World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:2, 2015

Semantic Indexing Approach of a Corpora Based On Ontology

Mohammed Erritali

***Abstract*—**The growth in the volume of text data such as books and articles in libraries for centuries has imposed to establish effective mechanisms to locate them. Early techniques such as abstraction, indexing and the use of classification categories have marked the birth of a new field of research called "Information Retrieval". Information Retrieval (IR) can be defined as the task of defining models and systems whose purpose is to facilitate access to a set of documents in electronic form (corpus) to allow a user to find the relevant ones for him, that is to say, the contents which matches with the information needs of the user. This paper presents a new semantic indexing approach of a documentary corpus. The indexing process starts first by a term weighting phase to determine the importance of these terms in the documents. Then the use of a thesaurus like Wordnet allows moving to the conceptual level.

Each candidate concept is evaluated by determining its level of representation of the document, that is to say, the importance of the concept in relation to other concepts of the document. Finally, the semantic index is constructed by attaching to each concept of the ontology, the documents of the corpus in which these concepts are found.

***Keywords*—**Semantic, indexing, corpora, WordNet, ontology.

I. Introduction

Due to the rapid growth in the volume of electronically stored information, the major problem which arises is to respond to a search query with relevant manner from a set of unstructured documents in a database called the corpus. This research problem is known as Information Retrieval (IR).

Information Retrieval can be defined as a set of techniques and tools dealing with access to information and its presentation, its organization and its storage [1], [2]. The term "information retrieval" is given by Calvin N. Mooers in 1948 for the first time in his thesis [3].

According to [18], an IRS is a set of computer programs that aims to select relevant [9] information that meets users' needs expressed in the form of queries. Lancaster cited in [19] notes that an IRS does not inform the user on the subject of his research .it simply reports the existence or non-existence of documents relating to his request.

From the above definitions, we can deduce that a user translates its needs in a structured way as a query that it transmits to the information retrieval system.

This one has as a main task to return to the user the maximum of relevant documents in relation to his need (minimum of irrelevant documents). For this, the information search system connects the available information (the corpus documents) and the requirements of the user (the user query).

In the literature we find several representations of the process of information retrieval [16], [17], [20]-[22] which show that the mapping of the information contained in a corpus on the one hand, and information needs of users, on the other hand, is done through two mechanisms: indexing and retrieval (search).

This operation is provided through a process known as the process in U as shown in Fig. 1.

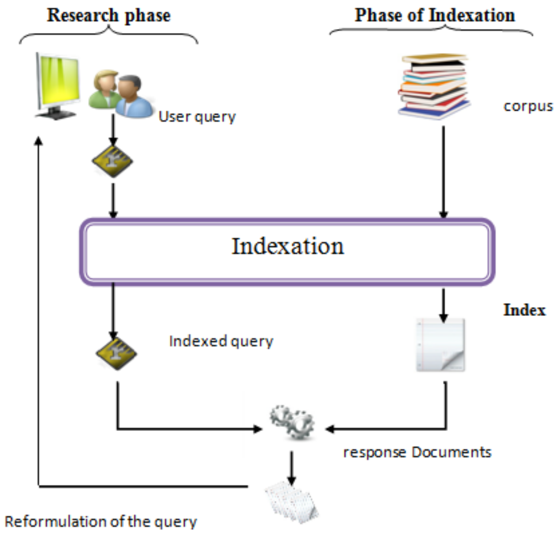


Fig. 1 Process of Information Retrieval

These IR systems are mainly based on statistical approaches and linguistic approaches of bottom level. These approaches take into account only the lexical or syntactic level, of the textual content of documents to identify words allowing finding documents that meet the needs of the user. The question that arises now is: How to find the rare and hidden document that contains concretes information? How to select documents with information relevant to specific goals?

In this context, the use of ontologies for the expansion of user queries can be a solution to remedy effectively to these problems [27], [28]. On the one hand, ontologies provide resources generally in the form of semantic relations [11], [12] in order to identify the meaning of words in the query and process this query to "expand" the search field. Moreover, they constitute a shared framework (common vocabulary) that the different actors can mobilize [28], [29] to approximate the language of the queries and the documents.

M. Erritali is with TIAD laboratory, Computer Sciences Department, Faculty of sciences and techniques, Sultan Moulay Slimane University, Beni-Mellal, BP: 523, Morocco (phone: +212670398838; e-mail: m.erritali@usms.ma).

In our work we will present a semantic indexing method based linguistic ontology built from Wordnet [30] [31].

II. Automatic Indexing Process

Automatic indexing is a completely automated process that is charged to extract words that characterize the document [14]. Automatic indexing is a set of automated processes on a document which are: segmentation, removal of empty or stop words, stemming or radicalization of words, and weighting.

1) The Tokenization (Segmentation)

The tokenization is also called segmentation. It consists to divide the text into elementary tokens. This is an operation which "locates" strings surrounded by separators (white space, punctuation), and identifies them as words.

2) Elimination of Empty Words (Stop Words)

Stop words (empty words) are prepositions and conjunctions. Elimination of empty words reduces the index, then we gain in storage space, but also the no treatment of empty words reduces the execution time of a System of information retrieval [8]. Seen that reducing the number of terms increases the performance, some systems consider, too, such as empty words some verbs, adjectives and adverbs. There are two techniques to filter out empty words:

- ✓ The use of a predefined list of stop words (also called anti-dictionary / stop-list).
- ✓ Count the number of occurrences of words in a document collection. Followed by striking words which their frequency exceeds a certain threshold.

3) Normalization of Index Terms

Normalization is a process that allows grouping the morphological variants of words as a single base. Its goal is to keep in the indexing language, the forms of representative words, which offers considerable gain of storage memory and an effective research. The normalization is based on one of two procedures: Stemming or lemmatization. [13].

a) Lemmatization

Lemmatization is used to group the words of the same grammatical category and transform them to their canonical form called lemma (e.g. different forms of a verb are transformed to infinitive) [10], [7]. This technique is based on the use of software and resources on lemmatization namely: TreeTagger, WinBrill and LEFFF.

Some lemmatizers can treat multiple languages (e.g. TreeTagger treats the English and German languages).

b) Stemming

Stemming transforms a word to its root. A stemmer seeks the root of a word based on its shape and the desired language. For example in French: "écologie, écologiste, écologique" are stemming by one word: "écologie" [13] [8].

In the literature there are several algorithms that are used in stemming as the algorithm of Lovins [23], Paice / Husk [24] algorithm and Porter [25] algorithm.

Snowball [26] is another Stemming tool which was invented by Martin Porter (the creator of the Porter algorithm). There are Snowball stemmers for various languages (French, English, Spanish …)

Experiments have shown that the Stemming and lemmatization significantly increase the search performance for morphologically rich languages such as French and Italian [13].

4) Weighting of Terms

To measure the importance of a word in a document indexing uses the concept of weight. The weighting is to assign a weight to terms of indexing and search. This weight is used to specify the relative importance of words represented in the documentation with respect to those identified in the request. The weighting consists to answer the question if all terms have the same importance and how to assign a weight to the extracted terms?

In general, the weighting formulas used are based on the combination of a local weighting factor quantifying the local representation of the word in the document [4], and a global weighting factor quantifying the overall representation of the term with respect to the collection of documents [10] [2].

✓ Local Weighting

Local weighting is used to measure the local representation of a term. It takes into account the local information of the term in relation to a given document. It indicates the importance of the term in this document. This weighting is generally measured by the frequency of the term $\mathrm{t_j}$ (term frequency, denoted $\mathrm{tf_{ij}}$ in the document $\mathrm{d_i}$considered.

✓ Global Weighting

The global weighting is based on the idea that a term does not distinguish the documents from each other during the search if it is distributed uniformly in all documents in the collection. Thus, this term does not have any discriminatory power. Therefore, the terms that appear in few documents are discriminating and weights are assigned to them. This weighting is expressed by the inverse document frequency $\mathrm{idf_j}$ of a term $\mathrm{t_j}$ in the collection. It is generally defined by:

$$\mathrm{idf_j} = \log(\frac{\mathrm{N}}{\mathrm{n_j}})$$

where N is the number of documents in the collection; $\mathrm{n_j}$ is the number of documents indexed by the term $\mathrm{t_j}$.

Salton [6] has defined a weighting formula tf * idf by:

$$w_{ij} = tf_{ij} * idf_j = tf_{ij} * \log(\frac{N}{n_j})$$

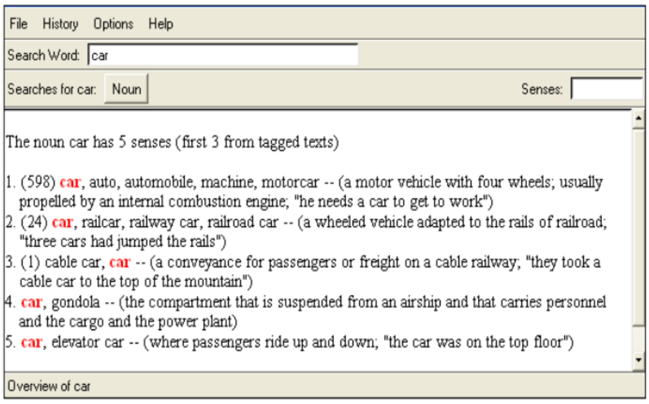
The measure tf * idf is a good approximation of the importance of a term in the document collections composed of document with homogeneous sizes. However, for collections containing documents of varying sizes, words in longer documents appear frequently with very high weight than those in shorter documents. So the longer documents are more likely to be selected [15], [2], [13].

III. Construction of a Linguistic Ontology

The term ontology is used in the field of the Semantic Web and refers to a structured set of concepts in a particular field of knowledge [5], [7].

As part of treatments using texts written in natural language, as the case of information retrieval is therefore necessary to determine the set of synonyms (candidates labels) to uniquely define a concept.

We propose in this phase a process to determine all possible labels of a concept. This process, based on the WordNet thesaurus, to make a match between the paths according to the relationship "is" ("ISA") in the ontology and the paths of synsets in the thesaurus. Version 2.0 of WordNet [30], [31] defines the common name 'car' with five synsets as shown in Fig. 2.



Search Word: car

Searches for car: Noun

Senses:

The noun car has 5 senses (first 3 from tagged texts)

1. (598) car, auto, automobile, machine, motorcar -- (a motor vehicle with four wheels; usually propelled by an internal combustion engine; "he needs a car to get to work")
2. (24) car, railcar, railway car, railroad car -- (a wheeled vehicle adapted to the rails of railroad; "three cars had jumped the rails")
3. (1) cable car, car -- (a conveyance for passengers or freight on a cable railway; "they took a cable car to the top of the mountain")
4. car, gondola -- (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
5. car, elevator car -- (where passengers ride up and down; "the car was on the top floor")

Overview of car

Fig. 2 The different synsets of the word car

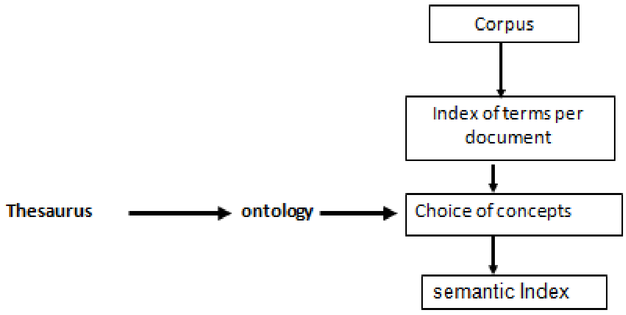


Fig. 3 Process of Semantic Indexing

In our work, the built ontology is a hierarchy of concepts each represented by a term (a label) and a set of synonyms for this term. Some concepts are linked by specific / generic relation, or part / all relation.

IV. Process of Semantic Indexing

The objective of the process is to build a semantic index of a corpus based on a linguistic ontology as illustrated in Fig. 3.

The indexing process includes the following steps:

- First, for each of the documents we build an index of terms with their respective frequency (number of occurrences)
- Then the ontology allows determining all concepts associated to the terms previously acquired. This allows building the index by linking the document to the ontology concepts that it contains, with a weight equal to the cumulative frequency of its synonyms.

V. Pairing Document / Query

The treatment associated to pairing involves two stages:

- Indexing of the query in the same way as the documents of the corpus for the determination of concepts which gives a result as shown in Table I.

TABLE I
Indexing of the Query

Concept 1	Concept 2	Concept 3
2	1	3

- To match the representation of documents with that of the query using scalar product presented by:

$$sc\,[j] = sc\,[j] + req\,[i] * tf\,[j]\,[i].$$

where tf is the weight of each concept for document j

In this step as the value of the similarity measure (scalar product) is greater as the document is more relevant to the query.

VI. Conclusion

In this article, we presented a process for indexing a documentary corpus. This process helps to build a semantic index based on a linguistic ontology built from WORDNET. After the construction of the index of each document that contains the terms with their frequency, we seek to identify the concepts of these terms.

For each concept obtained we calculate its weight which is the accumulated weight of its synonyms in the document. Finally, these concepts are associated with a set of documents in which they are located.

References

[1] Ricardo B Y., Berthier R N. Modern information retrieval, ACM (Association for Computing Machinery).
[2] Baziz, M. (2005). Indexation conceptuelle guidée par ontologie pour la recherche d'information (Doctoral dissertation, Toulouse 3).
[3] Mooers, C. N. (1948). Application of random codes to the gathering of statistical information (Doctoral dissertation, Massachusetts Institute of Technology).
[4] KARBASI, S. Pondération des termes en Recherche d'Information (Doctoral dissertation, Toulouse 3).
[5] Harrathi, F. (2009). Extraction de concepts et de relations entre concepts à partir des documents multilingues: approche statistique et ontologique.
[6] Salton, G. (1969). A comparison between manual and automatic indexing methods. American Documentation, 20(1), 61-71.
[7] Mallak, I. (2011). De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information (Doctoral dissertation, Université Paul Sabatier-Toulouse III).
[8] Aouicha, M. B. (2009). Une approche algébrique pour la recherche d'information structurée (Doctoral dissertation).
[9] Barry, C. L. (1994). User-defined relevance criteria: an exploratory study.JASIS, 45(3), 149-159.

[10] Boubekeur-Amirouche, F. (2008). Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets (Doctoral dissertation, Université de Toulouse, Université Toulouse III-Paul Sabatier).

[11] Roussey, C. (2001). Une méthode d’indexation sémantique adaptée aux corpus multilingues. Institut National des Sciences Appliquées de Lyon Lyon, Ecole Doctorale Informatique et Information pour la Société.

[12] Azzoug, W. (2014). Contribution à la définition d’une approche d’indexation sémantique de documents textuels.

[13] Porter, M. F. (1980). An algorithm for suffix stripping. Program: electronic library and information systems, 14(3), 130-137.

[14] Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1995, November). New retrieval approaches using SMART: TREC 4. In Proceedings of the Fourth Text Retrieval Conference (TREC-4) (pp. 25-48).

[15] Brini, A. H. (2005). Un modèle de recherche d'information basé sur les réseaux possibilistes (Doctoral dissertation, Toulouse 3).

[16] Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM), 7(3), 216-244.

[17] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

[18] Tebri H. Formalisation et spécification d’un système de filtrage incrémental d’information. Thèse de doctorat de l’université Paul Sabatier, Toulouse, 2004.

[19] V.Rijsbergen C. J. Information Retrieval. Department of Computing Science University of Glasgow.

[20] Iadh O. Un modèle d'indexation relationnel pour les graphes conceptuels fondé sur une interprétation logique, Thèse pour obtenir le grade de Docteur de l'Université Joseph Fourier, 1992.

[21] Piwowarski B, Denoyer L, Gallinari P. Un modèle pour la recherche d’information sur des documents structurés. 6es Journées internationales d’Analyse statistique des Données Textuelles. LIP6, PARIS – France, 2002.

[22] Denos N. Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application. Thèse pour obtenir le grade de Docteur de l'Université Joseph Fourier-Grenoble I, 1997.

[23] http://www.comp.lancs.ac.uk/computing/research/stemming/Links/lovins.htm

[24] http://www.comp.lancs.ac.uk/computing/research/stemming/Links/paice.htm

[25] http://tartarus.org/martin/PorterStemmer/

[26] http://snowball.tartarus.org/

[27] Guarino, N., Masolo, C., & Vetere, G. (1999). Ontoseek: Content-based access to the web. Intelligent Systems and their Applications, IEEE, 14(3), 70-80.

[28] Fabien GANDON, « Ontologie Engineering : a Survey and a Return on Experience », rapport de recherche INRIA, (Mars 2002).

[29] Bachimont, B. (2000). Engagement sémantique et engagement ontologique: conception et réalisation d’ontologies en ingénierie des connaissances.Ingénierie des connaissances: évolutions récentes et nouveaux défis, 305-323.

[30] Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. Indexing with wordnet synsets can improve text retrieval. In Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, pages 38-44, Montreal, Canada, (1998).

[31] http://wordnet.princeton.edu/wordnet/download/

M. Erritali obtained a master's degree in business intelligence from the faculty of science and technology, Beni Mellal at Morocco in 2010 and a Ph.D. degree in Computer Sciences from the faculty of sciences, Mohamed V Agdal University, Rabat, Morocco in 2013. His current interests include developing specification and design techniques for use within Intelligent Network, data mining, information Retrieval, image processing and cryptography.

He is currently a professor at the Faculty of Science and Techniques, University Sultan Moulay Slimane, and also a member of the TIAD laboratory.