Unstructured Data Analysis for Marketing Decisions in Agri-food Sector

Roberto PAIANO, Adriana CAIONE, Anna Lisa GUIDO, Andrea PANDURINO

Salento University, Engineering Innovation Department via Monteroni Lecce, 73100, Italy

roberto.paiano@unisalento.it, adriana.caione@unisalento.it, annalisa.guido@unisalento.it, andrea.pandurino@unisalento.it

and

Monica FAIT, Paola SCORRANO

Salento University, Economic Sciences Department via Monteroni Lecce, 73100, Italy

monica.fait@unisalento.it, paola.scorrano@unisalento.it

ABSTRACT

The increase in business, the intense competition between companies, make the marketing an operation and a management mechanism essential for enterprises.

On the other hand, social networks are expanding rapidly and can help companies in collecting customers' feedback, establishing a brand presence, observing the way their brands are discussed and perceived.

This paper focuses on the advantages of analysing data extracted from unstructured sources for marketing purpose and shows the results obtained in a case study in the agri-food sector.

The innovative architecture proposed is realized through the integration of some existing java based tools and of one or many ontologies, in order to retrieve more frequent concepts from unstructured sources, suggest links of articles and images, detect the language used in the sources, suggest other concepts related to the research and filter the results obtained from the elaboration of the unstructured sources.

Keywords: Knowledge extraction, ontology, unstructured sources, marketing intelligence, agri-food sector.

1. INTRODUCTION

In the actual economic environment, characterized by sudden changes, globalization and intensification of competition in any sector, knowledge of cognitive and behavioural elements, that characterize the purchasing process [1] [2], is a fundamental aspect for companies in order to gain a competitive advantage.

To allow the market entry and get good and consistent profits, it is essential to anticipate and beat the competition. To do so, companies need to adopt innovative and automated techniques to study customers' needs, tastes and interests and to contribute additionally to the results obtained using the consolidated marketing tools.

The existing management systems allow an optimal management of structured information (data readily identifiable, organized into structures and interpretable by software tools); the same can not be said for unstructured information (data without an identifiable structure). Some examples of unstructured data are images, video, documents, web pages and text. Data in web pages, represented in a mark-up language such as HTML, are considered to be unstructured [3]. Documents are readable by humans, but difficult to be processed by traditional software tools. The difficulty depends on the intrinsic ambiguity of natural language. In enterprise and institution environment, it has become primary the use of techniques that can extract information of interest from lots of documents that must be managed daily.

Particular relevance assumes the ability to understand, highlight and extract most relevant unstructured content (Sentiment Analysis) to transfer value to business processes. The exploitation of unstructured information allows the monitoring of the target market (market sensing) and the acquisition of data useful for designing marketing activities (market insight) [4]. A study on the expansion of social network in marketing is done in [5] where are examined advantages and risks of the social network marketing, its potential success for business and the possibility to promote products and services via the social network platform.

In recent years, techniques for Information Extraction (IE) have been increasingly developed. Many are the models of text mining proposed [6] to acquire relevant concepts from unstructured sources [7]. For example in [8] many tools are described and compared. The analysis and tests show the importance that may result from the integration of measurement and functionalities of different tasks in the perspective of providing useful analytic data out of text.

These tools use the Natural Language Processing (NLP), a type of semantic search that, trying to get closer to the mechanism of human learning, returns results containing concepts semantically related to each other. Often these tools and techniques are supported by ontology that is a formal, explicit specification of a shared conceptualization [9]. The introduction of ontologies comes as part of the semantic web, from the need to have a language for domain representation that allows expressing the meaning of the documents in the network. The peculiarity of the ontology is the use of common words and concepts to describe and represent the domain of interest; this makes it understandable and usable by people, applications, databases, etc. in order to share a common knowledge concerning any domain.

Semantics and ontologies maximize the value of the process of information and document management, automate and simplify the processes of analysis and classification of information and documents according to taxonomic rules constantly updated.

The paper [10] shows a framework that is useful for finding potential relationship between consumers and mobile OS market activities using ontology. It helps providers offer the right product to the right consumer, by retrieving important information.

Besides tools/algorithms integration, another aspect of great importance is the multi-language support. It is argued in [11] in order to make the approach of knowledge extraction independent of the language in which the text is written.

This paper makes a combined use of information extraction tools from unstructured sources, mostly from blogs, forums and social networks, and of ontologies that describe the specific domain, in order to extract useful information for companies.

The paper is structured as follows. Section 2 gives a brief description of the existing tools of knowledge extraction used within the proposed architecture that is explained in Section 3. We detail each module that composes the architecture and the integration of tools and ontologies. Section 4 presents, whit a use case in the agri-food section, how the software works. Section 5 exposes the results obtained from the test and the conclusions are given in Section 5.

2. TOOLS OF KNOWLEDGE EXTRACTION

In this section we analyse the existing tools of knowledge extraction used in the architecture.

AlchemyAPI

AlchemyAPI (*http://www.alchemyapi.com*) technology uses natural language processing and machine learning algorithms to extract semantic metadata from a text.

The API endpoints are oriented to perform content analysis of web pages accessible from the Internet, html pages or textual content.

Among the features available there is the possibility of extraction of entities, concepts, text categorization, extraction of relations, language detection, extraction of words, sentiment analysis, text extraction, etc., We analyse those which it is made use within the semantic marketing intelligence software.

Language Detection

AlchemyAPI provides functionality for the recognition of the language of a text, HTML page or web-based content. It identifies more languages than other services of text analysis, with extremely high accuracy rates.

Keyword Extraction

AlchemyAPI is able to extract keywords from a textual, HTML or web-based content. Statistical algorithms and natural language processing technologies are used to analyse data, to extract keywords that can be used to index contents, to generate tag clouds, etc.

This processing is supported in different languages, and also enables the foreign language content to be classified and labelled.

Author Extraction

AlchemyAPI is able to extract publisher information from web pages. If an article of news or of blog specifies an author, it attempts to extract it automatically.

Zemanta API

Zemanta (*http://developer.zemanta.com*) is a content suggestion engine for bloggers and other creators of information.

It analyses the user-generated content (for instance a blog post), using the natural language processing and semantic search technology to suggest images, tags, and links to related articles.

It suggests content from Wikipedia, YouTube, IMDB, Amazon.com, CrunchBase, Flickr, ITIS, Musicbrainz, MyBlogLog, Myspace, NCBI, Rotten Tomatoes, Twitter, and Snooth Wikinvest, as well as Blog of other users Zemanta.

Zemanta is a service that connects well-known databases in a single-point solution to detect other content.

The software uses the extraction feature of articles and images.

Jsoup API

It is a Java library (*http://jsoup.org*) for HTML content processing . Provides an API for extracting and manipulating data. Using DOM, CSS, jQuery-like methods, parses the HTML from a URL, file, or a string, finds and extracts data, uses DOM and CSS selectors, manipulates HTML elements, attributes, and text.

3. KNOWLEDGE EXTRACTION ARCHITECTURE

In this section we describe the architecture that we propose. It can be distributed into three main layers as shown in Fig. 1.

Fig. 1 Architecture layers

Presentation Layer manages the interaction with the user. It also shows the results obtained by the core elaboration modules of the marketing intelligence software, such as concepts represented in tag cloud or tables with links to articles and images related to the research.

Integration and Elaboration Layer integrates the existing java based tools described in previous section in order to extract the most important and frequent concepts in the unstructured sources selected and the occurrence of the keywords of interest. It also suggests other articles and images useful for the research.

All information about the researches done is stored in a database through the interaction with the underlying layer.

Database and Ontology Layer stores data of the users, unstructured sources and keywords and of the researches done by the user and the results obtained.

Exploding the three layers we show the modules that compose the architecture (Fig. 2).

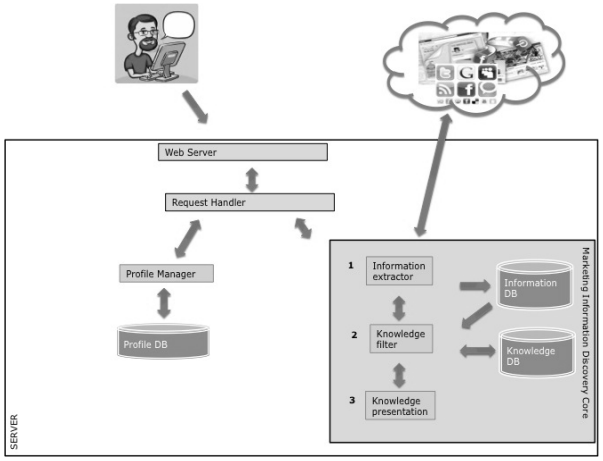

Fig. 2 Reference architecture

The core elements of the marketing intelligence system are:

1. Information Extractor: searches keywords within reference ontology in order to identify other related words. Searches the above words in the sources and calculates the occurrences. Uses the sources and keywords to search for other sources linked to the first;

2. Knowledge Filter: filters the extracted information based on the ontology;

3. Knowledge Presentation: displays this information to the user.

We now describe which modules compose the layer described before and we explain the role of each module that constitutes the architecture introduced.

Presentation Layer

The layer is a bridge between the user and the system. It is composed by:

- *Request Handler:* it receives the requests of the user. It interacts with the profile database in order to retrieve information of the user, or to store data related to him. On the other side, it forwards the requests to the module of marketing information discovery. Both this module of information discovery and the profile manager reply to the Request Handler with information of the authenticated user and of the outputs of the elaboration. The module, then, displays this information in a readable form for the user (for instance a tag cloud of relevant concepts or a table containing articles and images related to the research).

- *Profile Manager:* it is an intermediate level between Request Handler and Profile DB. It makes queries to the database to retrieve user's information and to store other ones.

Integration and Elaboration Layer

The table below summarizes the functionalities of each tool used and integrated in this layer.

API	*Use*
Alchemy	Extraction of key concepts from a textual content; Identification of the language used in the source; Extraction of the authors of the post
Zemanta	Tip of images and articles related to research
Jsoup	Extraction of pure content from web pages, providing as input the url of the source to be analysed; Extraction of links within the pages, useful for navigation.
WordReference	Translation of keywords in the language of the source

These functionalities are integrated in the modules:

- *Information Extractor:* it uses and integrates the existing java based tools, Jsoup, AlchemyAPI and WordReference, with the aim of analysing sources and keywords selected/entered by the user in the system. The inputs can be provided in different languages, in order to carry out researches related to countries and languages different from their own and get useful results. It is possible through the use of AlchemyAPI and WordReference. The module, indeed, detects the language of the sources and keywords, and if there are differences, it translates the latters in the language of the first ones.

- *Knowledge Filter:* it interacts with the Knowledge Database in order to extract other words or concepts related to the research that will be processed by the Information Extractor.

- *Knowledge Presentation:* it prepares the output to be presented to the user in the form of tag cloud or table of articles and images (extracted by Zemanta API), research results, etc.

Given as input one or more unstructured sources, one or more keywords, and the degree of the navigation for each source, the software can generate different outputs: a tag cloud, images and articles related to search.

Below (Fig. 3) we outline the logic and the integration of tools in order to generate a tag cloud containing the input keywords and the obtained concepts related to the domain, with the highest number of occurrences.

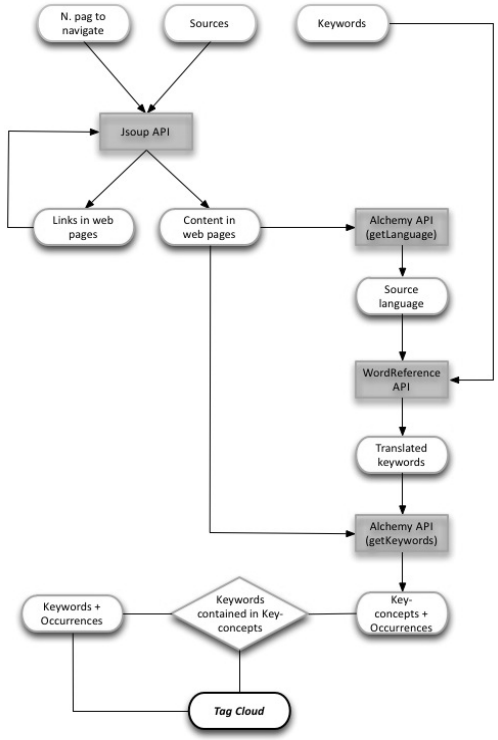

Fig. 3 Logic and integration of tools to generate a tag cloud

Through Jsoup, it is possible to extract pure content from a web page, given the address of the same. Besides extracting the content of the input sources, are extracted links (of number equal to that indicated by the user) contained in the page. Just as the first ones, these sources are processed through Jsoup to extract only the content.

The system checks whether it is possible to combine words entered with other semantically related by asking sector-specific ontology to extract other concepts connected to the first ones.

The content previously extracted from the source is passed to AlchemyAPI that detects the language used. It also identifies the language of the keywords. This information is used with WordReference API to translate keywords in the language of each source.

For each source the system (through AlchemyAPI) proceeds with the extraction of most recurrent concepts. The number of occurrences in the text of each concept is calculated.

In addition the system checks the presence or absence of the keywords selected/entered by the user. If the API has already extracted them, the value of its occurrences is increased; otherwise they are added to the list of words extracted. This list is used to create a tag cloud visible to front end.

The research results can be filtered on the basis of the ontology. Given the keywords, the system extracts the concepts related to them and the words that are not in this set are excluded from the result.

Zemanta API suggests contents such as images and articles (Fig. 4). This suggestion is done on the web in real time.

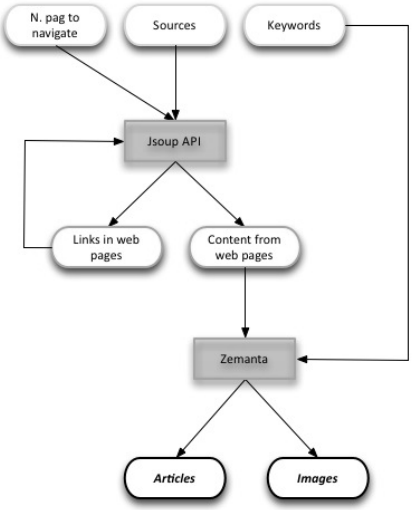

Fig. 4 Logic and integration of tools to obtain articles and images

Database and Ontology Layer

It is the layer of information storage. It includes two type of database:

- *Profile Database:* it stores users' information. Not only personal information, also it keeps track of the relations to transactions carried and to the users' unstructured sources.

- *Information Database:* it stores unstructured sources from which to extract the data and the keywords to search. These sources can be stored in shared manner or linked to the profile of a user. The database also contains an historical of the researches performed by the users and the obtained results.

The layer integrates also one or more ontologies related to the sector of interest. An ontology, for example relative to the wine sector, was found on the web (*www.w3.org/TR/owl-guide/wine.rdf*), and reused. It can be integrated with other information over.

The ontology contains descriptions of hierarchies and categories of foods and wines (Fig. 5), along with restrictions concerning the association of particular instances.

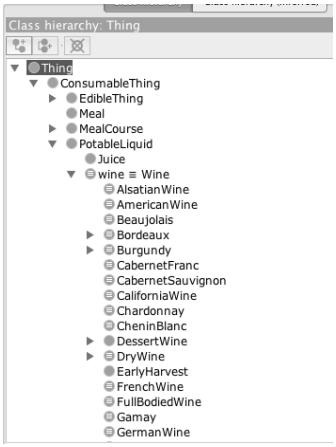

Fig. 5 Wine Ontology

In the proposed architecture, the respective module is:

- *Knowledge Database:* is the ontology of the specific agri-food field. The ontology is used in two direction:

  - Before processing sources and keywords, to suggest other words related to the research;
  - After processing sources and keywords, to filter the output deleting the words not related to the domain.

4. USE CASE IN AGRI-FOOD SECTOR

In this section we are interested in explaining with an example how the software works.

The software testing is carried out in collaboration with the Economic Sciences Department of the University of Salento. The focus fell on the olive sector.

A phase of identification of the input variables, such as sources and keywords, preceded the software test phase. We proceeded by identifying a foreign market more interesting, such as Australia, one of the new areas of consumption, and then we continued with the search of unstructured sources. The latter were classified by type (blogs, forums, producer, etc.) to identify individuals (bloggers, users, distributors, etc.) to extract information. The sources were provided by the Economic Sciences Department in response to a search on google.au preferring blogs, forums, reviews, magazines, web sites of producers of olive oil in Australia and the websites of the major exporters of oil in Australia.

As regards the choice of the keywords, we took those representing the three main categories of oil: *olive oil*, *virgin olive oil*, *extra virgin olive oil*.

The database has been populated with the detected sources and the keywords.

The experiment involved the sources of three major competitors in the Australian market (respectively Australian, Spanish and Italian competitors).

Primarily the sources and keywords relating to oil producers in the Australian market (Fig. 6) have been selected and processed.

http://oliveoilandlemons.com/; http://blog.fab.com/post/54003547378/typuglia-adding-character-to-olive-oil; http://coffeebeings.blogspot.it/2013/04/is-your-extra-virgin-olive-oil-real.html; http://dinersjournal.blogs.nytimes.com/2013/04/18/judging-the-worlds-olive-oils/?_r=0; http://extravirginblog.com/; http://oil-live.com/; http://olivegazette.blogspot.it/; http://oliveoil-olivenoel.blogspot.it/; http://oliveoilchic.blogspot.it/; http://openeuropeblog.blogspot.it/2013/05/good-news-commission-bottles-it-on.html; http://www.amazingoliveoil.com/olive-oil-blog.html; http://www.economonitor.com/dolanecon/2013/02/04/why-olive-oil-good-for-the-body-is-becoming-bad-for-the-pocketbook/; http://www.loveandoliveoil.com/; http://www.oliveoiltimes.com/olive-oil-basics/world/richard-gawel/10605; http://www.oliveoiltimes.com/resources; http://www.pacificsunoliveoil.com/blog/; http://www.pbs.org/food/features/kitchen-careers-love-and-olive-oil-food-blogger/; http://www.tastespotting.com/tag/olive+oil; http://www.thedieline		virgin olive oil; olive oil; extra virgin olive oil

Fig. 6 Sources and keywords related to oil producers in Australia

The output obtained is shown in the image below (Fig. 7). It contains also the keywords selected by the user and the words related to the first ones, extracted by the ontology.

Fig. 7 Tag cloud oil producers in Australia

Similarly, we selected sources and keywords for Italian and Spanish exporters of oil in Australia (Fig. 8 and Fig. 9 respectively).

http://verolio.com/; http://verolio.com/grove-test-page/; http://verolio.com/production/; http://www.basile.com.au/; http://www.basile.com.au/products-and-brands/brands.html; http://www.basile.com.au/products-and-brands/products.html; http://www.basile.com.au/recipes-and-tips/recipes.html; http://www.benfatti.com.au/our-producers.html; http://www.casagusto.com.au/olive-oils; http://www.internationalfinewines.com.au/Our-Producers.html		extra virgin olive oil; olive oil; virgin olive oil

Fig. 8 Sources and keywords Italian exporters of oil in Australia

http://www.olivesandoliveoilfromspain.com.au/cooking-with-olives; http://www.olivesandoliveoilfromspain.com.au/health-benefits-of-olives; http://www.olivesandoliveoilfromspain.com.au/olive-oil-products; http://www.olivesandoliveoilfromspain.com.au/spain-the-leader-in-olive-oil		virgin olive oil; extra virgin olive oil; olive oil

Fig. 9 Sources and keywords Spanish exporters of oil in Australia

The portions of the tag clouds are showed in the images in Fig. 10 for Italian exporters of oil and in Fig. 11 for Spanish exporters of oil.

Fig. 10 Tag cloud Italian exporters of oil in Australia

Fig. 11 Tag cloud Spanish exporters of oil in Australia

5. USE CASE RESULTS

In order to interpret and evaluate the results obtained through the use of the software, we have followed an approach for quantitative analysis software in order to identify words with greater occurrence.

Words can be classified into three main categories:

- Cognitive: words within the sphere of knowledge, perception, sensory attributes [12];

- Context: words that refer to extrinsic features (price, method of production, etc.) [13];

- Experiential: words that indicate the consumption occasions, the expected benefits [14].

Starting from the analysis of the tag clouds, shown in the previous section, it is possible to observe that both producers and exporters of olive oil in Australia, uses the three keywords (olive oil, virgin olive oil, extra virgin olive oil). *Extra virgin* concept is more recurrent in the web sites of Australian producers. Indeed, frequent words are *Extra virgin olive oil, Australian Extra virgin, Label extra virgin.* Italian and Spanish exporters use often the generic concept of *olive oil*.

Focusing on the three main categories, we indicate the words with greater occurrence in web sites of producers and exporters of olive oil in Australia.

Category	Australian Exporters	Spanish Producers	Italian Producers
Cognitive	natural olive oil, olive oil premium, olive oil refined, organic, pure olive oil acid, fresh oil, extra light, light	acid, fluidity, fragrance, fruity, low acidity, aromatic, pure olive oil, light olive oil	acid, fluidity, fragrance, fruity, low acidity, aromatic, pure olive oil, light olive oil delicious,

	olive oil, flavoured extra virgin		amazing, interesting, different
Context	environmentally friendly principles, ecologically sustainable production, good business practice, consistent quality yields, modern groves	olea europea quality, quality control pet bottles	
Experiential	fine Italian food, Italian food Australia, olive oil recipes, cooking effect, food safety, good diet plan		

We can observe that Australian exporters emphasize on aspects related to naturalness, purity and organoleptic qualities of the olive oil. On the other side, both Italian and Spanish producers accentuate the sensory attributes, in particular the flavour by Spanish producers.

Concerning the category of context, both the producers refer to standard element like quality and type of packaging. Australian exporters talk about modern cultivation and production of a technologically advanced.

At last there is a common thread in relation to the words of experiential, that is food and health benefits.

The analysis of the results leads to the affirmation that in addition to oil-exporting countries such as Italy and Spain, although local producers to satisfy the demand of consumption of the product. Therefore, the current operators and those who want to access this market must focus on enhancement and promotion of the production and consider the advent of new producers in Australia, the strategies initiated by non-traditional producers of olive oil and experiential and context elements on which implement marketing policies.

6. CONCLUSION

This paper proposes a software architecture of marketing intelligence that can extract, analyse and systematize unstructured information (current technological frontier) drawing from the web, in order to be able to support enterprises in the agro-food sector in identifying new markets and consumers' needs and trends. It goes beyond what is proposed in [10], integrating, in addition to ontologies specific for a sector, also existing tools of knowledge extraction and the management of the presentation of front end information to the user. The integration is useful because each tool has some strengths and weakness as described in [8] or has different characteristics.

It has provided useful results relative to the sector of Australian olive oil as the reference market, information on the product and on the actual operators in the sector, quantitative and qualitative data to help design marketing actions etc. This confirms the benefits for the business theorized within the paper [5].

Interesting is also the social aspect, a result of the sharing of information sources (at the discretion of the users) and the multi-language support that allows the use of unstructured sources and keywords in different languages.

The architecture can be used in any other scenario, not only in marketing, fundamental are the sources and keywords that are used as input for the software which will extract the information to be presented to the user.

7. REFERENCES

[1] J.P. Peter, J.C. Olson, **Consumer behavior and marketing strategy**, 4° ed., Irwin, New York, 1996

[2] M.R. Solomon, **Consumer behavior, buying, having, and being**, 6° ed., Pearson, Prentice Hall, New Jersey, 2004

[3] Pearson Education India, **Introduction to Database Systems**, cap. 17, 2008

[4] P. Kotler, K. L. Keller, F. Ancarani M. Costabile, **Marketing Management**, Pearson Education, 2012

[5] V. Bolotaeva and T. Cata, "Marketing Opportunities with Social Networks", **Journal of Internet Social Networking and Virtual Communities** [online], 2010

[6] M. Rajman and R. Besançon, "Text Mining - Knowledge extraction from unstructured textual data", **6th Conference of International Federation of Classification Societies**, 1998

[7] T. Wu, "Theory and Applications in Information Extraction from unstructured text", **Theses and Dissertations**, Paper 741, 2002

[8] A. Gangemi, "A Comparison of Knowledge Extraction Tools for the Semantic Web", **Proceedings of European Semantic Web Symposium (ESWC)**, Lecture Notes in Computer Science (LNCS), Springer, 26-30 May 2013, Montpellier, France, 2013

[9] R. Studer, R. Benjamins and D. Fensel, "Knowledge engineering: Principles and methods", **Data & Knowledge Engineering**, 1998

[10] Y. Nergis and A. Gülfem, "An Ontology-Based Data Mining Approach for Strategic Marketing Decisions", **The International Symposium on Management, Engineering and Informatics,** 2013

[11] D. Gerber and A.-C. N. Ngomo, "Extracting Multilingual Natural-Language Patterns for RDF Predicates", **Knowledge Engineering and Knowledge Management (EKAW)**, Lecture Notes in Computer Science, Springer, 8-12 October 2012, Galway City, Ireland, 2012

[12] J.-B. E. M. Steenkamp, "Conceptual Model of the quality Perception Process", **Journal of Business Research**, vol. 21, 1990

[13] G. Schamel, K. Anderson K., "Wine quality and varietal, regional and winery reputations: Hedonic prices for Australia and New Zealand", **The Economic Record**, vol. 79, n. 246, 2003

[14] A. Tragear, S. Kuznesof, A. Moxey, "Policy Initiatives for Regional Foods: Some Insights from Consumer Research", **Food Policy**, vol. 23, n. 5, 1998