Recent issues about textual information analysis in micro-marketing: methodological suggestions and some case studies

Recenti tendenze dell'analisi delle informazioni testuali nel micro-marketing: osservazioni metodologiche e alcuni casi operativi

Furio Camillo – Caterina Liberati
Dipartimento di Scienze Statistiche – Università di Bologna – furio.camillo@unibo.it

Riassunto: Nel lavoro vengono trattate alcune tematiche circa l'uso delle tecniche di analisi delle informazioni testuali per il moderno approccio al marketing one-to-one, ossia per il micro-marketing. Tale approccio implica il mantenimento, l'arricchimento e il monitoraggio del cosiddetto Customer Data Base, ossia di un archivio di dati relativi ai singoli clienti effettivi o potenziali. Le tecniche statistiche adottate dovranno riportare in generale il risultato di ogni analisi sul singolo record, generando dunque sul singolo cliente un arricchimento dell'informazione disponibile. In questo contesto vengono affrontati alcuni problemi legati al trattamento di informazioni testuali mediante la presentazione di casi di studio operativi. Viene inoltre proposto un approccio di data mining avanzato basato su metodi di tipo Kernel per il micro-marketing.

Keywords: text mining, CRM, one-to-one marketing, kernel, semantic

1. From marketing "of the production" to marketing "of the relationship"

In recent years we see the emergence of rapidly increasing technologies and analytical business intelligence systems that are being implemented to achieve organizational productivity, profitability, and customer services. These, now, are considered as a priority for many companies with large customer data bases to survive the stiff competition.
In particular, data collection strategies and technologies have changed from merely registration of information to a more structural form for data mining and analysis of customer-centric databases and panel data to address key business issues and delivery of integrated datasets for access via digital stream data, web-enabled applications, and modelling.
In this work we will show how data mining and micro data-mining are changing the research landscape since the methods and the data applications are ever growing and the next decade will see a revolution in an information based economy. As it is well known the new frontier of information processing is the fMRI ("Functional Magnetic Resonance Imaging") which has been used to test how product brands are presented in the brain. The goal of every consumer marketer is to have people identified with a brand, to develop the kind of loyalty that goes far beyond a utilitarian preference for, say, one kind of pickup truck over another (Newsweek, July 5, 2004, p. 46). Such an emerging field is now called "neuromarketing", or in the absence of fMRI, it is called the "micro-data mining."

Effective decision making requires ongoing feedback on the effectiveness of various marketing programs to help managers identify good vs. bad decisions taken in the past. If managers have such feedback systems to help them direct precious marketing resources towards initiatives that are more efficient, managerial decisions would be governed by a significant element of scientific rationality rather than subjective heuristics. Prediction models based on consumer behaviours with sufficient robustness can create such feedback systems, thereby increasing effectiveness of marketing strategies providing competitive advantage to many firms with customer-centric data base.

Indeed the orientation toward the customer-centric approach pushed many companies (private or public) to maximize their single customers' profitability tending to establish a relationship with each of the customers. Such relationship is divided in time series by interactions between the customer and the company and vice versa. If the interactions are built and managed well by the company, each side will tend to strengthen the business image and the confidence that the customer has towards the company. The other issue is to enrich the information the company has about the customer. Therefore, the classical literature on the subject discusses the possibility of triggering a virtuous circle that will make the customer more faithful with the hope that the profitability will increase for the company. If the relationship is managed badly, the business image will be damaged and tarnished.

Therefore, one-to-one marketing relationship passing from a transactional model to a more relational model has been advocated during the last few years in the literature by Peppers, Rogers, and Dorf (2000), where the differential features are distinguished in the two types of marketing strategies.

Table 1: *The two types of marketing strategies*

	IN THE TRANSACTIONAL MARKETING	**IN THE RELATIONAL MARKETING**
MARKETING FUNCTION	Marketing mix (4p)	Interactive Mix marketing
GOOD QUALITY DIMENSION	Technical quality	Functional quality
PRICE SENSITIVITY	High	Contained
MARKETING vs OTHER DEPARTMENTS interface	Limited or nonexistent	Strategic (discounted)

The increasing success of the customer based vision (or better human based) of the activity of companies' marketing and the simpler technology availability to be acquired is having the building and the use of the so-called Customer data base (CDB). CDB is the data processing structure of registration of the available information about the real and potential customers. The information can come from internal sources at the company and they therefore concern mostly on the behaviour of the individual with the company, or can come from sources: 1. outside; 2. the ad hoc surveys; 3. the inferable information for matching with other archives.

The design, the building, the management and the analysis of CDB also acquired a crucial role in the determination of the mechanisms which substantiate the company's internal and outside relations, as it is, according to the customer based vision, the only informative integrated source.

The perspective characteristic is that the technology and the analytical capacities, is taking companies and public operators always to set up the operating strategies "of

contact" and "monitoring" more than the activity at an individual level (customer relationship management (crm), direct marketing, etc).
Our work will focus on some new developments in data mining and micro data mining methodologies and modelling techniques related to textual data analysis strategies.
The paper will especially regard a few suggestions about the kernel approach introduction in text mining for the micro data mining. The proposed strategy for the statistical model selection is based on information measure of complexity (ICOMP) criterion developed by Bozdogan (2000) as our criterion or performance measure.
The use of ICOMP as a model selection criterion is that it combines a badness-of-fit term (such as minus twice the maximum log likelihood) with a measure of complexity of a model differently than AIC index of Akaike, or its variants, by taking into account the interdependencies of the parameter estimates as well as the dependencies of the model residuals. We operationalize the general form of ICOMP based on the quantification of the concept of overall model complexity in terms of the estimated inverse-Fisher information matrix (IFIM).
The paper shows 2 real case-studies about two main classes of technical problems: 1) the extension to the population of a sample result about a textual data analysis; 2) the use of the textual data analysis to approximate the semantic content of the e-trip of an e-commerce web portal user.

2. The future venue of market research and the text mining applied to an ethnographic project of Happiness-CRM

Recently, it has been recognised that doing societing rather than marketing is not just a way to play with words but it represents a necessary approach to the world of research as a whole. In other words, today, but especially in the future, companies will be interested in information about their clients not directly linked with the business. The idea is that not only the market researcher but also the final ‘reader’ (the company) can capture the key signs coming from the field, so that a company can have access to the researcher’s analysis but also to people’s world (e.g. photos, expressions, own words, doubts, feelings, material cultures, etc).
Therefore, in this passage from transactional marketing to relational marketing it is going towards a reality in which data are the main activity of any business.
The clients of research institutes are increasingly willing to attend interviews, groups and ethnographic research. Advanced companies ask their employees to leave their desks and explore ‘the field’, cool-hunting on their own products in order to grasp the ‘mood’ that surrounds their related scenario. This is particularly necessary when working on ethnographic projects, which ‘field’ is formed by the sum of feelings, sensitivities, moods and interpretation of gestures apart from words like in traditional one-to-one interviews or focus-groups.
In order to forefront these needs the research institute Future Concept Lab of Milan (FCL) developed a new method of analysing and presenting data that employs a *Digital Interactive Matrix* that allows the user to ‘navigate’ within the whole data collection extracted from the field. The method was designed in occasion of a European ethnographic research called ‘The Material Culture of Happiness’, a permanent research program created by FCL in order to provide an insight on day-to-day Happiness

experienced in 8 European countries namely Spain, France, UK, Italy, The Netherlands, Germany, Finland and Russia[1].
The research employs a cocktail of in-depth methodologies that combine psychological methods and with qualitative fieldwork. The research focused on young people (14-22) and mature adults (55-70) who have been asked to fill in a photo diary for a period of seven days and taking photos to the 'objects of their Happiness' (bring these people, places, products, etc) which therefore they recognise as meaningful to the building of their day-to-day well-being. Diaries have been followed by in-depth interviews with the respondents on the basis of ad-hoc designed discussion guides that took into consideration people's cultural backgrounds, life-stage, and the content of the diary.
The result is a collection of 1200 stories of Happiness reported in people's own words and through visuals, symbols and drawings. This collection of stimuli has been classified within the interactive matrix that combines social forces such as *Subjectivity, Sociability, Experimentation, Connectivity, Ethics,* with Ethno-behavioural variables that we defined as: *Domesticity, The Body Affair, Daily Responsibility, Leisure and Consumption, Commuting and Territory, Landscape and Nature, Extra Occasions*. The matrix is a tool of analysis that contextualise the events the respondents described faithfully reporting their own words and the images employed by respondents to describe the happy event.
Our research has been run on a non-statistical sample, according to the qualitative techniques of texts decoding. Even if the happiness stories are 1200, as usually happens in qualitative research, the extension of results to a representative population is not possible.
It doesn't essentially exist a statistical representative sample of the material and of collected information, because the approach to the problem is a qualitative one, and therefore the analysis of the collected texts is based on an accurate decoding project and on a classification of information which uses a one-to-one protocol reading of the texts.
Referring to a cognitive text reading, in other words to a typical process of the qualitative research, this story has been classified by FCL researchers in a grid with some interpretation keys. Our text mining target was to create statistical models of prediction using textual information (exogenous variables) for the classifying variables of qualitative nature (target variable) related to diaries' textual contents. The suggestion is to extend the cognitive interpretative model used by qualitative experts to a large quantity of textual material (the rules); this model would be so large that can be generated from a representative sample, according to the traditional terms of the samples statistic theory.
The texts automatic classification in predefined categories is one of the most frequent problems faced by contemporary text mining, in accordance with the computational statistics and Computer Science applied to automatic categorisation software (Yang 2000; Yang, Zhang, Kisiel 2003). The strategy we adopted is less inclined to the automatic classification and it uses factorial reduction techniques of collected texts for estimating a not parametric discriminant model on the factor scores (input variables) of a lexical correspondence analysis applied on the *stories-forms* matrix (Camillo, Tosi, 2004).
In Table 2 are showed the confusion matrixes of 9 non parametric discriminant models (nearest neighbour method) by which the stories have been reclassified into qualitative

[1] From "The Material Culture of Happiness" by Future Concept Lab – World Future Society 2004, in "Thinking Creatively in turbulent times", 2004, Ed. Howard F. Didsbury Jr, World Future Society – Bethesda, Maryland – U.S.A.

cognitive categories. The confusion matrix, classically, gives the rates of wrong and right reclassification of a discriminant model and in our case the results were good.

Table 2: *Confusion matrixes about the 9 qualitative cognitive categories prediction*

RESPONSABILITY

FROM\TO	YES	NO	
YES	**93%**	7%	100%
NO	9%	**91%**	100%
	20%	80%	100%

OBJECT RELATED

FROM\TO	YES	NO	
YES	**89%**	11%	100%
NO	12%	**88%**	100%
	34%	66%	100%

LEISURE

FROM\TO	YES	NO	
YES	**80%**	20%	100%
NO	6%	**94%**	100%
	67%	33%	100%

SUBJECT RELATED

FROM\TO	YES	NO	
YES	**87%**	13%	100%
NO	12%	**88%**	100%
	52%	48%	100%

MOVING

FROM\TO	YES	NO	
YES	**88%**	12%	100%
NO	7%	**93%**	100%
	16%	84%	100%

SPACE RELATED

FROM\TO	YES	NO	
YES	**86%**	14%	100%
NO	12%	**88%**	100%
	58%	42%	100%

SURPRISE

FROM\TO	YES	NO	
YES	**100%**	0%	100%
NO	4%	**96%**	100%
	8%	92%	100%

TEMPORALITY

FROM\TO	PRESENT	PAST	FUTURE	
PRESENT	**86%**	6%	8%	100%
PAST	0%	**94%**	6%	100%
FUTURE	0%	0%	**100%**	100%
	77%	10%	13%	100%

PRESENT

FROM\TO	YES	NO	
YES	**100%**	0%	100%
NO	6%	**94%**	100%
	12%	88%	100%

In order to make usable the results obtained the happiness research employed the psycho-text mining methodology called *Semiometrie.* [2]

The essential idea of this kind of use on 'external' texts consists in a research in the external text of semantic characteristics that connote more the semiometric axes, therefore, the interpretation of text special analysis can use the big values contrasts of mittle-European citizen.

In the table below is shown the correlation matrix among the happiness factors and the Semiometric axes. We have to point out that if we employ the happiness space to represent the Semiometrie space orthogonality of Semiometric axes is not guaranteed (Fig. 1). Therefore the collinearity between two Semiometric axes is a measure of how

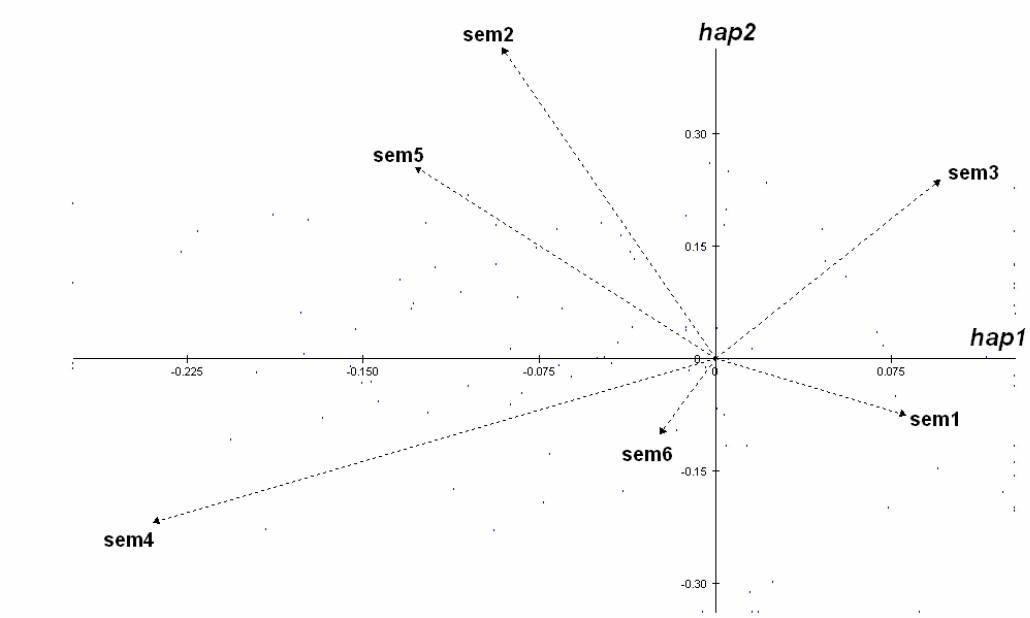
[2] Semiometrie is "a long list of words and thousand of people, in all Europe, are asked to give a mark (a score) more or less high depending on the agreeable or disagreeable characteristic of the single word" (Lebart, Piron, Steiner, 2003). The 210 selected words and properly declinated (substantive rather than verb or adjective; absence of article rather than determinative or indeterminative article) allow the reconstruction the psycho-cultural models that constitute the subconscious system of choice and of the identification of desires of European citizens. Semiometrie essential product is a set of semiometric axes defined from an analysis of the main components on matrix of agreement-disagreement score, given by interviewed people on the 210 evocative words. In particular, the six basic axes have been declined and interpreted, in importance of order, as follows:
1) methodological axis of participation; 2) duty-pleasure axis; 3) positive devotion to life-separation (pessimistic indifference) axis; 4) sublimation-materialism; 5) idealism-pragmatism; 6) humility-supremacy.

much large is the fusion effect determined by the usage of a specific external space (in our case the Happiness space).

Table 3: *Linear correlation: Semiometrie Axis (sem) vs. Happiness 2-axis (hap) on 106 common forms (Happiness and Semiometrie Corpora)*

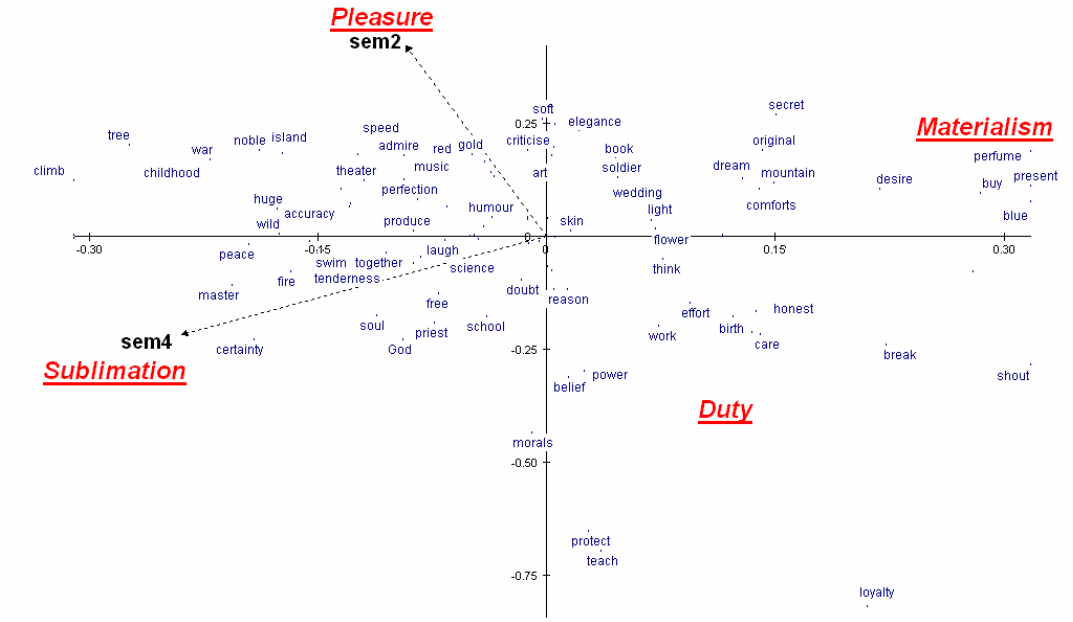
	sem1	sem2	sem3	sem4	sem5	sem6
sem1	1	-0.13834	0.51932	0.09244	-0.24664	0.12778
p-value		*0.1573*	*<.0001*	*0.346*	*0.0108*	*0.1918*
sem2	-0.13834	1	0.16395	-0.09968	0.25088	-0.05439
p-value	*0.1573*		*0.0931*	*0.3093*	*0.0095*	*0.5797*
sem3	0.51932	0.16395	1	-0.34645	0.20246	0.41653
p-value	*<.0001*	*0.0931*		*0.0003*	*0.0374*	*<.0001*
sem4	0.09244	-0.09968	-0.34645	1	-0.14917	-0.22842
p-value	*0.346*	*0.3093*	*0.0003*		*0.127*	*0.0185*
sem5	-0.24664	0.25088	0.20246	-0.14917	1	-0.11207
p-value	*0.0108*	*0.0095*	*0.0374*	*0.127*		*0.2527*
sem6	0.12778	-0.05439	0.41653	-0.22842	-0.11207	1
p-value	*0.1918*	*0.5797*	*<.0001*	*0.0185*	*0.2527*	
hap1	0.08095	-0.13809	0.09541	-0.23978	-0.12826	-0.0241
p-value	*0.4094*	*0.158*	*0.3306*	*0.0133*	*0.1901*	*0.8063*
hap2	-0.07593	0.61219	0.23943	-0.21962	0.25665	-0.10172
p-value	*0.4391*	*<.0001*	*0.0134*	*0.0237*	*0.0079*	*0.2995*

Figure 1: *The projection of Semiometric space onto Happiness space 2-axis*



In particular, in the Figure 2 is evident that from top to bottom of the factorial map deriving from the Lexical Correspondence Analysis applied to the Happiness textual corpus we move from *pleasure* to *duty* and from *materialism* to *sublimation*. In the map are written some significant words coming from both happiness diaries and Semiometrie vocabulary.

Figure 2: *The projection of Semiometric concepts and the significant words onto Happiness space 2-axis*



3. The *Glam on web* case: a semantic e-CRM system

Today more than in the past companies understand the value of collecting customer data which try to exploit in intelligent system for extracting interesting information.
The need of the business to get knowledge from data comes from demand to monitor own clients in order to preserve the relationship with its customers. This is extremely true when we consider e-commerce. Hence for the web world is developed the concept of e-CRM (electronic Customer Relationship Management) that is the electronic management of the relationship with the customer. e-CRM is the business ability to take care of its own clients through internet and tools purposely designed for the on-line world. In this scenario data mining tools for classification pattern are necessary to find out the most profitable segments of the customer database. Moreover, because of the feedback and the dialogue with the web customers is more simple to realize the business efforts to extract the pattern clickstream data are done to make predictions based on the past behaviour, for building deep customer profile that guide the one–to–one marketing strategies.
In our case study we analysed data and texts coming from Glam on Web portal[3] to discover if customers (buyers) and visitors of the site (no buyers) have the same behaviour. Therefore the target variable is the purchase made by the visitors that we can code with a binary string of values Y=[1,1,1,0,0……,0,1,1] with 1=buy and 0=no-buy; where the explicative variables are information coming from Customer Data Base (CDB) that is related to social-demographic data and texts clicked and/or concepts "married". A structural lexical analysis has been employed on textual descriptions of the

[3] The Glam on Web is projected to be a great emotional value site. It is composed by artistic pictures that "describe" products (or the product-lines) in materials and in features as a catalogue

402 garments photos. We think that users with a similar purchase-propensity have been submitted to the same emotional visual (therefore textual) stimuli. Following this construction we group the 402 garment photos into 7 classes according to a textual clustering process and in order to obtain the exogenous variables of our model we extract the factor scores for each user calculated on the lexical matrix *(users)*(1-7 cluster of objects)*. The factors scores coming from a lexical correspondence analysis (LCA) as a proxi of information relative to the clicked objects and of the semantic content (*the semantic basket*) of the e-trip of the user.
Hence the specification of such model is the follow:

Buy/no buy (1/0) = f (factor scores of the LCA)

Observing the confusion matrixes showed below (Table. 4) it's clear this model has good classification power. In fact the textual factors represent a stable reference space in which is always possible to add new objects without changing the set of predictors.

Table 4: *Results using Linear Discriminant Analysis*

Fisher DA

From/to	no buyer	buyer	Total
no buyer	88.5	11.5	100
buyer	57.69	42.31	100

In such frame is useful using a non-linear Discriminat Analysis which is capable to capture the nonlinear structures in the data: in fact texts are well represented with chi-square metric. It is known that the factor representation of data which use such metric could be non-linear. We propose to employ kernel-based methods as Kernel Discriminant Analysis which generally improve the LDA performance (Camillo, Liberati, 2006). The process illustrated in this work is really innovative: the usage of the text in order to profile customers represents an innovative method: in fact it allows business to exploit de-structured information in a strategic marketing project.
In this paper the authors propose a data mining tool for supervised classification pattern which is a nonlinear extension of LDA based on kernel functions: Kernel Discriminant Analysis (KDA). The main idea of the kernel method is that without knowing the nonlinear feature mapping or the mapped feature space explicitly, we can work on the feature space through kernel functions, as long as the problem formulation depends only on the inner products between data points. This is based on the fact that for any kernel function κ satisfying Mercer's condition (Cristianini Shawe-Taylor 2000) there exists a mapping Φ such that

$$< \Phi(a), \Phi(b) >= \kappa(a, b)$$

where <,> is an inner product in the feature space transformed by Φ (Burges 1998)).
The reformulation of DA in the feature space is very similar to the LDA case the decision boundary is obtained maximizing the ratio:

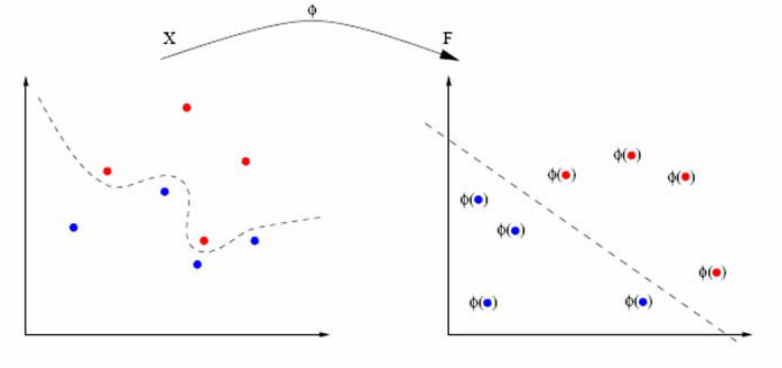
$$J(w) = \frac{w' S_B^{\Phi} w}{w' S_W^{\Phi} w}$$

where S_B^{Φ} S_W^{Φ} are Between and Within covariance matrixes in the feature space.
This maximization problem can be resolved by finding the leading eigenvectors of the $(S_W^{\phi})^{-1} S_B^{\phi}$. The number of the discriminant functions we obtain is equal to the number of groups minus 1. The discriminant scores are a linear expansion of the training patterns in the feature space:

$$\alpha = \sum_{i=1}^{n} w_i \phi(x_i)$$

Therefore the improvement employing this technique is we have decision functions which are linear in the feature space and correspond to nonlinear ones in the input space. (Figure 3).

Figure 3: *Decision boundary in the Input Space vs. Feature Space*



The use of KDA presents some subjective choices which are still open problems: the Kernel Function choice and its parameters.
In this paper we propose the use of Information Complexity theory (Bozdogan, 2000). In particular we referred to the index of complexity (ICOMP) as tool for the model selection in case of KDA (Bozdogan, Camillo, Liberati, 2006).

$$ICOMP\left(\hat{\Sigma}_W\right) = np \log 2\pi + n \log\left|\hat{\Sigma}_W\right| + np + 2C_{1F}(\hat{\Sigma}_W)$$

where n is number of objects, p is number of variables, C_{1F} is the measure of complexity of a covariance matrix $\Sigma_W = S^{\Phi}/n$ based on Frobenius norm.
Testing 150 models (50 for each different Kernel function: polynomial, RBF, Cauchy) we founded the best solution in the Cauchy kernel (width = 0.001).

Table 5: *Semantic basket model using KDA (Cauchy Kernel): confusion matrix*

KDA hibridizated with NNM (n=20)

From/To	no buyer	buyer	Total
no buyer	97.00	3.00	100
buyer	26.92	73.08	100

The improvements gained applying the KDA on the model are evident. KDA analysis is be able to overcome the limitations of a linear approach that in real case, as this one, are often important. Moreover choosing to adjust the discriminant scores obtained with a Nearest Neighbor process (with n=20) we got the best results showed above (Table. 5).
Obviously the problem with which we have to face when such method is applied is the generalization of the textual classification of the objects: in other words the allocation of a new object described by a different texts and concepts not expressed in the clusters found. To overcome this issue we propose use semiometric approach of text mining for codifying and analyzing visual web stimuli, according to a general socio-psychological landscape for the "words" interpretation. Semiometrie becomes hence a landscape of reference a scheme which allows the research to "read" words in term of values and/or concepts which belong to the Occidental society, so it is functional for all textual analysis.

4. Some suggestions and future developments

In showed case studies is clear that the micro data mining requires the use of techniques which allow to extend the result of a data analysis on a sample to the whole population, or better of an analysis which produces an enrichment result of the record individual's of the customer data-base.
The use of textual information for this enrichment could be crucial for the success of direct actions on each customers. In this sense techniques deriving from the Kernel approach of data analysis can be decidedly useful, associated with tools such as the model complexity measures aiming to the the best model selection.
In a future vision we think that, about the specific role performed in such contexts by the analysis of the available textual information, the further methodological developments on which focus our attention are two: 1) intensify the use of tools as Semiometrie for the building scenarios devoted to a semi-automatic interpretation and generalization of the results of a text mining analysis; 2) tweak and evaluate algorithms which exploit the kernel approach potentialities already in the explorative step of factorial reduction (usually according to a Lexical Correspondence Analysis).
About this latter statement it seems interesting to point out that recently it has been specified in a kernel frame the Binary Correspondence Analysis respecting the fundamental properties of such method.(Picca. D, Curdy B., Bavaud F., 2006).

References

Bozdogan H. (2000), "Akaike's Information Criterion and Recent Developments in Information Complexity", *Journal of mathematical and Psychology*

Bozdogan H., Camillo F., Liberati C. (2006); "On the Choice of the Kernel Function in Kernel Discriminant Analysis Using Information Complexity", *accepted to Proceeding Cladag2005 – Parma*

Burges C., (1998) A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*

Camillo F., Liberati C. (2006) - e-CRM, web semantic propensity models and micro-data-mining: an application of Kernel Discriminant Analysis to the Glam on Web case – Paper accepted for *JADT 2006* (Besancon, 19-21 April 2006).

Camillo F., Tosi M. (2004) - Approcci semiometrici di posizionamento mediante Text Mining. In E. AURELI CUTILLO, SERGIO BOLASCO. Applicazioni di analisi statistica di dati testuali. ISBN: 88-87242-59-3. ROMA: La Sapienza.

Cristianini N., Shawe-Taylor J. (2000), *Support Vector machines,* Cambridge University Press

Evans C, Marks R. (2001) - *Probing the subconscious using Semiometrie* – Admap

Lebart L., Piron M., J.F. Steiner (2003) – *La sèmiométrie* – Dunod Parigi

Picca D., Curdy B., Bavaud F. (2006) – Non linear correspondence analysis in text retrieval: a kernel view. Paper accepted for *JADT 2006* (Besancon, 19-21 April 2006)

Yang Y., Zhang J., Kisiel B. (2003) – *A scalability analysis of classifiers in text categorization* – Proceedings of SIGIR 2003 – Toronto Canada

Yang Y. (2000) – *An evaluation of statistical approaches to text categorization* – Kluwer Academic Publishers