Texto! *octobre* 2008, vol. XIII, n°4

Interpretive Latent Semantic Analysis

Christian Mauceri

Compagnie IBM France

Tour Descartes – Paris La Défense 5

mauceri@fr.ibm.com

***Abstract*— In this paper we propose to address a recurrent issue of LSA: the difficulty to interpret factors. We use an Interpretive Semantics approach to rethink LSA and propose a new framework for textual semantic analysis taking into account the necessary tradeoffs between human workload and efficiency.**

I. INTRODUCTION

Interpretation is a human activity not mechanizable. From an interpretive point of view reading a text is not reading a sequence of sentences as a computer program reads a sequence of instructions, it is rather an interpretative trajectory, a succession of operations that allow one to assign one or more meaning to a linguistic sequence. Successful endeavours like Google search engine aim at capturing traces of interpretations; hyperlinks for Google, buying behaviours for Amazon. Yahoo's early success was first due to interesting human classifications of sites by subjects. The Amazon Mechanical Turk[1] is a sign there are clearly things which can be done only by humans: those requiring interpretations. Capturing interpretations in text is a major issue. Latent Semantic Analysis is not an exception of this state of fact. We present here an Interpretive Latent Semantic Analysis (ILSA) of texts.

In the first section we briefly present the Interpretive Semantics theory and its link with the Gestalt theory. The second section describes our interpretive approach of LSA. We show in the third section how ILSA can be used to give an interpretation of "The heart of darkness". The fourth section concludes by a discussion on the method and further developments.

II. INTERPRETIVE SEMANTIC, PASSAGES AND KEYNESS

Interpretive Semantics (IS) is a linguistic theory founded by François Rastier [1] [2]. It is a "second-generation" synthesis of European structural semantics, developed in the wake of Ferdinand de Saussure.

Interpretive Semantic use uses minimal units of sense called semes to describe structural relations within a corpus. Semes are not used to describe isolated words but rather semes are defined as sets of words related to them. For instance, instead of describing a priori 'chair' as {/furniture/, /for sitting/,etc…}, /furniture/ is described by the set {'chair', 'closet', 'table', 'sofa'}. Semes depend on the context, they are not universal truth; they are defined by the act of reading. In Interpretive Semantic, reading is the result of an interpretation, an operation specifying the meaning of a text. An interpretation can add or remove semes to words, depending on the context.

A central concept in Interpretive Semantic is the notion of isotopy. An isotopy is the effect produced by a seme recurrence in a text. An isotopy analysis produces a list of words having some contextual semes in common. Isotopies are given a priori and come before the seme definition. The reader looks for isotopies he/she supposes to exist in the texts.

In Interpretive Semantics, the notion of passage [3] is preferred to the notion of sentence. On the signifier plan the passage is a text extract, a substring (in a computer program a text is a string) between two spaces or punctuations. On the signified plan, it is a fragment pointing to contexts on the right and on the left which can be close or distant; it's a textual extension of the Saussurian sign taking contexts into account. A priori, there is no logical definition of the passage even if it is something people easily understand and widely used in the rhetoric/hermeneutics tradition.

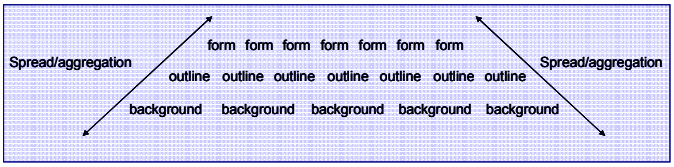

Fig. 1 Gestalt view of passages

[1] from the Wolfgang von Kempelen automaton dressed in Turk which was in fact controlled by an excellent chess player dwarf

From a Gestalt point of view passages [4] can be represented as forms drawn on a background (Fig. 1)

A. Outlines are substrings on the signifier plans, usually words, which are pivots of salient cooccurrences within passages. A statistical filter allows asking an interpreter to select salient pivots among a reduced list of substrings.
B. We can ask, in a similar way, an interpreter to validate forms as clusters of passages gathered according common outlines.
C. The background is seen as the set of the outlines, their repetitions bear isotopies. Relevant clusters of outlines hence draw forms on this background. On the signifier plan these forms describe semantic compounds.

III. INTERPRETIVE LSA

The passage Gestalt view sketched in the previous section fits nicely with Latent Semantic Analysis. Fig. 2 roughly depicts Latent Semantic decomposition [5] of a set of n documents seen as vectors in the Salton Vector Space model of documents. A Principal Component Analysis (PCA) of this set of documents reduces the dimensionality of the original space by projecting the documents on their first principal factors: those exhausting the maximum information.

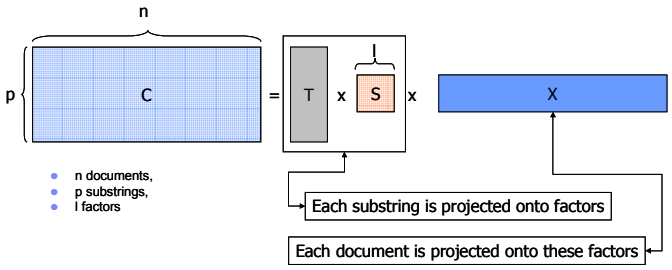

Fig. 2 LSA model

In such a model two documents can be seen as close even if they do not share any words. This overcomes the limitations of the traditional Salton model and makes LSA very attractive. However it is not always easy to explain the factor meaning but worse, depending on word weighting and selection, the model can produce spurious factors.

Many improvements of the initial model have been proposed such as Probabilistic Latent Semantic Indexing (PLSI) [6] and Latent Dirichlet Allocation (LDA) [7]. These models are based on the assumption that words in documents are generated by hidden latent topics, in other words each document is a mixture of topics. We cannot enter in the detail of these models but let's notice they do not really address the interpretation problematic: in short, they assume the model by itself can detect topics as salient collocations in the training set. On the contrary the model we propose focuses on the significance and interpretability of the factors by capturing interpretation[2].

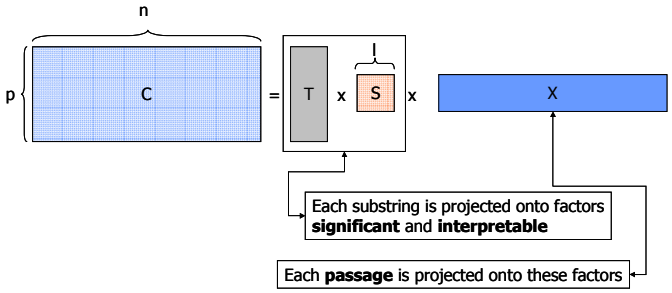

Fig. 3 ILSA model

Our model is presented in (Fig. 3), instead of considering documents we consider passages and we want our factors to be significant and interpretable, saying that they reflect a human point of view on a corpus of documents. By corpus we do not mean any random set of documents but a set of documents sharing characteristics of genre and discourse. We aim at capturing the human point of view on these data. This problematic is very general and

[2] We do not say human interpretation as from our point of view there is no such think as mechanical interpretation

corresponds to a semiotic view on data as depicted in (Fig. 4) where the semiotic vehicle/value[3] is governed by a higher order duality between guarantor and point of view [8].

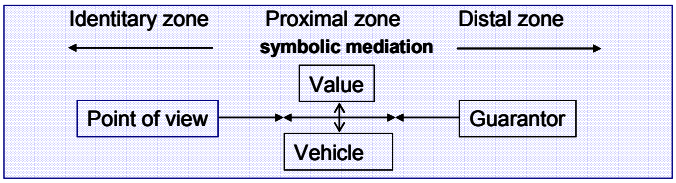

Fig. 4 A semiotic view of data

In the case we are interested in, the guarantor determines the corpus and the values we want to reflect by the use of LSA factors is a global point of view on corpus excerpts[4]. This is a very important point in our approach as in such corpus passages share regularities of content and expressions.

The passages we consider here are roughly determined by paragraphs in a document but more sophisticated descriptions could be considered. They are described by relevant collocations, saying collocations reflecting a human point of view. In order to alleviate the burden of collocation selection we filter them by the exact Fisher test; the null hypothesis being that the two words of a collocation appear together by chance. We only take into account the collocations when we can reject the null hypothesis with high confidence level[5].

This test has long been considered as impossible to compute for textual collocation because words in texts follow a Zipf distribution leading to contingency tables with very large margins and log likelihood ratio (LLR) [9] was preferred to test confidence level.

A x B	**A**	**Not A & B**	**Margin**
B	x	b – x	B
A & Not B	a – x	n – a – b + x	N – b
Margin	a	n – a	N

Table 1 Cooccurrence contingency table

In (Table 1) we show the contingency table for two words A and B cooccurring in a passage where word A appears ‘a’ times, word B, ‘b’ times and ‘n’ is the corpus size in words. The probability the two words appear ‘x’ times together is given by the formula below:

$$f(x)=\frac{\binom{a}{x}\binom{n-a}{b-x}}{\binom{n}{b}}=\frac{a!b!(n-b)!(n-a)!}{n!x!(a-x)!(b-x)!(n-a-b+x)!}$$

This probability is now computable thanks to the Lanczos approximation of the Gamma function and for the reasons exposed by Moore in his paper [10] we prefer to use the exact Fisher test rather than LLR.

We can consider the passages cooccurrence matrix filtered by this probability retains only cooccurrences with a high p-value and weighting them by their probabilities. By basing the LSI decomposition on this matrix rather than on the raw cooccurrence matrix we obtain more significant factors.

Shawn Martin [11] showed that it is possible to use, for a given kernel, the Gram/Schmidt orthonormalisation process to compute an Approximate Principal Component Analysis. An APCA uses points of the data set to define the factors.

[3] A generalisation of the signifiant/signifier duality
[4] A passage being the duality between an excerpt and the value given by the point of view
[5] a p-value of 0.95

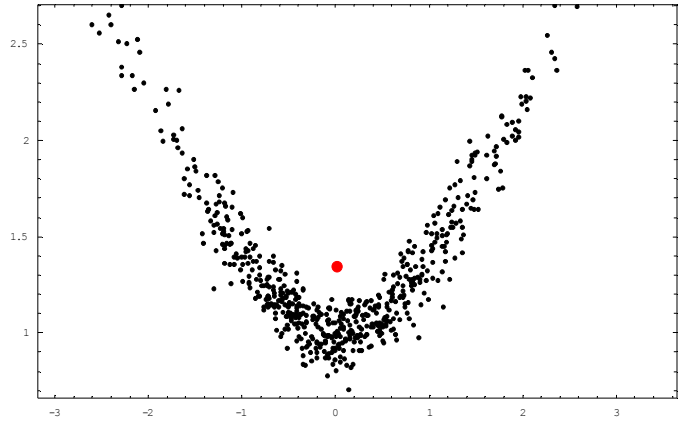

Fig. 5 Parabola with Gaussian noise

To give the flavour of the method let's consider the parabola in (Fig. 5) the red dot is the centre of gravity of the data cloud. The blue dots in (Fig. 6) belong to the data cloud in (Fig. 5) and are the best approximation of the principal axis of the data cloud.

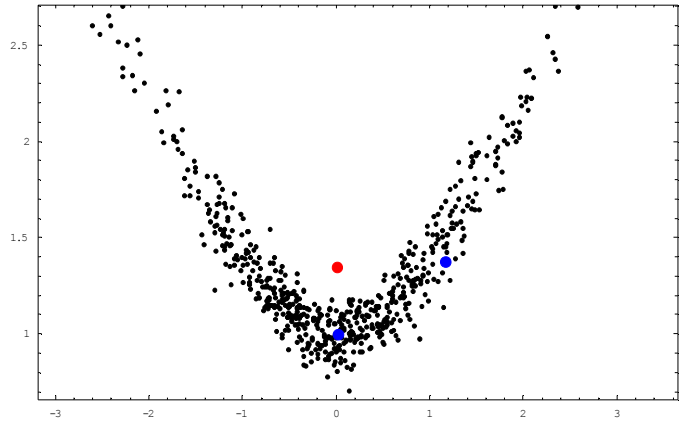

Fig. 6 Parabola and APCA

Cooccurrence matrix B filtered by the exact Fisher test we described previously can be seen as a graph whose nodes are words. If we consider the mapping:

$$\Phi(b_i)=\begin{pmatrix} 0 & . & 0 & b_{i,1} & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & . & 0 & . & 0 & . & . & . & . \\ b_{i,1} & . & . & b_{i,i} & . & . & . & . & b_{n,1} \\ 0 & . & 0 & . & 0 & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & . & 0 & b_{n,1} & 0 & . & . & . & 0 \end{pmatrix}$$

of the columns of matrix B, then : $k(b_i,b_j)=\langle\Phi(b_i),\Phi(b_j)\rangle$ [6] is a kernel on the nodes of the cooccurrence graph represented by matrix B. Two nodes are orthogonal if they are not directly linked. A very simple APCA for this kernel and the data clouds of the cooccurring words can be computed as follow:

[6] $\langle .,. \rangle$ is the Frobenius scalar product.

A. Select the node having the higher sum of this kernel with its neighbours as the principal axis
B. Remove it from the graph and go to previous step.

The projection of a word on the factors is straightforward: it is the corresponding value between the two nodes in matrix B. At this point we have partially reached our two initial goals: factors are significant because of the exact Fisher test filtering and they are interpretable as they correspond in fact to words in the corpus.
The last step is to check if the factors allow meaningful clustering of passages. It is that ultimate step which can allow us to carve interpretations of the corpus according to the previously computed factors.

Traditional hermeneutics retained three elements of understanding, three subtleties: understanding, interpretation, and application. Gadamer's endeavor [12] has been to show the impossibility to separate understanding and interpretation. For him, understanding is the understanding of the meaning, but the meaning is not immanent to work or things; it springs from the meeting of text and reader, of work and the public. The meaning springs from the interpretation; it seems to lead to the impossibility to assess the relevance of any interpretation. The answer of Gadamer is to refocus hermeneutics on what it considers its fundamental problem, "the application," in order to reestablish the authority of tradition.

Interpretation proceeds from the structure of the anticipation of the understanding, which is concretized in the hermeneutical circle: The understanding of the particular results from the general and the general from the particular in a constant movement of projection and readjustment. It is impossible to understand the meaning of the whole without understanding the meaning of its parts, but the meaning of the parts depends on the whole they form. What is presupposed of the analyzed whole determines the initial understanding of its parts, but in turn, the refinement of this understanding and even its challenge modify what was presupposed at the beginning and so on.

Clustering tools show and represent structures or relations between objects people cannot embrace due to their numerous and complex meanings. They allow parts identification of the whole according to the a priori of the analyst expressed by the coding of the data. The identification of these parts and the relations they maintain modify the analyst's prejudice: the initial coding can be changed accordingly and, therefore, part identification. Thus, the handling of a clustering tool is going, by a play of projections and readjustments, to materialize the understanding of the data, to materialize their interpretation, to materialize the hermeneutical circle.

The clustering method we use is based on kernel density as described in [13]. The kernel, here, is the cosine of the passages seen as vector in the principal factor space. It is possible to modify the clustering according to the interpretive process previously described by modifying the vocabulary taken into account for the collocation computation or by removing spurious collocations. This very constrained way to interact with the system is a guarantee against too subjective an interpretation; it is a way to take into account the hermeneutic application which was, as we saw previously, Gadamer's main concern.

Therefore, clustering is a bridge between LSA, as we have presented it so far and IS. Indeed as noticed by Yarowski [14] statistically significant collocations are often meaningful so we can expect that the principal factors of the collocation network are good indicators of the meaning of the passages they describe and clusters based on these descriptions strengthen these indications. From an IS point of view these clusters can be seen as semic compounds which are stable groups of semes not necessarily lexicalized. So it is not excessive to make an analogy between seme recurrences and word recurrences. Hence to a similarity induced by the projection on factors of significant cooccurrences it is possible to match a similarity induced by seme cooccurrences. It is a linguistic explanation of why documents which do not share common words can be similar in LSA.

IV. An application

"The heart of darkness" [14] by Joseph Conrad is a relatively short novel of about 39000 words. It inspired Francis Ford Coppola's movie "Apocalypse now". It is the tale of the mission of a man, Marlowe, to find another, Kurtz, in the Congo's innermost wilderness. But more deeply, it is about Man confronting his fears, insanity and death, it is a reflection about civilization and the primitive, a reflection about the hypocrisy of the hollow men (the pilgrims) coming from Europe to Congo motivated by greed and pretending to bring progress to the savages.

We have chosen a novel rather than a more traditional corpus because we want to emphasize interpretation and one can read a short novel to assess the quality of an interpretation rendered by passage clustering.

The electronic version of the novel we used is coming from the Guttenberg project. The excerpts corresponding to passages are strings at least fifty characters long starting and ending by a dot, an exclamation mark, a question mark or paragraph delimiter. For instance the two first passages taken into account in the novel are:

- "The Nellie, a cruising yawl, swung to her anchor without a flutter of the sails, and was at rest."
- "The flood had made, the wind was nearly calm, and being bound down the river, the only thing for it was to come to and wait for the turn of the tide."

This very raw definition led to segment the text in 1430 substrings we used as passage vehicles.

We used a classic stop words list in order to remove grammatical words our very raw processing is unable to take correctly into account[7]. We added to this list the words which can produce spurious collocation despite the statistic filter, for instance: 'seemed', 'made', 'came', 'little'. We could filter them in a less brutal way indeed, but the related treatments are beyond the scope of this paper. The initial vocabulary taken into account was constituted of words occurring at least three times[8]. It took about half an hour to read the 1698 selected words and mark those we did not want to keep in the cooccurrence network.

We kept the collocations having a p-value above 99% and projected them on the first 150 factors. After having been projected on these factors we clustered the passages in such way that the density of their cosines in a cluster was above 0.01 and their inter-cluster density below this same threshold. After some trials we removed some other pivots from the original list as they produced spurious factors, for instance: 'fellow' or 'bit'. Finally after 2 hours of analysis and retrials we got a clustering of ten clusters we are going to present succinctly. But let's first describe some indicators we commonly use.

As recommended by Gerard Salton [16] we normalize the vectors representing the passages in order they belong to the unity sphere. For cluster C we note G_C its centroid and therefore the cluster density is given by $\langle G_C, G_C \rangle$ where $\langle .,. \rangle$ is the usual scalar product. Similarly for two clusters C_1 and C_2 $\langle G_{C_1}, G_{C_2} \rangle$ is their inter-cluster density. If p is a vector representing a passage then its average similarity with the cluster C is $\langle p, G_C \rangle$ and the ratio $\frac{\langle p, G_C \rangle}{\sum_{p' \in C} \langle p', G_C \rangle}$ represents its belonging grade to the cluster, its contribution to the cluster.

We note $G_C = \left(f_{C,i}\right)_{1 \le i \le m}$ where $f_{C,i}$ is the projection of the centroid on the i[th] factor.

So for the i[th] factor we can define two indicators for a cluster C, the specificity and the intensity.

For a clustering □ the specificity is given by: $\frac{|C| f_{C,i}}{\sum_{C' \in \Pi} |C'| f_{C',i}}$, indeed if the i[th] factor only occurs in the cluster C this indicator reaches its maximum 1 and if it does not occur in the cluster C it is equal to 0.

For a clustering □ the intensity is given by: $\frac{\sum_{p \in C} f_{p,i}}{|C|}$, indeed if the i[th] factor only indexes the passages of the cluster C and only them this indicator reaches its maximum 1 and is equal to zero if it does not occur in any passage of the cluster.

Finally we borrow from Gerard Salton [16] the notion of discriminant value. For Gerard Salton a term is discriminant if when it is removed the global similarity of the documents increases. In our case, a factor is discriminant if it lowers the inter-cluster density.

Below are the clustering most discriminant factors.

Factor	Discriminant value
work	913,207
she	829,359
Kurtz	793,207
splendidly	781,742
talked	623,711
bank	572,883
cry	544,254
human	515,027
flames	509,344
clear	495,426

Notice we have kept the word 'she' as we considered it has a strong semantic value in the context of the novel: the black woman in Africa and Kurtz's spouse in Europe.

[7] This is not a mandatory step, we could indeed take into account more sophisticated string processing which would not lose text linearity and rehabilitate the important semantic role these words could play

[8] Because we did not want to spend too much time in vocabulary reading

Because we cannot detail all the clusters in this paper we have decided to present the three biggest clusters. Each cluster is described by its first five factors of highest contribution along with its passage of highest contribution.

This first cluster is dedicated to Kurtz which is not a surprise. This is confirmed by the first passage and by the very high specificity of the factor 'Kurtz' indicating it occurs almost only in this cluster with quite a high intensity.

Size: 140

Passage	**Contribution**
They had given up Kurtz, they had given up the station; Kurtz was dead, and the station had been burnt--and so on--and so on.	0,015

Factor	**Contribution**	**Specificity**	**Intensity**	**Discrimination**
Kurtz	47,15	0,858	0,337	793,207
bank	25,615	0,733	0,183	572,883
year	11,104	0,438	0,079	407,293
idea	9,228	0,582	0,066	162,668
half-caste	8,817	0,384	0,063	422,738

The second cluster is formed around the notion of color despite the name of the first factor. The first passage gives a good explanation of the reason behind. In Conrad novels and in this one in particular colors play an important role. The black and the white are obviously in opposition; the black woman in Africa and the white woman in Europe, for instance. By aggregation the 'human' factor plays an important role.

Size: 138

Passage	**Contribution**
He broke off. Flames glided in the river, small green flames, red flames, white flames, pursuing, overtaking, joining, crossing each other--then separating slowly or hastily.	0,013

Factor	**Contribution**	**Specificity**	**Intensity**	**Discrimination**
flames	25,089	0,742	0,182	509,344
Old	21,168	0,728	0,153	392,539
black	17,007	0,617	0,123	457,566
Neck	16,499	0,842	0,12	113,629
Human	11,212	0,436	0,081	515,027

This third cluster is, by far, the most interesting one as its first passage gives the key of the novel, the theme of the policeman and the butcher. By opposition to the primitive, civilized countries have policeman and butchers:

- ". . . Here you all are, each moored with two good addresses, like a hulk with two anchors, a butcher round one corner, a policeman round another, excellent appetites, and temperature normal--you hear--normal from year's end to year's end."

The three first factors confirm this important point.

Size: 134

Passage	**Contribution**
You can't understand. How could you?--with solid pavement under your feet, surrounded by kind neighbours ready to cheer you or to fall on you, stepping delicately between the butcher and the policeman, in the holy terror of scandal and gallows and lunatic asylums--how can you imagine what particular region of the first ages a man's untrammelled feet may take him into by the way of solitude--utter solitude without a policeman--by the way of silence--utter silence, where no warning voice of a kind neighbour can be heard whispering of public opinion? These little things make all the great difference.	0,016

Factor	**Contribution**	**Specificity**	**Intensity**	**Discrimination**
Solitude	25,065	0,841	0,187	260,383
Utter	24,262	0,855	0,181	218,062
policeman	21,639	0,81	0,161	243,816
Correct	14,183	0,622	0,106	311,836
Grass	13,043	0,608	0,097	282,137

The selection of clustering parameters is obviously crucial and more stable and homogeneous clusters corresponding better to semic compounds could be obtained using a higher density threshold. We used a low threshold in order to reduce the number of clusters and quickly assess the computed factors.

V. Conclusion

We have shown there is a deep relation between Latent Semantic Analysis which is a mathematical model and a linguistic theory, Interpretive Semantics. To our knowledge it is a first of a kind and opens new avenues to LSA. Indeed we have only shown how LSA can be explained using IS concepts and the model we have developed is still very rudimentary even if it can capture very interesting relations on real texts no matter the natural language used see [17] for an analysis of a French text. But beyond that we also showed that interpretation remains fundamental when analyzing text and that what we can call Interpretive Latent Semantic Analysis (ILSA) can be a model of choice for Computer Aided Interpretation of texts.

In the future we plan to use more efficient string processing techniques emanating from bioinformatics like suffix trees to detect passages and collocations not necessarily based on words.

REFERENCES

[1] François Rastier, *Meaning and textuality*, Toronto Buffalo: University of Toronto Press, 1997.
[2] Francois Rastier, Marc Cavazza, and Anne Abeille, *Semantics for description,* Center for the Study of Language and Inf., 2001.
[3] François Rastier, *Passages,* Texto ! [Online], Available: http://www.revue-texto.net/index.php?id=64, 2008.
[4] François Rastier, *Formes sémantiques et textualité,* Texto ! [Online], available: http://www.revue-texto.net/index.php?id=532, 2006.
[5] Deerwester, Dumais, Furnas, Harshman, Landauer, *Using Latent Semantic Analysis to improve access to textual information,* CHI's 88, 1988.
[6] Thomas Hofmann, *Probabilistic Latent Semantic Indexing,* SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.
[7] David M Blei, Andrew Y Ng, Michael I Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 2003.
[8] François Rastier, “Sémantique du web vs semantic web ?”, Texto ! [Online], Available: http://www.revue-texto.net/index.php?id=1729, 2008.
[9] Ted Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence*, Computational Linguistics, 1993.
[10] Robert C. Moore, *On Log-Likelihood-Ratios and the Significance of Rare Events*, Conference on Empirical Methods in NLP, 2004.
[11] Shawn Martin, *An Approximate Version of Kernel PCA, Conference on Machine Learning and Applications*, 2006.
[12] Hans Georg Gadamer, *Truth and method,* 2nd edition (January 1989), Continuum International Publishing Group, 1960.
[13] Christian Mauceri, Diem Ho, *Clustering by Kernel Density*, Computational Economics, 2007.
[14] David Yarowsky, *Unsupervised word sense disambiguation rivalling supervised methods*, Association for Computational Linguistics, 1995.
[15] Joseph Conrad, Heart of Darkness, ebook edition (January 2006), Project Gutenberg, [Online], Available: http://www.gutenberg.org/ebooks/526 , 1899
[16] Gerard Salton, C.S. Yang, A vector space model for Automatic indexing, Communication of the ACM, 1975.
[17] Christian Mauceri, Isotopie et indexation, Texto ! [Online], Available: http://www.revue-texto.net/index.php?id=122, 2008.