Web Information Retrieval System for Technological Forecasting

Raúl Montiel[1], Luis Lezcano Airaldi[1], Fabián Favret[2], Karina Eckert[2]

[1]*Universidad Tecnológica Nacional, Facultad Regional Resistencia, Resistencia, Chaco, H3500CHJ, Argentina.*
luislezcano@frre.utn.edu.ar; raulmontiel@gmail.com
[2]*Universidad Gastón Dachary. Posadas, Misiones, 3300, Argentina.*
{fabianfavret, karinaeck}@gmail.com

Abstract

Technological Forecasting and Competitive Intelligence are two different disciplines that, used together, provide the organizations with an invaluable analytic tool for the environment and the competing companies' behavior. This kind of technology can be used for extracting useful information to make strategic decisions. This paper describes a Web mining system which gathers the users' information requirements through a series of guided questions, constructs various search keys with the answers and uses them to perform a continuous search and analysis process by means of several web search engines and different information retrieval algorithms to score the relevance of the documents obtained. These documents are later presented to the user as an aid in the decision making process. After the description, the system was tested in several scenarios and the obtained results are shown and discussed.

Keywords: Web mining algorithms, technological forecasting, competitive intelligence, information retrieval.

Received 29 December 2016 / Revised 20 March 2017 / Accepted 04 April 2017

1. Introduction

The global growth of an organization's environment, the new technologies and the large amount of information available to make a decision, have risen the need for experts in handling this data to evaluate this changing environment, the competitors, the market evolution and the technology related to the organization. The innovation process that many organizations have today represents an opportunity for strategic change to deal with these factors.

The Decision Making Process (DMP) requires a detailed analysis of a considerable amount of information, which is often not available in its entirety at the right moment. The current challenge that organizations face is finding new forms of competition, putting them to test in critical grade to position themselves in the market, articulating with other environment's actors, promoting the continuous control and exercising the capability to monitor and forecast the environment variables.

To innovate in the market, a steady flow of internal and external knowledge is needed. Nowadays there are tools and methods that allow organizations to face the innovation challenge. These are known as Technological Forecasting and Competitive Intelligence (TF and CI) techniques [1–3]. These processes aim to constantly track and alert about technological advances and competitors' behavior in order to be able to adjust the strategies of the organization. This kind of information requires adequate applications that help the decision maker. These applications are known as Decision Support Systems (DSS) and are used in the DMP. A good reference about DSS can be found in [4–6].

When using the Internet as a data source, large amounts of unstructured data need to be analyzed to extract useful information for the organization. Web Mining (WM) [7,8] is the set of methods and algorithms for extraction and discovery of information, patterns and relationships in web resources.

Unlike other search systems, where there is a question-answer interaction with the user, the system presented in this article performs a continuous search of information about a specific topic, that is, it keeps exploring, analyzing and presenting new and better results to the user during a determined period of time. In order to achieve this, it has two defined parts: (1) definition and compilation of user's information requirements and (2) continuous search for resources on the web using web mining techniques. Both processes work coordinately to achieve consistency in the results. The idea behind the requirements task is to guide the user in defining the subject of the search and to generate various search keys that will be taken as a starting point for the web mining process. The second task aims to find web resources and send them back to the user for evaluation in order to help them in the DMP.

This article is structured as follows: Section 2

presents an introduction to Technological Forecasting, Competitive Intelligence and Web Mining. Section 3 describes the proposed system. Section 4 shows the tests performed and the results obtained. Finally, in Section 5 some conclusions and the current and future work are exposed.

2. Preliminaries

2.1. Technological Forecasting and Competitive Intelligence

Globalization leads countries to seek strategic tools in order to remain competitive in a global market where only stronger competitors survive. Therefore, organizational mechanisms and strategies directed to stay competitive and find new markets are necessary [1]. Techniques for TF and CI are strongly desirable in this context. They are used to observe what is happening on the market and with the main competitors, with the latest technological advances and research trends. TF and CI have become a mainstay to create new products or services, define marketing strategies and improve customer service. Their main goals are to identify opportunities and threats, to make strategic decisions and to achieve better competitiveness in organizations that make use of it [2,9].

TF is an organized and systematic way to capture and analyze information on science and technology from outside sources, to make it timely knowledgeable for decision making [10,11]. It includes all methods that attempt to anticipate and understand the characteristics, management and potential effects of technological changes, especially those related to invention, innovation and use [12]. There are several TF methods which can be categorized as follows [13,14]: trend analysis, expert's opinion, monitoring methods, modeling and simulation methods, statistical methods and modeling scenarios [3,15–17].

CI is a discipline that includes collection, analysis, interpretation and dissemination of information in the competitive environment in which companies move, acquiring a strategic value over competitors and the market in general, through a systematic and ethical process indispensable for the DMP [17–19].

CI is a set of ethical, legal and systematic methods that an organization can use to collect valuable information about their competitors, making a timely, specific and defined analysis. CI basically aims to get information about the activities of their environment and anticipate the future behavior of competitors, suppliers, customers, markets, products and services [1,20]. CI consists of three stages [21]: information gathering, information extraction and information contextualization. The goal of the use of a CI Information System can be defined by three aspects: improve the competitiveness of the company, accurately predict the changing environment and provide timely support for strategic decision making [22].

TF and CI are tools that complement each other well, have great similarities and become very useful in anticipating the environment in time to improve the competitiveness of an organization. However, there are certain differences: TF involves obtaining the most relevant information and analyzing the environment, while CI is a step in the management process of the information obtained, with particular emphasis on issues such as the adequacy of the display format for decision making and analysis of the results obtained by its use. Clearly, each tool has its own strengths, achieving better results when used together [1,23]. TF and CI have become essential for organizations that have processes of research, experimental development and innovation, allowing them to generate new projects, with a significant reduction of the risks that may be caused by activities in this area [10].

2.2. Web Mining

Nowadays, most organizations generate large volumes of data. Consequently, it becomes necessary to count on certain techniques to transform this data into useful knowledge. Among these techniques, machine learning, statistical analysis and visualization can be highlighted and are used in the Data Mining (DM) area.

Basically, DM is focused on the discovery and automatic extraction of knowledge in large volumes of data. Currently, the largest and most unstructured data source is the web. The web has many unique characteristics [7] such as (a) the existence of different types of data (images, audio, video and text), (b) information on web pages is varied, dynamic and noisy, (c) a small amount of information is linked, (d) it is also about services which allow people to interact with their sites, e.g., purchasing products, paying bills, and filling in forms.

The set of techniques to automatically discover and extract information found on the Web [7] is known as Web Mining (WM). The WM process can be formally defined as "the overall process of discovering information or potentially useful and previously unknown knowledge through data from the Web" [24].

WM tasks are classified into three categories: Web content mining, Web structure mining and Web usage mining. The first extracts or mines useful information or knowledge from Web page contents. The second, web structure mining, discovers useful knowledge from hyperlinks, representing the structure of the Web. This model may be useful to

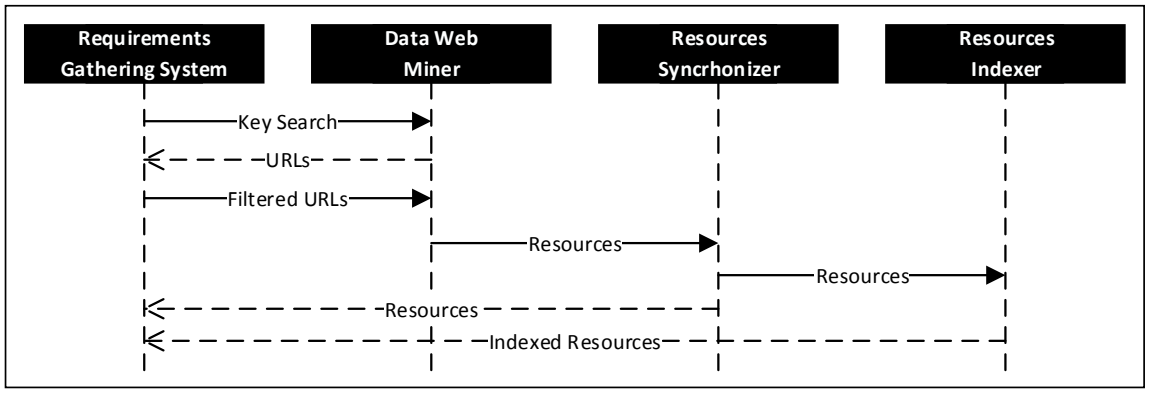

Figure 1. Interaction between the system modules

classify or group documents. Finally, web usage mining refers to the discovery of user access patterns from Web usage.

The process of WM is not trivial and is divided into several stages, such as Information Retrieval (IR), Information Extraction (IE), the Natural Language Processing (NLP) and Text Mining (TM).

A web search requires different techniques: IR is a field of study which helps the user find the necessary information within a large collection of text documents [24]. IR basically consists of searching a set of documents relevant to the query made by the user. It usually uses a ranking algorithm related to the relevance of the document being evaluated. In a more general definition, IR studies the acquisition, organization, storage, retrieval and distribution of information [7].

An IR System (IRS) includes a query analyzer that interprets plain text and separates a text string into individual words, eliminating those that are considered irrelevant to the query. It uses mathematical models to assign a score to each document to create an ordered index, ranking from the best to the worst score, considering their level of relevance. When the content analysis starts, the first documents obtained will boast the best scores [7].

3. Proposed System

In order to support the search of resources on the Web for later use by the TF and IC processes, an IR model based on WM techniques is proposed. This type of application requires the implementation of certain features to achieve its design goals. They are described below.

The system allows searching for information about a specific subject. To achieve this, it is divided into various subsystems or modules: the Requirement Gathering System (RGS), which helps the user in the explanation of the requirements, the Data Web Miner (DWM) module, which explores the Web looking for information about the subject, the Resource Synchronizer (RS) module, which synchronizes all the documents found between these two modules, and the Resource Indexer (RI), which constructs an index of the resources gathered in the repository. A view of the interaction between the different modules can be seen in Fig. 1.

The RGS sends the generated search keys to the DWM. The DWM sends queries to many search engines and returns the list of URLs found. Then, the RGS filters this list removing the duplicates and assigning each URL a score based on its position in the list and the times it has been repeated. After this process it sends the URLs back to the DWM, where it starts to crawl the web looking for resources. The RGS and the RI begin to synchronize and to index all the documents found by the DWM. In the next section a detailed description of functionality is presented.

3.1. Requirements Gathering System (RGS)

The RGS is a web application that performs the user-system interaction, and it is based on a previous model described in [25] Its main goal is to guide the user in answering a series of questions and characteristics about the subject in which they are interested. With these answers, the RGS constructs several search keys [25]. These search keys consist of a set of words concatenated with logical operators and reserved words taken from the answers provided by the user and from predetermined relationships between the questions. These operators are carefully chosen to match those used in the most common search engines. These keys are later used to obtain the URLs that initiate the mining process [25].

Once the RGS generates the search keys, it is sent to the DWM through a web service which returns several lists of URLs (one for each search engine). The DWM uses four different engines: Google, Bing, Intelligo and Mslmx Excite. Then, the RGS applies a ranking algorithm that operates as follows: it receives a list of URLs for each engine, then it combines these lists in a single one by removing duplicates, and assigning each URL a weight value

relying on its position in the list and the amount of times it appears on all the lists. Finally, the resulting list is sorted in consideration of the weight and sent to the DWM to begin the web crawling process. As the DWM finds resources in the web, these are synchronized in a repository in the RGS. The resources can be HTML pages, PDF files, office documents, videos, images, etc.

3.2. Data Web Miner (DWM)

The DWM has a web service interface that receives the requests from the RGS and starts the web mining process. The actions that the DWM component performs can be summarized in two steps:
(a) It receives one or more search keys from the RGS and sends these keys to the four search engines mentioned in Section 3.1. Then, it returns a list of the 10 first URLs for each one of these engines.
(b) It receives a merged and sorted list of URLs from the RGS as explained in Section 3.1 and begins the crawling process, exploring each URL for different types of resources, such as HTML pages and PDF documents, etc. During this process, the crawler looks up every link in the web sites and explores them to discover more resources. Every document is downloaded and analyzed in order to be assigned a weight value and, consequently, a ranking position considering its importance and relevance to the subject. This ranking process is explained in Section 3.5.5. The ranked documents are shown in the user interface and are then saved in a repository that is synchronized with the RGS. In this way, the user can later view and explore the files.
It is important to remark that the crawling process iterates over the pool of documents obtained applying four different algorithms to measure their relevance: Weighted Approach, C-Rank, Vector Space Model and Okapi. These algorithms are explained in detail in Section 3.5.

3.3. Resources Synchronizer (RS)

The use of web services allows the web mining process (DWM) to be separated from the user interface application (RGS). These two components run in two different locations, so it is necessary to synchronize all resources downloaded by the DWM with the RGS, including updates and removal of files. For this purpose, an open source file synchronizer and sharing server called ownCloud is used. It provides access to stored files via a web interface and desktop and mobile clients using the WebDAV protocol [26].

3.4. Resources Indexer (RI)

The documents downloaded and synchronized to the RGS need to be indexed in order to let the user perform all kinds of queries on them, like searching for specific terms, filtering by relevance, type of file, etc. To support this functionality, the open source indexing server Apache Solr was used. Solr can index many different file formats and extract metadata from them. It also provides a REST-like interface to access the service from other applications. This module is currently in development.

3.5. Algorithms

One of the main goals of the DWM is to find resources on the web and determine their relevance. For this to be accomplished, four algorithms were implemented, each with a different approach and techniques as described below.

3.5.1. Weighted Model Approach

The weighted model approach is characterized using a domain dictionary. It consists of three stages: first, all documents go through a pre-processing task. In most cases this process does not depend on the ranking method applied, but is included, as it contributes to better results. The pre-processing task consists of three steps: stop-word removal, stemming and tokenization [27].
The second stage consists of calculating the frequency of all terms for each document and building a domain dictionary applied to a determined topic. In this case a topic is related to the search key.
In the last stage all terms contained in the document are obtained and a verification is carried out to prove their existence both in the domain dictionary and the search key. For those terms that accomplish this, the frequency of the term is added to an accumulator called keyword hit count. On the other hand, the frequency of the term is added to the accumulator's positive hit count only if the term exists in the domain dictionary. In another case, the frequency of the term is added to an accumulator called negative hit count only. The final score is calculated as follows:

$$DRi = \frac{[Keyword_{Hit(i)} * \propto] + [Positive_{Hit(i)} * \beta] + [Negative_{Hit(i)} * \gamma]}{Keyword_{Hit(i)} + Positive_{Hit(i)} + Negative_{hit(i)}} \quad \text{(Eq. 1)}$$

where DRi represents a relevance score for document i, $Keyword_{Hit(i)}$ is the accumulation of the frequency of all terms that exist in the search query and in the domain dictionary, $Positive_{Hit(i)}$ is the accumulation of the frequency of all terms that exist only in the domain dictionary, and

$Negative_{Hit(i)}$ is the accumulation of the frequency of all terms that do not exist in the domain dictionary and search query. The variables $\propto$, β and γ have the value of $1; 0.75; 0.5$ respectively.

3.5.2. C-Rank Model

The C-Rank model is oriented to content collaboration of documents which have similar content. It is based on the idea that an author of a Web document adds links to other documents in order to complement the incomplete information (except advertising banners). Thus, a Web document linking to several documents proves to have value in that particular topic.

A C-Rank of a term in a page is defined by the sum of both the relevance score of the page to the term and a portion of the contribution score of the term to other pages.

$$CR_t(\mathrm{p}) = \lambda R_t(p) + (1-\lambda)\sum_{d}\sum_{q\in D(p,d)} \alpha_t^d(p,q) R_t(q) \quad \text{(Eq. 2)}$$

Eq. 2 show the C-Rank calculation, where $CR_t(p)$ represents the C-Rank of a term of t in a document p. $R_t(p)$ is the relevance score of the document p for the term t.$\alpha_t^d(p,q)$ is a percent contribution of term t in the document p for the document; $R_t(q)$ is a relevance score of the document q for the term t.

It is important to note that, unlike other collaboration models like PageRank, this one calculates the collaboration score with other documents using only the search key terms and not just by having a relationship [28].

3.5.3. Vector Space Model

Vector Space Model is perhaps the best known and most used IR model. It uses a function of similarity between the document and the query, and builds a ranking of retrieved documents [7]. Each document in the collection is represented by a t-dimensional vector, where t is the cardinality of the set of terms in the corpus of documents, and each element of the vector has a weight of the term associated with that dimension. Thus, any document and query may be represented by a vector in this vector space. Both assigning weights to terms in the documents and the similarity calculation can be performed in different ways [8].

Since in this model a document is represented as a vector of weights, the first step is to obtain these values. The weights are acquired by calculation of frequencies, usually using TF*IDF scheme, where TF is the frequency of the term t_i in the document d_j and IDF is the inverse of the number of documents in the collection where t_i appears. This scheme proposes establishing a relationship between the frequency of a term within a document and its frequency in the documents of the collection. The weights of each element of the vector are not only *0* and *1*, as in the known Boolean Model, but may be any value [7,8]. The second step consists of assigning a relevance-representative weight to each document. To achieve this, the model measures the similarity between the query vector and the document vector. The angle between two vectors is used as a measure of the divergence between the vectors and the cosine similarity is the most popular measure of similarity (since it has the useful property that is *1* for identical vectors and *0* for orthogonal vectors).

With the similarity values obtained from each document in the collection, the documents are sorted according to their relevance to the user's query [7,8]. The following equation shows how the measure of similarity for each document is calculated in the collection:

$$Sim\,(Q, D_i) = \frac{\sum_{j=1}^{t} w_{q_j} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{q_j})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}} \quad \text{(Eq. 3)}$$

Where $D_I = w_{di1}, w_{di2}, \ldots, w_{dit}$, is the vector representing a document. $Q = w_{q1}, w_{q2}, \ldots, w_{qt}$, is the vector representing the query. t, is the number of terms in the collection. w_{ij}, is the weight of term j in document i. w_{qj}, is the weight of term j in the query q.

3.5.4. Okapi Model

This system is based on a probabilistic model and incorporates the terms frequency and length of documents for calculation; its weighting scheme is called BM25 and, along with TF*IDF, it is one of the most effective. Both are considered benchmarks for the development and evaluation of new models and new weighting schemes. BM25 is a variation of TF*IDF scheme using a probabilistic model. It is based on the concept of "bag of words" (vocabulary of the document's terms) instead of vectors, as in the Vector Space Model [7,29]. The calculation for the relevance value of each document is presented below.

$$Okapi\,(d_j, q) = \sum_{t_i \in q, d_j} \ln\frac{N - df_i + 0.5}{df_i + 0.5} \times \frac{(k_1+1) f_{ij}}{k_1\left(1 - b + b\frac{dl_j}{avdl}\right) + f_{ij}} \times \frac{(k_2+1) f_{iq}}{k_2 + f_{iq}} \quad \text{(Eq. 4)}$$

Eq. 4 shows the Okapi formula used to calculate the relevance value, where k_1 (between *1.0 - 2.0*), b

(usually *0.75*) and k_2 (between *1 - 1000*) are parameters; t_i, is a term, f_{ij} is the raw frequency count of term t_i in document d_j. f_{iq}, is the raw frequency count of term t_i in query q. df_i, is the number of documents that contain the term t_i. dl_j, is the document length (in bytes). N is the total number of documents in the collection and *avdl,* is the average document length of the collection.

3.5.5. Rank Position Calculation

To combine the results of the four algorithms described above, a method called Reciprocal Rank was used. This method takes two or more lists of ranked documents and produces a unified list with the combination of documents according to its position in each list. This approach provides better effectiveness than the scores used separately and better performance of the overall system [7] [30]. When the same document is returned by more than one algorithm, it has a greater chance of being more relevant. When this happens, the inverse of its position in the lists are added together to obtain the new score. **Eq. 5** shows the calculation of the score r for each document d_i using its position in list *j*.

$$r(d_i) = \frac{1}{\sum_j \frac{1}{d_{ij}}} \qquad \text{(Eq. 5)}$$

After this iteration, the documents are sorted according to its new relevance score in a single list. This process is later repeated when new documents are found, in which case, the first set of documents is also a ranked list for the next iteration.

4. Simulations and Test

To assess the performance of the proposed model, three experiments have been designed. In each of them an expert user performs a search related to their field of specialty using the system. After a period of 24 hours, they proceed to evaluate the first twenty results returned by the system using a quantitative scale, assigning each document a score from 1 to 5, where 1 means not relevant and 5 stands for very relevant. Two of these test cases are related to the tea production industry and the remaining case relates to software development.

4.1. Gourmet Tea Case

This project called "Gourmet tea" involves searching for information about this specific type of tea. The purpose of the search was to find papers, patents, market information, tea countries, trends, tea trading, etc. Academic and governmental entities were specified as preferred sources of information. Terms such as "advertising" or "ads" were excluded from the search process. The search keys generated by the RGS are shown in Table 1.

Table 1. Search keys for the Gourmet Tea Case.

Gourmet Tea
sell OR buy AND Gourmet Tea AND global trading AND gourmet tea countries AND offer and demand AND gourmet tea marketing channels AND trends and characteristics AND export volume AND import volume AND tea industries AND patents OR paper OR cite -promotions -ads -advertising
buy AND Gourmet Tea AND global trading AND gourmet tea countries AND offer and demand AND gourmet tea marketing channels AND trends and characteristics AND export volume AND import volume AND tea industries AND patents OR paper OR cite -promotions -ads -advertising
sell AND Gourmet Tea AND global trading AND gourmet tea countries AND offer and demand AND gourmet tea marketing channels AND trends and characteristics AND export volume AND import volume AND tea industries AND patents OR paper OR cite -promotions -ads -advertising
sell OR buy AND Gourmet Tea AND global trading AND gourmet tea countries AND offer and demand AND gourmet tea marketing channels AND trends and characteristics AND export volume AND import volume AND tea industries AND paper OR cite -promotions -ads -advertising
sell OR buy AND Gourmet Tea AND global trading AND gourmet tea countries AND offer and demand AND gourmet tea marketing channels AND trends and characteristics AND export volume AND import volume AND tea industries AND patents -promotions -ads -advertising

The search keys displayed in Table 1 are generated using different methods which combine the answers provided by the user with predefined logical operators. These methods generate one or more keys depending on the user's answers. The first method simply uses the main subject of search to capture the essence of the requirement. This is the first key. The second method combines the selected actions (e. g. buy or sell) with the different information sources selected (e.g. papers or patents), generating the remaining four search keys.

A summary of the results obtained by the system, along with their corresponding score is shown in Table 2. It can be observed that more than 50% are very relevant results respect to the search subject.

Table 2. Evaluation of the first 20 results for Gourmet Tea.

Relevance	Score	Resources	Percentage
Very relevant	5	12	60%
Relevant	4	1	5%
Moderately relevant	3	2	10%
Slightly relevant	2	1	5%
Not relevant	1	4	20%
Total		**20**	**100%**
Avg. relevance		**3.8**	

The resources assessed with score 5 consist of tea

shops, e-commerce sites for gourmet tea and gourmet food from USA, Japan, Brazil and the UK. Two of them focus specifically on gourmet tea (mygourmettea.com and thegourmettea.com.br). It is also worth mentioning the sites of the United Kingdom Tea Association and the USA Tea Association. This is an interesting issue because the user wanted information about countries that produce or import gourmet tea and their marketing strategies, which can be seen in the results obtained.

The result with score 4 is a site of gourmet coffee but it also contains gourmet tea-related products like accessories, gifts, etc. This result is related to the subject but is not very relevant for the user.

The results with score 3 consist of a document of Swedish market of food products, imports and exports and distribution channels, another document of horticulture development which includes tea among other herbs and vegetables.

The result with score 2 is an e-book about organic agriculture trends, which is not directly related to organic tea production.

The results with score 1 consist of the site of a university, a company's cloud services, an article about the climate inversion and the WikiLeaks site. These were considered irrelevant to the subject.

The average score of relevance for this project was 3.8, which goes from moderately relevant to relevant in the scale. This score can be considered as a good measure for the entire set of results obtained.

4.2. Organic Tea Case

Organic tea is a specific type of tea, and the purpose of the search was to find papers, patents, information about certifications conditions for organic tea, exporting and importing countries, tea industry, etc. The search keys generated are shown in Table 3.

The search keys shown in Table 3 are generated combining the user's answers as described in the previous experiment.

A summary of the results obtained by the system, along with their corresponding score is shown in Table 4. It can be seen that high relevant resources (5 and 4) correspond to 75% of the results.

The results with score 5 consist of a site with real cases of tea certifications from several countries, two e-commerce sites for tea and its related accessories, different types of drinking tea, recipes, etc.; an article explaining the organic tea business and marketing, two sites for selling tea, a blog with over 300 articles about organic tea, PDF documents about organic tea, coffee and cocoa and one about the chain of value of tea in Nepal. These resources have a high relevance to the user, given their search for information about certification conditions in different countries, market trends and companies that produce and commercialize tea.

Table 3. Search keys for the Organic Tea Experiment.

Organic Tea
sell OR buy AND Organic Tea AND trends AND global trading AND offer and demand AND certification conditions AND value added AND exporting countries AND importing countries AND customs offices and tea industries AND patents OR paper OR cite -promotions -ads –advertising
buy AND Organic Tea AND trends AND global trading AND offer and demand AND certification conditions AND value added AND exporting countries AND importing countries AND customs offices and tea industries AND patents OR paper OR cite -promotions -ads –advertising
sell AND Organic Tea AND trends AND global trading AND offer and demand AND certification conditions AND value added AND exporting countries AND importing countries AND customs offices and tea industries AND patents OR paper OR cite -promotions -ads –advertising
sell OR buy AND Organic Tea AND trends AND global trading AND offer and demand AND certification conditions AND value added AND exporting countries AND importing countries AND customs offices and tea industries AND paper OR cite -promotions -ads -advertising
sell OR buy AND Organic Tea AND trends AND global trading AND offer and demand AND certification conditions AND value added AND exporting countries AND importing countries AND customs offices and tea industries AND patents -promotions -ads -advertising

Table 4. Evaluation of the first 20 results for Organic Tea.

Relevance	Score	Resources	Percentage
Very relevant	5	9	45%
Relevant	4	6	30%
Moderately relevant	3	1	5%
Slightly relevant	2	4	20%
Not relevant	1	0	0%
Total		**20**	**100%**
Avg. relevance		**4.0**	

The results with score 4 consist of a site with information about organic tea and herbs, spices, bath and body oils, etc.; another result is a blog by Daniel Giovannucci with publications about rural development, agro enterprises, sustainability and agricultural standards, organic and fair trade. It contains publications for trends in organic tea; two sites with multiple combinations of organic tea with other ingredients; another result is an article from a recognized newspaper about organic tea and its market, and last, there is a document of value-added standards in the North American food market. These resources are related to the tea industry and they are also relevant for the user.

The results with score 3 consist of a link to Amazon with several types of organic tea, a document about sustainability impact assessment for the agriculture sector, a guide about business with Russia and a document of regulation of GMO crops and foods in Kenya. These results do not correspond to the

requirements specified by the user but they contain useful specifications about organic production that the user deemed somewhat relevant.
The results with score 2 are two links from Wikipedia, one about tea and another about the coffee market. Evidently, this information is irrelevant to the search subject.
The average score of relevance was 4.0, which corresponds to relevant in the scale. This can be considered a good result for the whole project.

4.3. Xamarin Case

Xamarin Studio is a tool for developing multiplatform mobile applications. The purpose of the search was to find videos and tutorials about this tool. The search keys generated by the RGS are shown in Table 5.

Table 5. Search keys for the Xamarin Case.

Xamarin
Xamarin AND Videos AND Tutorial

In this case the user skipped several questions of the RGS. Therefore, only two of them were generated using the same methods described in the previous experiments.

Table 6. Evaluation of the first 20 results for Xamarin.

Relevance	Score	Resources	Percentage
Very relevant	5	3	15%
Relevant	4	5	25%
Moderately relevant	3	4	20%
Slightly relevant	2	0	0%
Not relevant	1	8	40%
Total		**20**	**100%**
Avg. relevance		**2.75**	

A summary of the results obtained by the system, along with their corresponding score is shown on Table 6. Contrary to what happened in the two previous cases, the percentage of relevant resources (5 and 4 scores) strongly decreases.
The results with score 5 consist of a video posted by someone who has a collection of videos about coding with Microsoft technologies. In the video he explains the structure of a project through an example. The second result is another video from the CodeGeek magazine which has a set of videos about different technologies for developers. This video is almost 4 hours long and explains different aspects of a Xamarin project. The third result is a list of 100 videos explaining Xamarin for Android, iOS and the use of components. These three results are very relevant to the topic searched.
The results with score 4 consist of several links from the Xamarin official site (www.xamarin.com), which has guides and tutorials for beginners and a few results are tutorials from other educational sites.
The results with score 3 consist of forums, reviews, blogs, and marketing videos. The results with score 1 were tutorials about subjects not related to Xamarin.
In Table 6, it can be observed that most results are relevant for the user (score 4 – 5), but it also shows that there are several irrelevant documents (score 1). Documents with somewhat medium relevance correspond to an ambiguous situation. Therefore, they require a deeper analysis of the sites they link to, to assess the relevance of the results as a whole. This is done by the crawler in a later stage of the process.
The average score of relevance for this project was 2.75, which is in the middle between slightly relevant and moderately relevant in the scale. With this project the system didn't perform well, but this can be attributed to the great number of irrelevant results.

5. Conclusions and Future Work

This article introduces a Web Mining-based system used to find relevant information based on user's requirements. The system has four modules that exchange information. This architecture allows low coupling, decentralization of its main functions and distributed implementation by different teams, geographically separated.
In order to obtain precise requirements, the user is guided using questions (relevant characteristics, sources, type of information, geographical zone, commercial features and undesirable terms to exclude) about a specific topic. In this way the collection of requirements becomes a very simple and easy task. This information is resumed in several search keys used to start the search process.
The model uses four search engines to obtain the first resources. These links are processed and merged into a single list to obtain a ranking of relevance. Every link is analyzed using C-Rank, Vector Space and Okapi algorithms which determine the importance of the resource based on the user's requirements. These resources are presented to the user, classified in terms of their relevance.
In order to test the system, three scenarios were used: Gourmet Tea, Organic Tea and Xamarin. Analyzing these three cases, based on the evaluations made by the test experts, a good system's performance can be observed. The large number of documents with high scores of relevance reflects the system's adequate function. It is worth noting that the first two test cases produced overall more relevant results than the third. Also, in these

cases, the user entered answers to all the questions in the RGS resulting in larger search keys but also more specific ones, which in turn conducts to more relevant resources being returned by the system. Therefore, it can be stated that the way the system builds and uses the search keys and performs the continuous analysis of web documents provides better results than manually extracting information using the search engines in the traditional manner.

Finally, the proposed model is not a common search-response system; instead, it is a system that performs continuous exploration of resources on the Web. In this way, it returns more relevant results as the crawling progresses.

Currently, two topics have been covered in the project: (a) the resources indexer, whose main goal is to improve the results visualization and, therefore, simplify the user-system interaction; (b) the analysis algorithm. In order to obtain more reliable results, the semantic analysis will be considered in the next algorithms implementation.

Acknowledgments

This work has been supported by the Projects UTN4058 of National Technological University (UTN-Argentine) and "Modelos de Análisis de Información para la Toma de Decisiones Estratégicas" Gastón Dachary University (UGD-Argentine). Authors are especially grateful for the collaboration of UGD's engineering students Matías Barboza, Victor Alvarenga and Leandro Witzke.

References

[1] M.I. Ramírez, D. Escobar Rua, B. Arango Alzate, Vigilancia Tecnológica e Inteligencia Competitiva, Gestión de Las Personas Y Tecnología. 13 (2012) 238–249.

[2] B. Arango Alzate, L. Tamayo Giraldo, A. Fadul Barbosa, Vigilancia Tecnológica: Metodologías y Aplicaciones, Gestión de Las Personas Y Tecnología. (2012) 154–161.

[3] S. Madnick, W. Woon, A. Henschel, A. Firat, Technology Forecasting using Data Mining and Semantics, Cambridge, MA, 2008.

[4] J.P. Shim, M. Warkentin, J.F. Courtney, D.J. Power, R. Sharda, C. Carlsson, Past, Present, and Future of DecisionSupport Technology, Decision Support Systems. 33 (2002) 111–126. doi:10.1016/S0167-9236(01)00139-7.

[5] D.J. Power, R. Sharda, Decision Support Systems, in: Springer Handbook of Automation, 2009: pp. 1539–1548. doi:10.1007/978-3-540-78831-7_87.

[6] C.W. Holsapple, Decision Support Systems, in: Encyclopedia of Information Systems, International Thomson Business Press, New York, 2003: pp. 551–565.

[7] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2011. doi:10.1007/978-3-642-19460-3.

[8] F. Johnson, S. Kumar Gupta, Web Content Mining Techniques: A Survey, International Journal of Computer Applications. 47 (2012) 44–50. doi:10.5120/7236-0266.

[9] R. Barainka, Modelos de Vigilancia e Inteligencia Competitiva, BAI Agencia de Innovacion, 2006.

[10] J.M. Durán, M.M. Martínez, J.V. Triano, La Vigilancia Tecnológica en la Gestión de Proyectos de I+D+i: Recursos y Herramientas, El Profesional de La Informacion. 15 (2006) 411–419.

[11] J.C. Vergara, La Vigilancia Tecnológica antes y después de la UNE166006 : 2006 Ex, PUZZLE: Revista Hispana de La Inteligencia Competitiva. 5 (2006) 37–41.

[12] A.K. Firat, W.L. Woon, S. Madnick, Technological forecasting–A review, Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology. (2008).

[13] V. Coates, M. Farooque, R. Klavans, K. Lapid, H.A. Linstone, C. Pistorius, et al., On the Future of Technological Forecasting, Technological Forecasting and Social Change. 67 (2001) 1–17. doi:10.1016/S0040-1625(00)00122-0.

[14] T.J. Gordon, J.C. Glenn, Futures research methodology, The Millennium Project, 2003.

[15] F.J. Parent, J.K. Anderson, P. Myers, T. O'brien, An examination of factors contributing to delphi accuracy, Journal of Forecasting. 3 (1984) 173–182.

[16] R.R. Levary, D. Han, Choosing a Technological Forecasting Method, Industrial Management. 37 (1995) 14–-18.

[17] T.J. Gordon, A simple Agent Model of an Epidemic, Technological Forecasting and Social Change. 70 (2003) 397–417. doi:10.1016/S0040-1625(02)00323-2.

[18] P. Escorsa, P. Lázaro, Intec. La Inteligencia Competitiva. Factor Clave para la Toma de Decisiones Estratégicas en las Organizaciones., Comunidad de Madrid Consejería de Educación Dirección General de Universidades e Investigación Fundación madri+d para el Conocimiento, 2007.

[19] P.T. Gibbons, J.E. Prescott, Parallel Competitive Intelligence Processes in Organisations, International Journal of Technology Management. 11 (1996) 162–178.

[20] R.G. Vedder, M.T. Vanecek, C.S. Guynes, J.J. Cappel, CEO and CIO perspectives on competitive intelligence, Communications of the ACM. 42 (1999) 108–116. doi:10.1145/310930.310982.

[21] T. Hiltbrand, Learning Competitive Intelligence From a Bunch of Screwballs, Business Intelligence Journal. 15 (2010).

[22] I. Anica-Popa, G. Cucui, A Framework for Enhancing Competitive Intelligence Capabilities using Decision Support System based on Web Mining Techniques, International Journal of Computers, Communication & Control. 4 (2009) 326–334.

[23] A. Hidalgo Nuchera, G. León Serrano, J. Pavón Morote, La Gestión de la Innovación y la Tecnología en las Organizaciones, Ediciones Pirámide, 2002.

[24] V.H.E. Jeria, Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa, 2007. decsai.ugr.es/Documentos/tesis_dpto/100.pdf.

[25] L. Lezcano Airaldi, F. Sa, M. Karanik, L. Wanderer, Modelo de Recopilación de Requerimientos para Vigilancia Tecnológica e Inteligencia Competitiva (VTeIC), in: 3er Congreso Nacional de Ingeniería Informática Y Sistemas de Información (CONAIISI), 2015.

[26] OwnCloud, OwnCloud self-hosted file sync and share server, (n.d.) www.owncloud.org.

[27] S.S. Bama, M.S.I. Ahmed, A. Saravanan, Enhancing the Search Engine Results through Web Content Ranking, International Journal of Applied Engineering Research. 10 (2015) 13625–13635.

[28] D.-J. Kim, S.-C. Lee, H.-Y. Son, S.-W. Kim, J.B. Lee, C-Rank and its variants: A contribution-based ranking approach exploiting links and content, Journal of Information Science. 40 (2014) 761–778. doi:10.1177/0165551514545429.

[29] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search, Information Retrieval. 82 (2011) 944. doi:10.1080/14735789709366603.

[30] R. Nuray, F. Can, Automatic ranking of information retrieval systems using data fusion, Inf. Process. Manag. 42 (2006) 595–614. doi:10.1016/j.ipm.2005.03.023.