International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

Model for semantic processing in information retrieval systems

Ph.D Roberto Passailaigue Baquerizo[1], MSc. Hubert Viltres Sala[2], Ing. Paúl Rodríguez Leyva[3], Ph.D Vivian Estrada Sentí[4]

[1]Canciller Universidad Tecnológica (ECOTEC)
Guayaquil, Ecuador
[2]Departamento de Práctica Profesional,
Universidad de las Ciencias Informáticas,
La Habana, Cuba
[3]Departamento de Soluciones Informáticas para Internet,
Universidad de las Ciencias Informáticas,
La Habana, Cuba
[4]Departamento Metodológico de Postgrado
Universidad de las Ciencias Informáticas,
La Habana, Cuba

Abstract - *The processing of information with semantic annotation allows to identify the intention of search of the users 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 recover the most relevant information among all the information available on the web. In the model, three components (Trace-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 web sites to extract information and perform semantic annotation by applying different information processing techniques. The processing component analyzes the preferences of the user 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, the metrics of precision and completeness were used to demonstrate the quality and relevance of the information retrieval with semantic annotation.*

Key Words: Semantic Web, information retrieval, processing, relevance, semantic annotation, similarity

1. 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 enabled a large volume of web content to be generated. The information available on the web is dispersed, poorly structured or invisible to the common user, making it difficult to access information of high quality and value to the user. In this context, users when they access the Internet are overwhelmed by information overload and do not easily and quickly obtain the information that best suits their needs, limit their experience in the use of information retrieval systems.

There are more than a trillion websites on the Internet and every day there is an exponential increase in the amount of information available. Generating new opportunities and different 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 it, users rely on information retrieval systems (IRS) to find what they are looking for.

Information retrieval systems using different tools, methods and techniques retrieve public information from the web for later analysis, selecting and ordering the most relevant information for the user's needs.

Among the main sources of information are the component repositories, databases and search engines that allow to simplify and group relevant information, using certain concepts of information organization. The main objective of an SRI as proposed in [1] is to satisfy the user's need for information in a natural language query specified through a set of key words (see figure 1), which help identify the most relevant to the user.

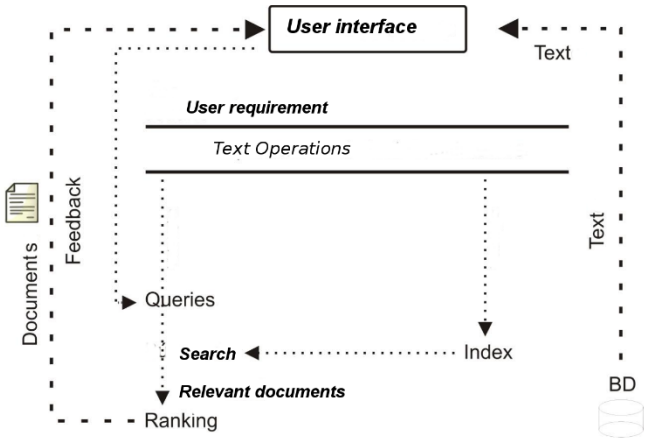

Figura-1: Source information search process [2]

Several authors in [3], [4], [5], suggest that the Information Search and Recovery has as main objective to provide relevant information to the user to satisfy their information need. Within the BRI, five main activities are defined (locating, selecting, interpreting, synthesizing and communicating information) to guide the process of obtaining information tailored to the user's needs. These five activities are covered in the three main components of a search engine today (crawler, indexer and processor).

During the information retrieval process, traditional search engines generally use techniques that determine relevance by matching keywords in documents and do not analyze the relationships between the implicit meaning of keywords and the document. For them it is necessary to carry out a process of identifying the user's intention behind the question asked and adjust the result to the context of the question.

Several authors argue that the semantic retrieval of information improves the quality and relevance of the information shown to users, since it uses natural language processing techniques, uses ontologies to identify the context and the relevance is established by the semantic similarity of the query And indexed documents.

1.1 Semantic retrieval of information

The Semantic Web is changing the way of obtaining information on the Internet, it is one of the technologies that have generated the most impact for Internet users because of the quality of the information they get. Berners-Lee in [6] 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 Allowing data stored on the Web to be intelligently processed by the machines, making it easier for people to search, integrate and analyze available information.

The semantic web has as principle the processing of information automatically by the use of artificial intelligence using a great variety of algorithms. It also aims to understand the need expressed by the user in a query performed and provide the search for meaning, identifying and providing reliable information. To perform the semantic search semantic search engines are used that 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 [7] is associated to the following characteristics:

- ✓ Performs field searches.
- ✓ Has ability to extend query terms using synonyms or related words.
- ✓ Identifies named entities, such as company names, organizations or individuals that are used with that meaning in the search process.
- ✓ Uses grouping techniques to construct categorizations of content on which to search or 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 queries and even factual questions, for which concrete answers are obtained [7].

The characteristics discussed above demonstrate the semantic web's possibilities in retrieving information where a user expresses in natural language his or her search intention and the searcher analyzes and selects the information adjusted to that need. In the context of the Cuban web where technological limitations difficulty the information retrieval process to solve this problem it is necessary to employ the retrieval of semantic information.

1.2 Information retrieval on the Cuban web

In Cuba there are more than 6 thousand websites hosted under the .cu domain with varied information. In order to access the information stored on the Cuban web, users use different information retrieval systems but do not always obtain relevant information, mainly due to:

- ✓ Heterogeneity of sources of information.
- ✓ Quality of information.
- ✓ Visibility of information.
- ✓ Accessibility of information.

In addition to the above mentioned elements another factor that affects the information retrieval is the use of systems that use algorithms to calculate the similarity by words, where the semantics of the information is not analyzed. An analysis of the systems that determine the similarity by keywords showed the following deficiencies:

- ✓ Difficulty understanding the user's need expressed in natural language.
- ✓ Low accuracy of results because the similarity of keywords is enhanced.
- ✓ Sensitivity of the results against the exact terms introduced.
- ✓ Selection of the information by the relevance of the positioning of the website.

The above difficulties show little exactitude and accuracy in the information retrieval process and decrease the user experience when performing a search for information. These deficiencies coupled with the need to provide users with high-quality information raises the need to develop an information retrieval system with semantic annotation that allows the selection of information that is more adjusted to the needs of users and thereby improve their experience in the Cuban web.

1.3 Semantic search of information

The semantic web is an extension of the current web, several authors [2] [6] [7] [8] [9] [10] suggest that information can be efficiently obtained by integrating, automating and reusing data using various techniques to Improve the relevance of the information collected. Semantic searches provide relevant results by understanding the need for user information expressed in natural language.

According to Redondo in [8] the aim of semantic search is to improve the accuracy of the search by understanding the user's intention when making a query and the contextual meaning of the data in the knowledge source. Semantic search predicts what the user explicitly expresses (search intent) and adjusts their need (context) to available information by selecting the most relevant one for the user.

Information retrieval systems focus their implementation on understanding search using query processing, extracting knowledge from data sources, adjusting user preferences, and calculating relevance. The model proposed in the research is based on the retrieval of relevant information for the user using semantic technology.

2. METHODOLOGY

In order to obtain relevant information for users, a computational model is implemented that allows the processing of the information available semantically. In the model, the three main components (Tracking-Indexing, Processing and Presentation) are considered; which will identify the need for user information through the processing, selection and subsequent publication of the retrieved information. Figure 2 presents the components that support the process of searching and retrieving information on the web. Each of the three components is described below.

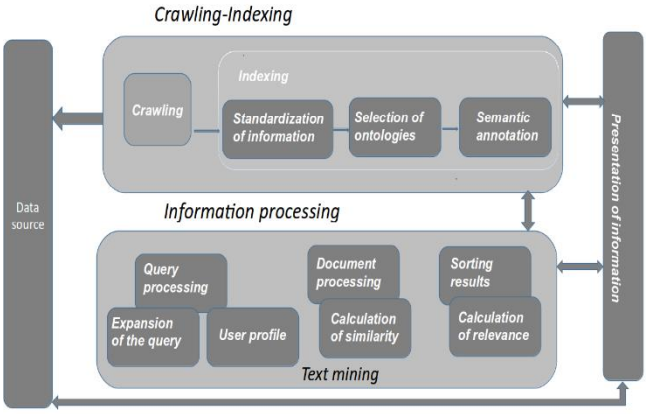

Fig - 2: Computational model for the semantic processing of information (own elaboration)

2.1 Tracking and indexing component

The crawling and indexing component allows the identification of available web sites, as well as retrieving and storing information from each web page for further processing and presentation to users when making a query.

The crawlers are in charge of exploring the web identifying the pages that have been created or updated to continue updating its index of information. After tracking different metadata (url, content summary, links, keywords, language) are stored that are used to extract knowledge using semantic web techniques.

2.1.1 Tracking the web

The crawl process starts with a list of links to websites provided by previous crawls or sitemap; The greater the number of links the 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 carry out the Nutch tracing in a distributed way using the policies of selection, re-visit, courtesy and parallelization that allow a thorough search. The crawler configuration determines which sites to crawl, how often and how many pages to scan on each site (Google, 2017).

2.1.2 Indexing the web

After performing the tracking process, each web page is analyzed to identify the main elements and then store the information and create an index of contents that allows to improve the information retrieval process. In the indexing process, the information tracked is standardized, defining the necessary metadata for the processing of the information.

Subsequently, the knowledge graph is generated by extracting from each page the content according to the context through the use of a general ontology and a specific one according to the category of the web page. Solr and Apache Jena use different techniques and algorithms to extract the implicit knowledge of web pages.

Solr implements the vector space model and uses an inverted file system to create the index; In addition to performing the normalization process has multiple analyzers and can define own analyzers [11]. For semantic reasoning the information uses Apache Jena which provides an API for reading, writing, extracting and processing RDF graphs. It also has an inference engine to reason about ontologies and to perform queries with SPARQL specification. In addition, the algorithm CF-IDF (concept frequency - reverse document frequency) is used for the creation of the index based on the annotations made, which according to [12] and [9] improves the information retrieval process.

2.2 Processing component

It is responsible for processing and analyzing texts in natural language by associating each sentence of a text with a semantic representation based on an ontology with thousands of words, where words are categorized according to the different meanings they have and where the relationships between them are defined.

Gruber defines an ontology in [13] as "an explicit specification of a conceptualization" that allows to add 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 words and add a natural meaning to it. The use of Ontologies makes it possible to improve the natural language processing of the query performed by the user and the information collected by the crawlers on the web.

2.2.1 Query processing

Users when accessing an information retrieval system formulate the questions in natural language. In order to understand the intention behind the question asked, different techniques need to be processed and applied to identify the user's need for information. The query processing has as main objective the disambiguation of the terms entered by the user generating as output a triplet in RDF format.

2.2.2 User Profile Processing

It allows you to generate and update the user profile according to your implicit and explicit preferences using various elements (categories selected in your profile, search history and user location) to get better results when a user performs a search.

2.2.3 Calculation of similarity

In order to determine the similarity between the query performed by the user and the information indexed in the searcher, the results of the query processing, the user profile processing, and the relevance index of the semantic annotation performed during the storage process of information.

The similarity is determined using Levenshtein's algorithm for short texts and the cosine function.

2.2.4 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 [14] [15] is used to determine the relevance coefficient according to the user profile, the query and the semantic similarity index.

The relevance coefficient obtained is used to order the results and show a number of elements that can be assimilated by the user.

2.3 Presentation component

Employing 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 search and an advanced search that comply with the principles of user-centered design. In the simple search the user enters the question and shows the most relevant results. Advanced search allows

the user a greater level of customization of 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 words:** return results that specifically contain all the words in the criterion
- ✓ **With the exact phrase:** returns results that specifically contain the exact phrase entered in the search criteria
- ✓ **Site:** allows you to search for results by defining the websites or domain

2.4 Validation of the model

In the evaluation of the proposed model, we used the Precision and Completeness metrics that allow us to check the quality of the results obtained. For the validation an experiment was designed on the information published on the Cuban web. In the experiment we analyzed the results provided to the questions formulated by the users using an SRI without semantic processing and the proposed model.

The precision values obtained were 8.3 and exhaustiveness of 8.5, corroborating that the retrieval of information with semantic annotation improves the retrieval of information. In addition, an expert consultation was conducted where the concordance showed a high level of satisfaction with the application of the proposed model.

The evaluation using the metrics and the expert consultation demonstrates the quality, relevance and relevance of the 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 systems of retrieval of semantic information.

3. CONCLUSIONS

The analysis of the information retrieval process identified as the main deficiencies the overload of information, heterogeneity of information sources and interoperability that greatly hinder the adequate processing of available information.

The use of a component for the tracing-indexing, processing and presentation of the information allowed to retrieve relevant information for the users.

The calculation of the relevance using the semantic similarity allows to improve the information retrieval process.

The validation of the model using the metrics of Precision and Completeness and the consultation of experts allows to check the quality of the obtained results.

REFERENCES

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

VUOTTO, A.; BOGETTI, C. y FERNÁNDEZ, G. Application of TF-IDF factor in the semantic analysis of a documentary collection, biblios, 015, vol 60, p. 1-13.

SALTON, G. y MCGILL, M. Introduction to ModernInformation Retrieval. McGraw-Hill, Inc., 1983.

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/197/175

MARTÍNEZ MÉNDEZ, F. J. Recuperación de información: modelos, sistemas y evaluación. Murcia, KIOSKO JMC, 2004. 106 p.

BERNERS-LEE,T. et al. "The semantic web," Scientific american, vol. 284, no. 5, pp. 28-37, 2001

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

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

GARCÍA MORENO, C. "Desarrollo de un modelo para la gestión de la I+D+i soportado por tecnologías de la Web Semántica" ,2015.

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.

MONTERO PUÑALES, E. M. y PLACENCIA SALGUEIRO, A. Sistema de recuperación y análisis de información para

investigadores del Instituto Investigativo ICIMAF. INFO 2016, 2016, p 2-15.

[1] 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.

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

[15] PASSAILAIGUE Baquerizo, R., et al. Algorithm for calculating relevance of documents in information retrieval systems. International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 3, Marzo. 2017. e-ISSN: 2395-0056.

BIOGRAPHIES

Ph.D Education, Canciller University ECOTEC, Ecuador

MSc. En Informática. Universidad de las Ciencias Informáticas, La Habana.

Ph.D Computing, Adviser postgraduate, Habana, Cuba

Ing. en Informática. Jefe de departamento de Soluciones Informáticas para Internet. Habana