Toward a Taxonomy of Concepts using Web Documents

Rim Zarrad
RIADI Laboratory
Team of research SIIVA
University of Tunis El Manar
Tunis, Tunisia
zarrad_rim@yahoo.fr

Narjes Doggaz
URPAH laboratory
Faculty of Sciences of Tunis
University of Tunis El Manar
Tunis, Tunisia
narjes.doggaz@fst.rnu.tn

Ezzeddine Zagrouba
RIADI Laboratory
Team of research SIIVA
University of Tunis El Manar
Tunis, Tunisia
ezzeddine.zagrouba@fsm.rnu.tn

ABSTRACT
Due to the rise of the Web and the need to have structured knowledge, an interesting line for research is the formalization of ontologies and the creation of conceptual taxonomies from Web documents. The traditional methods for ontology learning and especially those extracting domain concepts from a textual corpus often privilege the analysis of the text itself, whether they are based on a statistical or linguistic approach. In this paper, we propose an approach which differs from the traditional ones since it uses information on the document structure to extract relevant information. Our approach studies each material form in the text in order to extract the most relevant concepts constituting the ontology related to a given field. The concepts are obtained by analyzing the occurrences of the candidate terms in the titles and in the links belonging to the documents and by considering the used styles.

Our approach has been experimented on a French corpus of Web documents related to the medical field. Primary results are encouraging and seem to validate our approach. We present also in this paper a new method for the extraction of the hierarchical links between the concepts of the ontology. The taxonomic links are established in three phases: a linguistic step is based on the canonical syntactic structure of the extracted concepts, the second step consists in applying lexico-syntactic patterns which convey the hyperonymy relation and the third step analyzes the hierarchy of the titles in each document to extract taxonomic relations.

Keywords

1. INTRODUCTION
With the growing number of users and information sources available on the Web, the need for developing methods allowing the extraction of the relevant information from HTML documents is obvious.

In fact, information systems have become very complex because of the large volume of multidimensional data they contain; the challenge is no longer to gather data but to extract and visualize relevant information. The collections of textual data make emerge functional needs (data retrieval, document categorization, indexing, constitution of lexical resources, creation of knowledge bases, data mining...). The data retrieval systems mainly aim to extract the semantics contained in documents. We try by these methods to summarize long texts in brief information on the texts significance. Due to the rise of the Web and the need to have structured knowledge, a big part of research concentrates on the formalization of ontologies [2]. Indeed, the huge quantity of resources easily available on the Internet requires the creation of conceptual taxonomies from Web documents which is an interesting line for research.

We distinguish three main ontology building methods: the methods based on the distribution of the linguistic units [3], the ones based on the linguistic approaches [24, 6], and those based on lexico-syntactic patterns [8, 21, 17]. These methods are based on the techniques of natural language processing.

The methods which are based on the distributional processing are robust and do not require preliminary knowledge on the field. They extract the classes of terms which are the concepts of the ontology and organize them in structured systems reflecting a conceptual hierarchy. However, these methods disregard the context of sentences which is necessary to have a precise interpretation of the semantic classes; they are not adapted to have a precise analysis of the corpus. Moreover, these methods tend to extract conceptual links of different natures which are difficult to dissociate without the contribution of an expert of the field. Indeed, it is not always obvious to deduce the relation between the syntags. The elements of the same class can be at the same time synonymous, antonyms, hyponyms or, simply, have a common semantic link. The terminological extraction only based on statistical criteria encounters an additional difficulty due to the various possible terminological variants to denote a concept. Finally, these methods are well performing on large corpora, they can not be applied on a small corpus. The use of such an approach shows many limits but it offers a comprehensive view which is necessary to apprehend a specialized corpus.

The linguistic approach does not consider the corpus in a comprehensive manner but locally. In other words, it is based on the properties of the language to extract the main concepts of the field. However, it is not reasonable to specify a linguistic approach for each new field of study.

The last approaches are based on lexico-syntactic patterns (rules describing a formed regular expression of words and grammatical
categories corresponding to the syntactic forms of the relation and its arguments). These patterns characterize the semantics of the relation. In this lexico-syntactic ontology learning approach the text is scanned for instances of distinguished lexico-syntactic patterns that indicate the relation of interest. Thus, the underlying idea is very simple: Define a regular expression that captures re-occurring expressions and map the results of the matching expression to a semantic structure, such as taxonomic relations between concepts. Although the relations extracted have a high precision, the production is often weak because of the complexity and of the diversity of the patterns who can express the same relation.

In [25], we have proposed an approach to extract the main concepts of the astronomy field. It differs from the traditional ones since it uses information on the document structure. By assigning to each material form a weight, the candidate terms are filtered in order to keep only the relevant terms of the field. It is based on a new measure $\text{CR}_ICF$ which filter the candidate terms to obtain the concepts of the studied field. This measure removes the blank words which occur in the corpus and studies the documents’ structure in order to extract the main concepts of the field. The experiments of the new approach are conducted on the medical field. We present also in this paper a new method for the extraction of the hierarchical links between the concepts of the ontology.

This paper is organized as follows: In section 2, a variety of methods for ontology learning which are based on documents’ structure are described. In section 3, we present the general approach by giving the different steps leading to the taxonomy of concepts. The concept refiner step is described in details in section 4. We give some details on the used corpus and we give some experimental results in section 5. In section 6, we present the current work concerning the extraction of taxonomic links between the obtained concepts. Finally, we give our conclusion and we present some perspectives.

2. RELATED WORKS

Ontology or taxonomy engineering is primarily concerned with the definition of concepts and relations between them. The traditional methods of building taxonomy of concepts from textual corpus focus on the analysis of the text itself. In the last decade, several approaches try to build ontologies by analyzing the material formatting of the texts. Indeed, the visual properties of texts are not just an ornamentation of the text but constitute an important component implied in the significance.

The work in [12] concerns the automatic ontologies learning from databases specifications in XML format. The approach initially consists in studying the XML tags and their hierarchy to extract the concepts and the hierarchical and non hierarchical links between them. Moreover, the remainder of the text is analyzed using lexico-syntactic patterns. The used corpus contains very short and synthetic texts so it is poor in terms of links between concepts. This approach has some inconsistencies in the level of the hierarchical and conceptual links. Moreover, the same concept may be classified in different levels of the hierarchy.

The study made in [14] on a corpus of XML documents corresponding to the specifications of a geographical database generates a taxonomy of concepts. This approach exploits only the visual structure of the text (style, bold characters, underlined, framing...) by making hierarchies of the text elements according to their visual structure. The intervention of the expert is necessary to validate the field concepts.

Building ontologies from semi-structured data uses both traditional data mining and Web content mining techniques. For example, in [13], the authors describe a clustering procedure to extract the concepts of a field. This approach is based on studying the structure of Web pages in order to build a database table. It uses the $\text{TF} \times \text{IDF}$ measure [11]. The Term Frequency $\text{TF}(w,d)$ is the number of times the word $w$ occurs in document $d$. The Document Frequency $\text{DF}(w)$ corresponds to the number of documents in which the word $w$ occurs at least once. The Inverse Document Frequency $\text{IDF}(w)$ can be calculated from the Document Frequency $\text{DF}(w)$ by

$$\text{IDF}(w) = \log \left( \frac{\text{ID} \text{f} \text{l} \text{o} \text{r}}{\text{D} \text{f}(w)} \right)$$

Where $\text{ID} \text{f} \text{l} \text{o} \text{r}$ is the total number of documents of the corpus. The IDF measure indicates the opposite frequency of document. It is a measurement of the importance of the term in the whole corpus. It aims at giving a more important weight to the least frequent terms, considered as discriminant.

Among the methods based on the document structure, several ones focus on the titles belonging to the document. The method presented in [7] was evaluated on a corpus of Web documents in the domain of Agriculture. It uses two complementary approaches to extract the main concepts in the agricultural field and subsequently generates the whole ontology. The first approach utilizes the structure of phrases appearing in HTML headings while the second uses the hierarchical structure of the HTML headings for identifying new concepts and their taxonomical links between seed concepts and between each other. An analysis [15] was made on a corpus of web documents containing scientific articles, articles of Wikipedia, journalistic articles, emails, mailing list and discussion forums. The results are interesting when analyzing terms which appear in the titles. The presence of the common noun in the title is considered as primordial. The named entity appears in 45% of the titles and 44% of the titles studied in their analysis contain adjectives [15].

The titles have a great place in the text. They are textual objects which are visible by typographical properties (bold and/or underlined and/or italic and/or numbered) or dispositional ones (presence or not of tabulation, line feeds before or after enumerations, classification). These properties enable the titles to play an important role in the level of the material organization of the text. The authors in [9] consider that the titles inside the documents constitute the relevant “main doors” for a selective nonlinear navigation in the text. The titles do not present only a material face but also a semantic one related to the fact that the titles are composed of lexemes and syntactic units carrying significance. They participate in the learning of the semantics of the document. According to [10], the titles are at the same time In and Out of the text. In since they constitute a part of the text content, they fulfill certain discursive functions; Out because they have this particular status to be distinguished from the body of text, they form a visual textual organizer. The titles are thus entities which have their own functions and which are distinguished clearly from the other parts of the text.

Two approaches were implemented for the automatic recognition of the titles and the links between titles and textual contents. The
first approach is based on rules and patterns of recognition. It aims to find in other documents occurrences of text snippets that are close to the extracted patterns. It characterizes the titles, studies the links between them by considering their level in the document hierarchy and finds all the links between the titles and the texts. The second approach attributes titles to text snippets. It is based on training techniques for the automatic segmentation of text. This approach aims to automatically associate titles to text snippets.

Moreover, it is obvious that the text snippets which are emphasized (bold, italic, etc) are more important than those which have not such marking. Indeed, the styles used in the text have a big impact on its comprehension. Today, the new tools related to computer science allow creating pages on the World Wide Web by offering a multitude of typographical marking techniques. These markings describe the terms considered to be the most important in the texts.

3. GENERAL APPROACH

Our approach aims to extract the main concepts of a given field and detect the hierarchical links between them by analyzing the text structure in the corpus. In Fig 1, we present the different steps to find taxonomy of concepts from a corpus of several web sites. The corpus is built using Google Web API library. The preprocess step is then realized using the Tree-Tagger tool, the HTML Parser and by generating a stop list to remove the blank words. After extracting the candidate terms from the corpus, they will be filtered by analyzing their occurrences in the titles and in the links belonging to the documents and considering the used styles. The hierarchical links are then extracted by combining a linguistic approach and another one based on the titles’ hierarchy in the documents.

3.1 Corpus Building and Pretreatment

The approaches of ontologies learning from text are generally based on the use of a corpus of texts and the application of the data mining techniques. This corpus must be representative of the field for which we try to build ontology. With these techniques, we inject in ontology the knowledge contained in the texts by extracting concepts and links from the texts. In our case, we use a corpus of HTML documents which is obtained using Google Web Search API. Indeed, Google programmers offer this tool in order to extract the results of a request. A list of keywords describing the field followed by a list of search parameters are introduced. We obtain as result a list of URLs of the web sites which are the most representative of the field and which verify these search criteria.

The corpus pre-processor is performed in three steps that we detail below.

Lemmatization. In the pretreatment step, we use a tagger (Tree Tagger [20]) in order to associate to each instance of a word its grammatical category (adjective, verb, noun,...) and its canonical form. The taggers, also called morpho-syntactic positioners, are based on spelling and grammatical dictionaries to perform their analysis.

Parsing. A parser is a syntactic analyser which processes texts written in a natural language in order to extract information. In our approach, we use HtmlParser 1.6 which is a free java library for the extraction and the text processing of a web corpus. It extracts data from the tags of the HTML documents.

Removing stop-words. A general stop-list is used to locate the blank words in the corpus. Indeed, many words which occur frequently in the corpus have no informational content (articles, prepositions, conjunctions...).

3.2 Candidate Terms (CT) Selection

Candidate Terms (CT) selection consists in searching within the corpus the interesting terms which are candidates to be the concepts of ontology. The CT are linguistic units which qualify an object or a concept of the real-world. They can be simple terms but correspond in general to nominal groups called nominal syntags. The existing approaches of the CT extraction are classified in two categories: syntactic or statistical. The syntactic approach [4] analyzes the grammatical role of the words in these texts. Statistical approach studies the frequency of the words in the texts [6].

Our approach rests on the extraction of the candidate terms. They are terms of the corpus which are identified by specifying the canonical syntactic structure. We extract two types of syntags according to their structures.

- Class 1: we extract syntags composed by only one word, they can be either a “Name” or a “Named_Entity”
- Class 2: we consider syntags containing two words: they have as syntactic structure the sequence “Name Adjective” or “Named_Entity Adjective”.

The candidate terms which are extracted will be then filtered according to their appearance in the titles, styles and hyperlinks used in the corpus. This step is described in details in the next section.

4. CONCEPT EXTRACTION

After the CT selection, we proceed to filter the obtained CT in order to find the concepts of the studied field. For this, we were inspired by the TF*IDF measure [11] to compute the relevance of a syntagm S in the corpus. Indeed, we propose a new measure CR*ICF (CR_ICF) where CR, the Corpus Relevance of a candidate term S, corresponds to its relevance in all the documents of the corpus and ICF indicates the opposite frequency of the studied Corpus. The refinement step consists on maintaining only the CT having a value CR_ICF higher than a given threshold. The selected CT represent the relevant terms or concepts of the field.

4.1 Inverse Corpus Frequency (ICF)

Several methods [13, 18] focus on the traditional TF_IDF measure to extract relevant information from a set of textual documents. This measure reflects how important a word is to a document in a corpus. The IDF factor (Inverse Documents Frequency) gives a high weight to the terms which seldom occur in the whole corpus and removes those which occur in a high number of documents. Thus, it is not interesting to apply IDF on a domain corpus in order to find the concepts of the field, it can rather be used for indexation or in order to extract the key-words of a document. To avoid this problem, we define a new factor denoted by ICF (Inverse Corpus Frequency) which is a measurement of the importance of the term in other corpora C of different fields. If a candidate term which appears in our corpus occurs frequently in these corpora, it will be considered as general
word (i.e. the term is not specific to the studied field) and consequently, it will be removed. This measure aims at giving a more important weight to the least frequent terms in C considered as discriminant. The ICF measure of term $t_i$ is computed as follows:

$$ICF(t_i) = \log \left( \frac{|C|}{|C_i / t_i \in C|} \right)$$  \hspace{1cm} (2)

where $c_j$ is the $j^{th}$ corpus and $|C|$ is the total number of the corpora.

The IDF factor focuses on the number of documents of the studied corpus in which the CT appears. To illustrate our idea, we will present some examples where we explain how the IDF measure is not interesting in our case. Let consider, a medicine corpus MC and a candidate term $C_1 = "Spinal-cord"$. As $C_1$ occurs in many documents in the Medicine Corpus, the IDF value of $C_1$ is low and consequently, this CT which is one of the main concepts of the medicine field, will be removed and will not be considered as
a relevant term of the field. Besides, let consider the candidate term \( C_2 = \text{"example"}. This CT occurs rarely in our Medicine Corpus MC, so that, its IDF value is high and the general term “example” will be kept as a concept of the medicine field. We can conclude that the IDF factor is not adequate in case of extraction of concepts from a domain corpus, it is rather useful for the extraction of relevant terms of a document regarding the whole corpus.

To overcome this disadvantage, we define the ICF factor (eq 2). Our ICF factor, focuses on the occurrences of a CT in other corpora. Thus, a CT that appears frequently in corpora from different fields is considered as a general term and can not be considered as a concept of the studied field. This is the case of the candidate term \( C_2 \) which occurs frequently in the corpora. Moreover, a CT that has few occurrences in the corpora, will be considered as discriminant of the field in question and be selected as a concept. This is the case of the candidate term \( C_1 \), which has few occurrences in the corpora but is discriminant in the medicine field. Hence, it must be retained as a concept of the medicine corpus.

When we compute the ICF factor for each CT, we note that ICF (spinal-cord) > ICF (example). Indeed, the CT \( C_1 \) is a concept of the medicine field whereas \( C_2 \) is a general term and can not be considered as specific to the studied field. As a conclusion, we can affirm that the ICF factor is more suitable than the IDF not be considered as specific to the studied field.

**4.2 Corpus Relevance (CR)**

In [24], we have defined the Corpus Relevance of a candidate term \( S \) as follows:

**Definition 1:** Let be a corpus \( C = \{d_1, d_2, \ldots, d_m\} \) of \( m \) HTML documents \( d_j \), the Corpus Relevance CR of each syntagm \( S \) is the sum of its Normalized Relevance NR in all the documents of the corpus.

\[
CR(S) = \sum_{j=1}^{m} NR(S, dj) = \sum_{j=1}^{m} \left( \frac{R(S, dj)}{\sum_{k=1}^{n(j)} R(S, dj)} \right)
\]

where \( n(j) \) is the number of syntagms in the document \( dj \) and \( R(S, dj) \) is the relevance of the syntagm \( S \) in the document \( dj \). The denominator is the sum of the relevance of all the syntagms in this document.

The normalized relevance of a syntagm \( S \) in the document \( dj \) is used to avoid the problems related to the length of the document.

The relevance \( R \) of a syntagm \( S \) in a document \( dj \) is the sum of its Title relevance \( R_{title}(S) \) and its Style relevance \( R_{style}(S) \) in the document. In fact, we have focused on the documents structure in order to filter the candidate terms. The titles, the hyperlinks and the words in a bold or italic style generally represent elements carrying relevant information.

**Computing Title Relevance**

It is important to consider the titles and the subtitles of the documents for the extraction of the main concepts. Indeed, they generally contain the relevant terms of the field. Our approach extracts the titles belonging to the documents by analyzing the HTML tags and assigns to each title the level in the titles’ hierarchy in the document. Three factors are analyzed when extracting the titles of the documents: the title type, the title length and the title level. These factors are detailed in the rest of this section.

Among the extracted titles, we keep only those which could contain relevant terms and which are in coherence with the textual content in the document. Indeed, a title has sometimes an empty informational content. It is the case of the subtitles: Introduction, Conclusion, Abstract, … which occur frequently in professional documents and academic papers. The lexicon of the title is primarily used to indicate the textual type of the section which follows (it is an introduction, a summary, a conclusion.) and not its contents. However, if the syntagms belonging to the title occur in the rest of the document, there is implication of the title in the textual content [9]. For this reason, we propose to detect automatically the type of each extracted title by analyzing its occurrences in the whole document. Indeed, our approach checks for every extracted title if it appears in the rest of the text. If it not appears, it will be considered as not interesting and will be disregarded.

It is also remarkable to note that the syntagm relevance in a given title strongly depends on the length of this title. Thus more the number of syntagms in a title is low, more the syntagm relevance in this title is high. We call title_length the number of syntagms which belong to the title. As an example, we extract from a French medicine document the two following titles: "epidemiology" (épidémiologie) and "A training in France to the new jobs in care establishments" (Une formation en France pour accéder aux nouveaux métiers des établissements de soin). We note that the only syntagm (epidemiology) belonging to the first title is more representative than the five syntagms composing the second title which are less significant (training, France, job, establishment and care).

To consider the title length, we compute the relevance \( R_S \) of a syntagm \( S \) which appears in the title \( t \). It corresponds to the inverse of \( l(t) \) where \( l(t) \) is the length of the title \( t \). We apply our method on the two titles "epidemiology" and " training in France to the new jobs in care establishments" given in the example above, we obtain \( t_{epidemiology} = 1 \), \( r_{training} = 0.2 \), \( r_{France} = 0.2 \), \( r_{job} = 0.2 \), \( r_{establishment} = 0.2 \), \( r_{care} = 0.2 \). We can note that these results reinforce our intuition.

Finally, we calculate the title relevance of a syntagm in a document as follows:

**Definition 2:** The Title relevance \( R_{title}(S) \) of a syntagm \( S \) corresponds to its relevance in all the titles of the document. It is computed as follows:

\[
R_{title}(S) = \frac{1}{\sum_{i=1}^{l} i} \sum_{i=1}^{l} \sum_{j=1}^{n(i)} (i^* l(t'_j))^{-1}
\]

We give the weight \( 1/i \) to each level \( i \) in the hierarchy of the titles in a HTML document cannot exceed six levels.

**Computing Style Relevance**

The style with which the terms of a text are written indicates generally the relevance of this term in the document. The terms...
written with "bold" or "italic" style for example are more relevant than the terms which do not profit from such a marking. Moreover, the syntagms which appear in the hyperlinks belonging to the documents are generally more interesting than the terms in the rest of the corpus. We can easily detect the emphasized words (bold, italic) or terms belonging to the hyperlinks of the documents by analyzing the tags of HTML documents. For each candidate term extracted, we define four coefficients:

- \( f_{\text{Bold}}(S) \) corresponds to the Bold style frequency of a syntagm \( S \) in a HTML document.
- \( f_{\text{Italic}}(S) \) corresponds to the Italic style frequency of a syntagm \( S \) in a HTML document.
- \( f_{\text{url}}(S) \) corresponds to the hyperlink frequency of a syntagm \( S \) in a HTML document.
- \( f_{\text{Rest}}(S) \) corresponds to the appearance frequency of a syntagm \( S \) in the rest of the document.

**Definition 3:** We define the style relevance \( R_{\text{Style}}(S) \) of a syntagm \( S \) as follows:

\[
R_{\text{Style}}(S) = \sum_{\text{Format}} w_{\text{Style}} \ast f_{\text{Style}}(S)
\]

Where Format = {Bold, Italic, URL, Rest} and \( w_{\text{Style}} \) is a weight given to each style and it was fixed experimentally.

5. EXPERIMENTS

We have used a corpus of 80 HTML documents which are related to the medical field. When applying our method, we obtain the total number of 5411 candidate terms extracted from these documents: 2844 CT belonging to the class 1 (syntagms composed by only one word which is either a “Name” or a “Named_Entity”) and 2567 CT belonging to the class 2 (syntagms containing two words having as syntactic structure the sequence “Name + Adjective” or “Named_Entity + Adjective”).

After the extraction of the CT, the refinement of these CT is realized as follows:

- Removing CT belonging to a limited number of documents.

\[
|\{d_j; t_i \in d_j\}| \leq \alpha
\]

After experimenting with various \( \alpha \) values on a data sample, the best results were found for \( \alpha=3 \) and \( \alpha=2 \) for respectively the concepts with only one word and those containing two words.

- Removing CT which have a CR*ICF value less than a constant \( \beta \).

\[
\text{CR}(S_i) \ast \text{ICF} \leq \beta
\]

To use the ICF measure, we have used a reference corpus that is a sample of the general language. The candidate terms which occur frequently in this corpus will be pruned. The values of \( \beta \) were fixed at 0.1 and 0.15 for respectively the concepts belonging to the class 1 and 2.

To evaluate the obtained results, we have used the CISMeF catalog based on the thesaurus MeSH (Medical Subject Headings) [23]. It is a taxonomic hierarchy of medical and biological terms suggested by the U. S National Library of Medicine. This thesaurus contains about 41658 french terms which represent the basic concepts in the medicine field. It constitutes an invaluable tool for the specialists. We have also fixed \( w_{\text{Bold}}=3 \), \( w_{\text{Italic}}=2 \), \( w_{\text{url}}=2 \) and \( w_{\text{Rest}}=1 \). When running the developed application using 80 HTML documents as input, we have as result a total number of 54 concepts. We have verified if these concepts belong or not to the thesaurus Mesh. This step led to a retention of 30 concepts which corresponds to a precision of 55.56%.

The values of precision, recall are computed for each extracted concepts. The F-measure [19], commonly used in lots of data mining areas, corresponds to the harmonic mean of precision and recall. It is computed as follows:

\[
F - \text{measure} = \frac{2 \ast \text{precision} \ast \text{recall}}{\text{precision} + \text{recall}}
\]

An experimental study has been conducted in order to compare the ICF and IDF factors. For that, we compare the results obtained by the most used measure in the literature TF_IDF and those obtained using TF_ICF measure (the product of TF and our ICF factors). The measures are applied on corpora of different sizes. Among the extracted concepts, we have chosen 460 concepts having the highest values of TF_IDF and TF_ICF (400 for concepts belonging to the class 1 and 60 for ones belonging to the class 2). The results concerning the two types of concepts are given in Fig 2.

![Figure 2](image_url)

**Fig. 2.** Evaluation of one-word concepts (left) and two-word concepts (right) using TF_IDF and TF_ICF measures
We note that the TF_ICF measure gives more interesting results than TF_IDF ones. Indeed, the TF_IDF measure reflects how important a word is to a document in a collection or corpus. The IDF factor is based on counting the number of documents in the collection of documents which contain the term in question. The intuition was that a query term which occurs in many documents is not a good discriminator and should be given less weight than one which occurs in few documents [22]. This measure is not adequate in case of extraction of concepts from a domain corpus, it is rather useful for the extraction of relevant terms of a document regarding the whole corpus.

We have also done a comparison study between the results obtained using our measure Corpus Relevance (CR) and those obtained using the CR_ICF measure (see Fig 3). We obtain better results when using the CR_ICF measure. This can be explained by the fact that the use of ICF factor removes the candidate terms which occur the most in the reference corpus and than which are the less specific to the field of medicine.

In Fig 4, we give the results obtained using our measure CR_ICF and the TF_ICF measure. The results concerning the two classes of concepts are given in Fig 4. As we can note, the CR-ICF measure gives more interesting results than TF_ICF one. These results show the impact of considering the documents structure in the extraction of concepts by giving the most relevant ones. For the first class of concepts, it is remarkable to note that when the number of studied documents increases, the distance between the two measures increases too. Since the number of two-word concepts in the studied corpus is low compared to the number of one-word concepts, we can increase the number of documents in the corpus to obtain more interesting results.

![Fig. 3. Evaluation of one-word concepts (left) and two-word concepts (right) using CR and CR-ICF measures](image1)

![Fig. 4. Evaluation of one-word concepts (left) and two-word concepts (right) using TF-ICF and CR-ICF measures](image2)
6 Extraction of hierarchical links

In this section, we present our current work dealing with the extraction of hierarchical links between the obtained concepts. Several methods focus on the extraction of the taxonomic links. Although, the statistic ones are robust, they are applied on corpora of large size which is not the case of our corpus. Moreover, these methods do not really extract the taxonomic relations but propose a group of terms, from which an expert deduce the relations which could be synonymy, meronymy or other conceptual relations.

The results obtained by the linguistic approach and the lexico-syntactic patterns approach are relevant and have high precisions on any size of corpus. So that, we propose to use these two approaches in order to extract taxonomic links. Our method operates as follows:

6.1 Linguistic phase

There are two types of extracted concepts according to their structures. The first type is a syntagm composed by only one word, which can be either a “Name” or a “Named_Entity”. The second one is a syntagm containing two words which have as syntactic structure the sequence “Name + Adjective” or “Named_Entity + Adjective”. We can easily notice that there is hierarchical relation between the two types of concepts. Indeed, let S1 and S2 be two concepts, S1 is a a “Name” or a “Named_Entity” and S2 is “Name + Adjective” or “Named_Entity + Adjective”, it is trivial that S1 is a sub-concept of S2 if S2 is composed by the concept S1 followed by an adjective.

For example, our method generates two hierarchical links from the medicine corpus:

As we can note, the two concepts “respiratory infection” (infection respiratoire) and “bacterial infection” (infection bactérienne) are two types of infection and can be effectively considered as sub-concepts of the concept “infection”.

6.2 Projection of lexico-syntactic patterns in the corpus

The objective is to detect the hierarchical links between concepts of the corpus using lexico-syntactic patterns. This approach is based on the definition of a semantic relation then on the observation of sequences in corpora which convey the desired relation. In our case, this method extracts the syntactic contexts which “mark” the hyperonymy link (interesting relation for the hierarchical between a potential couples t1/t2 where the term t1 is more general than the term t2). To do this, we use the patterns defined by Marshman [16] which are specific to hyperonymy relation:

- \( SN_1 \text{ is a } SN_2 \) (\( SN_1 \text{ est un } SN_2 \))
- \( SN_1 \text{ is a kind of } SN_2 \) (\( SN_1 \text{ est un type / une forme / une sorte de } SN_2 \))
- \( SN \text{ for example / as } SN_1, SN_2, SN_3, \ldots \) (\( SN \text{ par exemple / comme } SN_1, SN_2, SN_3, \ldots \))

where \( SN \) and \( SN_i \) are Nominal Syntagms.

We have also extend this list by other patterns defined manually:

- \( SN \text{ such as } SN_1, SN_2, SN_3, \ldots \) (\( SN \text{ tel que } SN_1, SN_2, SN_3, \ldots \))
- \( SN \text{ in particular } SN_1, SN_2, SN_3, \ldots \) (\( SN \text{ en particulier } SN_1, SN_2, SN_3, \ldots \))
- \( SN \text{ especially } SN_1, SN_2, SN_3, \ldots \) (\( SN \text{ notamment } SN_1, SN_2, SN_3, \ldots \))

The real problem when applying these approaches is the weak production. In fact, the number of taxonomic links extracted is low compared to those generated by the statistic approaches. To improve the production when extracting the taxonomic links, we propose to extend the linguistic and the lexico-syntactic patterns phases by considering the structure of the documents.

6.3 Analyze of the titles’ hierarchy in documents

The extraction of the hierarchical links can be established by analyzing the hierarchy of the documents titles.

According to [9], “The nesting or the parallelism of titles of sub-sections belonging to a given section reflects relations of subordination or juxtaposition existing between these sections”. The text can be then considered not like a linear succession of blocks, but like a structure of elements of high level which include other elements. In our approach, if a concept S1 is the only syntagm which appears in title 1 and S2 is also the only syntagm belonging to title 1.1 (subtitle of title 1) then the relation is-a can be established between S1 and S2. Our approach focus on the number of syntags belonging to the titles to extract taxonomic links. When the number of syntags is high in the titles, it is difficult to extract these relations between these syntags.

For example, from a document which deals with diarrhea, we extract the two titles:

- H1: diarrhea (diarrhée)
- H2: chronic diarrhea (diarrhea chronique)

whereas H2 is a subtitle of H1.

Our method generates a hierarchical link between the syntagms belonging to the titles.

The text can be then considered not like a linear succession of blocks, but like a structure of elements of high level which include other elements.
Indeed, chronic diarrhea is a type of diarrhea and can be considered as a sub-concept of this disease.

7 Conclusion and Future Work

In this paper, we have presented an approach which focuses on the documents’ structure in order to extract the main concepts of a given field. Indeed, the material form of the documents provides interesting information on the semantics contained in the texts. We have defined a new measure denoted CR-ICF in order to filter the extracted CT. The CR factor is based on the occurrences of the CT in the titles, the links and the used styles in the documents, whereas the ICF one is based on the occurrences of the CT in other corpora in order to check if the CT is a general term or it is specific to the studied field. We obtain as result the most relevant terms (concepts) of the field. The taxonomic links are then established by combining three methods: the first one is based on the canonical syntactic structure of the extracted concepts, the second one searches in the corpus the couples of concepts which convey the hyperonymy relation by applying lexico-syntactic patterns and the third methods analyzes the hierarchy of the titles in each document to extract taxonomic relations. We have evaluated our approach on a web corpus related to the medical field and the first results concerning the extracted concepts seem promising. The step of taxonomic links extraction is being implemented.

As future work we aim to:

(i) Increase the corpus size in order to improve the precision and recall values. Considering the fact that the field on which we work is wide and contains a huge number of concepts, we will probably obtain more interesting results when the number of documents in the corpus is high.

(ii) Make an experimental study on the extraction of taxonomic links in order to obtain taxonomy of concepts. The evaluation of the results can be done by comparing the extracted links with those given by thesaurus mesh [23].

(iii) Extend the work performed by extracting the functional relations (non hierarchical relations) to lead to ontology.

7. REFERENCES


