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 context of the information. 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

Your goal is to simulate the appearance of papers published in *IEEE conference proceedings* [1], with changes to style of the author-institution-email sections, as shown here. Any questions should be sent to the technical committee chair, email can be found in LACCEI's MyReview submission site.

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 Deco, Reyes and Bender (2012) 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.

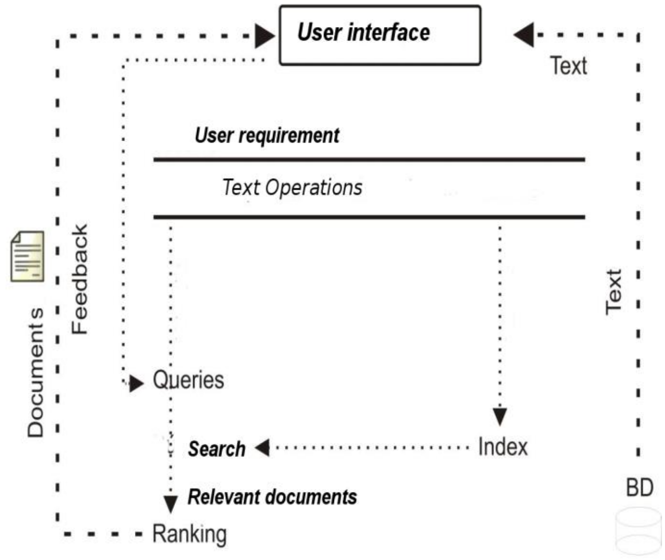

Figure 1: Information search process – source: Vuotto, Bogetti y Fernández, 2015

Authors such as Salton and Mcgill (1983), Gonzalo, et al. (2017) and Martínez (2004) 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

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.308
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

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. Berners-Lee, et al. (2001) 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. Berners-Lee, et al. (2001) 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 Martínez, et al. (2010) 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 (Martínez, et al., 2010).

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 Vuotto, Bogetti and Fernández (2015), Berners-Lee, et al. (2001), Martínez, et al. (2010), García (2015) and Redondo (2017) 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 Redondo (2017) 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.

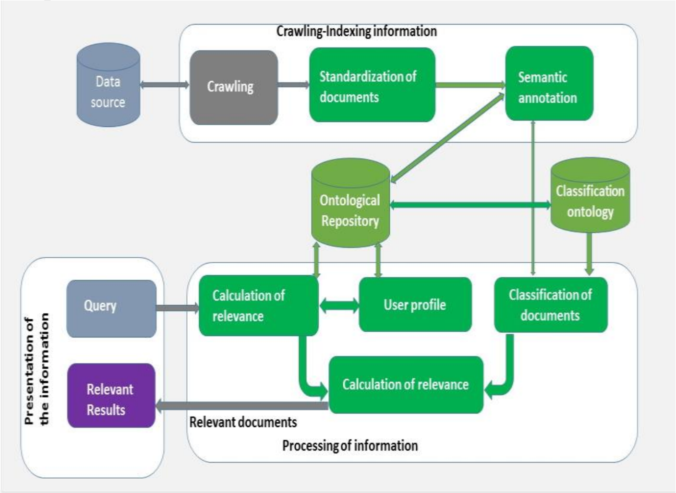

Figure 2: Semantic information retrieval (own elaboration)

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 (Goossen, et al., 2011, García, 2015).

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 Gruber (1983) 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.

User profile

It allows to generate and update the user's profile according to their implicit and explicit preferences using several elements (selected categories in their profile, search history and user's location) to obtain a better result when a user performs a search.

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.

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 Baquerizo (2017) 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.

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

To validate the information retrieval process with semantic annotation, quantitative and qualitative methods were used. An experiment was designed and the Accuracy and Exhaustivity metrics were applied. In the experiment, the result given to the questions asked by the users was analyzed using an SRI without semantic processing and the proposed model. The results obtained for Precision and Comprehensiveness were superior to 0.8, demonstrating that the information retrieval process with semantic annotation improves the quality of the results.

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

[3] Deco, C.; Reyes, N. y Bender, C: Recuperación de Información en Bases de datos no estructuradas, XIV Workshop de Investigadores en Ciencias de la Computación, 2012

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

[5] Gonzalo, C.; Codina, L., 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

[6] Goossen, Frank, 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.

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

[8] Martínez Méndez, F. J. Recuperación de información: modelos, sistemas y evaluación. Murcia, KIOSKO JMC, 2004. 106 p.

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

[10] 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

[11] 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.

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

[13] 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.