RESEARCH ARTICLE OPEN ACCESS

Use of Artificial Intelligence And Web Scraping Methods To Retrieve Information From The World Wide Web

Marco Scarnò
(Cineca, SuperComputing Application and Innovation group), Master degree in Statistics and Economical Science,
Yakob Seid (FAO), Master degree in mathematical statistics, This work was performed in Rome, Italy*
Corresponding author: Marco Scarnò

ABSTRACT
The World Wide Web could represent a valid substitute for the traditional way of acquiring information. However, using this platform implies new challenges that are derived from the intrinsic nature of the information attained, which is expansive, sparse and mostly unstructured. We investigated the possibility of structuring data from different websites through web scraping techniques. Moreover, we exploited what is offered by some web search engines to progressively create queries that enabled us to select the most useful information. To s uch purpose, we used a strategy that simulated the human behaviour obtaining timely price statistics to be included in the Agricultural Market Information System developed by FAO.
Disclaimer: The designations employe d and the presentation of material in this paper do not imply the expression of any opinion whatsoever on the part of the Food and Agricultural Organization of the United Nations (FAO) concerning the legal or development status of any country, city or area of its authorities, or concerning the delimitation of its frontiers or boundaries. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO in preference to others of a similar nature that are not mentioned. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of FAO

I. INTRODUCTION

The Agricultural Market Information System (AMIS)[1] is an interagency platform to enhance food market transparency and policy response for food security.Launched in 2011 by the G-20 Ministers of Agriculture following the global food price hikes in 2007-2008 and 2010, AMIS brings together the principal trading countries of agricultural commodities. It assesses global food supplies, focusing on wheat, maize, rice and soybeans and provides a platform to coordinate policy action during times of market uncertainty. The platform aims to address the gaps and shortcomings of existing informative systems by using innovative methods and new technologies. In pursuit of this objective, the application of web scraping techniques was seen a possible way to get timely price information. The emergence of Big Data has made it necessary to consider a new paradigm of collecting data. In addition, is apparent that a new analytical strategy isneeded in order to identify the most relevant sources offered by the World Wide Web.Some institutions, like the Italian National Statistical Institute[2], are developing or implementing a system to retrieve timely price statistics through the web. However, the experiences observed from these institutions often require a clear and precise structure of the source of information, namely identified web sites in which the data are located in specific positions. On the contrary, we attempted to identify more generic sources of information through the web and to incorporate some additional values like statistical analysis, comments and evaluations of a phenomenon that might contribute to observed price variations.

II. MATERIALS AND METHODS

Web scraping is a computer software technique for extracting information from the Internet. It is mainly used to transform the information collected from unstructured sources to structured data (typically found in documents written in HTML language). It is implemented by automatic procedures that permit access to the content of one or more websites; this process simulates human behaviour when browsing for a specific purpose.
Despite the proficiency of the technique, an automatic approach cannot ensure correct recognition of the relevant information when dealing with a number of diverse websites characterized by a specific, often unique, structure. The results, however, improve when a special schema is applied on the web pages to note the elements; this is typically the case when the websites are aggregators of information (like the major search engines).When the structure of a website is not known a priori, the

information can be gathered by post-processing the scraped texts; this can be done by applying text mining and machine learning methods[3].

The first step of this work entailed identifying three web sites with fixed structures from which the prices of four key crops (maize, rice, soybean and wheat) can be downloaded. The following sites were selected:

1. https://www.numbeo.com/cost-of-living/: the largest database of users' contributed data about cities and countries worldwide. It provides current and timely information on world living conditions, including costs of commodities, housing indicators, health care, traffic, crime and pollution;
2. http://www.indexmundi.com/: a data portal that collects facts and statistics from multiple sources; and
3. http://www.fao.org/giews/food-prices/tool/public/#/home: a web-based tool for the analysis and dissemination of domestic prices of basic foods mainly in developing countries.

Because of the static structures of these websites, an application to download the underlying information on the websites was developed; the information was then organized in statistical data sets that could be automatically analysed or just stored, allowing the user access to a database of prices.The next challenge involved dealing with the underlying dynamics of the web; new, more relevant websites can be created and the internal structures of those selected can be changed. Moreover, the websites selected may not necessarily be the best suited for this study because other application platforms on the Internet may contain more useful information. It has to be noted that the information may not just be available on HTML pages, but could be found on such platforms as Excel files, PDF files and PowerPoint presentations

Typically, the "human" approach for conducting a proper search involves using the possibilities offered by the most common web search engines. Unfortunately, this is not a suitable method because of the difficulty in identifying the proper terms to be used in the searches and the need to thoroughly verify all the possible results. To overcome these issues, an "intelligent metasearch engine"[4] was built, which learns the users' preferences. Specifically, the system retrieves the information from the World Wide Web and, through a simple interaction with the user, is able to properly rank the results and derive the terms that would make the most efficient queries.A metasearch engine is a tool that can execute the same query on different search engine. Moreover, it formats and ranks the gathered data before presenting the results to the users. For this report, the search engines used were Google, Bing and DuckDuckGo[5].As discussed, the two main issues in this study are the content of the queries to be submitted and further ranking of the results. Specifically, should the term "price" be included in the search about the commodities of interest or would some other words be more appropriate? This is a common problem when a user refines the search in a series of consecutive steps by adding or deleting terms in the query based on the results of the previous queries.

To simulate such behaviour, a mechanism was implemented capable of identifying the discriminant terms characterizing the results that were found of interest by the user; some of those terms were used to perform new queries and others were used to rank the results.

This is an iterative approach in which the user plays the role of the trainer. Making the selections involves a learning process with the final result be carried out by the intelligent metasearch engine.

1. The steps characterizing the system are the following:
2. A series of initial queries are executed on some web search engines;
3. The retrieved URLs (usually several hundred) are ranked by considering specific indexes that verify the presence of some terms in their content;
4. The user selects the URLs that are most representative of his or her interests;
5. Two "virtual" documents are created; the first with all the terms deriving from the selected web sites, the other with the terms from the not selected ones;
6. A statistical method is used to identify the terms that are more discriminant for the selected URLs and believed relevant;
7. The most significant terms (m), are selected and added to the original queries, while $m+k$ are added to list of terms that contribute to the ranking of the results;
8. The new queries are executed and the process is repeated starting from step 2.

We verified that, starting from the second iteration of the full process, the first ranked results are able to discover useful sources on the Internet, well representing the interest of the users.

Moreover, the structure of the query and all the learned terms can be saved, making it possible to execute the same search in another time. In this case, however, the results may be different, but they will be enhanced by the results of the previous learning steps.Concerning the queries executed in the first step, a series of standard queries was considered and submitted to Google, Bing, and DuckDuckGo. In particular we referred to: "world food price AND cereals AND wheat AND maize AND rice AND

soybean"; or "world food price AND (cereals)OR(wheat)OR(maize)OR(rice)OR(soybean) ".Each web search engine provided a number of URLs (HTML pages or other types of files); the results were then accessed and their content was classified as a textual document (i.e. as a series of terms). Symbols and strange characters were then deleted from the text and a specific algorithm, based on a part of speech analysis, was applied to transform each term in its lemma (i.e. in its canonical form, dictionary form or citation form).The ranking of each URL was then evaluated by considering the relative occurrences of the terms used in the queries (in their lemmas form), as they were found in each treated URL. The final results were then presented to the user; to start the learning process, the user must select the URLs believed to be the most relevant. This action led to the establishment of two virtual documents, one contains the lemmatized terms referred to on the web sites of interest and the other consists of the web sites that were not deemed to be of interest. From the analysis of these text a "term-document" matrix can be produced (as in Table 1), containing the frequencies of each lemma as in each of the two documents.

Lemma	**Document of interest (frequencies of term)**	**Document not of interest (frequencies of term)**
Area	1	961
Assistance	1	16
Average	2	288
Barley	2	235
Base	1	68
Butter	1	43
Cereal	3	1095
Change	2	901
…	…	…

Table 1: example of the term-document matrix

It has to be noted that the different lemmas characterizing the matrix are usually equal to tens of thousands; moreover, each of the document can be characterized by different totals of terms. Both these factors have to be properly considered when introducing a methodology able to identify the discriminant terms.

To solve these issues, the Correspondence Analysis (CA) technique[6] is applied.

2.1 The identification of the relevant terms by means of the Correspondence Analysis

The CA allows for the projection of both the rows and the columns of a frequencies table in a subspace (in this case dimension 1) in which it is possible to identify the correspondences between the projected objects. The technique introduces a metrics that takes into account the different totals of the rows and of the columns, namely the different frequencies of each term and their total identified in each virtual document.

In this case, two qualitative variables are considered; the first variable is the "generic lemma", which assumes values from 1to *n*, while the second variable represent the two options: in a web site of interest (1) or not (2). This can be represented in a symbolic matrix:

$$\begin{bmatrix} x_{11} & x_{12} \\ \dots & \dots \\ x_{i1} & x_{i2} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{bmatrix} \quad (1)$$

In (1) the generic quantity x_{i1} represents the number of times the lemma *i*was found in the document of interest. Let:

- $x_{oj} = \sum_{i=1}^{n} x_{ij}$, with *j=1,2*;
- $x_{i0} = \sum_{j=1}^{2} x_{ij}$, with *i=1,…,n*;
- $x_{oo} = \sum_{j=1}^{2} x_{oj} = \sum_{i=1}^{n} x_{i0} = \sum_{j=1}^{2} \sum_{i=1}^{n} x_{ij}$

A more typical representation of (1), when considering the CA, derives from the possibility to consider the ratio between the real and the expected (in case of independence) frequencies. In particular:

$$\begin{bmatrix} \frac{x_{11}}{\sqrt{x_{10}x_{01}}} & \frac{x_{12}}{\sqrt{x_{10}x_{02}}} \\ \dots & \dots \\ \frac{x_{i1}}{\sqrt{x_{i0}x_{01}}} & \frac{x_{i2}}{\sqrt{x_{i0}x_{02}}} \\ \dots & \dots \\ \frac{x_{n1}}{\sqrt{x_{n0}x_{01}}} & \frac{x_{n2}}{\sqrt{x_{n0}x_{01}}} \end{bmatrix} = \boldsymbol{F} \quad (2)$$

Now the following quantities are considered:

- $C_1 = \sum_{i=1}^{n} \left(\frac{x_{i1}}{\sqrt{x_{i0}x_{01}}} \right)^2$
- $C_2 = \sum_{i=1}^{n} \left(\frac{x_{i2}}{\sqrt{x_{i0}x_{02}}} \right)^2$

In particular C_1 and C_2 are, respectively, the sum of the squares of the values in the first and in the second columns of the matrix $\boldsymbol{F}$.

CA is based on the eigen-decomposition of the matrix $\boldsymbol{D} = \boldsymbol{F}^T\boldsymbol{F}$; in particular: $\boldsymbol{D\lambda} = \boldsymbol{\lambda V}$, where $\boldsymbol{\lambda}$represents the vectors of eigenvalues and $\boldsymbol{V}$ the matrix of eigenvectors.

Later in this paper all the related formulas will be deeply showed; in this context, it can be considered that the eigenvalues of $\boldsymbol{D}$ are*1*and ϕ^2, with the relation: $\phi^2 = 1 - C_1 + C_2$.
It has to be noted that ϕ^2 is the Phi-square index, i.e. a measure of association for two binary variables; it was introduced by K. Pearson[7] and assumes values in the interval [0:1], where *0=no association*, *1=maximum dependency*.
Without considering the trivial eigenvalue *1*, the coefficients of the eigenvector associated to ϕ^2 are:

$$v_1 = \sqrt{\frac{1 - C_1}{1 - \phi^2}}$$

$$v_2 = -\sqrt{\frac{C_1 - \phi^2}{1 - \phi^2}}$$

By means of the coefficients of the eigenvector, it is possible to evaluate the projections of the two types of virtual documents (W_1, W_2) and of the *n* lemmas. This allows for the introduction of a condition that, if satisfied, helps in identifying the lemmas that can be associated with the document of interest (or to the document that is not of interest).

$$\frac{x_{i1}}{x_{i2}}\sqrt{\frac{x_{02}}{x_{01}}} > \sqrt{\frac{C_1 - \phi^2}{1 - C_1}} \quad (3)$$

Moreover, the discriminant terms can be ranked accordingly to their representativeness by considering the value of the expression:

$$I_i^r = \left(v_1 \frac{x_{i1}^2}{x_{i0}x_{01}} + v_2 \frac{x_{i2}^2}{x_{i0}x_{02}}\right) \quad (4)$$

As a result of (3) and (4), a list of terms can be selected; between these, the most significant *m* can be considered and added in the queries, while *m+k* can be introduced in the evaluation of the ranking of the results. It has to be noted that, in this case, we considered *m*=3 and *k*=7. This led us to extract *10* terms in each iteration of the learning process. Obviously, the terms that were associated with the not significant results were examined; this was done to eventually delete some of the terms that were significant in a previous step but not in the current one.Finally, it could be noted the ranking of resulting URLs obtained after each learning step can take into account the presence in their texts of the selected and collected terms; for each of these, specific weights can be introduced, depending on, for example, the level of representativeness, or the actual iteration, etc.

2.2 A real example

To implement all the strategy we used the software ADaMSoft, an Open Source general-purpose software written in Java for data management, data analysis, ETL, etc., that integrates, between others, methods and libraries to parse HTML pages or to interpret the results of queries submitted to a web search engine[8].
The website https://www.numbeo.com displays, in its pages, the prices of several commodities for many countries. It is characterized by a static structure that, once selected a country of interest, permits to retrieve an HTML file in which the information are clearly identified by specific tags (i.e. specific elements that characterize the HTML language). For instance, it is possible to consider the results obtained for Italy, as in Figure 1.

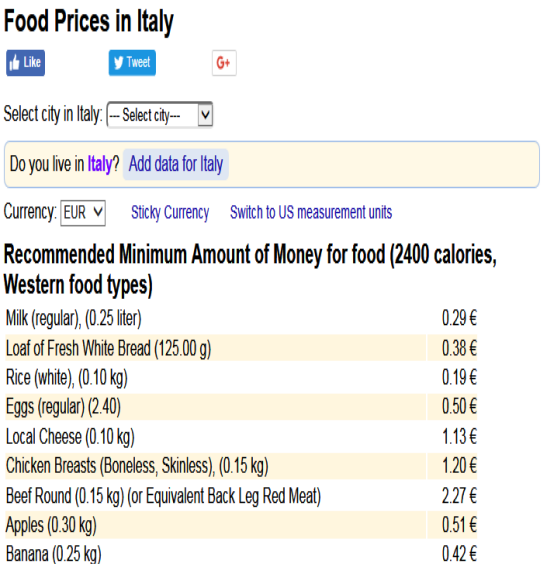


Fig. 1

The application that we developed permits to the user the selection of a Country (Figure 2) and, then, displays all the prices in a sheet (Figure 3)that can, eventually, be saved in an Excel file.

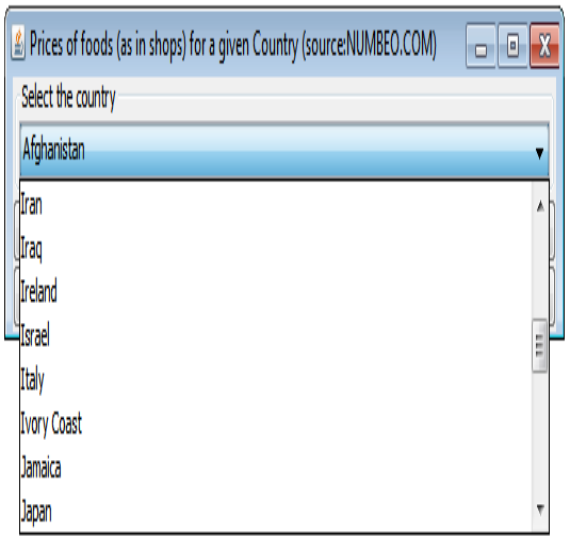


Fig. 2

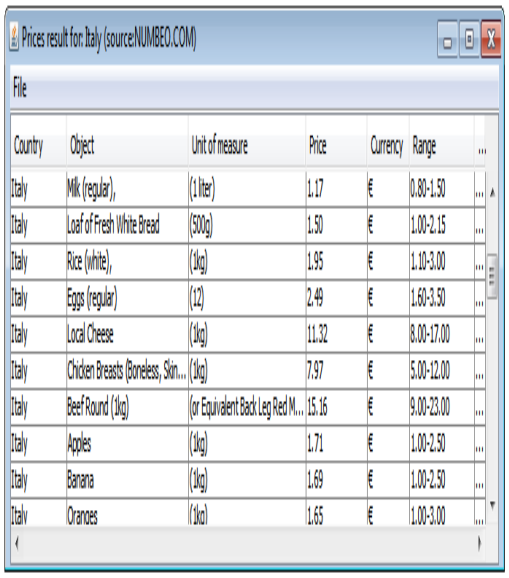


Fig. 3

Concerning the results obtained with the metasearch engine, it is possible to consider that the execution of the initial queries resulted in 183 different URLs, each characterized by different types of files, such as PDF, HTML pages and Doc. A quick analysis of their content led to the selection of 5 results as those of interest (Figure 4).

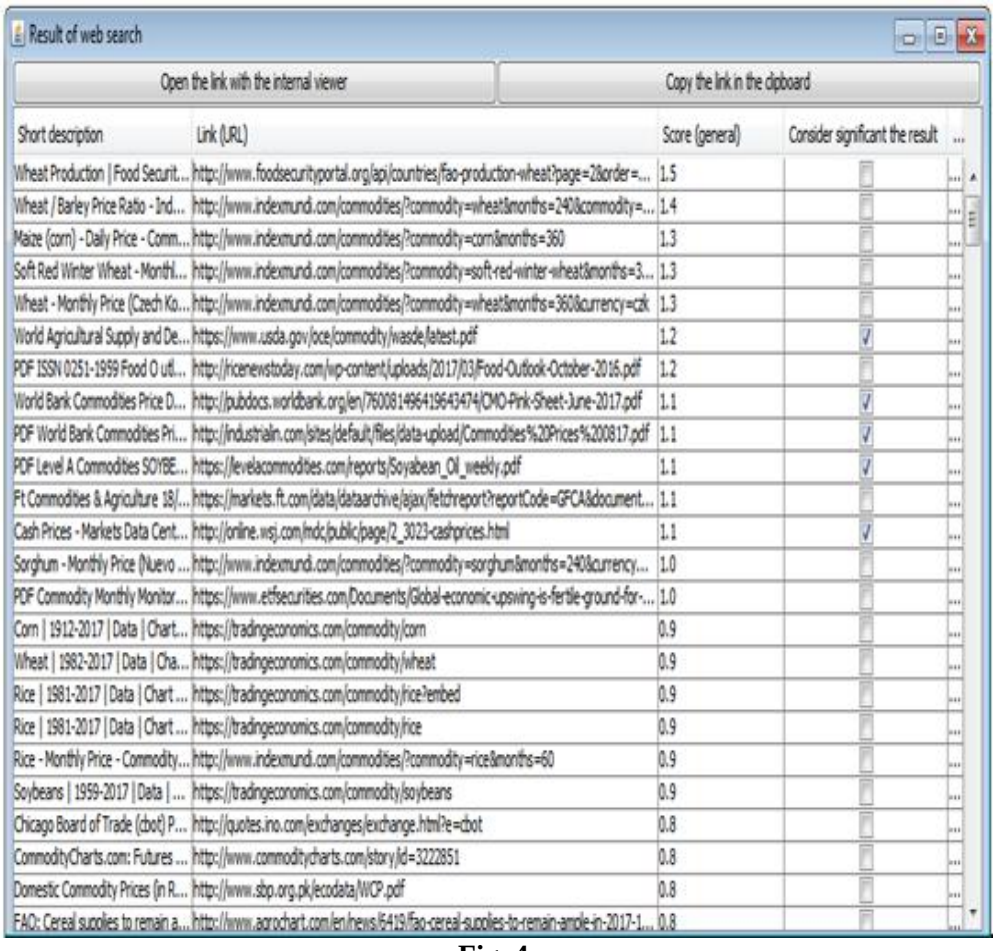
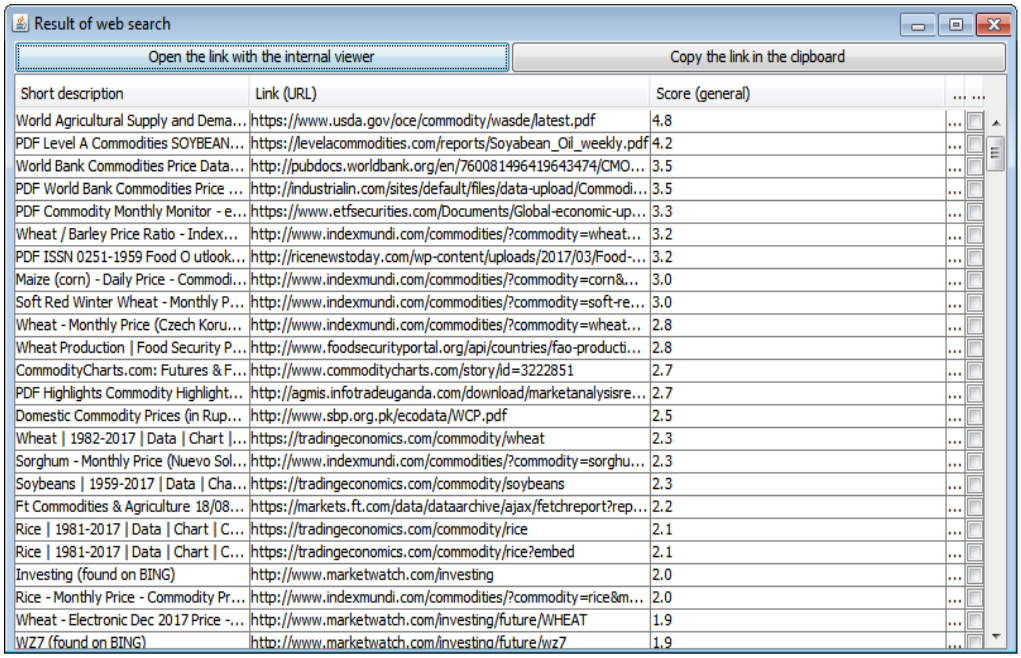


Fig. 4

From the learning process, the terms identified as being more discriminant were as follows: “July”, “intra”, “advice”, “avg” and “qtl”. It has to be noted that the term “July” refer to the month before the one in which the example query was executed (i.e. “August”). The next execution of the queries, enriched with the learned terms and based on the new ranking system, gave the results as shown in Figure 5.


Result of web search

Open the link with the internal viewer | Copy the link in the clipboard

Short description	Link (URL)	Score (general)
World Agricultural Supply and Dema...	https://www.usda.gov/oce/commodity/wasde/latest.pdf	4.8
PDF Level A Commodities SOYBEAN...	https://levelacommodities.com/reports/Soyabean_Oil_weekly.pdf	4.2
World Bank Commodities Price Data...	http://pubdocs.worldbank.org/en/760814964196436474/CMO...	3.5
PDF World Bank Commodities Price ...	http://industrialin.com/sites/default/files/data-upload/Commodi...	3.5
PDF Commodity Monthly Monitor - e...	https://www.etfsecurities.com/Documents/Global-economic-up...	3.3
Wheat / Barley Price Ratio - Index...	http://www.indexmundi.com/commodities/?commodity=wheat...	3.2
PDF ISSN 0251-1959 Food O utlook...	http://ricenewstoday.com/wp-content/uploads/2017/03/Food-...	3.2
Maize (corn) - Daily Price - Commodi...	http://www.indexmundi.com/commodities/?commodity=corn&...	3.0
Soft Red Winter Wheat - Monthly P...	http://www.indexmundi.com/commodities/?commodity=soft-re...	3.0
Wheat - Monthly Price (Czech Koru...	http://www.indexmundi.com/commodities/?commodity=wheat...	2.8
Wheat Production | Food Security P...	http://www.foodsecurityportal.org/api/countries/fao-producti...	2.8
CommodityCharts.com: Futures & F...	http://www.commoditycharts.com/story/id=3222851	2.7
PDF Highlights Commodity Highlight...	http://agmis.infotradeuganda.com/download/marketanalysisre...	2.7
Domestic Commodity Prices (in Rup...	http://www.sbp.org.pk/ecodata/WCP.pdf	2.5
Wheat | 1982-2017 | Data | Chart |...	https://tradingeconomics.com/commodity/wheat	2.3
Sorghum - Monthly Price (Nuevo Sol...	http://www.indexmundi.com/commodities/?commodity=sorghu...	2.3
Soybeans | 1959-2017 | Data | Cha...	https://tradingeconomics.com/commodity/soybeans	2.3
Ft Commodities & Agriculture 18/08...	https://markets.ft.com/data/dataarchive/ajax/fetchreport?rep...	2.2
Rice | 1981-2017 | Data | Chart | C...	https://tradingeconomics.com/commodity/rice	2.1
Rice | 1981-2017 | Data | Chart | C...	https://tradingeconomics.com/commodity/rice?embed	2.1
Investing (found on BING)	http://www.marketwatch.com/investing	2.0
Rice - Monthly Price - Commodity Pr...	http://www.indexmundi.com/commodities/?commodity=rice&m...	2.0
Wheat - Electronic Dec 2017 Price -...	http://www.marketwatch.com/investing/future/WHEAT	1.9
WZ7 (found on BING)	http://www.marketwatch.com/investing/future/wz7	1.9

Fig.5

It is possible to verify that the resulting ranking was modified; for instance, the first URL identified contains specifically the PDF in which are prices of the commodities of interest. This web site was 6th in the first query.

2.3 Conclusion and future development

This system was shown to be effective for the users in identifying the information, sparse on the web, and monitor the prices of specific commodities.The iterative learning of the words constitutes a way by which hundreds of results can be ranked according to the user's personal interest; this can be different among the users.

This approach requires a "manual" intervention, which will become progressively less necessary with training and more use of the system.Moreover, the application of the Correspondence Analysis (adapted for the two dimensional problem) results an efficient way to review many terms. This made it possible to identify the discriminant terms in a reasonable time and to carry out the programme on a normal personal computer.

Obviously some aspects still need to be investigated, specifically those related to the use of learnt terms in the queries (should they be simply added together or combined one by one with the initial queries?), and the acceptable amount of these learnt terms to be considered as significant, according to the index of representativeness.However, we believe that the "intelligent" metasearch engine can also be adapted for more general cases, i.e. when the user wants to apply a method that is more effective in retrieving information from the World Wide Web.

2.4 Appendix: Mathematical adaptation of the CA in the *nx2* case

Given the matrix as in (2), Correspondence Analysis considers the eigen-decomposition of the matrix $\boldsymbol{D} = \boldsymbol{F}^T\boldsymbol{F}$ that, in our case, has dimension 2x2. In particular:

$$\boldsymbol{D} = \boldsymbol{F}^T\boldsymbol{F} = \begin{bmatrix} C_1 & K \\ K & C_2 \end{bmatrix}$$

It is well known that the first eigenvalue of the matrix used in the CA is equal to *1* (this is because the matrix is centred); moreover, the sum of the other eigenvalues correspond to the Phi-square index of the contingencies table represented by matrix itself. In a 2x2 matrix, it is possible to consider that:

$$\lambda_1 + \lambda_2 = trace[\boldsymbol{D}] = C_1 + C_2 = 1 + \phi^2$$

Given the propriety that the determinant of a matrix is equal to the product of its eigenvalues, it is possible to derive that:

$C_1C_2 - K^2 = \phi^2$, so that: $K = \sqrt{C_1C_2 - \phi^2}$

To find the eigenvectors $\boldsymbol{V}$ it is possible to consider the system of equalities:

$$\begin{cases} v_1(C_1 - \phi^2) + v_2\sqrt{C_1C_2 - \phi^2} = 0 \\ v_1\sqrt{C_1C_2 - \phi^2} + v_2(C_2 - \phi^2) = 0 \end{cases} \quad (5)$$

From (5) it is possible to derive:

$$v_2 = -v_1\frac{C_1 - \phi^2}{\sqrt{C_1C_2 - \phi^2}} \quad (6)$$

Avoiding to substitute (6) in (5), because this will lead to the trivial solution $v_2 = v_1 = 0$ (the eigenvector defines a line passing for [0:0]), we can select the eigenvector having a norm 1; this implies:

$$1 = \sqrt{v_1^2 + v_2^2} = \sqrt{v_1^2 + v_1^2\frac{(C_1 - \phi^2)^2}{C_1C_2 - \phi^2}} \quad (7)$$

From (7) it follows that:

$$v_1 = \sqrt{\frac{C_1C_2 - \phi^2}{(C_1 - \phi^2)^2 + C_1C_2 - \phi^2}}$$

$$v_2 = -\frac{C_1 - \phi^2}{\sqrt{C_1C_2 - \phi^2}}\sqrt{\frac{C_1C_2 - \phi^2}{(C_1 - \phi^2)^2 + C_1C_2 - \phi^2}} = -\frac{C_1 - \phi^2}{\sqrt{(C_1 - \phi^2)^2 + C_1C_2 - \phi^2}}$$

By referring to the equality:$C_1 + C_2 = 1 + \phi^2$ it will be possible to simplify the above relations:

$$v_1 = \sqrt{\frac{1 - C_1}{1 - \phi^2}}$$

$$v_2 = -\sqrt{\frac{C_1 - \phi^2}{1 - \phi^2}}$$

The coefficients of the eigenvector permits to evaluate the projections of the two types of web sites (W_1, W_2) and of each of the *n* lemmas. In particular:

$$Proj(W_1) = v_1\sqrt{\phi^2\frac{x_{00}}{x_{01}}}$$

$$Proj(W_2) = v_2\sqrt{\phi^2\frac{x_{00}}{x_{02}}}$$

$$Proj(Term_i) = v_1\frac{x_{i1}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{01}}} + v_2\frac{x_{i2}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{02}}}$$

Given those projections, it is possible to determine which term most closely characterizes the web sites of interest. To this purpose, the scalar product (i.e. the cosine of the angle) between the projections of the *i*-th generic term and of each web site can be considered:

$$\cos(\theta_{i1}) = \frac{Proj(W_1)Proj(Term_i)}{\|Proj(W_1)\|\|Proj(Term_i)\|}$$

$$\cos(\theta_{i2}) = \frac{Proj(W_2)Proj(Term_i)}{\|Proj(W_2)\|\|Proj(Term_i)\|}$$

If $\cos(\theta_{i1}) > \cos(\theta_{i2})$ then the *i*-th term characterizes the web sites of interest. It can be considered that the scalar product of a one-dimensional vector can assume just three values: *1*, *0*, *-1*. In details:

$$\cos(ab) = \frac{ab}{\|a\|\|b\|} = \frac{ab}{\sqrt{a^2}\sqrt{b^2}}$$

A value of *-1* occurs when one value between *a* and *b* is negative (but not both). In our case, the *i*–th term will be associated to the web sites of interest when:

$$\cos(\theta_{i1}) = \frac{Proj(W_1)Proj(Term_i)}{\|Proj(W_1)\|\|Proj(Term_i)\|} > 0$$

This implies:

$$v_1\frac{x_{i1}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{01}}} + v_2\frac{x_{i2}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{02}}} > 0 \Rightarrow$$

$$\Rightarrow \sqrt{\frac{1 - C_1}{1 - \phi^2}}\frac{x_{i1}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{01}}} - \sqrt{\frac{C_1 - \phi^2}{1 - \phi^2}}\frac{x_{i2}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{02}}} > 0 \Rightarrow$$

$$\Rightarrow \sqrt{\frac{1 - C_1}{1 - \phi^2}}\frac{x_{i1}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{01}}} > \sqrt{\frac{C_1 - \phi^2}{1 - \phi^2}}\frac{x_{i2}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{02}}} \Rightarrow$$

$$\Rightarrow \frac{x_{i1}}{x_{i0}}\sqrt{\frac{x_{00}}{x_{01}}\frac{x_{i0}}{x_{i2}}\frac{x_{02}}{x_{00}}} > \sqrt{\frac{C_1 - \phi^2}{1 - \phi^2}}\sqrt{\frac{1 - \phi^2}{1 - C_1}} \Rightarrow$$

$$\Rightarrow \frac{x_{i1}}{x_{i2}}\sqrt{\frac{x_{02}}{x_{01}}} > \sqrt{\frac{C_1 - \phi^2}{1 - C_1}} \quad (8)$$

The condition (8) gives a simple rule that, if satisfied, terms that characterize the web sites of interestcan be identified. Moreover, not all the terms that satisfy the (8) can be considered significant; an interesting index that may help in solving this issue is the level of representativeness of the projection of each term in the first, and unique, axe; it is evaluated by considering:

$$I_i^r = \frac{x_{i0}}{x_{00}}\frac{\left(Proj(Term_i)\right)^2}{\phi^2} = \frac{1}{\phi^2}\left(v_1\frac{x_{i1}^2}{x_{i0}x_{01}} + v_2\frac{x_{i2}^2}{x_{i0}x_{02}}\right)$$

This index makes it possible to rank all the terms identified as discriminant for the web sites of interest; it will be then possible to consider just the first *k* characterized by the higher values.

III. REFERENCES

[1]. More details are available at www.amis-outlook.org/amis-about/en/

[2]. Polidoro F., Giannini R., Lo Conte R., Mosca S., Rossetti F Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. Statistical Journal of the IAOS 31 (2015) 165–176

[3]. Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarnò M., & Summa, Internet as data Source in the Istat survey on ICT in enterprises. Austrian Journal of Statistics, 2015, 44 (2)2, 31-43

[4]. Berger, S, Great Age Guide to the Internet, 2015, Indianapolis, IN,USA, Que Publishing

[5]. More details available at: https://duckduckgo.com/
[6]. Benzécri, J.P., L'Analyse des Données. Volume II, L'Analyse des Correspondances. 1973, Paris, France, Dunod
[7]. Clifford, W.K. , The Common Sense of the Exact Sciences, with prefaces by Karl Pearson and Bertrand Russell,1946, New York, Kopf
[8]. For further information refer to http://adamsoft.sourceforge.net

Marco Scarnò. "Use of Artificial Intelligence And Web Scraping Methods To Retrieve Information From The World Wide Web." International Journal of Engineering Research and Applications (IJERA) , vol. 08, no. 01, 2018, pp. 18–25.