 4 for 4 classes (very complicated)

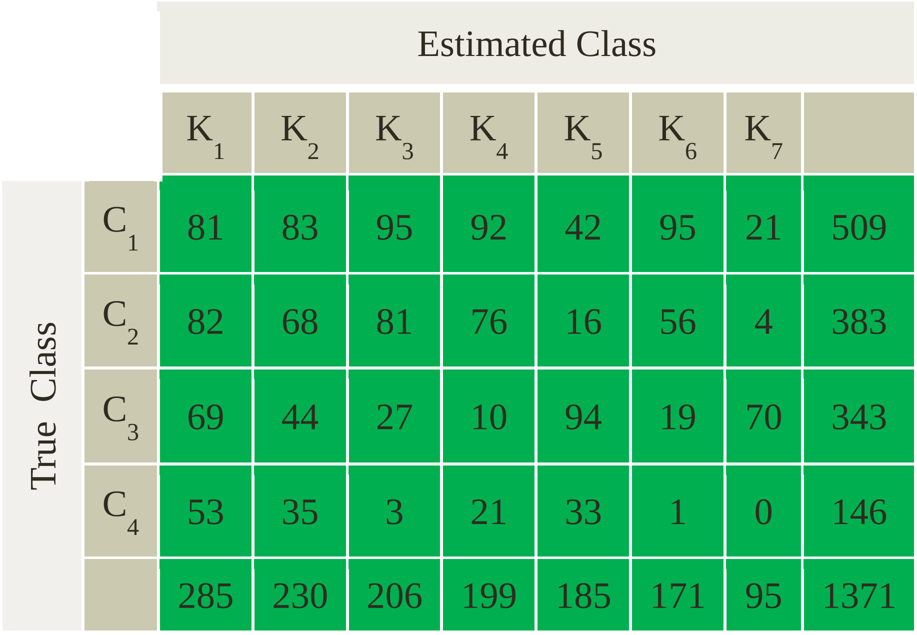
In the Figure 12 it is presented a more complex situation where the elements of the matrix are so big and so close in values that is difficult to establish a clear association. The textbooks treating contingency matrix (Fleiss,1981),(Japkowicz & Shah, 2011) uses *kappa criterion* to establish if the hypothesis that assume the existence of a significant difference between the data in the matrix and a uniform distribution (Witten et al., 2011) could be consider acceptable. In (Volovici, 2016) I used a measure of the relative importance of the value a_{ij} in the cell of the matrix

$$\Delta_{ij} = \frac{a_{ij} - f_{ij}}{\sqrt{f_{ij}}}$$



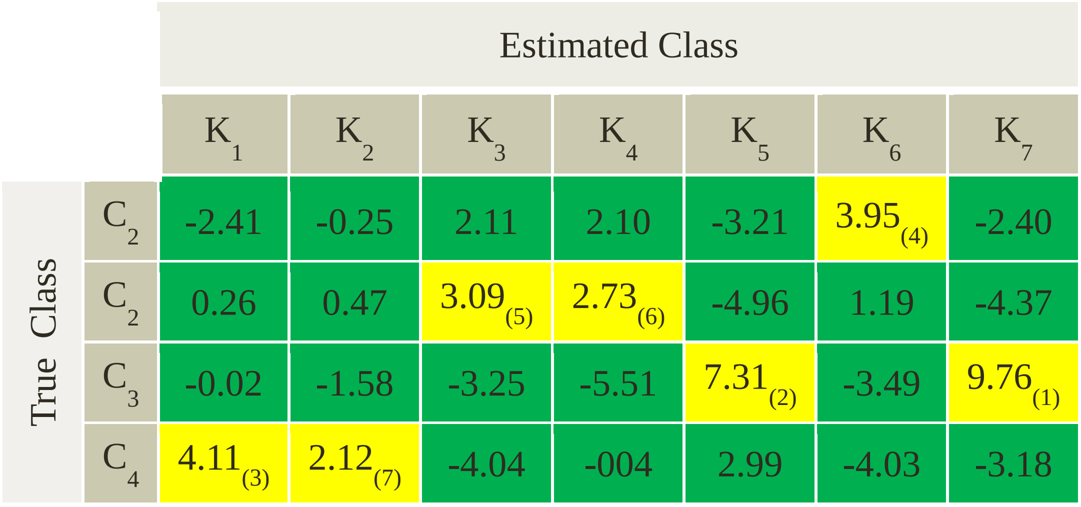
		Estimated Class			
		K_1	K_2	K_3	K_4
True Class	C_1	1.53	-5.43	2.73	2.44(3)
	C_2	3.12(2)	1.88	-3.64	-3.3
	C_3	-3.43	2.03	1.41(4)	0.58
	C_4	-3.55	6.28(1)	-2.46	-0.82

Figure 13: Example 4: assignment 0f estimated classes to real/true classes



		Estimated Class							
		K_1	K_2	K_3	K_4	K_5	K_6	K_7	
True Class	C_1	81	83	95	92	42	95	21	509
	C_2	82	68	81	76	16	56	4	383
	C_3	69	44	27	10	94	19	70	343
	C_4	53	35	3	21	33	1	0	146
		285	230	206	199	185	171	95	1371

Figure 14: Hypothetic estimation of 7 classes from 4 original classes



		Estimated Class						
		K_1	K_2	K_3	K_4	K_5	K_6	K_7
True Class	C_2	-2.41	-0.25	2.11	2.10	-3.21	$3.95_{(4)}$	-2.40
	C_2	0.26	0.47	$3.09_{(5)}$	$2.73_{(6)}$	-4.96	1.19	-4.37
	C_3	-0.02	-1.58	-3.25	-5.51	$7.31_{(2)}$	-3.49	$9.76_{(1)}$
	C_4	$4.11_{(3)}$	$2.12_{(7)}$	-4.04	-004	2.99	-4.03	-3.18

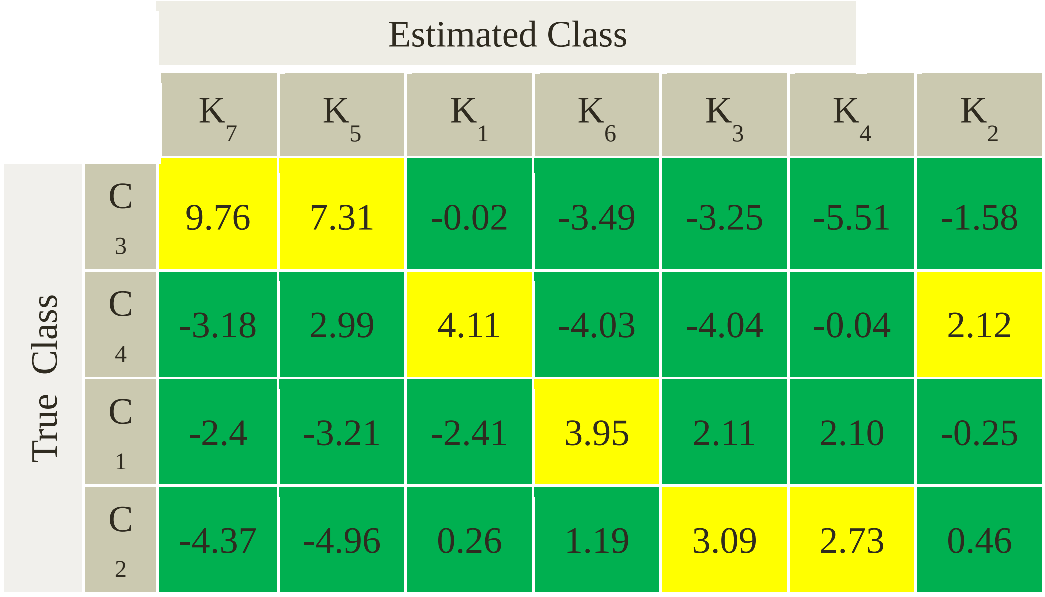
Figure 15: The Association Matrix after the application of the proposed method

The method proposed in (Volovici, 2016) could be applied for classifying objects from n classes in a different number m of classes, usually higher.

After the publication, the method was applied by Luiz Henrique Barbosa Mormille in (Mormille, 2018). In Figures 14 and 15 I applied the method for a hypothetic classification of objects obtained from 4 classes to 7 estimated classes.

It is clear that it is not straightforward obtaining the best m=7 yellow cells capable of maximizing the sum of True Positives where m is the minimum from the number of columns and rows, more exactly the minimum number of classes (original or estimated) because we want to force as much as possible 1-to-1 associations. In the present case there are not so many possible combinations, but for the scalability of the method for higher numbers it is necessary to use knowledge from another area in Computer Science: Dynamic Programming.

Finally, it is necessary to rearrange the matrix in a form as much as possible similar with one with the selected cells on the principal diagonal for the square matrices (Figure 16).



		Estimated Class						
		K_7	K_5	K_1	K_6	K_3	K_4	K_2
True Class	C_3	9.76	7.31	-0.02	-3.49	-3.25	-5.51	-1.58
	C_4	-3.18	2.99	4.11	-4.03	-4.04	-0.04	2.12
	C_1	-2.4	-3.21	-2.41	3.95	2.11	2.10	-0.25
	C_2	-4.37	-4.96	0.26	1.19	3.09	2.73	0.46

Figure 16: The final form of the Association Matrix

User Profile and similarities

The searching for relevant documents is realized using a query representing a list of words. Unfortunately information offered for the searcher is scarce because it is very difficult to establish what is relevant and what is not only on the basis of few words. And any information that could be used to discover what is the need of a specified user. A possible solution is to create for every user a **profile**. A method used is to create a model of the behavior of the user. In the same way as each document is stored in the form of a vector of frequency of words from a list of words (dictionary), each user will be remembered (Morariu, 2008) by a list of behaviors (classes of document searched in the past and jumps from a class to another class similar to the hyperlinks used in browsing the web).

A new idea that emerged from the field of Data Mining is to create **Recommendation systems**. These types of systems are also named *recommender systems*. Recommendation system using previous search of the user try to find *association rules* (Alpaydin, 2016) like: “people who search this item also access that item”. In a way, the researchers are trying to create *a generative model* to predict what documents will be relevant in the next search of the user.

These systems are similar with Information filtering systems with a notable application: email spam filters. For filtering must be used methods or information extraction. Methods for that task use principles of Machine Learning. The user shoul be able to rank the satisfaction related to the relevance of the retrieved documents after the search. Another idea is to use methods that imply **Reinforcement Learning** for creating the generative model of the user.

Plagiarism detection

The automatic systems used for plagiarism detection are based on counting similarities. More precisely, it is realized a search for identical list of successive words presented in the analyzed document and every document from a collection. The documents are selected on the basis of the representation on the form of the vector of frequencies of words. The problem uses methods of classification.

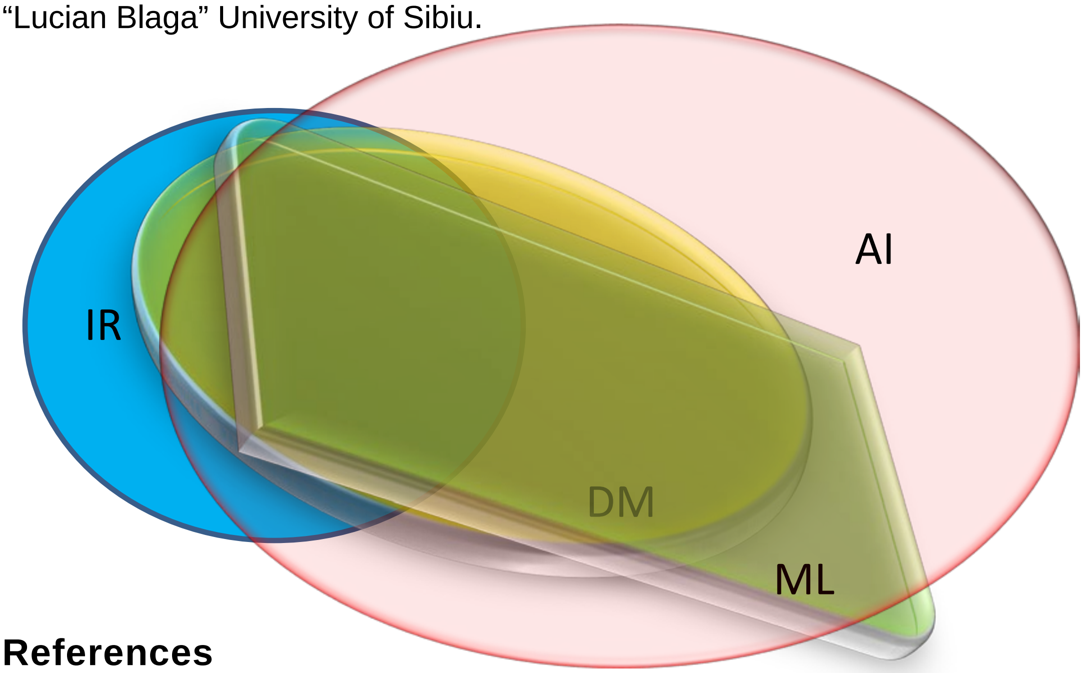
But in the present it is not possible to detect plagiarism if the text is copied in form of a translation from another language. It will be important to have reliable automatic translation tools from Machine Learning. In the future it will be very significant to extract knowledge fron the field Natural Language Processing.

The process of writing implies using other sources, other documents. It is impossible to not have borough ideas, concepts or propositions from other documents. For this reason it is an imperative need for the management of knowledge in the form of counting links between documents, ideas, concepts.

In the present we are waiting **Semantic Web** based on *ontologies* (relations among terms) (Crețulescu & Morariu, 2012) (Morariu, 2008). Because it is important, which was the source of inspiration it is important to represent the time in the networks creating the web, suggest nd work trying to integrate results from the study of **causality** (Pearl et al., 2016).

Conclusions

This paper presents an overview of relations between different fields related with Information Science that is possible to help solving important problems in Information Retrieval. It was a good opportunity to describe some of the research subjects in progress at the Department of Computer Science and Electrical Engineering from "Lucian Blaga" University of Sibiu.



References

Crețulescu, R. and Morariu, D. I. (2012). Text Mining: tehnici de clasificare și clustering al documentelor, *Editura Albastră*, Cluj-Napoca.

Morariu, D. I. (2008). Text Mining Methods based on Support Vector Machine*, Matrix Rom*, Bucuresti, 2008.

Fletcher, R. H. ; Fletcher, S. W. and Fletcher, G. S. (2014). Clinical epidemiology: the essentials, *Wolters Kluwer/Lippincott Williams & Wilkins Health*, 5th ed edition.

Volovici, M. R. and Volovici, D. (2013). Tehnici moderne de găsire a informațiilor căutate*, Editura Universității „Lucian Blaga" din Sibiu*.

Loftus, G. R. and Loftus, E. F. (1987). Essence of Statistics, Second edition, *Alfred A. Knopf Series in Psychology*, New York, 1987.

Alpaydin, E. (2016). Machine Learning: the New AI*, The MIT Press*, Cambridge, Massachusetts.

Anderberg, M.R. (1973). Cluster Analysis for Applications. *Academic Press*, 1973.

Fleiss, J.L., (1981). Statistical Methods for Rates and Proportions. *Wiley Series in Probability and Statistics*. Wiley.

Japkowicz, N. and Shah, M. (2011). Evaluating Learning Algorithms: A Classification Perspective. *Cambridge University Press*, New York, NY, USA.

Van Rijsbergen, C. J. (1979). Information Retrieval. *Butterworth-Heinemann*, Newton, MA, USA, 2nd edition.

Witten, I.H.; Frank E. and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann Publishers Inc*., San Francisco, CA, USA, 3rdedition.

Volovici**,** D. (2016). **“**Evaluation of Classication in More Than Two Classes”, *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences IJASITELS,* Vol. 6, Issue 2.

Mormille, L.H.B. (2018). *L*earning Probabilistic Relational Models: A Novel Approach, *Disertacao (Mestrado)*, Ecola Politecnica da Univrersidade de Sao Paulo, Sao Paulo.

Causal Inference in Statistics: a Primer*,* John Wiley & Sons Ltd.