

LISTENING INSTEAD OF ASKING

HOW BLOGS CAN PROVIDE A NEW WAY TO BETTER UNDERSTAND MARKET TRENDS

François Laurent
Alain Beauvieux

INTRODUCTION

About 70 million blogs have been created worldwide with around 9 million in France. An extraordinary new field of investigation is offered by the blogs as millions of instances of electronic data can be processed by software programs directly.

This paper will explain how to study what is said on blogs thanks to two kinds of tools: the first used to access the sources and collect any relevant information from the blogs, comprehensively and continuously; the second to analyse this collected information and in particular carry out semantic analysis.

WEB 2.0

Blogs are part of the new Internet usually called Web 2.0, as Tim O Reilly has named it.¹ What is Web 2.0? There are various ways of answering the question.

The easiest definition, and certainly the nearest to reality, could certainly be that if Web 1.0 was the web of companies and administrations, etc., then Web 2.0 is *the new Web that individuals now build as part of their daily routine*. Web 2.0 is defined through the spontaneous actions of individuals.

Web 1.0 was built with sites like *Le Journal du Net* and *Au Féminin* whose projects and processes do not essentially differ from those of the *New Observer* or *Prima*: journalists speaking with authority, almost officially, rarely suffering any contradiction from readers – readers with whom it is possible to express an opinion

in clearly identified places: the forums replace the letters to the editor.

Web 2.0 grants the same place to all bloggers: everyone can express themselves independently of financial status or technical capability. Some organisations cheat a little, not hesitating to pay writers for blogging at their site: *Engadget*, for example, which has an audience comparable to that of CBS: CBS News, is part of a commercial enterprise: *Weblogs Inc*.

At the opposite, the sixth most popular, and therefore arguably influential, French blog according to *Technorati*² is *Clea Cuisine*³: “Clea is 26 years old, she makes green cakes, eats bio and adores ‘les tartines.’”; the 18th, according to Wikio, is: *La marmite de Cathy*.

However, Web 2.0 is not only blogs. More generally, it offers new spaces for spontaneous expression from consumers, social networks such as *You Tube*, *My Space*, or community portals such as *Ciao.fr*.⁴

A METHODOLOGICAL PROBLEM

Each year in October, a national sample of 10,000 households representative of the French population aged 15 years and over elects the Products of the Year. The site⁵ shows the 41 prize winner for 2007: sausage, yoghurt, toothpastes, shampoos, detergents, and so on.

All these products were launched in hypermarkets and supermarkets during the twelve months preceding the contest; the jury made up of consumers makes their decision “according to two criteria: product attractiveness

and usage value". Not innovation – even if the products are recent and certain brands maintain the ambiguity: we can read on the Procter & Gamble site: "The Great Price Marketing Innovation recognises the most innovative products since 1988".⁶

The toothpaste category rewards Signal Care Freshness, the "first toothpaste in the world" to use the so called "Core&Core" patented technology (which) gives a feeling of extreme freshness, has an anti-bacterial action and helps to reinforce tooth enamel".⁷

Ciao "is an on line community of several million members who give their opinions and rank millions of products and services for other consumers benefit. Ciao makes available the opinions of consumers not influenced in any way and information on the prices of hundreds of retailers products on a day to day basis making it one of the most detailed sources of "shopping intelligence" on the Web".⁸

Its impossible to resist to temptation of reading what *Ciao* members say about the toothpaste elected Product of the Year: they rank it "103 out of 113 at the *Ciao* hit parade: *Your favorite toothpastes*", which seems less wonderful.

Two kinds of opinion are available: the first tend to be long and structured, appear very positive and strongly endorse the product:

"Hello everybody

"What a pleasure to have a mouth in perfect condition, and even more so, when my breath is so perfect! [...]

"I just wanted to write a few lines for a great toothpaste.

"I kiss you".

Others are shorter and significantly more critical, not to say ironic:

"Good, Ok, a toothpaste is useful above all for cleaning teeth, but me, I also want to have fresh breath: with this new Signal it's missing".

Or: *"The texture of this toothpaste is weird, gelatinous, yuk! When you put it on the brush it tends to slip off into the sink, not exactly practical. It tastes more like a*

drug than a toothpaste, it also slips around your mouth and ends up on your gums: result, you always have to retrieve it and then brush hard for a long time to properly clean your teeth; it freshens your breath, but not for long".

First observation: the previous example should not be interpreted as a criticism of *Signal*, many other brands elected *Products of the Year* fare no better in the eyes of *Ciao* members who, for example, criticise very harshly *Tampax Compak Fraich* from *Procter & Gamble* but at the same time all vote for the company's *Pampers Active Fit* nappies.

Second observation: the research was conducted by *Worldpanel* and could not be suspected of irregularities or mistakes; the overall methodology of the selection should not be called into question either. Admittedly many products are not represented simply because their manufacturers do not wish to take part.

The real questions are elsewhere and they are of two types.

Firstly of a sociological nature: the selection of the *Products of the Year* is based on the idea that the innovation constitutes a cornerstone of marketing; and that an organisation's capacity to continually launch new products is good for brand reputation. In pursuit of this it is logical to seek the votes of the 10,000 panelists constituting the jury. But does the axiom preserve its universality?

Secondly of a methodological nature: who is actually telling the truth? Are people who answer the public opinion polls really representative of the studied population? Are their answers identical to those of non respondents? Or on the contrary do Internet users who, as a result of Web 2.0, can now comment freely merely constitute a rebellious or naturally controversial community?

Is it the case that the people who agree to answer surveys and those who comment on Web 2.0 constitute nothing more than two micro populations lost within a dumb ocean?

THE MARKET RESEARCH CHALLENGE

Two sources of information are available.

The first one is traditional: the *Products of the Year*; the other one is a Web 2.0 site: *Ciao*. These sources are at complete opposite ends of the spectrum regarding the value of innovative products.

We do not as yet have all the necessary elements to decide between them: however, having argued in favour of the latter, we can, at this time, provide complementary information that highlights failings in some traditional areas of market research!

First of all, we have to question the legitimacy of a survey which necessarily finds each year, and for forty categories of products, really appealing innovations ... when many other analyses underline the precarious nature of the challenge.

For instance according to company *XCT*, one new product in every two is a failure in the two years following its introduction to the market; and *Ernst & Young* and *ACNielsen* estimate that "43% of the "true new products" launched in Europe are a failure in the 12 months which follow their launch".⁹

Other researchers are more alarmist, according to Jean-Claude Andréani: "in 10 years, the % of new product launch failures will have risen from 40% on average for FMCG products, durable products and services, to 95% in the USA and 90% in Europe".¹⁰

I suggest comparing all these results collected at meetings conducted in Paris from qualitative groups dedicated to using new purchasing processes, and during which time the question of innovation and the Products of the Year had spontaneously been mentioned:

"Personally I bought the product elected Product of the Year...

"But it is just marketing, these brands".

Concerning the quantitative studies we would highlight some simple facts that are too often overlooked.

Telephone surveys, for example: with home phones it is only possible to reach four out of five French people - and 9% have adopted VoIP: As with cell phones there is no directory for internet phones.

However, the most serious problem concerns the percentage of absolute non respondents: how many people hang up as soon as the automated call asks them to answer a survey? We generally consider that an average of five to seven calls are necessary for one to be useful! But these four to six others, what would their answers have been? Impossible to know!

The same problem concerns face to face surveys: how many people quicken their pace in the street even before the interviewer has a chance to approach them?

We'll never know what somebody could say who systematically refuses to speak...

Qualitative research has its own problems, not better, not worse: we could evoke, for example, the "professionalism" of consumers participating in focus groups: last year in France, a huge discussion started within the profession leading ... to nothing.

And we always meet the same faces in focus groups; we always hear the same comments: "*Hang on, today I was not entitled to join, why?*"

With the previous statements it is not our intention to discredit any one profession: before highlighting new approaches it's important to stress that market research is a social not exact science.

And it's not because the results we observe from analysing the Blogosphere contradict them that they are necessarily false.

The definition is the following: "the collective power of the small sites that make up the bulk of the web's content"; "an essential part of Web 2.0 is harnessing collective intelligence"; "designed for "hackability" and remixability" ¹¹.

Studying what is said on Blogs requires two kinds of complementary tools:

- A tool *to access the sources and to collect* any relevant information from the blogs, *comprehensively and continuously*.
- A tool capable of analysing this collected information, and in particular, able to carry out some *semantic analysis*.

BUILDING A RELEVANT COLLECTION

“Long tail” vs. “blogs of authority”

The four million blogs (except teenagers) constitutes a group of observers who can be put into two main categories: the most active contributors, who often are an authority in their field, and the huge majority of occasional contributors, the latter commonly referred to as the “long tail”.

The first category can be identified starting with simple criteria such as the “ranking” of the blog by the principal search engines which correctly calculate its visibility and recognition within the Blogosphere by combining the number of page views and hyperlinks that point towards the blog from other sites.

In a way, surprisingly, the number of authoritative blogs is quite low: one can regard it as being a few tens of blogs, seldom is it in the hundreds. For example, the number of French blogs having a coefficient “ranking” higher than 5 in the field of new technologies is about sixty. It is a question of trying to confirm that the representation of these blogs is sufficient to guarantee complete information and subject coverage within the Blogosphere.

In other words, to believe that “everything that is not known as in the authoritative blogs does not have a value” is, from our point of view, very restrictive: a quick visit to any section of blogs very often shows that the authoritative bloggers often comment on opinions that are already known but that seldom generate any spontaneous reactions contrary to the long tail blogs. To limit the perimeter of the collection plan to just the

authoritative blogs risks eliminating essential data for later analysis.

Accessing the blogosphere

It is easy to identify and index the list of authoritative blogs by criteria of their ranking and the specific addition of known addresses, and subsequently to index them and maintain the listings on a regular basis. On the other hand, it is unrealistic to adopt this approach with the blogs that form the “long tail”, such a volume requires the data processing resources that only the largest of Internet search engines can offer.

From this point of view using their indexing capabilities is an interesting option, the frequency of indexing that they are capable of applying to the Blogosphere is more than sufficient for the studies considered. On the other hand the definition of the Blogosphere is too broad and covers sites that are not always relevant. For example the first result using the query term “Organic foods” in GoogleBlogs (United States), is the site “www.agricultureinformation.com” which describes itself as the “worldwide agriculture business portal”.

On French blogs using the query term “GMO” (GMO in French), the first three results given by GoogleBlog (F) are newspapers (“www.estrepublikain.fr”, “www.cyberpresse.ca”, “www.lejisl.fr”)!

The way in which we chose to deal with this challenge was to develop a tool for post-processing which makes it possible to qualify what is a blog and what is not. This qualification is carried out by analysis of the contents and the structure of the blog.

Thus, a comprehensive set of criteria were defined which makes it possible to recognise, for example, the existence of an author with a description of his or her profile. The advantage of this approach is that it makes it possible to design, evolve and change the criteria in order to add conditions, making it possible to take into account the characteristic of blogs that may have previously not been included and thus to optimise the process overall.

Currently, from a question submitted to a blog search engine on a given subject, and after filtering by the dedicated module, the error rate, i.e. of sites considered as blogs in error, is about 7%.

This rate is itself is open to discussion insofar as these sites are often, in extreme cases, something between real blogs and mere places of expression that experts in all instances would have regarded as blogs.

To express centers of interests

The approach taken consists of selecting all blogs identified as being relevant with regard to a question on a particular subject using the results provided by Internet engines and then removing those sources which are not blogs. This approach thus implies that the formulation of the question used for the selection is sufficiently broad so as not to exclude any key information from the subsequent process of analysis.

Indeed, the objective of the process is “to observe rather than to ask”. It is therefore essential not to build in barriers or filtering of information in the construction of the analysis interface.

The choices made in the selection criteria are generally very broad (name of the product or category of products for example) to guarantee fully comprehensive coverage.

The technology used allows a combination of Boolean requests and constraints exploited by fuzzy logic algorithms to define a research perimeter that is very broad yet highly focused on the required subject by taking into account the “meaning” of the results referenced.

Compromise “coverage” vs “precision” (noise, silence)

To seek absolute perfection is unrealistic and it’s clear that an effective process of control must be employed by the user in order to confirm or cancel the choices made by the tool.

In the same way it is possible to define more or less restrictive search criteria, the variation of these parameters implying on one hand a larger coverage but with

the likelihood of more “noise” returned in the results or on the other a higher degree of accuracy with the risk of potentially missing certain elements. This “coverage-precision” compromise depends on the subject (the kind of the vocabulary, the popularity of the subject, etc.) and can be optimised by the user.

Specific research and recurring collection

Two operations are generally necessary when building a data collection plan: *searching*, i.e. the ability for a user to find at any particular time all current or existing information that answers a specific question; and *collecting*, which is a recurring operation making it possible to identify, at regular time intervals, any new information that has been published on a given subject.

In both cases, the user will have the necessary tools, a search engine and a collect engine, allowing him or her to build a data collection plan. The process of timed collection makes it possible to follow the evolution of a trend or event by organising, in successive and correctly dated subsets, the relevant data which in turn then allows far more accurate processes of analysis to be applied to it subsequently.

Extraction of relevant information

The last challenge is related to the nature of blogs: the data is generally very poorly structured, unstructured even, the significant pieces of information often being swamped in a flow of auxiliary information which it is essential to eliminate. This absence of structure (the complete opposite to data exploited by data mining tools) constitutes an important technological barrier. A specific module makes it possible to extract the parts of the blog considered to be significant by analysis of its topology and its content. In this case, only the initial post is retained, any subsequent comments being removed. Indeed, the comments that follow an initial post are generally not very reliable and are often the medium chosen by agencies that specialise in distributing “sponsored noise” (hype or buzz). The peripheral elements of the blog are eliminated, only the initial contribution is preserved.

STRONG TECHNOLOGICAL CONSTRAINTS

Discovering vs. monitoring

To analyze data originating from blogs requires a technology capable of processing textual data. It is clear that data-mining tools cannot be used except for creating an image of a preliminary structure according to a predefined model of each blog. This approach would consist of applying a reading “grid” to elements of the blog which would in fact negate completely the idea of “observing rather than asking”. Given this the best approach is to use tools capable of Natural Language Processing.

Two established schools of thought exist: the linguistic approach which necessitates, in advance, the building of linguistic components such as semantic dictionaries, networks, conceptual graphs, linguistic cartridges, etc. before any processing can take place; and the statistical approach which exploits primarily information theory algorithms.

In fact neither approach is really satisfactory for analysis of the blogosphere: the linguistic approach, in addition to high levels of maintenance and prohibitive development costs, is dependent on the fact that the user models in advance what they seek. It is adapted for use in relatively stable semantic environments for monitoring and follow-up operations. However it is intrinsically unsuited to the process of observation where the user can be involved in a process of discovery, not necessarily in possession of any particular prior subject knowledge or wanting to try and model something about which they have little or no knowledge.

The statistical approach is often regarded as insufficient also as soon as the volume of data being exploited falls below a certain level. We noted that for the subjects studied the number of blogger contributions of real value is often less than 100, making the statistical approach impractical.

To highlight the significant elements

Our approach is based on a mixture of both linguistic and statistical technology and primarily consists of

determining the principal concepts and utilising them on a hierarchical basis. A concept can be thought of as any list of words corresponding to a valid syntactic construction and satisfying a set of criteria such as their frequency of appearance, allowing their importance to be determined within the text in which they reside.

This weighting makes it possible to treat concepts on a hierarchical basis, each blog contribution being characterised by a tree structure. This approach known as “document signature” or “document ADN” is usually exploited by information retrieval and competitive intelligence software and makes it possible to obtain very high quality results in new semantic fields.

ANALYSING THE COLLECTED DATA

After being identified by a Collect or Search engine, each blog is processed in order to extract the principal contribution and then subsequently to calculate its signature. On completion of these processes, the user has a collection of data across which multiple processes of analysis are possible.

Extracting key concepts: Determining from all the selected contributions which are the principal concepts. This calculation is carried out directly starting with the signature of each contribution. As a result it is possible to know, looking across all contributions, which expressions are most frequently used.

Extracting named entities: In parallel to the signature calculation process, it is possible to apply rules to each contribution in order to extract elements of data meeting these rules: names of people, locations, organisations and companies, etc. This operation, known as extraction of named entities, well tried and tested in information retrieval solutions, allows the user to quickly identify within a collection of data the people associated with or quoted in the contribution.

Volume and Centering: the process of acquisition and processes described previously makes it possible to create, for each contribution, specific metadata: date, source, contributor, criteria of selection, key concepts. This metadata corresponds to the structured data of

PART 1 / CREATIVE APPROACHES AND NEW METHODS

each contribution to which statistical algorithms can be applied: in particular, it is easy with this approach to study the evolution of a concept over a period of time and to qualify its relative importance, an operation called Centering. For example if a concept is the name of a product it is possible to identify whether the product is simply being mentioned or the subject of a dedicated contribution, with all the intermediate values.

Clusterisation of data: the signature of each document makes it possible to easily apply clusterisation algorithms and thus to identify subsets of contributions having common concepts. The clusterisation calculation is produced by algorithms used in data-mining tools and is particularly robust. It is possible, according to the required application, to modify the principal parameters and hence obtain more or less dense clusters.

A calculation of density and centrality (links, inter-clusters) is then made for each cluster which is then, in turn, positioned on a graph around two axes. (See figure 1.)

The analysis of this graph is particularly useful because it makes it possible to identify which are the strong topics or, just as importantly, the newly visible or “unexpected” topics.

The most central clusters gather together and represent those contributions amongst whom the most commonly shared concepts are present (centrality). The importance of their density directly translates the coherence of this subset. Conversely, the clusters which are dense but not very central thus not very related to the other clusters are regarded as being “unexpected” compared to the others. They convey coherent ideas (density) but are detached from the majority. These clusters are often studied in great detail as “weak” signals can be a sign of new but important emerging trends or events; the probability and possibility that these weak signals become strong signals is one of the primary tasks of the analyst.

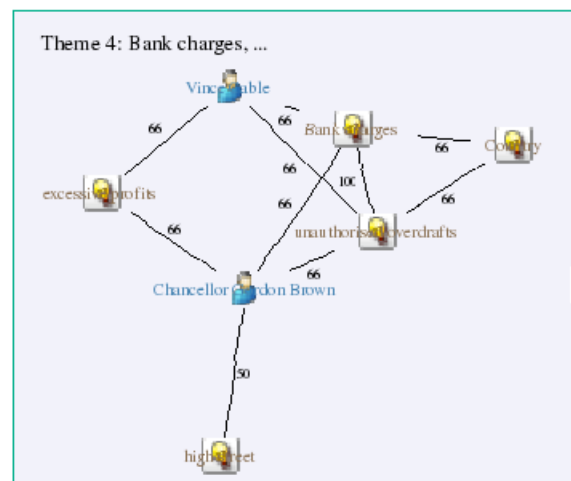
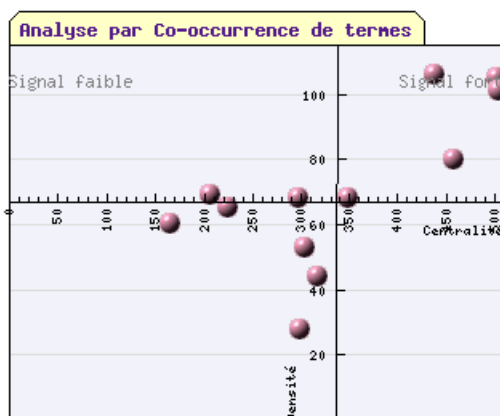
PROVEN EXPERIMENTAL RESULTS

This methodology was applied to three cases studies and led to experimental results validating this approach. The three experiments are described below.

When launching a new product

In the case of this new yoghurt, we studied the reactions of consumers through the blogs. We showed that the principal extracted concepts were the same key messages communicated direct by the manufacturer

FIGURE 1
CARTOGRAPHY OF THE REACTIONS OF BLOGGERS TO THE FINANCIAL RESULTS OF ENGLISH BANKS



themselves: duration of consumption (“six weeks”), originality of the formula based on natural ingredients (“vitamin E”, “green Tea”), the correlation between “health and beauty”. Moreover, we could appreciate the difference in sensitivity between two categories of contributors: consumers and specialists. In the latter case, the scientific basis of the new product appeared to be essential in particular through concepts such as “clinical study”. We note that a similar study undertaken using traditional methods led to the same conclusions.

When communicating the annual results of large banks

We compared three categories of messages: the first being those made by the banks themselves through their press releases and statements made by their executives, the second being those from the financial press and the third being the opinion of consumers published in the blogs. A study highlighted the gap that existed between the institutional communication, often mirrored precisely by the specialised press, and the reactions of customers. Two very different types of vocabulary were evident: one relating to the performance and profitability, the other consisting of words like “vampires, robbers” highlighting a clear gap in the communication strategy of these particular establishments. It is worth noting that these reactions were outside the framework of normal consumer associations, originating from individuals who would not normally be associated with having any type of militant or vociferous tendency – the same contributors being perfectly able to detail the delivery of their new kitchen just having highly criticised their banker!

Following a campaign on the topic “Without aspartame”

At the end of 2006, there was widespread surprise at the advertising campaign of a well known French company in the food sector promoting their diet products as being guaranteed “without aspartame” when in fact this sweetener has been used in the production of diet products for many years. We noted that this strategy was based on a rumour conveyed exclusively via blogs where aspartame is presented as being “carcinogenic”. This

rumour is regularly disputed and contradicted by official organisations; each month new drugs which contain the sweetener come to market with all the necessary approvals from the authorities concerned. The subject was covered by the press where the majority of newspapers took the position of the official organisations. A precise study of the blogosphere, out of the mainstream media field, showed that the rumour was propagated by isolated groups. This identification was made possible by a calculation of correlation between the names of the contributors and the people who were quoted in these contributions. As a result it was easy to highlight that this was an attempt to associate the sweetener with its manufacturer, a key player in the world of the genetically modified organisms (GMO), and thus create the impression that it was “bad food”. These groups, vehemently opposed to GMO, concluded that aspartame carried with it supposed carcinogenic risks and in an ideological mix fed the rumour. Taking into account the marginality of this communication and its lack of long term influence amongst the general public, it became less important for the competitors of this particular company to take a position on the subject. We also note that this marginality and the absence of any real basis for the rumour led the manufacturer in question to very quickly drop their “without aspartame” message.

A development on a large scale

Based on convincing experimental results, these methodology and tools are currently being evaluated by several market studies institutes who are currently building offers intended for advertisers.

In the societal field, within the French Government’s Information service, an “Observatory of the public expression” was created implementing a platform similar to that described in this article. From this point of view, the recent French presidential elections provided a very important field of investigation, being able to measure the importance of blogs as a place of debate and exchange of opinions. The principal protagonists were not mistaken there and the blogs of the two principal candidates were the subject of

a very elaborate communication strategy, completely new in France.

These studies are too recent to draw general conclusions. It's clear, however, that the approach lends itself very well to the analysis of well defined areas of research such as consumer reaction following a product launch. On the other hand, it is more complex to implement it in relation to more general problems which are harder to describe with a defined vocabulary. For example, there are multiple ways of speaking about the "pleasure of food" which represents a key concept for many companies but whose precise definition remains difficult to describe.

CONCLUSION: LISTENING INSTEAD OF ASKING?

Blogs are undeniably a medium that allow tens of millions of individuals to express their views spontaneously. From this point of view the interest of blogs as an Observation platform lies in the long tail, blogs of any real authority were too few in number and mainly reflected agreed opinion. They are not in themselves a source of original information, especially when compared to the press.

Interpretation and analysis of that spontaneity is, for the first time, manageable on a large scale as the medium that supports these exchanges is electronic and therefore allows them and their content to be processed using appropriate software programs. We can consider a new field of research that is not based on a pre-defined set of questions and makes no attempt to pre-judge consumer opinion. On the contrary it allows observation without restriction, the analysis being made subsequently. The difference is significant because this new approach represents an infinitely more conducive way of listening and allows for analysis across a spectrum that was previously not possible.

From this point of view, the increasing interest of organisations and advertisers for analysis of the blogosphere is indicative of a trend to integrate understanding of this area with traditional marketing processes. Much remains to be done, especially in terms of being able to evaluate what credibility to give a blogger whose contributions may in fact be representative of just purely

ideological views. The case study on aspartame clearly shows that the rumour this sweetener was carcinogenic was purposely created by pressure groups and in no way spontaneous. The blogosphere is a reflection of society at a given moment its analysis taken in context as representing trends and events in society. A panel of selected consumers is likely to be more subtle but perhaps give less information about emerging consumer thought and potential changes in purchasing and decision behaviour.

Footnotes

1. From: timbl@info.cern.ch (Tim Berners-Lee).
2. <http://www.technorati.com>
3. <http://clairejapon.canalblog.com>
4. <http://www.ciao.fr>
5. <http://www.produitsdelannee.com>
6. <http://www.fr.pg.com>
7. <http://www.produitsdelannee.com>
8. <http://www.ciao.fr>
9. <http://www.emarketing.fr>
10. Jean-Claude Andréani : *Marketing du produit nouveau*, in *Revue Française du Marketing* n° 182, 2001.
11. <http://www.oreillynet.com>

References

- Beauvieux, Alain* et al. (2006). *Référencement des outils d'Intelligence économique par les usages*, in *Regards sur l'Intelligence économique*, May.
- Beauvieux, Alain* et al. (2007). *L'Intelligence économique coté métiers*, in *Regards sur l'Intelligence économique*, Juin.
- Beauvieux, Alain* et al. *L'entreprise dans l'économie de la connaissance*, Livre blanc du G.F.I.I. – www.gfii.asso.fr
- Laurent, François*. (2006). *Les études marketing – Village Mondial*.
- Laurent, François*. (2005). *La grande mutation des marques high tech – Village Mondial*.
- Laurent, François*. (2007). *Web 2.0 et les études marketing*, in *Marketing Magazine*, Février.

PART 1 / CREATIVE APPROACHES AND NEW METHODS

Laurent, François. <http://www.marketingsdead.net>, various papers.

The Authors

François Laurent, ConsumerInsight, France; Chairman, Adetem (National Association for Marketing), France; and Chairman, Irep (National Association for Advertising), France.

Alain Beauvieux is Chief Executive Officer, AMI Software, France.