Information retrieval with semantic annotation

Hubert Viltres Sala, Paúl Rodríguez Leyva,
Juan Pedro Febles Rodriguez and Vivian Estrada Sentí

June 9, 2019

Information retrieval with semantic annotation

Hubert Viltres Sala[1], Paúl Rodríguez Leyva[2], Juan Pedro Febles[3], Vivian Estrada Sentí[4]

[1] *Informatic Science University,La Habana, Cuba, hviltres@uci.cu, febles@uci.cu*
[2] *Informatic Science University, La Habana, Cuba, pleyva@uci.cu" vivian@uci.cu*

Abstract*– The processing of information with semantic annotation allows to identify the intention of search of the user and to adjust the result according to the information context. The present research proposes a model for the retrieval of information with semantic annotation that allows to help the user to retrieve the most relevant information among all the information available on the web. In the model, three components (Crawler-Indexing, Processing and Presentation) are developed, that allow identifying the need for user information through the processing, selection and subsequent publication of the retrieved information. The crawling and indexing component allows the identification of available websites to extract information and perform semantic annotation by applying different information processing techniques. The processing component analyzes the user's preferences and processes the query performed to calculate the similarity of the indexed information. Subsequently the results are sorted according to the relevance to show in the Presentation component a quantity of information that can be assimilated by the users. For the validation of the proposal we used the metrics of precision and exhaustiveness that allowed to demonstrate the quality, relevance and relevance of the information retrieval with semantic annotation.

Keywords-- Semantic Web, information retrieval, relevance, semantic annotation, similarity

I. Introduction

The development of society, the emergence of technologies and tools to improve access to information and the rapid growth of the Internet in recent years, has made it possible to generate a large volume of web content. The information available on the web is scattered, is poorly structured or invisible to the common user, making it difficult to access high quality information and value for the user. In this context, users when they access the Internet feel overwhelmed by the information overload and do not quickly and easily obtain the information that best suits their needs, limiting their experience in using an information retrieval system. There are more than a billion websites on the Internet and every day the amount of information available increases exponentially. Generating new opportunities and dissimilar challenges for users when they try to obtain relevant information. Due to the large amount of information available on the Internet and the difficulty of assimilating them, users rely on information retrieval systems (SRI) to find what they are looking for.

Information retrieval systems through the use of different tools, methods and techniques retrieve public information from the web for further analysis, selecting and ordering the most relevant information for the user's need. Among the main sources to obtain information are repositories of components, databases and search engines that allow to simplify and group relevant information, using certain concepts of organization of information. The main objective of an IRS as set out in [1,2,3,4] is to satisfy the need for information raised by a user in a query in natural language specified through a set of keywords (see Figure 1), which help identify the most relevant information for the user.

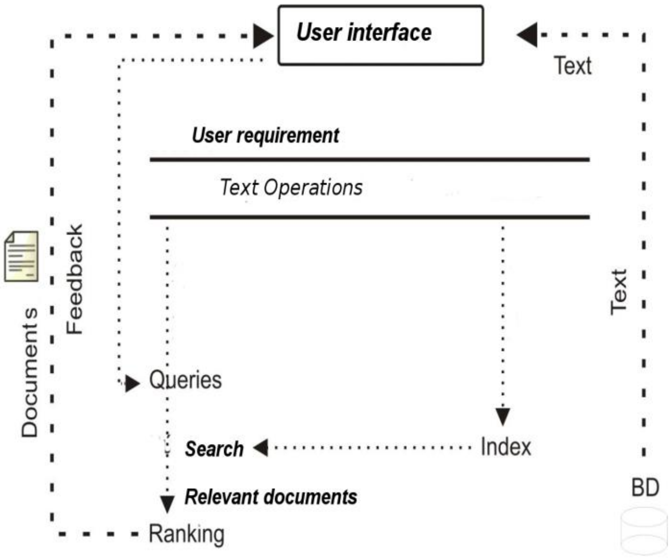

Figure 1: Information search process – source: Vuotto, Bogetti y Fernández, 2015

Authors such as Salton and Mcgill (1983), Gonzalo, et al. (2017) state that the Search and Recovery of Information has as main objective to provide relevant information to the user to satisfy his need for information. Within the BRI five main activities are defined (locate, select, interpret, synthesize and communicate the information) to guide the process of obtaining information adjusted to the user's need. These five activities are contemplated in the three main components of a search engine currently (crawler, indexer and processor).

During the process of information retrieval, traditional search engines generally use techniques that determine the relevance of the coincidence of the keywords in the documents and do not analyze the relationships that exist between the implicit meaning of the keywords and the document. Several authors suggest that the semantic recovery of information improves the quality and relevance of the information shown to users, since it uses processing techniques in natural language, uses ontologies to identify the context and relevance is established by the semantic similarity of the query and the indexed documents.

The Semantic Web is changing the way to obtain information on the Internet, it is one of the technologies that has generated the most impact for internet users due to the quality of the information it obtains. In [8] defines the Semantic Web as "... an extension of the current Web, in which information has a well-defined meaning, facilitating computers to work better in cooperation with humans" and its main objective has been to allow data stored on the Web can be processed by machines in an intelligent way, making it easier for people to search, integrate and analyze the available information.

Semantic information retrieval

The Semantic Web is changing the way to obtain information on the Internet, it is one of the technologies that has generated the most impact for internet users due to the quality of the information it obtains. In [8] defines the Semantic Web as "... an extension of the current Web, in which information has a well-defined meaning, facilitating computers to work better in cooperation with humans" and its main objective has been to allow data stored on the Web can be processed by machines in an intelligent way, making it easier for people to search, integrate and analyze the available information.

The principle of the semantic web is the processing of information automatically through the use of artificial intelligence using a wide variety of algorithms. It also aims to understand the need expressed by the user in a query and provide the search for meaning, identifying and providing reliable information. To carry out the semantic search, semantic search engines are used, which are "information retrieval systems that understand the user's need and analyze the information available on the Web through the use of algorithms that simulate understanding or understanding".

The general functioning of a semantic search engine in [9] is associated with the following characteristics:

- Allows searches by fields.
- It has the ability to extend the terms of the query by means of synonyms or related words.
- Identify named entities, such as names of companies, organizations or people, that are used with that meaning in the search process.
- Use grouping techniques to build content categorizations on which to search or to group key terms. This is the case of tag clouds that show the key terms of a website according to its importance.
- Detects relationships between search terms and words that appear in content based on knowledge models represented through ontologies.
- It offers the possibility of using natural language to express questions and even factual questions, for which specific answers are obtained [9].

The aforementioned characteristics show the possibilities of the semantic web in information retrieval where a user expresses in natural language his intention to search and the search engine analyze and select the information adjusted to that need. In the context of the Cuban web where the technological limitations difficulty the information retrieval process to solve this problem it is necessary to use the recovery of semantic information.

Information retrieval on the Cuban website

In the Cuban web there are more than 6 thousand websites hosted under the .cu domain with varied information. Users to access information use different tools, which does not always recover the relevant information, mainly due to:

- Heterogeneity of information sources.
- Quality of the information.
- Visibility of information.
- Accessibility of information.
- Difficulty in understanding the need of the user expressed in natural language.
- Little accuracy of the results because the similarity of the keywords is enhanced.
- Sensitivity of the results against the exact terms introduced.
- Selection of information due to the relevance of the positioning of the website.

The aforementioned difficulties show little precision and accuracy in the information retrieval process and diminish the user's experience when searching for information. These deficiencies coupled with the need to provide users with high quality information raise the need to develop an SRI with semantic annotation that allows selecting the information that best suits the needs of users to improve their experience on the Cuban web.

Semantic information retrieval

The semantic web is an extension of the current web, authors such as [5], [8], [9], [10] and [11] state that it allows obtaining information efficiently through the integration, automation and reuse of data using various techniques to improve the relevance of the information collected. According to [11] the objective of the semantic search is to improve the accuracy of the search by understanding the intention of the user when making a query and the contextual meaning of the data in the source of knowledge. The semantic search predicts what the user explicitly expresses (search intent) and adjusts its need (context) to the available information by selecting the most relevant one for the user. The model proposed in the research is supported on the basis of retrieving relevant information for the user using semantic technology by understanding the intention to search, extraction of knowledge from data sources, adjustment of user preferences and calculation of relevance.

METHODS

In order to obtain relevant information for users, a semantic information processing mechanism is implemented. The proposal covers the three main components of the SRI (Crawling-Indexing, Processing and Presentation). Figure 2 shows the components that support the process of searching and retrieving information on the web. Next, each of the three components is described.

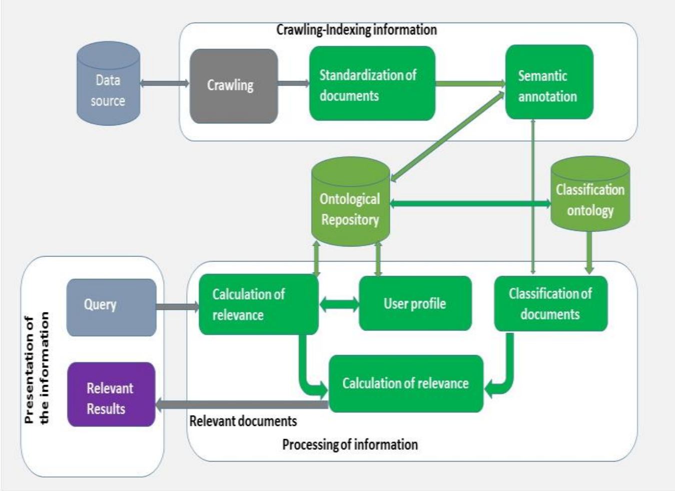

Figure 2: Semantic information retrieval

Component crawling and indexing

The crawling and indexing component allows identifying the available websites, in addition to retrieving and storing the information of each web page for further processing and presentation to users when making a query. The crawlers are responsible for exploring the web identifying the pages that have been created or updated to continue updating their index of information. After being tracked, different metadata (url, summary of content, links, keywords, language) are stored and used to extract knowledge using semantic web techniques. The crawling process begins with a list of links to websites provided by previous crawls or by sitemap; the higher the number of links the better the crawling process will be. During this process special attention is given to new websites, changes to current websites and broken links.

The crawler analyzes each page, downloads its content and identifies new links to continue the process on a recurring basis. It is used to perform the Nutch crawling in a distributed way using the selection, re-visit, courtesy and parallelization policies that allow a thorough crawling. The crawler configuration determines which sites to crawl, how often, and how many pages to explore in each site.

After performing the crawling process, each web page is analyzed to identify the main elements and then store the information and create an index of contents that allows improving the process of information retrieval. In the process of indexing, the tracked information is standardized by defining the metadata necessary for the processing of the information. As tools to perform information processing, Solr and Apache Jena are used, which use different techniques and algorithms to extract the implicit knowledge of web pages. For the semantic reasoning of the information Apache Jena is used that provides an API to read, write, extract and process RDF graphs; you also have an inference engine to reason about ontologies. Additionally, the algorithm CF-IDF (Frequency of the concept - inverse document frequency) is used to create the index based on the annotations made [10, 12].

.

Semantic annotation

The semantic annotation process has as input a document and the domain ontology. To make the semantic annotation, the terms of the document are extracted and identified and are related to concepts in the ontology. The semantic annotation proposal is based on the modification of [13] methodology, which consists of identifying and extracting the terms; associate the terms to a concept and store the document annotation using domain ontologies and WordNet.

To calculate the semantic index, the TF-IDF algorithm is used to determine the frequency of appearance of ontology concepts in the document [2, 3, 14]. In the calculation of the relevance between the concepts and the document, the relations of hierarchy of the concepts are analyzed and a semantic similarity metric based on the path between concepts is applied (see figure 3).

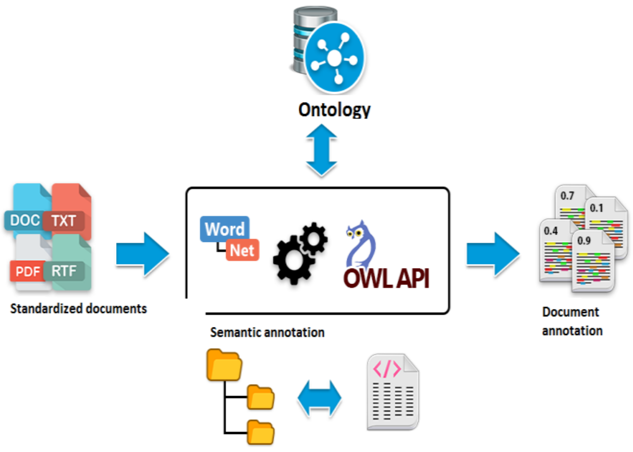

Figure 3: Semantic annotation

When applying TF-IDF to a document, its semantic index is generated, formed by a vector that contains all the concepts with their respective value. For each concept, its index is determined and stored in the document in the concept-value format.

Information processing component

The component processes information in natural language by associating each sentence of a text with a semantic representation using an ontology as a basis. In [15] an ontology is defined as "an explicit specification of a conceptualization" that allows adding a sense to the information that needs to be processed. It consists of 5 components (concepts, relationships, functions, instances and axioms) that describe the relationships of the words and add a natural sense. The use of Ontologies makes it possible to improve the processing in natural language of the query made by the user and the information collected by the crawlers on the web.

Expansion of the query

Users when they access an information retrieval system formulate the questions in natural language. In order to understand the intention behind the question, it is necessary to process and apply different techniques to identify the user's need for information. The main objective of the processing of the query is the disambiguation of the terms entered by the user, generating as output a triplet in RDF format.

To improve the information retrieval methods of combination of conditions are used by adding new words to the query made. This method makes it easier to understand the user's need and enrich the query. In the consultation expansion, different approaches are used [16, 17] that use techniques based on knowledge, corpus and relevance to recover information. Among the main proposals to expand a query are those related to techniques that determine the similarity between texts, terms and concepts.

In the present investigation (see figure 4) the query expansion technique is applied based on the concepts of a domain ontology and the preferences of the users. To disambiguate the terms entered by the user and adjust the search results to their preferences.

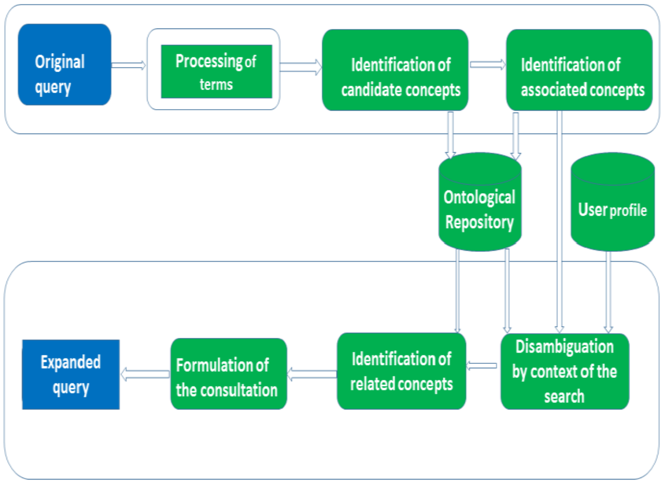
Original query
Processing of terms
Identification of candidate concepts
Identification of associated concepts
Ontological Repository
User profile
Expanded query
Formulation of the consultation
Identification of related concepts
Disambiguation by context of the search

Figure 4: Semantic query extension

To identify the relationship between the terms of the query and the concepts in the domain ontology, the cosine similarity algorithm is applied. In the process of identifying the concepts, phrases, simple concepts and concepts related to the analyzed term are obtained. Next, the semantic similarity between the concepts within the ontology is calculated to identify which concept best represents the term introduced by the user. To determine the semantic similarity between the concepts, the semantic similarity between two concepts is applied by means of the relationship counting method [16] that identifies similar concepts.

The proposed method concatenates the concepts related to the original terms through a logical operator (AND, OR, NOT) to retrieve relevant information. A new semantically enriched query is obtained as a result.

User profile

Obtaining the user's profile in the information retrieval process allows them to better understand their intention to search and adjust the results obtained to their preferences [2, 3, 15]. The proposed method allows to generate and update the user's profile according to their implicit and explicit preferences by combining several elements (categories selected in their profile, search history and location) to accurately meet the needs of the user. As a result, the user's preferences profile is obtained based on their browsing history and domain ontology concepts. These elements make it possible to better understand the user's search intention.

To create the user profile, the selected categories and the queries made are analyzed. The method has as input the categorized queries, the categories of the documents and the associated concepts. As output, the user's search profile according to the category (PBUC) and the profile of concepts associated with searches (RPUD). The user's profile is updated according to their browsing history and the constant update allows to personalize the search results and improve the RI process.

Calculation of similarity

To determine the similarity between the query made by the user and the indexed information in the search engine, the results of the query processing, user profile processing and the relevance index of the semantic annotation made during the storage process are used information. The similarity is determined using the Levenshtein algorithm for short texts and the cosine function.

The processing of documents is a fundamental element to determine the relevance and quality of the information provided to users. According to [18] and for the purposes of this research, quality is defined as the ability to meet the needs of a user and is measured in the level of precision and accuracy in the results provided.

To improve processing, techniques and algorithms are applied to group documents and facilitate the retrieval of information. The method proposed in the present investigation allows categorizing documents and grouping them according to a set of predefined categories.

To perform the categorization process, the concepts annotated in the document are analyzed and associated with the predefined categories. A document with the annotations made is obtained from the indexing-tracking process; The concepts obtained from the document are analyzed to determine their relationship with the categories in the ontology and identify which one best represents it. Each category has a set of concepts that are represented in the domain ontology used to improve the document classification process.

In the categorization process, different algorithms are used to categorize the information. To determine the similarity between the concepts associated with the category and the concepts noted in the document, the cosine similarity is used (equation 1).

$$\mathrm{sim}(c,d) = \frac{\sum W_{t,d} \text{ x } W_{t,c}}{\sqrt{\sum_t W^2{}_{t,d}} \text{ x} \sqrt{\sum_t W^2{}_{t,c}}}$$

The concepts of documents are represented by the vector d= $c_{1,d}$ $c_{2,d}, \cdots, c_{k,d}$, the categories by c= ($co_{1,d}$, $co_{2,d}, \cdots, co_{k,d}$), sim(c,d) is the relevance of the category for the document, ISC the semantic index of the concept for the document.

After calculating the similarity, the predominant category is selected and stored in the structure of the document. The stored document contains the relevance percentage for each category (see table 1).

Table 1: Indexes by category of document

Category	*Environment*	*Politics*	*Culture*	
percentage	0.6	0.6	0.6	0.6

Calculation of relevance

After obtaining the semantic similarity, we proceed to calculate the relevance to show the most relevant information for the user. In this process the algorithm proposed in [1, 2, 4] is used, which determines the coefficient of relevance according to the user profile, the query and the semantic similarity index. The coefficient of relevance obtained is used to order the results and show a number of elements that can be assimilated by the user.

To retrieve relevant and customized results, a method is proposed that integrates the search categories of the user profile and the categories of the documents in a relevance algorithm. The search categories are obtained from the user's preferences profile and the document category of the semantic categorization process (see table 2, 3 and 4).

Table 2. Queries made and categories to which they belong.

Query	*Culture in the world*	*Nature in the Amazons*	*The Malaicyan tiger in danger*	*The ocean and the wild life*	*The exterior politic*
Categoy	Culture	Environment	Environment	Environment	Politics

Table 3. Indexes by category consulted.

Category	*Environment*	*Politics*	*Culture*
index	0.6	0.2	0.2

Table 4. Query categories.

Category	Environment	*Politics*	*Culture*
index	0.6	0.2	0.2

The expression (1) is applied to two scenarios in which the relevance $R(u,q,d)$ differs in its values:

$$R(u,q,d) = \alpha\, SCD(q,d) + \beta\, RCD(q,d) + \gamma\, RPUD(u,d) \quad (1)$$

where *u* is the user's search profile, *q* is the query inserted by the user, *d* is a document to calculate the relevance, *SCD (q, d)* in [0,1], *RCD (q)* in [0,1], *RPUD* <= 1 and $\alpha + \beta + \gamma = 1$. Notice that R (d) <=1, being 1 the maximum relevance value for a document.

The parameters used in (1) are the following:

- **SCD:** Similarity between the user's query *q* and the document *d*. To calculate this value, it is proposed to use the cosine formula considering in the Vector Space Model.
- **RCD:** A matching function between the user's query *q* (*QC*) and the document *d* (*DC*), expressing the relevance with regard to their categories.
- **RPUD:** A matching function between the profile of the user *u* (*USP*) and the document *d* (*DC*), expressing the relevance with regard to their categories.

The relevance calculation for each document is defined by executing a series of steps that are described in two procedures (Rule 1 and Rule 2). These steps constitute the functioning of the algorithm proposed for the calculation of relevance.

Rule 1

```
Input:
USP – Set of (category, value) of the user profile.
DC – Set of (category, value) associated to a document.
Output: RPUD

$I = \{c \mid (c,*) \in USP, (c,*) \in DC\}$
if $I == \emptyset$ then RPUD = 0
else
    $(A,B) = \arg\max_{(a,b)} \{b \mid (a,b) \in USP, a \in I\}$
    $(A,C) \in DC$
    if $B > C$ then RPUD = C    // Example 1
    else RPUD = B               // Example 2
```

Rule 2

```
Input:
USP – Set of (category, value) of the user profile.
DC – Set of (category, value) associated to a document.
Output: RPUD

$I = \{c \mid (c,*) \in USP, (c,*) \in DC\}$
if $I == \emptyset$ then RPUD = 0
else                                        // Example 3
    $AB = \{(a,b) \in USP \mid a \in I, b = \max\{v \mid (*,v) \in USP\}\}$
    $(*,B) \in AB$
    $C = \max\{c \mid (a,c) \in DC, (a,*) \in AB\}$
    if $B > C$ then RPUD = C
    else RPUD = B
```

As a result of applying the algorithm, a list of documents that may be relevant for the user is obtained. Several documents can share the same similarity value and the user preferences profile is used to order the results. The algorithm returns a list of documents ordered by relevance to the user.

Presentation of the information

Using the user experience techniques, the system interface is designed where the user can perform the query and obtain the results. The information retrieval system has a simple and advanced search that complies with user-centered design principles. In the simple search the user enters the question and the most relevant results are shown. The advanced search allows the user a higher level of personalization of the results using one of the following filters:

- With any of the words: returns results that contain one or some of the words in the search criteria.
- With all the words: return results that specifically contain all the criteria words.
- With the exact phrase: returns results that specifically contain the exact phrase entered in the search criteria.
- Site: search results by defining the website or domain.

Validation of the information retrieval process

In the validation of the information retrieval process with semantic annotation, quantitative and qualitative methods were used. An experiment is designed and applied to evaluate the IR process, where 30 users who frequently use an IRS were selected. Users are assigned a preference profile and a search query. A process of selection of 290 documents was carried out and the method of semantic annotation and categorization was applied. For each query, the relevant documents were identified and users were asked to insert the queries in the Orion Recovery System (without semantic processing and with semantic processing).

The users selected the documents that in their opinion were relevant and recorded the data. To process the data, the metrics defined in [2] are used to determine the relevance of the search results obtained. The result of applying the metrics is shown in Table 5, the precision value (P) and Recall (E) improves considerably in relation to information retrieval before applying semantic processing. The results obtained for Precision and Recall were greater than 0.8, allowing to corroborate that the process of information retrieval with semantic annotation improves the quality of the results.

Table 5. Experiment results.

P sin procesamiento	P con procesamiento	E sin procesamiento	E con procesamiento
0.50	0.84	0.45	0.80

The precision values obtained were acceptable, corroborating that the recovery of information with semantic annotation improves information retrieval. Additionally, an expert consultation was conducted where the agreement showed a high level of satisfaction with the application of the proposed model. The evaluation using the metrics and the consultation of the experts demonstrates the quality, relevance and relevance of information retrieval with semantic annotation. Allowing to adjust the most relevant results to the needs of the user, increasing their experience in the use of semantic information recovery systems.

CONCLUSIONS

- The analysis on the process of information retrieval allowed identifying the main overloads of information overload, the heterogeneity of information sources and interoperability that make the adequate processing of available information difficult.
- The use of a component for the crawling-indexing, processing and presentation of information allowed retrieving relevant information for users.
- The calculation of relevance using semantic similarity allows improving the process of information retrieval.
- The validation of the model using the Accuracy and Comprehensiveness metrics and the consultation of experts allows to verify the quality of the results obtained.

REFERENCES

[1] Baquerizo, R. P., et al. Algorithm for calculating relevance of documents in information retrieval systems. International Research Journal of Engineering and Technology. 2017, 4(3). pp. 1-5.

[2] Viltres Sala, H. et al. Procesamiento Semántico de información en Sistemas de Recuperación de Información. Revista Cubana de Ciencias Informáticas, 2018, vol. 12, no 1, p. 102-116.

[3] Ghaleb Alshaweesh,O. Intelligent Personalized Approaches for Semantic Search and Query Expansion,Thesis Degree of Doctor, Faculty of Engineering & Information Technology,University of Technology Sydney, 2019

[4] Rodríguez Leyva, P, et al. Modelo computacional para el desarrollo de sistemas de recuperación de información. Revista Cubana de Ciencias Informáticas, 2018, vol. 12, no 1, p. 173-188.

[5] Vuotto, A.; Bogetti, C. y Fernández, G. Application of TF-IDF factor in the semantic analysis of a documentary collection, biblios, 2015, vol 60, p. 1-13.

[6] Salton, G. y Mcgill, M. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., 1983.

[7] Gonzalo, C. et al. Recuperación de información centrada en el usuario y SEO: Categorización y determinación de las intenciones de búsqueda en la Web. [Consultado el: 15 de enero de 2017] Disponible en: http://journals.sfu.ca/indexcomunicacion/index.php/indexcomunicacion/article/download/197/1

[8] Berners Lee, T. et al. “The semantic web,” Scientific american, vol. 284, no. 5, pp. 28-37, 2001

[9] Martínez-Fernández,J. L. et al. Búsqueda semántica a través del Procesamiento de Lenguaje Natural, 2010 p. 2-3.

[10] García Moreno, C. "Desarrollo de un modelo para la gestión de la I+D+i soportado por tecnologías de la Web Semántica" ,2015.

[11] Redondo, S. ¿Qué es la búsqueda semántica y por qué me debe importar? [Consultado el: 15 de marzo de 2017] Disponible en: http://www.senormunoz.es/SEO-MARBELLA/que-es-la-busqueda-semantica-y-por-que-me-debe-importar

[12] Goossen, F. et al. News personalization using the CF-IDF semantic recommender. En Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011. p. 10.

[13] Rodríguez García, M. A., et al. Creating a semantically-enhanced cloud services environment through ontology evolution. Future Generations in Computer Systems, 32, 2014, p 295–306.

[14] Otero García, E. N. Descubrimiento de grafos en datos enlazados para la anotación semántica de documentos. Thesis Degree of Doctor, Universidad de Santiago de Compostela, Galicia, España, 2017.

[15] Gruber, .T. R. “A Translation Approach to Portable Ontology Specifications”. Knowledge Acquisition, 5(2), 1993. pp.199-220.

[16] Niño Zambrano, M. A., et al. Modelo Semántico de Expansión de Consultas para la Búsqueda Web (MSEC). Revista UIS Ingenierías, 2013,11 (1), pp.1-10. ISSN 2145-8456

[17] Peng, L; Lai, Ming-ming; Zhang, X. Research on Semantic Information Retrieval Model of Bamboo & Rattan Domain Based on Query Extension. En Journal of Physics: Conference Series. IOP Publishing, 2019. p. 052093.

[18] Pesántez Peñafiel, C. Y Modelo de gestión por procesos basado en la Norma ISO 9001: 2008 aplicado a la Empresa Compufácil. Tesis de maestría. Universidad Politécnica Saleslana, Quito, Ecuador, 2016.