Universidad Nacional de Colombia

Facultad de Ingeniería

Maestría en Ingeniería - Automatización Industrial

BIOPROSPECTUS: Information fusion and search to support bioproduct development

Master Thesis

Director:
Fabio González Ph.D. & M.Sc
Departament of Systems and Industrial Engineering

Co-Director:
Emiliano Barreto Ph.D. & M.Sc
Departament of Bioinformatics

Advisors:
Maria Teresa Regueiro M.Sc
Departament of Bioinformatics

Raul Ramos Pollán Ph.D. & M.Sc
Departament of Systems and Informatics

Proposer:
Samier Said Barguil Giraldo
C.C. 1.010.179.793

Area:
Computer Science

Research Line:
Information retrieval

UNIVERSIDAD NACIONAL DE COLOMBIA

Contents

Resumen

Proper use and exploitation of biodiversity resources (bioprospecting) depends upon knowledge of different organization levels (molecular, cellular and ecosystem) of biologic and genetic resources. This relies on appropriate systematic capacities to explore different information resources such as scientific literature, compound databases, medical and biological ontologies, among others.

This paper presents a prototype computational system (BIOPROSPECTUS) to support bioprospecting natural products from Colombian biodiversity with biological activity. Bioprospectus is a knowledge base information retrieval (IR) system that integrates and analyzes information from multiple sources and domains, provides query suggestions based on an expert curated ontology and offer result exploration capabilities on top of a metasearch engine.

The system was developed using information exploration (IE) and natural language processing (NLP) techniques over a set of raw scientific articles, integrating additional information from a collection of external knowledge bases. Users may express their information needs by combining textual and semantic keywords in a query that can be refined through domain specific knowledge structured as an ontology. For the evaluation two quantitative measures were taken and compared to a reference system. Based on the results, the system holds great promise providing a technological foundation to identify new bio-products from the colombian biodiversity targeting sustainable development and added value of our biodiversity.

Acknowledgement

To my parents.

Chapter 1

Introduction

This chapter includes a deep explanation the Bioprospecting process as well as the project objectives, the contributions achieved during the project development. Finally, the whole document structure described.

1.1 Problem Definition

Bioprospecting can be understood as a new form to give use to the biodiversity, through the search or systematic exploration of biological sources with potential for economic exploitation by developing new compounds and products. In mega-biodiverse countries, sustainable use of biodiversity becomes a great opportunity to improve their competitiveness and a way to contribute to the country's development, based on the opening of new markets for high value added products.

Addition of value occurs as far as raw materials found and extracted from a given territory can be complemented with innovative development processes. Theses process are part of the value chain known as sustainable biocommerce and refers to the production, processing and commercialization of goods and services derived from biodiversity under environmental social and economic sustainability criteria. Proper use of biodiversity resources depends upon knowledge of different organization levels (molecular, cellular and ecosystem) of biologic and genetic resources. To achieve this, it is necessary to count on an appropriate systematic capacity to carry out exploration of biodiversity (bioprospecting), which will allow a better understanding of biodiversity and the production of high added value products. Given the great opportunities based on the biodiversity potential of megadiverse countries, it becomes necessary to improve bioprospecting capacities through the use of modern biotechnology techniques (omics and bioinformatics), as the first link in the value creation chain.

However an effort to developing a prototype system able to identify, integrate, catalog, model and extract relevant information from different biodiversity information sources is a big step, as this type of systems have been seldom reported so far in the literature. Therefore, this paper presents a first approach of a Bioprospeting system named Bioprospectus.

Bioprospectus is a specific information retrieval system able to integrate scientific literature with a large expert?s curated knowledge base. In the system, relationships between plants, biological activities or chemical compounds reported in the scientific literature can be easily identified as

well as the reference to the set of articles than contains the information. This it useful because speeds up the first steps of the bioprospecting process because it is well known that searching for scientific literature can be a time-consuming activity and more in the bioprospecting tasks where comprehensive search results are needed.

1.2 Objectives

1.2.1 Main Objective

The principal aim of this study is to design and to implement a prototype of a bioprospecting system for the exploration and automatic integration of different information sources related with medicinal plants in order to show their possible medical/pharmaceutical uses.

1.2.2 Specific Objectives

1. To identify, to catalog and to model information sources like inputs of the bioprospecting prototype.
2. To develop a module for information extraction, able to automatically identify and collect relevant information from the different information sources.
3. To develop a prototype information retrieval system that supports information searching and exploration tasks associated with bioprospecting processes.
4. To evaluate the performance of the prototype system taking into account the usability, information source integration, scalability and relevance of the presented results.

1.3 Results and Contributions

The results and contributions of this work can be summarized as follows:

1. A completely functional knowledge based information retrieval system to allow the Bioprospeting process. This system is available online for public usage at http://srs.ibun.unal.edu.co/Bioprospectus.
2. The Bioprospectus software was built using open source software technologies. The source code as well as the Documentation is available online for public usage at https://sbarguil.bitbucket.io/documentation
3. An experts curated ontology to support the Bioprospecting process. This ontology can be easily manipulated and complemented for additional needs. The ontology is available for download at https://sbarguil.bitbucket.io/documentation.
4. An oral presentation as part of the main sessions in the SolBio International Conference And Workshops 2016 on "Bioinformatics and Computational Biology for Innovative Genomics" with the work named ""Bioprospectus: Biodiversity data integration and search to support bioprospecting of the industrial uses of plants".

5. A publication about the Bioprospectus Software architecture and evaluation: Barguil, S., Suarez, K., Gonzalez, F. "Bioprospectus: Biodiversity data integration and search to support bioprospecting of the industrial uses of plants" in The Knowledge-Based-Systems of ELSEVIER. *Under Revision*

1.4 Document Structure

This Document is organized as follows:

- Chapter 2 introduce all the theoretical bases and concepts involved in this work related to the bioprospecting, information retrieval and information exploration process.
- Chapter 3 presents the method and describes in detail the process of data acquisition and ontology creation as well as the Bioprospectus prototype and other tools used for this work.
- Chapter 4 shows the experimentation and results obtained.
- Chapter 5 contains the conclusions and discuss some future lines of work.

Chapter 2

Background

This chapter focus on the core concepts of the project. It includes a deep explanation of the concepts related with Bioprospecting as well as a collection of the studies done in this area. By the other hand, relevant concepts of information retrieval and information exploration are formally described, in order to contextualize the reader with the steps followed for the project development.

2.1 Bioprospecting

According to the World Health Organization (WHO) bioprospecting is the process of discovery and commercialization of new products based on biological resources [1]. In other words, Bioprospecting is a new name for an old practice: it refers to the industrial product developments based on plants, traditional knowledge, and microbes culled from the biodiversity-rich regions in the world most of which reside in the so called less developed and developing countries [2]. According to this, and taking into account the large biodiversity Colombia, has a great amount of raw materials for the Bioprospecting process. Moreover, if we already have an important ethnic tradition among indigenous people on the use of medicinal plants which nowadays is culturally recognized as a reliable solution to some health problems [3].

In general, Biodiversity offers three fundamental sources of inspiration to the modern scientist: chemicals, genes, and designs. Fields of applications include pharmaceutical development, agrochemistry, cosmetics (chemicals), development of recombinant pharmaceutical proteins, enzymes, agricultural biotechnology (genes), architecture, mechanical engineering and sensor technology (designs) [4]. However, the main focus of the Bioprospecting process is the pharmaceutical industry. Thus the scientific research on medicinal plants generally originates medicines in shorter time, usually with lower costs and more accessible to the population. In most cases the raw materials used in manufacturing medicines are imported which in different places increase costs of medicines radically [5] [6].

Consequently, the sense of bioprospecting process applied on the pharmaceutical industry is to try getting the medicines closer to the population by improving the way how a nation explodes its natural resources. Due to that the objective of the process lies in "ethical resource acquisition" related to the promotion of sustainable development and biodiversity conservation strategies [2]. Thus, some public and private companies have been working in order to develop a biodiversity databases and bioprospecting tools such as:

- Atlas de la Biodiversidad de Costa Rica (CRBio) http://www.crbio.cr/. This is a private research institute, who has a high influence in areas like inventory and monitoring, conservation, biodiversity informatics and bioprospecting. In the last topic INBio has been a pioneer institution in the study of chemical substances and genes present in plants, insects, marine organisms and microorganisms, which may be utilized by the pharmaceutical, medical, biotechnology, cosmetics, food and agricultural industries [7].

- Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) in Mexico http://www.conabio.gob.mx/. This is a public research institute. The Institute is oriented in the promotion of the scientific knowledge of the biologic biodiversity as well as their conservation and appropriate use [8].

- Center for Biodiversity Conservation http://www.amnh.org. This center is part of the American Museum of Natural history in United States and has the mission of mitigate critical threats to global biological and cultural diversity. This center applies information technologies to collect, organize and analyze biological and environmental data from expeditions, remote sensing, natural history collections, modeling and databases.

- State of the World's Plants of the Royal Botanical Garden http://stateoftheworldsplants.com. This site provides, for the first time, a baseline assessment of our current knowledge on the diversity of plants on earth, the global threats these plants currently face, and the policies dealing with them. It has information about New plant discoveries, Plant genomics and research related with useful plants for different industry sectors.

2.1.1 Natural Medicine and Biodiversity

Medicinal plants have played an important role in the treatment of human diseases. Its use has been recorded since 3000 BC in Egypt and other countries in North Africa and the Middle East. In Chinese traditional medicine is an important area of study today, thanks to the effect of interaction between the compounds of different plants. In India, Ayurveda (since 2000 BC) is an ancient tradition to describe the methods for using herbal therapy [6].

In a scientific context Phytomedicines, whose origin are medicinal plants, are used in the treatment of various injuries. According to the World Health Organization (WHO) the Phytomedicines are finished medicinal products whose active ingredients are made from plants, other plant material, combinations between those or plant preparations. For plant material are considered: juices, resins, vegetable oils and other substances of a similar nature [5] [6].

Approximately 30% of total sales of medicines worldwide are based on products derived from natural resources [9]. Besides, the process of design and development of a new medication has become complex and most important than that It consumes many financial resources, which is, according to several authors, a bottleneck [10]. As a matter of fact companies has explored various alternatives to overcome this impasse screenings from the use of high throughput screening or combinatorial chemistry, and they do not have obtained substantial improvements. Technological advances, such as the bioinformatic predictions and information of molecular biology and genomics, have been risen since 2002 obtaining about 20 new structures with biological activity. The approximate cost of these developments was estimated in 2006 in a number about 800 million US dollar [11]; currently the most expensive steps are clinical studies because according to some specialists the probability

of approval of medications in development process does not exceed 19% [12].

An important case of study is related with antibiotics because in the past decade only seven new chemical entities (NCE) have been approved for the treatment of infections caused by bacteria and from these, a considerable number came from Natural products [13] [6]. The advancement of knowledge on the human genome has changed the way that medicine works ceasing to be an "art" to start being considered as a science. The increasing individual knowledge of the variations between humans make one of the basic principles of Ayurveda (defined by Charak truth more than 4000 years ago) as: "every individual is different from another and should be considered as a separate entity similarly as there are variations in the universe " [5] [6]. This postulate has given rise to a new discipline named Ayugenomics [9] that basically in the sense of designing specific drugs for each individual based on knowledge of natural products through holistic knowledge integrated into a single database, to collect all available information on natural resources and combining the strengths of traditional knowledge, experimental information chemistry, from the HTS and genomic data. Nevertheless, there exist three main problems in the design and development of a new drug as well as in the generation of new functional prototypes, (namely time, financial constraints and toxicity) [14].

According to this, scientist have been using chemical compounds mainly called secondary metabolites extracted from medicinal plants. Those compounds are derived from biological processes and are categorized into groups according to the chemical class they belong to, for instance alkaloids, terpenoids, phenolics, tannins among others [15]. Consequently, the secondary metabolites can be used as:

- As Drug precursors: Which can be obtained by processes involving chemical modification of an intermediary. These chemicals have primarily large and varied use. For example, they are used in the synthesis of plastics, pharmaceuticals, cosmetics, perfumes, detergents or aromas.
- As Drugs prototype: This is the group that has structurally related compounds that have a pharmacological activity. These prototypes can be modified to improve their pharmacokinetic and pharmacodynamic properties
- For pharmacological tests: Pharmacological tests help researchers to elucidate the mechanism of action and its interaction with intracellular signaling pathways and biological mechanisms related with diseases.

Derived from the opportunity of synthesize the secondary metabolites from natural plants a significant number of small biotechnology companies specialized in the identification of natural products have been risen in the last years. For those companies Colombia is considered as a megadiverse country extremely rich in species, ecosystems and diversity of geo-climatic conditions from tropical to temperate, from desert regions to the Andes [16] which constitute the perfect location to carry out their investigations. In fact, It is necessary to develop tools that allow the relation between various information sources with different origins as well as the process of search with the aim to increase the number or the quality of the products.

2.2 Information Retrieval

The amount of available data has grown exponentially in the last decades, ranging from unstructured plain text files to structured databases, passing by repositories that contains multimodal data, such as text, images, audio, video among others. This brings us a great opportunity in the sense that we can-not only explode and relate all the information gathered in different data sources but also It gives us the opportunity to create adequate tools to achieve better information exploration results.

One of the main problems as a result of the of the data increase is How can you handle this data to find something useful and to satisfy a particular information need?, this is the question that information retrieval (IR) tries to answer. Information retrieval is the area of Computer Science that study the problem of search and access to information, their main goal is to formulate the models and to develop the tools that allow users to solve his information requirements in a precise and efficient way [17] [3].

However, information exploration is a broader activity that is carried out in a variety of ways by different people. For instance, browsing through books, library shelves, or hypermedia documents all are formats of information exploration activity [18] According to this, one of the main applications of information exploration techniques based on data analysis methods (i.e. classification, Principal Component Analysis, correspondence factorial analysis, Products analysis) is related to the annotation and curation of biomedical documents.

In the recent years, several works has been developed with the aim of probe different techniques of Information Exploration (IE) and Information retrieval (IR) over the scientific literature. Some of the works are centered in the analysis of the citations relations between publications, this analysis can be very helpful in the systematic retrieval of scientific literature during the literature revision of a particular topic [19] Recently, Some other works, have created models to optimize the retrieval process capturing the important characteristics of the research paper and the constituent bibliographic references and citations helping the user to make faster and efficient decisions on the relevance and usefulness of papers [20].

2.2.1 Information Retrieval and Search

In order to describe the behaviour of a system that allows the user to make information retrieval, we present the functional blocks that according to the literature can be needed in Figure 2.1. Those blocks accomplish the functions related to the Information extraction (IE). That ask is related with automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction. Each block accomplish the following functions:

Data acquisition

One of the remarkable differences between information retrieval and information exploration is related with the target orientation. That is because the information retrieval search methods are

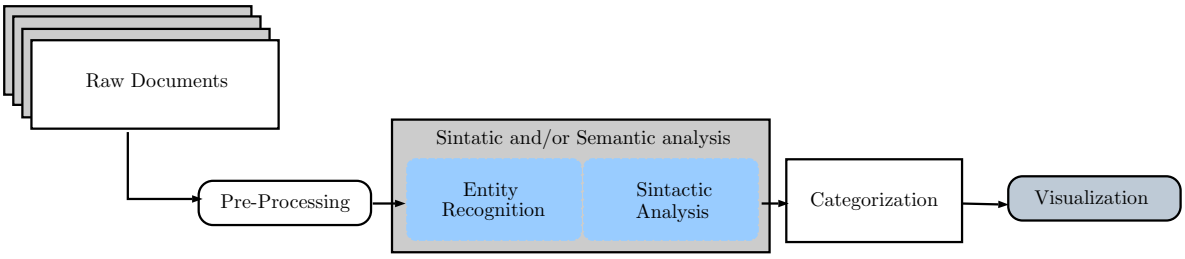

Figure 2.1: Information retrieval follows a semi-starndarized process in order to extract information from structurated text and presents it to the user

based on queries while the information exploration is based on browsing and browsing is distinguished from querying by the absence of a definite target in the mind of the user. In that way, the data acquisition process for a information exploration system may combined different sources and data structures in order to implement an unstructured associative network that satisfies the user information goals[18].

Because of that, the development and improvement of information extraction systems usually requires the existence of available resources like reference corpus (manually annotated text resources), thesaurus or ontologies [21] [22]. In the case of biomedical event extraction, various sources have been compiled, including:

- Gene Ontology: The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases. The GO Consortium (GOC) has grown to incorporate many databases, including several of the world's major repositories for plant, animal, and microbial genomes [23] .

Information Extraction

One of the main requirements of a good information extraction system is a rich feature representation. Most event extraction systems present a complex feature set based on tokens, sentences, dependency parsing trees and external resources, described as follows:

- Tokenisation consist of dividing the input text up into its constituent units (words and sentences) or its tokens. Sentence segmentation is requires classification systems since only certain classes within a sentence are considered when identifying interactions. Normally, Tokenisation splits the retrieved abstracts into sentences including titles of each paper. This can be done for example, using simple regular expressions, to identify sentence boundaries. Word segmentation is done by dividing the token (i.e. sequences of words) according to parentheses, numbers and Greek letters, and filtering tokens such as stop-words and biomedical terms. Some tokens are completely ignored, since these are considered to be non informative (of, the, and, in, punctuations and symbols) [24] [25].
- Sentences based features intend to reflect general characteristics of the sentence where the target token is present. Features are provided to reflect the number of tokens present on each sentence [26].

- Dependency parsing provides information about grammatical relations involving two words, extracted from a graph representation of the dependency relations in a sentence. Commonly used features include the number or type of dependency hops between two tokens, and the sequence or ngrams of words, lemmas or part of the speech (POS) tags in the dependency path between two tokens [27] [28].

In the implementation of the information extraction process we must deal with a particular set of challenges related to the NLP, such as:

- Tokenization
  - The system must deal special characters like: *John's*, *a state- of-the-art solution*
- Normalization
  - The system must match ***U.S.A.*** and ***USA*** automatically.
- Steaming
  - The user may wish different forms of a root to match: *authorize*,*authorization*
- Stop Words
  - The user may (or not) omit very common words : *the*,*a*,*or*,*and...*

Further optimization can be achieved by adding biomedical knowledge to the feature set (External sources). To provide this knowledge, dictionaries of specific domain terms and trigger words are matched in the text and the resulting tags are used as features. Thus, the tokens that are part of a matched term contain a feature that reflects such information.

Syntactic Analysis - Semantic Analysis

Automatic term Mapping Automatic Term Mapping (ATM) is a task consisting of identifying mentions of terms within a text and associate them with an entry within a terminology . This terminology can be previously defined using an ontology or a knowledge database. ATM is also the last step in a complex process called Automatic Term Identification. These steps are: Term Recognition, Term Classification and Term Mapping . For Term Recognition (TR) we have to identify which portions of the text contain domain concepts. For instance if we have the following portion of text we have to discriminate the text portions related to the biodiversity related domain [29].

Term Classification consists of assigning these text portions to a broad category label. However classification is not enough to establish the identity of a text portion. Term Mapping determines the relationship between a text portion and a well-defined concept contained, as our case, in an ontology. As describes there are several challenges that must be addressed before a term is successfully associated with an entry within the ontology. These challenges can be the lexical variations of a term, term synonymy and term homonymy.

In the recent years, several works has been developed in the last years with the aim of probe different techniques of Automatic term Mapping in the Biological domain, because this a challenging topic

and is a perfect way to systematically explore and access the great amount biomedical publications available. Some of the works publish until now has explore emerging techniques to access biological resources through extraction of entity names (NER)2.2.1 and relations among them [30], some of the has been working in the disambiguation problem [29][31][32], because looking the reports of biological entities in the scientific domain differences in spelling can be found, as well as a more complex set of name variations the plants can has more than one scientific name (common name or synonyms).

Other set of works has been related with the identification in text of the abbreviated forms in the text [33][34], detect this abbreviations is a challenge which is particularly critical in the case of the biomedical text due to its richness in acronyms and shortened form of words. However is well known that the issue with the abbreviations comes from the fact that they are constantly changing and are very hard to detect because its generation does not follows any pattern, so a significant effort is required to maintain terminological resources up to date [35]. However, some rosources as Snomed-CT is one of the most complete terminological resources available within the biomedical domain and it has an extensive (more than 7000) set of acronyms and abbreviations of their biological concepts [36].

Despite the large number of researches done in this area, there are very few available tools to perform ATM. One of the most popular and the gold standard is MetaMap. Metamap was developed at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text [37]. Inspired on the MetaMap algorithm, another tool named Term-mapper was developed recently. Term-mapper is an ATM software which was developed with the aim of have a tool that offers a good balance between complexity and speed, support Spanish and English, and can be easily extended to work with any knowledge source [38].

Name Entity Recognition Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, among others. Named-entity recognition in the biomedical domain is generally considered to be more difficult than in other domains, for several reasons:

- The amount of information: there are millions of entity names in use and new ones are added constantly [39]. Therefore, a dictionary or a machine learning system capable of capture all the entities is hardly feasible.
- In the biomedical field hardly ever you reach a consensus on the name to be used for a given entity or even regarding the exact concept defined by the entity itself. So the same name or acronym can be used for different concepts [39].

In general, name entity recognition (NER) techniques can be categorized into four types: Lexicon-based or Dictinonary-bused, Rule-based, Machine learning, and Hybrid techniques. Machine Learning (ML) is a key technique in NER, while hybrid techniques usually comprise a combination of machine learning and other methods. Recently, there has been an increased interest in applying supervised ML models to undertaking NER tasks [].

The main features of the NER techniques are:

- In the Dictionary-based methods a dictionary containing key words with their corresponding classes is used to identify and assign an entity [40]. A dictionary or gazetteer is a collection of words or phrases referring to a particular entity. A dictionary can contain instances of one type like locations, or several types. Dictionaries of names will be used as features in trained approaches or a way to identify all candidates in a given text. Dictionaries provide powerful features that improve the performance of the NER and NEC systems. Usually, a word appearing only in one of the several typed dictionaries will have higher probability to be of that type in which dictionary it appears. [40].

- The Rule-based methods commonly follow some manually defined linguistic patterns, which are then augmented with additional constraints based on word forms and syntactic categories to generate better matching precision. The main advantage of this kind of approaches is that most of them require little computational effort. A particularly challenging aspect of rule-based NER in practice is domain customization in order to produce accurate results in new domains [41]. However, implementing a domain-independent approach for rule-based NER typically requires a significant amount of manual effort to:

  - Identify the explicit semantic changes required for the new domain (e.g., differences in entity type definition).
  - Identify the portions of the (complex) core annotator that should be modified for each difference
  - Implement the required customization rules without compromising the extraction quality of the core annotator.

- Nowadays Machine Learning-Based named entity recognition is the most recent and successful approach developed. Thats is because the Machine Learning techniques have a high innovation potential and its features can be easily extended to different domains. In addition the core learning algorithms usually are language independent which supports multi-linguality also they allow novel combinations with relational learning approaches as well as close relationship to currently developed Machine Learning-approaches of reference solution [42] [43].

  The ML base method could be done by using two types of algorithms. Firstly, unsupervised learning which training only needs very few seeds and very large un-annotated corpus because based on this seeds the system can extract and generalize patterns from the context of the seeds. In addition, it can use the patterns to further label the corpus and to extend the seed set (This is called Bootstrapping). Secondly, the supervised Learning approach where training is based on a very large and available annotated corpus and usually is based on statistical methods such as Hidden Markov Models (HMMs) [44], Maximum Entropy Markov Models (MEMMs) [45], Conditional Random Fields (CRFs) [46] [47] and Support Vector Machines (SVMs)[48].

  In addition to purely supervised learning, which depends on the amount and quality of annotated data, semi-supervised approaches have also been proposed [49] [50]. Wang et al combined labeled data with large amounts of unlabeled data, using a rich representation based on semantic features (such as walk subsequence features and N-gram features, among others) and a new representation based on Event Feature Coupling Generalization (EFCG). EFCG is a strategy to produce higher-level features based on two kinds of original features:

class-distinguishing features (CDFs) which have the ability to distinguish the different classes and example-distinguishing features (EDFs) that are good indicating the specific examples. EFCG generates a new set of features by combining these two kinds of features and taking into account a degree of relatedness between them.

The majority of the published systems that extract biomedical relations vary in the way they extract and process the information. For instance, It can extract syntax information which usually consists of associate each token with a particular grammatical category based on its definition and context (POS) tags or other information described in a syntax theory, such as Combinatory Categorical Grammars (CCG) [51] or Government and Binding Theory [25]. Syntactic information is usually incorporated via a parser that creates a syntactic tree. However, it can extract semantic information which consists of specific domain words and patterns. Semantic information is usually incorporated via a template or frame that includes slots for certain words or entities [52] [53] [54] [22].

In contrast, some studies have revealed the possibility to have a balanced approach which uses more equal amounts of syntax and semantic parsing. Syntax parsing takes place first, often resulting in an ambiguous parse. More than 100 parses can be generated for a single sentence [55]. Semantic analysis is then applied to eliminate the incorrect syntactic parse trees and further identify domain words such as proteins and genes. In this fashion, systems combine the flexibility of syntax parsing with the precision of semantic analysis [56].

Information Categorization

After the syntactic analysis or the semantic analysis, a number of hypotheses can be generated. However, some of those hypotheses are not worthy of investigation [57]. For that reason, how to define the notion of relevance in order to score a document appropriately is an important definition. In the last years, many different retrieval models have been proposed and tested including:

- **Vector Space model**: The meaning of a document is conveyed by the words used in that document. Thus, documents and queries can be mapped into a term vector in the Hilbert space, where each dimension of the vector represents the tf-idf for one term.

  - Term-Frequency ($tf_{t,d}$): Represents the frequency of the term t in the document d. How many times t occurs in d.
  - Inverted Document Frequency (idf_t): The document frequency represents the number of documents that contains a particular term t. So we can define the inverse document frequency of the set of N documents as:

$$idf = \frac{N}{df_t} \tag{2.1}$$

    This messure shows the term specicity, in a document collection.

  This way tf-idf shows a balance between the local term frequency and the global document. Once calculated the tf-idf matrix for the documents and the query, a rank can be generated according to the lowest normalized inner product between the documents matrix ($\vec{d}$) and the vector of the query ($\vec{q}$) [58]:

$$cos(\vec{d}, \vec{q}) = \frac{\vec{d} * \vec{q}}{\|\vec{d}\|\|\vec{q}\|} = \frac{\sum q_i d_{i\,i=1}^{\|V\|}}{\sqrt{\sum q_{i\,i=1}^{2\|V\|}}\sqrt{\sum d_{i\,i=1}^{2\|V\|}}} \tag{2.2}$$

- **Probabilistic IR models**: In Classic probabilistic models the basic idea is to rank the documents in a collection based on their probability of being relevant to the current information need. This is expressed as the conditional probability Pr(relevance|d), or the probability that the information needed is met given document d [59].

Okapi BM25 Is an example of a well known Probabilistic IR models. It works as a bag-of-words (BOW) retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. Actually BM25 is not a single model, but defines a whole family of ranking models [60].

Given a query Q, containing keywords $q_1, ..., q_n$ the BM25 score of a document D is:

$$BM25(D, Q) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{2.3}$$

where:

- $f(q_i, D)$ is q_i term frequency in the document D .
- $|D|$ is the length of the document D in words.
- $avgdl$ is the average document length in the text collection.
- k_1 and b are free parameters.
- $idf(q_i)$ was described in 2.1.

As compared with the above, models based on the probabilistic ranking principle have garnered more attention and achieved more success in past decades [60][61].

Information Retrieval systems Evaluation

Commonly a network performance evaluation, precision and recall are able to evaluate the performance of entity recognition. Precision is the fraction of retrieved entities which are relevant to the entity recognition. Recall in information retrieval is the fraction of the entities that are relevant to the entity recognition.

$$\text{Precision} = \frac{\text{Relevant Information} \cap \text{Retrieved Entity}}{\text{Retrieved Documents}} \tag{2.4}$$

$$\text{Recall} = \frac{\text{Relevant Entity} \cap \text{Retrieved Entity}}{\text{Relevant Entity}} \tag{2.5}$$

In hypotheses generation, there already exist researches about the effective chemical constituents for particular medical usages. Those research papers can be used to validate the generated hypotheses.

2.2.2 Information Exploration

Like we said in the previous section (2.2.1) one of the remarkable differences between information retrieval and information exploration is related with the target orientation. That is because the information retrieval search methods are based on queries while the information exploration is based on browsing and browsing is distinguished from querying by the absence of a definite target in the mind of the user. In that way, the development and improvement of information extraction systems usually requires the existence of available resources like reference corpus (manually annotated text resources), thesaurus or ontologies [21] [22].

In addition the information extraction process usually is developed offline while the information exploration is dynamic process which require a direct and constant interaction with the final user. Due to that the exploratory search (how is also known the Information exploration) usually have to deal with the following constraints:

- The user is unfamiliar with the domain of their goal (i.e. need to learn about the topic in order to understand how to achieve their goal)
- The user is unsure about the ways to achieve their goals (either the technology or the process)
- The user is or even unsure about their goals in the first place.

Consequently, exploratory search nowadays is a big task that covers a broader class of activities including information retrieval, evaluating, comparing, and synthesizing, where the system combines querying and browsing strategies to help the user in the learning and investigation process.[62]

Chapter 3

Proposed approach

The methods used to generate the proposed solution is described along the chapter, this is why the following topics:

- Set of system design principles, system architecture and components used in the automatic term mapping and information retrieval engine.
- The text corpus construction and consolidation which means how we selected and downloaded the set of articles used by the information retrieval prototype.
- Selection principles used for the evaluation and selection of the datasources used for the Ontology creation.

are described in detail.

3.1 Bioprospectus: System Architecture

This section will give an overview of the whole system. The system will be explained in its context to show how the system interacts with other systems and introduce basic functionality of it. It will also describe what type of users interact with the system and what functionality is available. At last, the constraints and assumptions for the system will be presented.

In general, the Bioprospectus system is conceived as a domain specific information retrieval systems that is aware of the particular concepts of interest when performing bio-prospection-related information search. Bioprospectus allows users to search a large database of scientific papers related to biodiversity and its industrial uses. Bioprospectus has a large knowledge base that encompasses plants, chemical compounds and biological activities. This knowledge base is used to enrich and complement the results produced by the information retrieval module allowing users to understand better the semantical content of their search results. Also, the system uses the knowledge base to support the formulation and refining of queries that better reflect the information needs for users.

3.1.1 System Perspective

The architecture of the prototype Bioprospectus was conceived as a service oriented architecture in which each component exposes a set of services through RESTful interfaces that use JSON as common idiom. These RESTful interfaces are consumed by a web application which is the face of

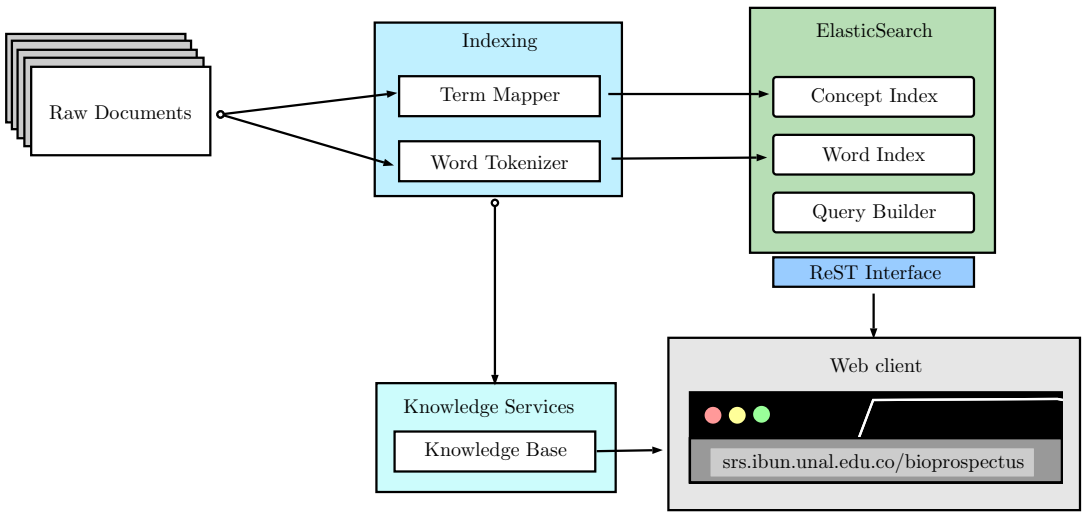

Figure 3.1: Bioprospectus system architecture is conceived as a service oriented architecture based on RESTful Interfaces between each fuctional block

the system. The main considerations followed in the design and implementation can be summarized as follows:

- Build the entire system using only open source technologies
- Build a system easily scalable that can cope with real scale problems
- Build a system that can be extended to work in many domains and with any knowledge sources
- Use a basic modern web architecture i.e. rich client which consume backend services

The whole system architecture is shown in Figure 3.1.

Each of components used in the system are described below:

- **Knowledge Services** is a Java application which offers a variety of services to access a knowledge source of concepts, terms and relations. It also maintains in memory the representation of the ontology hierarchy in order to allow fast queries over this structure, and also give access to precomputed values about the information content and similarity between concepts which are required by other components in different process.
- Full text and Documents search engine is supported on **Elasticsearch**[**?**] which is an open source search server and can be used as an efficient document store, this feature is exploited to store both the document and the semantic metadata, this way all this information is obtained when a document is retrieved as a part of the answer to a query.
- **Indexing** is performed using a WordTokenizer and Term mapper. Raw documents are processed in a two-steps process, Figure 3.3. In the first step we use Term mapper for establishing relationships between documents and concepts presents in the ontology (knowledge base). The result of this process is a set of documents with annotations named concept index. In our

solution the concept index includes not only the id of the concept present in the document but also all the IDs of more general concepts in the ontology (parents) of this particular concept. In the second step we generated a Vector Space Model, where documents are represented as vectors using TF-IDF. ElasticSearch allows us to build a representation for each document using the tokens and concepts that appear on the respective document. We have named this as Word Index.

- **Web Client** has been developed using technologies such as HTML5, CSS3, Bootstrap, Javascript, Jquery, and Backbone JS. It is a rich client which invokes through AJAX the RESTful services exposed by other components in the system and dynamically renders the appropriate views using Backbone.

Indexing and retrieval strategy

Bioprospectus indexing strategy consists of a two-step process, where two indexes are built using the raw documents and stored using ElasticSearch. Figure 3.3 depicts the complete indexing process. Raw documents consist of a series of JSON documents including several fields such as the title, authors, full text, abstract and DOI a sample pf the JSON file is shown in Figure 3.14. ElasticSearch is capable of building an inverted index for every field, however, we were only interested in indexing by title, abstract and full text. Afterwards, every field is tokenized at the word level (Word Tokenizer in Figure 3.3) and stored in ElasticSearch using an inverted index. Word Index supports free-text search at title, abstract and full text fields.

However, in the scientific domain, several terms can be attributed to the same concept, e.g., a plant can be recognized by its common name (Garlic), scientific name (*Allium Sativum*) or by its short name (A. Sativum). This problem has been previously described as Entity Recognition, in the case of Garlic, all the term mentions should be attributed to the same concept. Word Index fails at recognizing these semantic differences.

Thus, to support entity recognition a second index named concept Index were created. In the upper section of Figure {Indexing can be depicted the process for the construction of this index. For this purpose, we use as input the raw documents set and a knowledge source, in our case an ontology containing a set of concepts within the bioprospecting domain and the relationship between these concepts. Further details about this ontology are presented in subsection 3.3.

Then we follow a process known as Automatic Term Mapping, a sample of a short piece of annotated text is shown in Figure 3.4, where fragments of text are identified as entities and labeled correspondingly as a concept/category of the ontology. Term-Mapper [38] is a highly configurable ATM software, can scale to thousands or even millions of documents, is capable of dealing with English and Spanish documents and additionally, this tool can be easily extended to work with any knowledge source.

Taking this into account, Term-mapper fit into the requirements of the Bioprospectus system because we can use our own ontology using the proper terminology to support the bioproduct development. In addition to that we want to also exploit hierarchical relations present in the ontology, for the specific of a plant, if Term-Mapper encounters the species name of a plant, we are also interested in indexing more general concepts like its genus and family. We modified Term-Mapper to include this hierarchical information of a concept provided by the ontology into the ATM process.

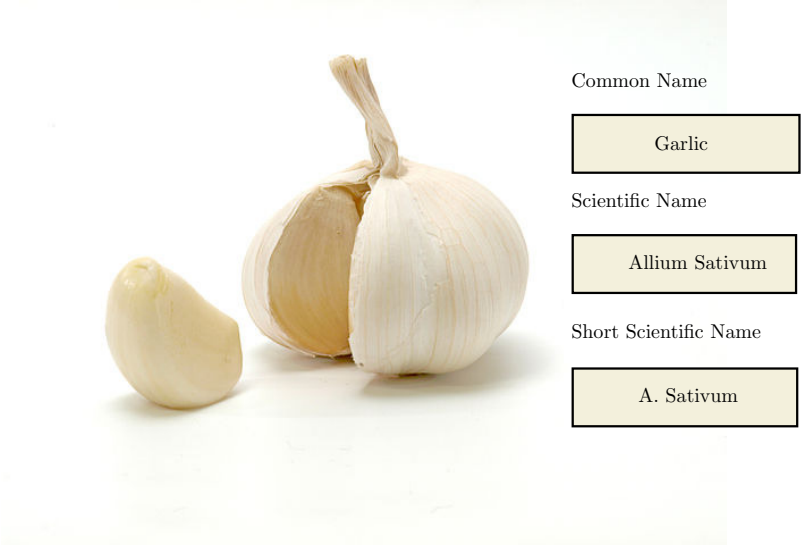

Figure 3.2: Name Entity Recongnition (NER) example for plants

Finally, we have a Concept Index, where every document is stored properly along with a related set of concepts, previously found by the Term-Mapper.

The indexing process described above allows us to exploit external information through the Concept Index and the ontology database, providing a semantic interpretability for the query. Both Word Index and Concept Index are stored in ElasticSearch using its API.

Query Resolving

After the construction of both indexes, It is necessary to deal with the user information needs and resolve the query to retrieve the articles, that better fit into the requirements. In the system, a query is composed of three fundamental components:

1. **Free-text search**: It is 100% handled by Word Index, allowing the system to retrieve documents over indexed fields such as title, abstract and fulltext. For free-text search, ElasticSearch performs a standard Vector Space Model retrieval process, obtaining as results a ranking of relevant documents, where the ranking is estimated by Okapi BM-25 weighting scheme.

2. **Filtering**: It is 100% handled by the Concept Index. Concept Index allows users can add to ontology concepts as filters to the query. Documents retrieved will be filtered whether they contain or not the given ontology concept. 3.1.2 section describes how Bioprospectus allows the user to include the information of the Concept Index in the web interface. Figure 3.5 depicts the whole Query Resolving strategy.

3. **Query expansion**: It is used to enrich the Free-text query. Two strategies were formulated to increase the precision of the retrieved documents:

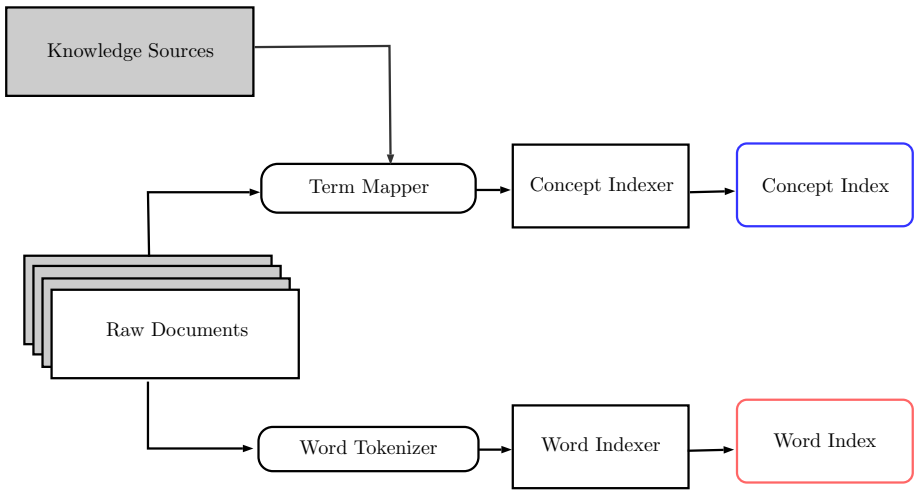

Figure 3.3: Bioprospectus Indexing Process must supports two concepts and term queries, thus, the indexing is based on Term Mapper and a Word Tokenizer

(a) Query Expansion 1 (QE1) is a strategy consisting of adding synonyms or descriptors of an ontology concept to the free-text search. Figure 3.6 shows a brief example for the concept Antibacterial.

(b) Query Expansion 2 (QE2) consists of adding leaf associations of a concept in the ontology, these associations represent examples of a given concept. Figure 3.7 shows a brief example for the concept Antibacterial.

Experimentally, we have found that QE1 performs better than QE2. The results are shown in the section 4.2.1.

3.1.2 System Functions

Bioprospectus is a knowledge base information retrieval (IR) system that integrates and analyzes information from multiple sources and domains, provides query suggestions based on an expert curated ontology and offer result exploration capabilities on top of a metasearch engine. Thus, within a web client, the users will be able to search and identify relations among plants, bio-activities and chemical compounds, as well as receive scientific articles indexed in the system as evidence of these relations in order to automate bioprospecting process.

The result of a search will depend on the type of query the user inputs. There are three types of query the user can use in the system.

Free text Query

A Free text query can be formulated typing one or more concepts in the search box of the system (Figure 3.8). For those keywords a keyword query is formulated using an `AND` operator and then the retrieval process is executed by the system following this steps:

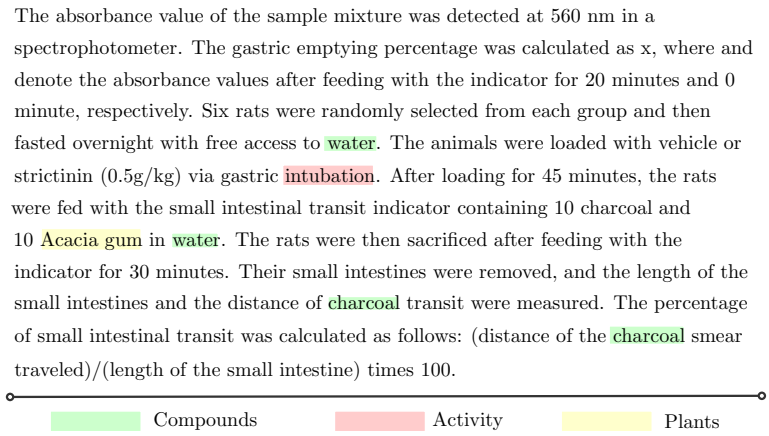

Figure 3.4: Sample of a scientific article marked with each of the categories used by the system: Plants, Compounds or Biological Activities

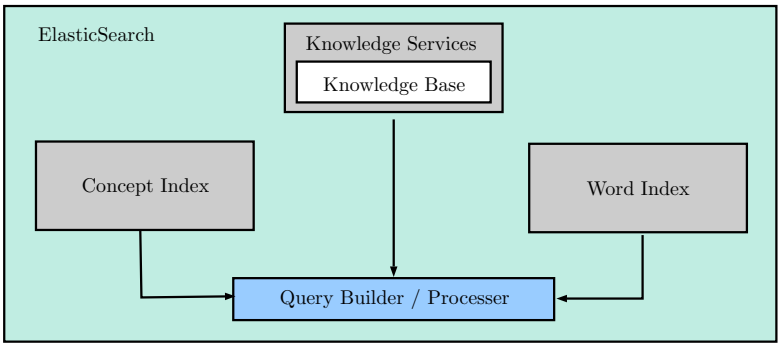

Figure 3.5: The query resolving process is supported by ElasticSearch, based on the index created in the offline processing

1. Keyword query is used to perform a standard Vector Space Model retrieval process, obtaining as results a ranking of relevant documents along with its semantic metadata. Those concepts are used as suggestions in the results window.

2. A list of documents will be displayed, sorted by their relevance. A list of concepts associated with the query will be displayed on the left pane of the results tab.

Concept Query

A concept query can be formulated using one or more concepts of any of the categories presented by the system through the navigation of the ontology three as shows in Figure 3.9 for the desire group of concepts in the main categories:

- A plant common/scientific name, genre or species.
- An bio-activity name.
- A chemical compound name.

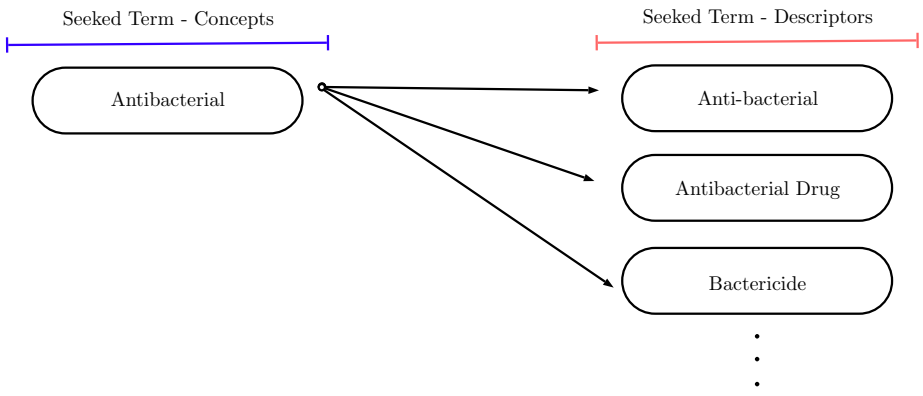

Figure 3.6: QE1: The query was formulated using the the seeked term plus all the synonyms (descriptors) associated to the term in the ontology.

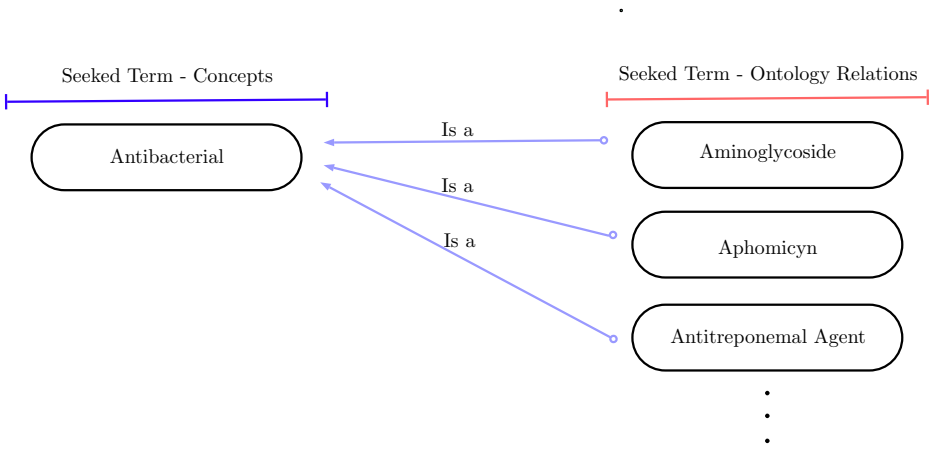

Figure 3.7: QE2: The seeked term was added plus all the leafs (down relationships) associated to the term in the ontology

or adding it directly from the suggestions when you type the in the search box (Figure 3.10).

For a concept query the system in the results window will update the list of filters to show how many concepts are present for each category, and also will show a list of articles related to the concepts present (as shows in Figure 3.10). The concept query has the ability to process boolean queries, using the operators +, - or Query Expansor as shows in Figure 4. The query expansion method adds to the terms used for the user to define their query the concepts related to this term in the ontology. This increase the information need and expand the volume of the related concepts, to get a most precise results.

Rifamycins *and* Spartium junceum

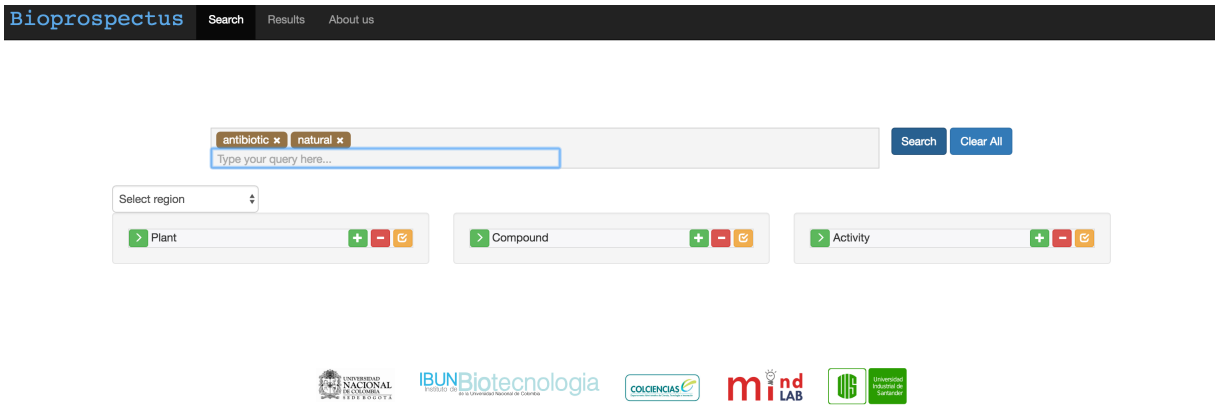

Figure 3.8: Bioprospectus Graphical User Interface supports free text typing, once the user start typing the system suggest related concepts

Query expansion

Free-text query can also be expanded using our knowledge source. The user can use a concept as query expander instead of a filter. In that case, all the related terms of a concept will be added to the free-text search query. Although these terms will be added as optional terms in the free-text search, i.e., their presence will only increase the relevance of certain documents containing those terms.

Figure 3.12 depicts also the process involved in the query expansion component. In this specific case, Bioprospectus allows the user to identify if the concept is added as a filter (left part of Figure 3.12) or as a query expander (right part of Figure 3.12). In the later case, synonyms of antibacterial were added to the free text query, while those concepts, such as Allium cepa are added as filters and they will have a bracket indicator, which shows the type of filter, MUST (+) or MUST NOT (-).
Once the user introduce their query the system will execute the following retrieval process:

1. Concepts are used to formulate a boolean query which result in a subset of documents that contains the concepts along with its semantic metadata, note that the documents obtained as results are a subset not a ranking. Those concepts are used as suggestions in the results window.

2. Simultaneously the free-text query is used to perform a retrieval process, obtaining as result a semantic ranking of relevant documents along with its semantic metadata. Those documents corresponds to the articles presented in the results window.

3. A list of suggested concepts is also retrieved. Figure 3.13 shows how the main interface presents the related concepts to a query. This allows users to filter their search with specific concepts relevant to their query

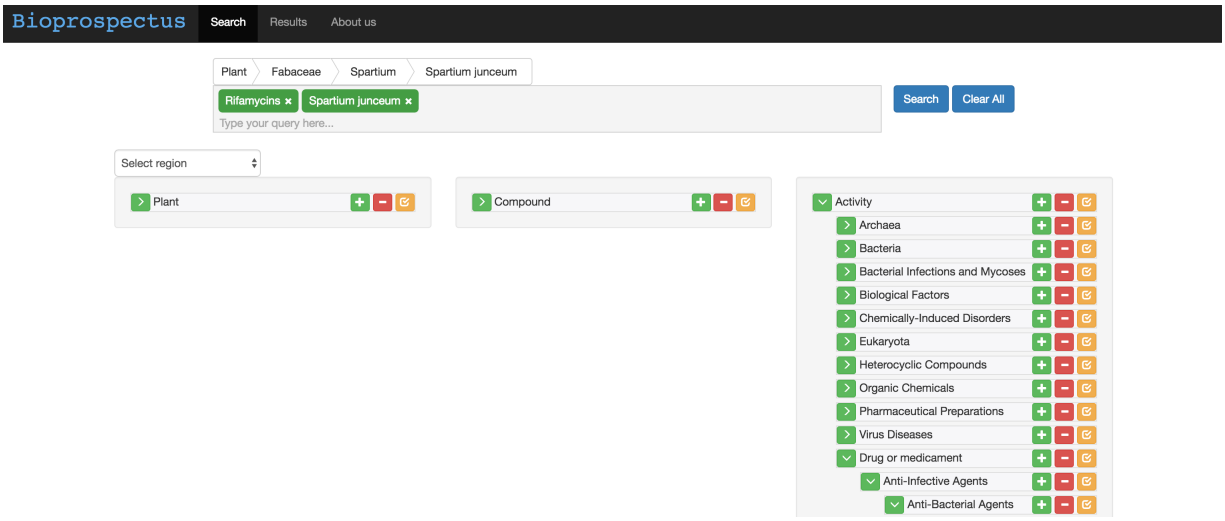

Figure 3.9: Bioprospectus Ontology Tree Exploration

3.2 Article acquisition and Corpus consolidation

In order to consolidate a proper text corpus useful for the bioprospecting process we have collected a set of more than 18K open access scientific articles. The set of articles has been collected using the ELSEVIER web API which allows journals and books published by Elsevier on ScienceDirect full-text platform [63] as well as the ScienceDirect and Scopus search engines.

The set of articles were collected taking into account the relevance of the journals on topics related with medicinal plants, phytomedicine, medicinal and aromatic plants among others. Some relevant journals used for the corpus consolidation are listed in Table 3.1.
Once the articles were collected and the articles corpus were generated using the following JSON standard format , an example of a scientific article processed and saved in the standard format is shown in the Figure 3.14.
Once, the complete set of JSON articles were stored for indexation using the Elasticsearch capabilities. As we have described before, during that process words and semantic metadata from each article are extracted form the articles.

3.3 Database selection and Ontology construction

In the context of knowledge engineering there are many types of resources which can be referred as knowledge sources (KS). In general a KS is a compendium of information about a general or specific domain but the type, structure and the level of detail in this information can vary from one source to another. In recent years the ontology - explicit formal specifications of the terms in the domain and relations among them - has been one of the most used knowledge sources as a fundamental part in the artificial intelligence applications development. In fact, Ontologies has become part of the core of the web applications and the search engines.

Formally, an Ontology is a formal representation of knowledge by a set of concepts within a domain and the relationship between these concepts and usually defines a common vocabulary for

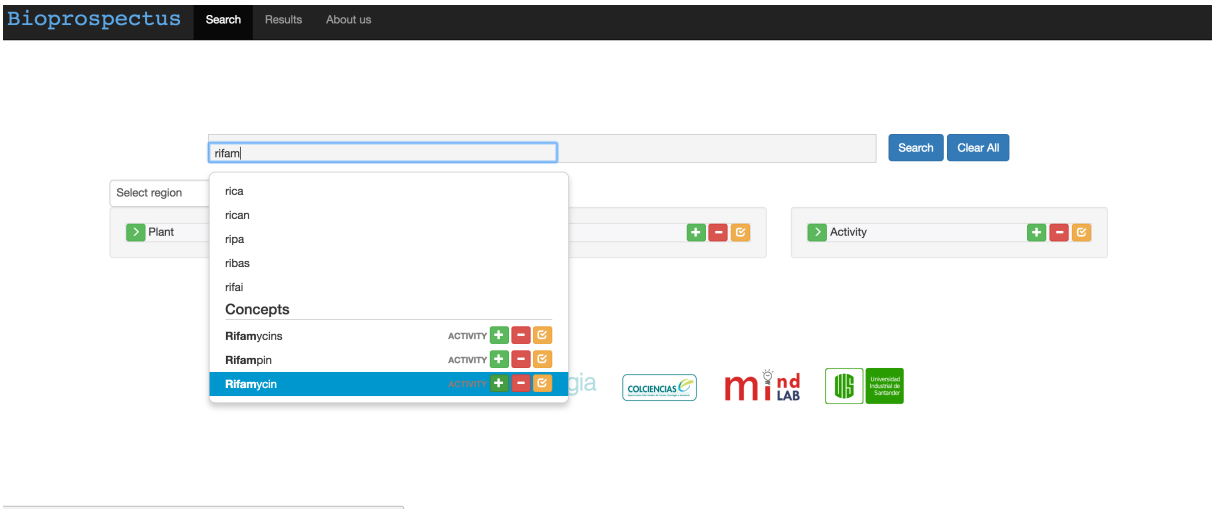

Figure 3.10: Bioprospectus Sugestion Box

Tabla 3.1: Set of Journals used for the Corpus cosolidation

Name	Publisher
Annals of Global Health	Elsevier
Journal of Applied Research on Medicinal and Aromatic Plants	Elsevier
Biotechnology Reports	Elsevier
Current Plant Biology	Elsevier
Electronic Journal of Biotechnology	Elsevier
Genomics, Proteomics & Bioinformatics	Elsevier
Journal of Medicinal Plants Research	Academic Journals
Journal of Herbs, Spices & Medicinal Plants	Taylor and Francis Online
European Journals of Medicinal Plants	Science Domain International
Phytomedicine	Elsevier
International Journal of Phytomedicine	Advanced Research Journals
Bioactive Molecules and Medicinal Plants	Springer

researchers who need to share information this particular domain.

Nowadays, the ontologies on the Web range from large taxonomies categorizing Web sites to categorizations of plants and their features, scientific articles or medical /biological concepts, among others. However, the amount of information available is high but not all can be relevant for the Bioprospectus platform domain, for that reason as part of the system development we have decided to build a referential ontology in order to:

- Establish a common language inside the information retrieval engine.
- Allow the user to extend/refine their search with relevant terms.
- Add better information exploration capabilities

In order to build the ontology we have defined some relevant sources of information according to

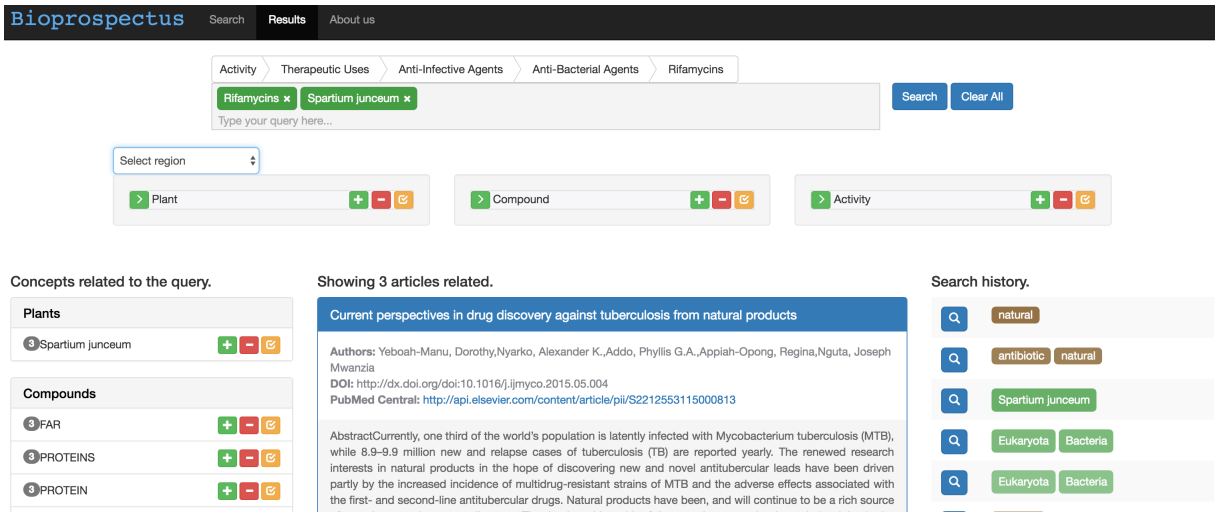

Figure 3.11: Bioprospectus Results Window

Tabla 3.2: Set of Public access database used for the ontology cosolidation

Database	URL
Chemical & Drug Information	http://chem.sis.nlm.nih.gov/chemidplus/
ChemSpider	http://www.chemspider.com/
ChemSynthesis	http://www.chemsynthesis.com/
eMolecules	https://www.emolecules.com/#?click=screening-compounds
Super Natural II	http://bioinf-applied.charite.de/supernatural_new/index.php
chEMBL	https://www.ebi.ac.uk/chembldb/
Dr. Duke's Phytochemical and Ethnobotanical Databases	http://www.ars-grin.gov/duke/
Pubchem	https://pubchem.ncbi.nlm.nih.gov/
Taxonomy NCBI	http://www.ncbi.nlm.nih.gov/taxonomy
GBIF	http://www.gbif.org/
SIB COLOMBIA	http://www.sibcolombia.net/web/sib/home
Institute of Biological Resources Research Alexander von Humboldt	http://www.humboldt.org.co/

each of the categories of the Bioprospectus platform. The data sources pre-selected on chemical compounds, plants taxonomy and/or bioactivities are listed in Table 3.2.

- Chemical Drug Information http://chem.sis.nlm.nih.gov/chemidplus/: It has about +400K chemicals including synonyms and structures, to find chemicals listed in the database of the National Library of Medicine (NLM).
- ChemSpider http://www.chemspider.com/: ChemSpider is a free database of chemical structures that provides quick access to more than +39M structures, properties and associated information. By integrating and linking compounds from 500 data sources. It allows researchers to discover a more complete picture of the chemical data freely available in one online search. It is owned by the Royal Society of Chemistry.
- Medical Subject Headings (MeSH) https://www.ncbi.nlm.nih.gov/mesh: Is a controlled medical vocabulary managed by the NCBI and is used foe indexing articles for the PubMed database.
- ChemSynthesis http://www.chemsynthesis.com/: It provides free access to synthetic chemicals with references that are not limited to a few magazines. Alongside these references

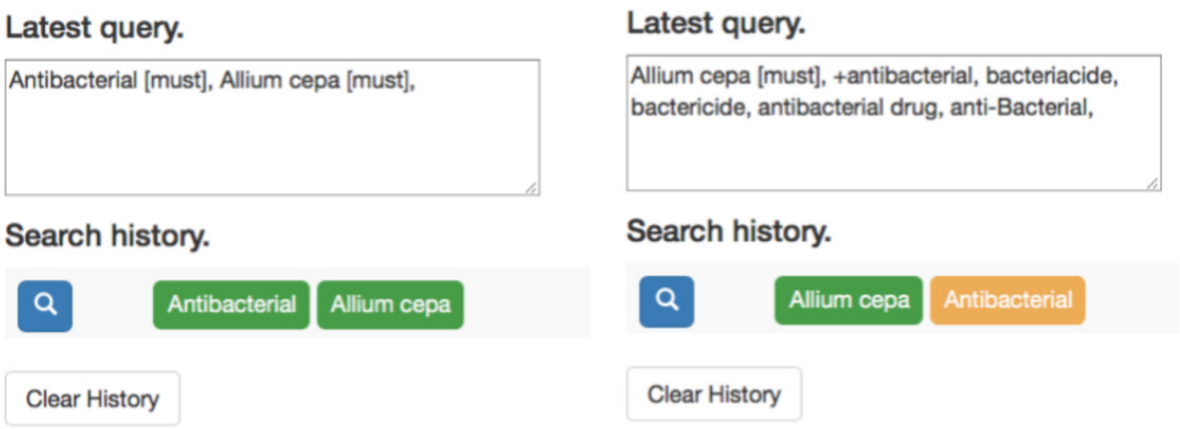

Figure 3.12: The Latest Query window shows the detail of how many items as weel as the boolean operator used in the query

synthesis also it contains physical properties of the substances listed . There are currently more than +40K compounds and over +45K synthesis references in the database.

- eMolecules https://www.emolecules.com/#?click=screening-compounds: This database has about +7M compounds. The search for compounds is performed by the exact structure, substructure and similarity, or by importing files or SMILES SD files.

- Super Natural II http://bioinf-applied.charite.de/supernatural_new/index.php: It is a database of natural products. Contains +325K natural compounds, including corresponding structures, physicochemical properties, toxicity and potential suppliers 2D information. The database provides information on the taxonomy and some medicinal properties of natural products.

- chEMBL https://www.ebi.ac.uk/chembldb/: Chembl is a database of small bioactive molecules, such as drugs . 2D structures contains, calculated properties (for example , log P , Molecular Weight , Lipinski parameters , etc.) and bioactivities (binding constants, pharmacology and ADMET data). The data are cured by primary scientific literature, and covers a significant fraction of the SAR and the discovery of modern drugs.

- Dr. Duke's Phytochemical and Ethnobotanical Databases http://www.ars-grin.gov/duke/: This database allows consulting several selectable lists of plants containing chemicals with a particular activity.

- Pubchem https://pubchem.ncbi.nlm.nih.gov: PubChem provides information on the biological activities of small molecules. PubChem is organized as three databases linked in the information retrieval system NCBI Entrez . These are substances PubChem , PubChem Compound and PubChem bioassays.

- Taxonomy NCBI http://www.ncbi.nlm.nih.gov/taxonomy: The NCBI Taxonomy database is a curated set of names and classifications for all of the organisms that are represented in-GenBank. When new sequences are submitted to GenBank, the submission is checked for new organism names, which are then classified and added to the Taxonomy database.

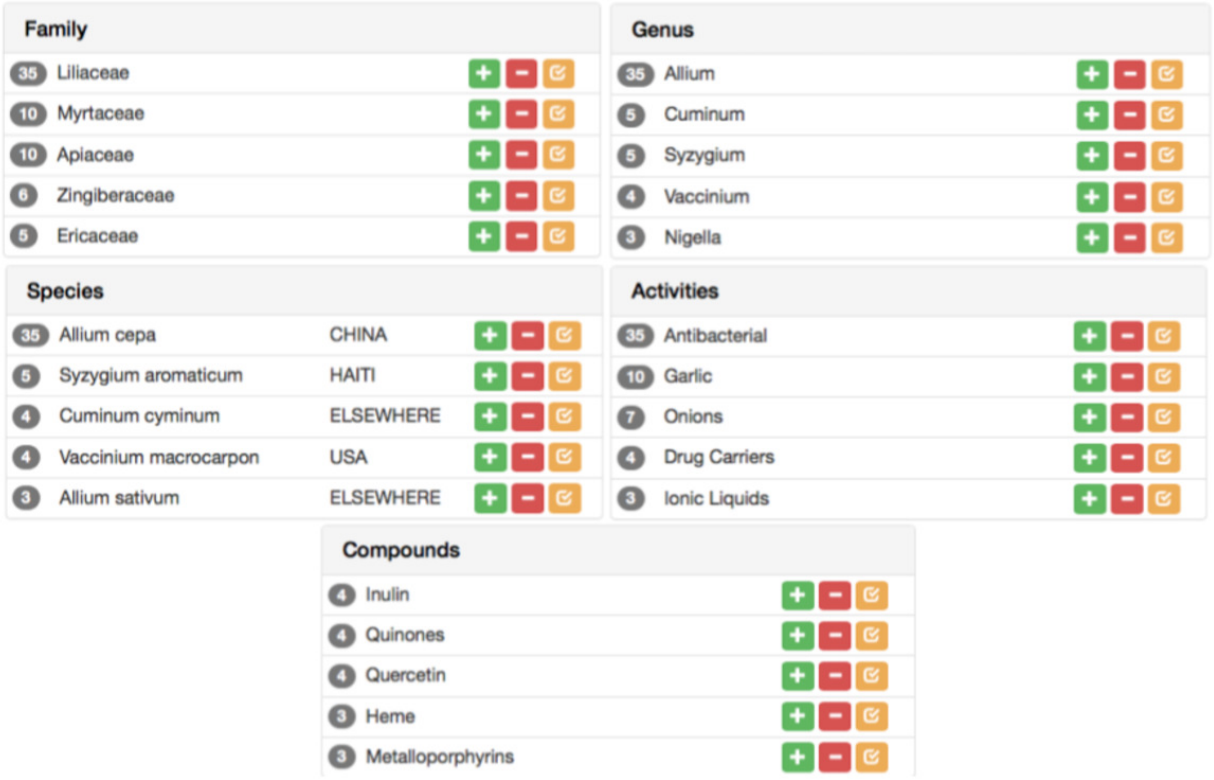

Figure 3.13: The Suggested Concepts window shows the detail of how many times a term apears in the articles retrieved in the lastest query

- GBIF http://www.gbif.org/ Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments. It allows access to anyone, anywhere to data on all types of life on Earth. It provides a single point of access to more than +570M records, freely shared by hundreds of institutions around the world, so is the database largest Internet biodiversity. The data accessible through GBIF relate to evidence of more than 1.6 million species, collected over three centuries of natural history and exploration including current observations of scientists, researchers and automated monitoring programs.

- SIB COLOMBIA http://www.sibcolombia.net/web/sib/home The SiB Colombia is a country initiative aims to provide free access to information on biological diversity in the country for building a sustainable society. This initiative facilitates the online publication of biodiversity data and access to a wide variety of audiences , supporting timely and efficient integrated management of biodiversity. The SiB Colombia , is a network of networks , responsible for providing tools for integrating , publishing and consult biodiversity information (data , metadata, reference sets and species fact sheets) , to make it more readily available to users. Thus, although access and use of information through a single platform allowed ensures that it is properly used and recognized authorship. The publication scheme of SiB Colombia is a free service that is supported by the tool IPT (Integrated Publishing Toolkit, for its acronym in English), a web open source application developed by Global Infrastructure Biodiversity Information Facility (GBIF) and it has been customized by the SiB Colombia to publish and

```
{"article": {
   "@doi": "http://dx.doi.org/doi:10.1063/2.1204305",
   "title": "Command functions of open loop galvanometer scanners with optimized duty cycles ",
   "abstract": "The paper approaches the problem of the command functions of galvanometer-based scanners (GS) that
   are necessary to produce the linear plus parabolic scanning function of the GS, which we have proved previously to
   produce the highest possible duty cycle (i.e., time efficiency) of the device. We have completed this theoretical aspect
   (which contradicted what has been stated previously in the literature, where it has been considered that the linear plus
   sinusoidal scanning function was the best) with the experimental study of the most used scanning functions of the GSs
   (sawtooth, sinusoidal and triangular), with applications in biomedical imaging, in particular in optical coherence
   tomography, demonstrating that the triangular function is always the best one to be applied, from both an optical and a
   mechanical point of view. In the present study the input voltage/command function which should be applied to the GS
   to produce the desired triangular scanning function (with controlled non-linearity for the fastest possible stop-and-turn
   portions) was determined analytically, in relationship with the active torque that drives the device. This command
   function is analyzed with regard to the specific, respectively required parameters of the GS: natural frequency and
   damping factor, respectively scan speed and amplitude. The modeling in an open loop control structure of the GS is
   finally discussed as a trade-off between using the highest possible duty cycle and minimizing the maximum peaks of
   the input voltage.",
   "authors": {
      "author": ["Duma, V.F."]
      },
      "@pmid": "",
      "@pmc_article_url": "http://api.elsevier.com/content/article/pii/S2095034915301641"}
}
```


Figure 3.14: Example of an article saved as JSON before processing

record resources (datasets with their associated metadata)

- Institute of Biological Resources Research Alexander von Humboldt http://www.humboldt.org.co/ Humboldt Institute 's mission is to promote, coordinate and conduct research that contributes to knowledge, conservation and sustainable use of biodiversity as a factor of development and welfare of the Colombian population. Networking works with multiple organizations with capacity to influence decision -making and public policy. As part of its functions, the Institute is responsible for conducting , on the mainland of the Nation , scientific research on biodiversity, including hydro-biological and genetic resources. It also coordinates the National Biodiversity Information System (SIB Colombia) and the establishment of the national inventory of biodiversity.

After the selection of the candidates data sources we have defined some technical parameters for comparison in order to defined our knowledge base and evaluate the relevance of complement this Knowledge source with data from different databases. The technical comparison is shown in the Table 3.3 for the biological Databases and Table 3.4 for compounds and Biological activities.
After this comparison and taking into account the amount of data available, the depth level of each database as well as its relevance in the field we have selected as the most appropriate database to be implemented in the system for each category:

- Taxonomic data are: GBIF.
- Chemical compounds: Pubchem and ChEBI
- Bioactivities: MeSH and ChEBI

Tabla 3.3: Comparison between the different biological datasources colected for the Ontology building

	Taxonomy	GBIF	NY-Botanical	ICN-UNAL
Data Type				
Data availability	Yes	Yes	No	No
Data Format	SQL	CSV	Web Interface	Web Interface
Version Control	Yes	Yes	-	Yes
Amount of Data	+1M	+570M	+7.2M	+520K
General Information				
Is there Parent-Son relationships?	Yes	Yes	Yes	Yes
Has It data not only in English?	Yes	No	No	Yes
Has It WorlWide Distributtion?	Yes	Yes	Yes	Yes
Biologic Information				
Has It biodiversity information?	Yes	Yes	Yes	Yes
Has It information of plant location?	Yes	Yes	Yes	Yes
Has It information of common names?	Yes	Yes	Yes	Yes

Once we have downloaded and processed each database we have collected the information and built the ontology with the following structure:

- Table Concepts has 58974 active and unique concepts
  - ConceptID: It is an unique identifier of each concept.
  - Fully Specified Name: Each concept has at least one Fully Specified Name (FSN) intended to provide an unambiguous way to name a concept. The purpose of the FSN is to uniquely describe a concept and clarify its meaning. The FSN is not a commonly used term or natural phrase and would not be expected to appear in the human-readable representation of a clinical record.
  - Category: describe each of the specified categories in the project (Plants, Compound or Activities)
  - Country: For the plant concept provides the specification of the country where the plant was reported.
- Table Descriptors has 61283 descriptors from which around 7000 corresponds to abbreviated forms this descriptors can be seen as synonyms or plants common names by which you can refer a concept.
  - DescriptorID: It is an unique identifier of each descriptor
  - ConceptID: Relates the Descriptor with the main Concept.
  - Prefered Term: The Preferred Term (PT) represents a common word or phrase used to name a concept in practice or in the literature. Only one of the available Descriptions can be designated as preferred. Depending on the computer application, the PT may be the Description selected for display, although each application is assumed to be allowed to independently determine the most appropriate display term for a given situation.
- Table Relationships includes 61510 relations between concepts
  - RelationshipID: It is an unique identifier of each relationship.

Tabla 3.4: Comparison between the different Chemical datasources colected for the Ontology building

	PubChem	CheMBL	CHeBI	ChemSpider	Dr. Dukes	MeSH
Data Type						
Data availability	Yes	Yes	Yes	Yes	Yes	Yes
Data Format	XML	SQL	OWL	Web Service	Web Service	SQL
Version Control	Yes	Yes	No	No	No	Yes
Amount of Data	+72M	+1M	+100K	+43M	+5.75M	+7M
General Information						
Is there Parent-Son relationships?	Yes	Yes	Yes	Yes	Yes	Yes
Has It data not only in English	No	No	No	Yes	Yes	Yes
Has It WorlWide Distributtion?	Yes	Yes	Yes	Yes	Yes	Yes
Biological Information						
Has It Biodiversity information?	No	No	Yes	No	Yes	Yes
Has It common names information?	No	No	Yes	No	Yes	Yes
Chemical Information						
Has It Chemical Structure Information?	Yes	Yes	Yes	Yes	No	Yes
Has It synonyms?	Yes	Yes	Yes	Yes	Yes	Yes
Has It of biological activities?	No	Yes	Yes	Yes	Yes	Yes
Has It SMILE information?	Yes	No	Yes	No	No	Yes

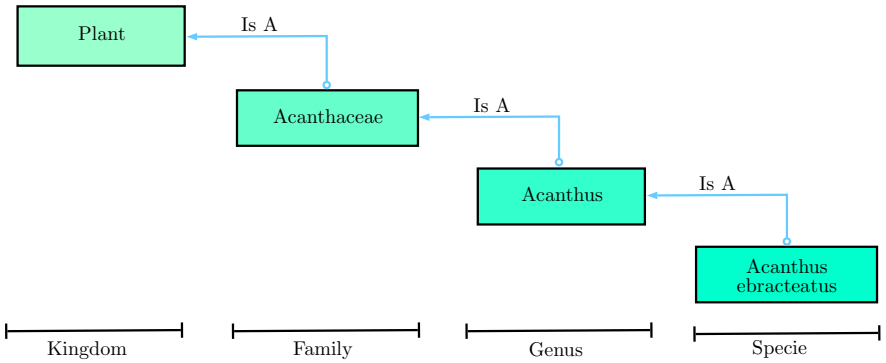

Figure 3.15: Example of Vertical relations created in the ontology

- ConceptID1: Relates the Descriptor with the first Concept.
- RelationshipType: There are several types of relationships described or modeled in the database:
  - Is A: The "Is a" relationship is used to create a hierarchical relationship between concepts, relating specific concepts to a more general category.
    For example:
  - Has: The "Has" relationship is used to create a horizontal relationship between concepts, relating specific concepts to a more general category.

- ConceptID2: Relates the Descriptor with the second Concept.

The relationships between tables in the ontology is depicted in Figure 3.17 (entity relationship diagram):

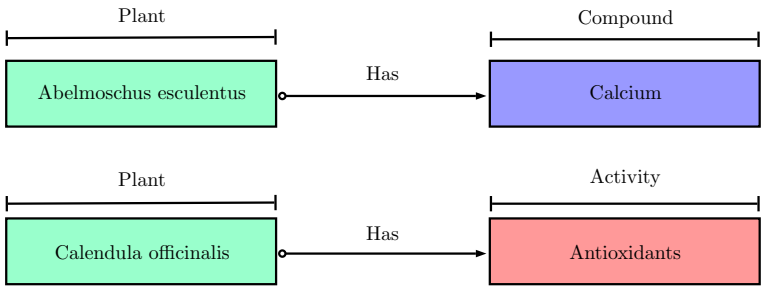

Figure 3.16: Example of Horizontal relations created in the ontology

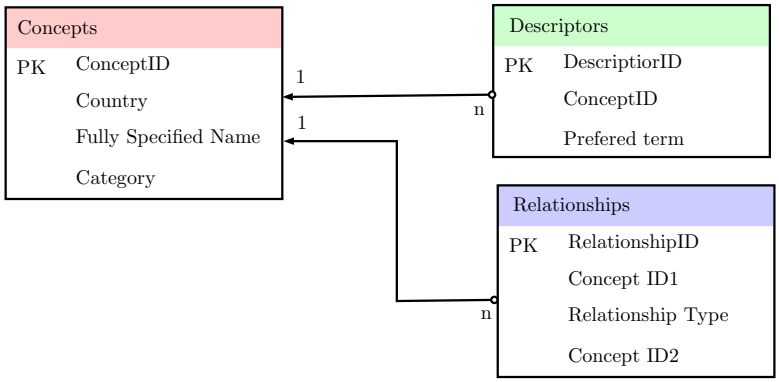

Figure 3.17: Entity relation diagram of the Ontology

3.3.1 System Performance - Use Case

Bioprospecting can be understood as a new approach to give use to the biodiversity, through the systematic exploration of biological sources with potential developing new compounds and products. One of the most interesting areas of research is related to find new therapeutic uses of the natural resources, particularly plants available in a Country. This application has a huge impact, because the pharmaceutical industry is trying to get the medicines closer to the population, decreasing their prices and increasing their coverage.

For the qualitative evaluation of the system we have started looking for the *Anesthesia* activity in humans, analysing the compound and the species of plants present in the scientific articles with this biological activity. This need arises because there is now a growing interest in natural analgesic derived from secondary metabolites of plants, their essential oils and flavonoids.

Once we have made the query 477 articles were retrieved. In those articles some compounds were listed: Hematoxylin, Methyl Ethers, Isoflurane, Ketamine and Xylazine. Between those concepts Hematoxylin is a compound extracted from the heartwood of the logwood tree and is commonly used in Histopathology. It has a non direct relationship with the anesthesia activity. Additionally, Xylazine is commonly used for sedation, anesthesia, muscle relaxation, and analgesia in animals such as horses, cattle and other non-human mammals. As we are looking for the compounds related to anesthesia an particularly in humans, those two concepts can be removed (marked as - in the query) an update the results.

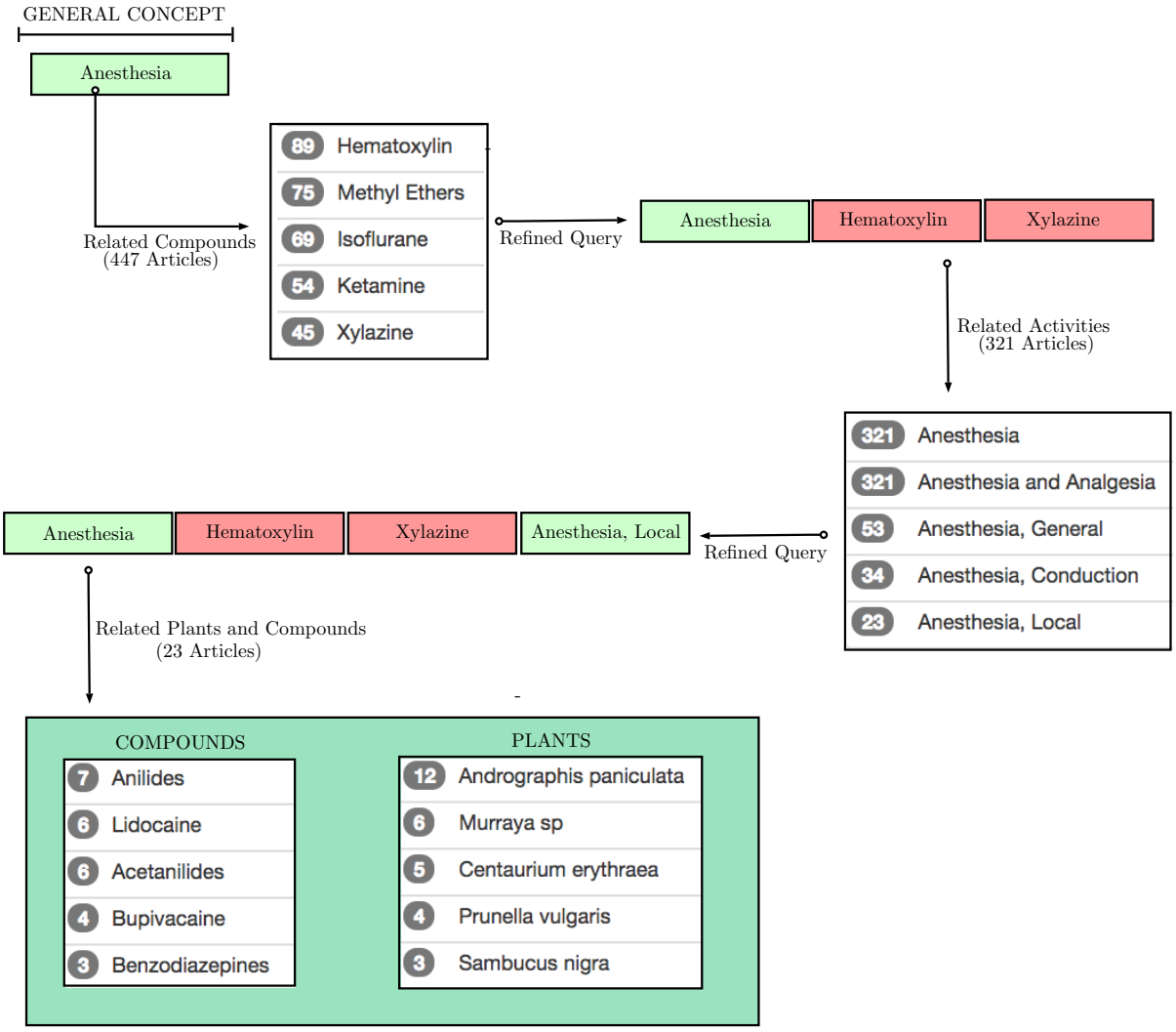

Figure 3.18: Quealitative evaluation use case

After this update, 392 articles were found. So we have decided to get deeper in the information exploration adding the term Anesthesia Local from the suggestions in the activitie's Box. With this additional filter, the amount of articles get reduced to 23. With this articles, in the specie's box some herbal species with a great anesthesic potential are reported. From this set, the 48% of the articles have relationship with *Andrographis paniculata*. This specie has a very wide tradition in the Indian Ayurvedic medicine and several studies have shown that *Andrographis paniculata* products have sedative effects [64].

In this set of species *Sambucus nigra* were also found, this plant has been widely used against viral diseases but does not has reported any biological use related with the Anesthesia. However, making a deeper analysis of the articles retrieved, we have found that the articles were the *Sambucus nigra* were present the *Andrographis paniculata* was also described. This behavior is justified because, the Bioprospectus systems index all the concepts present in an article but finding the relationship between those concepts is a functionality that can be explore in a second version of the system.

Chapter 4

Evaluation and Results

This chapter contains the details of the quantitative evaluations proposed to evaluate the system. We have followed the two standard measures used in the literature to evaluate this kind of systems. The performance of the system was compared against a reference system. The relevance or not relevance quantification was made by experts of the Biotechnology Institute from the National University of Colombia

4.1 Experimental Setup

In order to compare and evaluate the performance of the Bioprospectus system, a group of experts must generate binary relevance judgements of the articles retrieved by the system in a particular set of information needs. Two metrics: precision and Mean Average Precision were calculated.

4.1.1 Systems Performance

To measure the performance of an information retrieval system it is necessary to have at least one information need (query) along with their associated ranked results, this allows the experts to binary classify the documents in a collection as relevant or irrelevant and then compute the following measures.

- **Precision@N**: which corresponds to the proportion of the retrieved documents that are relevant over the total. The first N retrieved values are taken into account for the evaluation. We have set the value of N in a typical value of 50.
- **Mean Average Precision (MAP)@N**: Corresponds to average of the precision value obtained for the set of top N documents existing after each relevant document is retrieved, and this value is then averaged over information needs.

$$MAP = \frac{\sum_{q=1}^{n} P[n]}{n} \tag{4.1}$$

According to this, to evaluate the Bioprospectus system we have defined five (5) information needs to be seek in the system. The information needs were the following:

- IN-1. Plant AND Antibacterial

- IN-2. Plant AND Antiviral
- IN-3. Plant AND Antifungal
- IN-4. Plant AND Antiparasitic
- IN-5. Plant AND Disinfectants

To evaluate the performance of the retrieval system and integrate the knowledge base of the ontology developed, we have tested the five information needs with the two queries expansion QE1, QE2 described in the figures 3.6 and 3.7.

4.2 Results

This section contains all the results obtained for the quantitative evaluations of five information needs related to the Bioprospecting problem.

4.2.1 System Performance

The Bioprospectus system is a domain specific information retrieval systems that is aware of the particular concepts of interest when performing bio-prospection-related information search. Bioprospectus allows users to search a large database of scientific papers related to biodiversity and its industrial uses. Bioprospectus has a large knowledge base that encompasses plants, chemical compounds and biological activities. Bioprospectus additionally offers some extra capacities to facilitate the user experience: such as results highlighting and reports generation.

To evaluate the system. five (5) information needs and two (2) queries expansions were used for the system evaluation. For each information need we have calculated precision and MAP over the first 50 documents retrieved. The results for the QE1,QE2 and the reference system are shown in Figure 4.1 and 4.2 respectivelly.

The results for the QE1, evidence that retrieving articles using the synonyms of a particular term as an additional source of information and as a complement of the query results in a better formance of the Bioprospectus system against the reference system in four of the five informacion needs. By the other side, retrieving articles using the set of relations (QE2) of a particular term in the ontology increase the precision only in two of the five information needs compared to the QE1 results. In a priori, someone can believe that integrate more related terms to the query can generate better results, because you can expand your article universe, however, this evidence the tradeoff between precision and recall, because you can increase the number of articles but those could be worse related to the information need. Other example of this behavior occurs seeking the term IN2-Antiviral. Where the Bioprospectus system reach a precision of 92% using the QE1 while the QE2 was 72% and ScienceDirect precision is 42%. In this experiment, eighth synonyms of the antiviral concept were used to make a deepest inspection of the articles collected. However, in the QE2 twenty-nine (29) related concepts were used to refine the query generating a reduction in the precision index performance.

By the other hand, the highest performance was obtained for the IN1-Antibacterial using the QE1, beating the precision goten in the QE2 and in the reference system. In this case, the performance

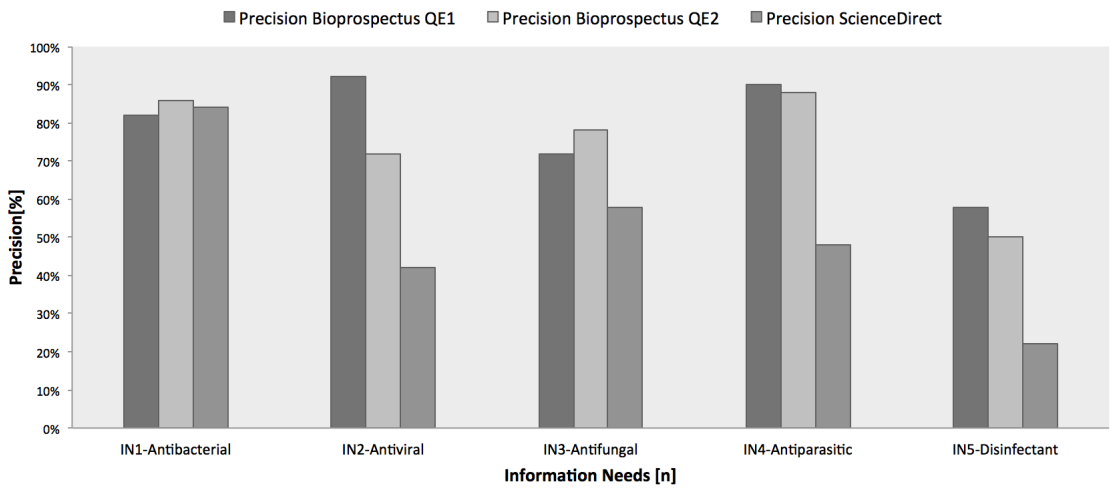

Figure 4.1: Comparison of the Precision for all the information needs between Bioprospectus and ScienceDirect for all the information needs

improvement could be related to the great amount of articles related with Antibacterial in the compendium and many of them related with chemical compounds with reported antibacterial effect.

The lower performance of the two systems was reached during the evaluation of the articles related to the concept IN5-Disinfectant. The Bioprospectus system reach 58% and 50% using four synonyms and the related terms respectively, while ScienceDirect 22%. The root cause of this performance, was related with the lower amount of articles which has relationships with the disinfection process. According to an exhaustive analysis of the compendium of 16k articles collected, just a 0.15% of the articles has information related with the disinfection process.

The curves precision-recall were generated for the five information needs. Figures 4.4 and 4.4 shows the comparison between the two systems for the Plant AND Antibacterial, IN-2. Plant AND Antiviral respectively. In those curves, It is shown that for the antibacterial concept both systems have a high precision-recall measure, which means that both systems are returning accurate results (high precision), as well as returning a majority of all positive results (high recall). However, in the case of the antiviral concept, the value of recall measure to the Science Direct is lower. This is directly related to a high false negative rate. In this case the system return many results, but most of its didn't satisfy the information need.

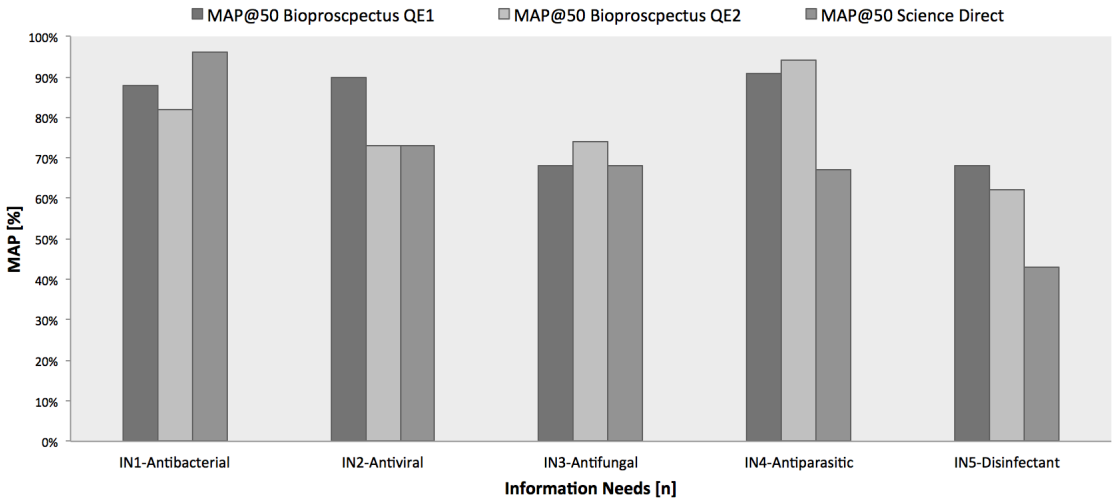

Figure 4.2: Comparison of the Precision for all the information needs between Bioprospectus and ScienceDirect for all the information needs.

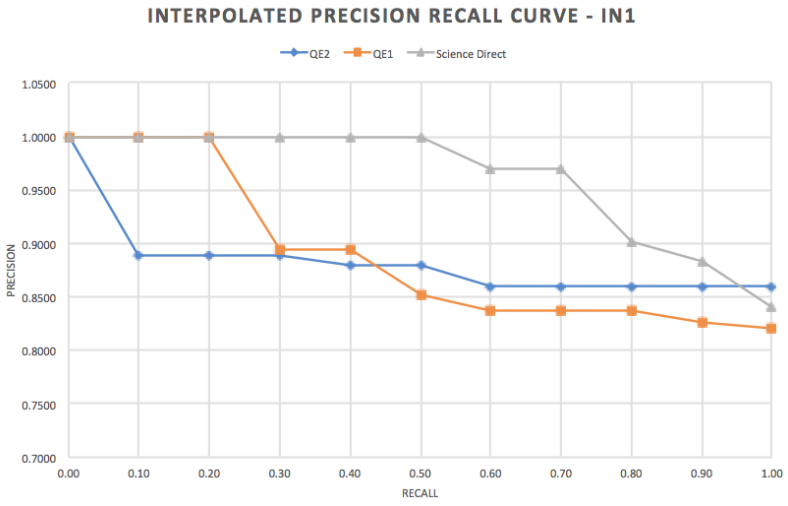

Figure 4.3: Precision-Recall comparison for Antibacterial information need. Bioprospectus QE1, QE2 and Science Direct were tested

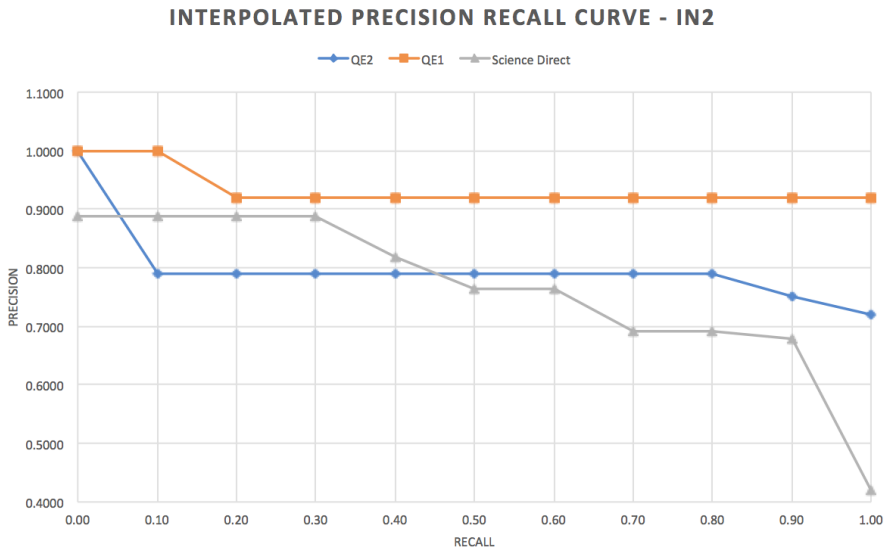

Figure 4.4: Precision-Recall comparison for Antiviral information need. Bioprospectus QE1, QE2 and Science Direct were tested

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This work presents a bioprospecting software system which integrates different information sources that collectively support the process of looking for potential industrial uses of biodiversity. The most important feature of the system is its ability to automatically identify concepts, which are relevant to the bioprospecting process, in a document database, relate those concepts to the information present in knowledge databases and show them to the user in a friendly and effective way.

The system was tested using two approaches. The first one, focus on extract performance measures such as: Precision and mean average precision (MAP), and compare them against the performance of a well known retrieval system (ScienceDirect).

During this evaluation, two queries expansion methods were used in the Bioprospectus system. With those queries expansion (QE) we want to integrate the knowledge ontology developed into the retrieval process and not only during the indexation. Thus, we have added to the seeked term in the QE1 the synonyms and in the QE2 the related concepts. In addition, five Information needs were tested in both systems. As the main result, the precision of the Bioprospectus system were better than the precision reach using ScienceDirect in four of the five queries tested. The better performance were reached seeking IN2-Antiviral using the QE1, in this case the system precision were 92%. However, the lower performance were reached in the IN5-Disinfectant were the system gets a precision of 50%. The main reason of this performance could be related with the lack of articles related to this particular topic.

In the second approach, we have shown a specific use case related to the bioprospecting task. The use case tested the system's usability as well as the pertinence of the concepts suggested by the system. For that, we have defined a general query and used the filters to explore in a deepest way the information presented by the system. As a result, we have formulated a query with five terms, and as a result we have found a well described pharmacological use of a *Fabaceae* specie (*Senna singueana*). In general, the system holds great promise providing a technological foundation to identify new bio-products from the Colombian biodiversity targeting sustainable development and added value of our biodiversity.

5.2 Future Work

As future work, we expect to generalize the automatic term mapping functions in order to find explicit relationships into the text and not only focus on the concepts occurrence.
Finding automatic relationships can automatically increase the set of concepts and relationships of the knowledge base and serve as evidence to establish the potential industrial use of different species of the Colombian biodiversity.

Chapter 6

Bibliography

[1] N. A. for Drug and F. C. Indonesia, "Trips, cbd and traditional medicines: Concepts and questions.," in *Report of an ASEAN Workshop on the TRIPS Agreement and Traditional Medicine*, National Agency for Drug and Food Control World Health Organization, February 2001. 2.1

[2] C. Hayden, *When nature goes public: The making and unmaking of bioprospecting in Mexico*. Princeton University Press, 2003. 2.1

[3] F. Crestani, "Application of spreading activation techniques in information retrieval," *Artificial Intelligence Review*, vol. 11, no. 6, pp. 453–482, 1997. 2.1, 2.2

[4] N. Mateo, W. Nader, and G. Tamayo, "Bioprospecting," *Encyclopedia of biodiversity*, vol. 1, pp. 471–488, 2001. 2.1

[5] H. Polur, T. Joshi, C. T. Workman, G. Lavekar, and I. Kouskoumvekaki, "Back to the roots: Prediction of biologically active natural products from ayurveda traditional medicine," *Molecular Informatics*, vol. 30, no. 2-3, pp. 181–187, 2011. 2.1, 2.1.1

[6] K. H. et al, "A systems approach to traditional oriental medicine," *Nat Biotech*, vol. 33, no. 3, pp. 264–268, 2015. 2.1, 2.1.1

[7] F. Bolaños, "Biodiversity of costa rica," 2017. 2.1

[8] E. García, "Comisión nacional para el conocimiento y uso de la biodiversidad (conabio). 1998," *Climas (clasificación de Köppen, modificado por García). Escala*, vol. 1, no. 1, p. 000, 2017. 2.1

[9] B. Patwardhan, "Ayugenomics-integration for customized medicine," *Indian J. Nat. Prod.*, vol. 19, pp. 16–23, 2003. 2.1.1

[10] S. Straus, "Herbal medicines - what's in the bottle?," *New Engl. J. Med*, vol. 347, no. 1997-1998, 2002. 2.1.1

[11] C. P. Adams and V. V. Brantner, "Estimating the cost of new drug development: is it really 802 million?," *Health Affairs*, vol. 25, no. 2, pp. 420–428, 2006. 2.1.1

[12] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of health economics*, vol. 22, no. 2, pp. 151–185, 2003. 2.1.1

[13] D. G. Brown, T. Lister, and T. L. May-Dracka, "New natural products as new leads for antibacterial drug discovery," *Bioorganic & medicinal chemistry letters*, vol. 24, no. 2, pp. 413–418, 2014. 2.1.1

[14] A. A. Lagunin, R. K. Goel, D. Y. Gawande, P. Pahwa, T. A. Gloriozova, A. V. Dmitriev, S. M. Ivanov, A. V. Rudik, V. I. Konova, P. V. Pogodin, D. S. Druzhilovsky, and V. V. Poroikov, "Chemo- and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review," *Nat. Prod. Rep.*, vol. 31, pp. 1585–1611, 2014. 2.1.1

[15] N. R. Farnsworth, O. Akerele, A. S. Bingel, D. D. Soejarto, and Z. Guo, "Medicinal plants in therapy," *Bulletin of the world health organization*, vol. 63, no. 6, p. 965, 1985. 2.1.1

[16] D. Armenteras-Pascual, J. Retana-Alumbreros, R. Molowny-Horas, R. M. Roman-Cuesta, F. Gonzalez-Alonso, and M. Morales-Rivas, "Characterising fire spatial pattern interactions with climate and vegetation in colombia," *Agricultural and Forest Meteorology*, vol. 151, no. 3, pp. 279–289, 2011. 2.1.1

[17] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. 2.2

[18] J. A. Waterworth and M. H. Chignell, "A model of information exploration.," *Hypermedia*, vol. 3, pp. 35–58, 2002-01-03 1991. 2.2, 2.2.1

[19] N. J. van Eck and L. Waltman, "Citnetexplorer: A new software tool for analyzing and visualizing citation networks," *Journal of Informetrics*, vol. 8, no. 4, pp. 802–823, 2014. 2.2

[20] A. S. Raamkumar, S. Foo, and N. Pang, "A framework for scientific paper retrieval and recommender systems," *arXiv preprint arXiv:1609.01415*, 2016. 2.2

[21] N. Moon and R. Singh, "Experiments in text-based mining and analysis of biological information from medline on functionally-related genes," in *Systems Engineering, 2005. ICSEng 2005. 18th International Conference on*, pp. 326–331, Aug 2005. 2.2.1, 2.2.2

[22] S. I. O'Donoghue, H. Horn, E. Pafilis, S. Haag, M. Kuhn, V. P. Satagopam, R. Schneider, and L. J. Jensen, "Reflect: A practical approach to web semantics," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, 2010 2010. 2.2.1, 2.2.1, 2.2.2

[23] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000. 2.2.1

[24] G. Grefenstette and P. Tapanainen, *What is a word, what is a sentence?: problems of Tokenisation*. Rank Xerox Research Centre, 1994. 2.2.1

[25] Y. Wang, Z. Yang, H. Lin, and Y. Li, "A syntactic rule-based method for automatic pathway information extraction from biomedical literature," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pp. 626–633, Oct 2012. 2.2.1, 2.2.1

[26] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech communication*, vol. 32, no. 1, pp. 127–154, 2000. 2.2.1

[27] J. Nivre, "An efficient algorithm for projective dependency parsing," in *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*, Citeseer, 2003. 2.2.1

[28] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič, "Non-projective dependency parsing using spanning tree algorithms," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 523–530, Association for Computational Linguistics, 2005. 2.2.1

[29] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of biomedical informatics*, vol. 37, no. 6, pp. 512–526, 2004. 2.2.1

[30] L. Hirschman, A. A. Morgan, and A. S. Yeh, "Rutabaga by any other name: extracting biological names," *Journal of Biomedical Informatics*, vol. 35, no. 4, pp. 247–259, 2002. 2.2.1

[31] K. B. Cohen, G. K. Acquaah-Mensah, A. E. Dolbey, and L. Hunter, "Contrast and variability in gene names," in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pp. 14–20, Association for Computational Linguistics, 2002. 2.2.1

[32] O. Tuason, L. Chen, H. Liu, J. A. Blake, and C. Friedman, "Biological nomenclatures: a source of lexical knowledge and ambiguity," in *Proceedings of the Pacific Symposium of Biocomputing*, no. 9 in 0, p. 238, 2003. 2.2.1

[33] H. Yu, G. Hripcsak, and C. Friedman, "Mapping abbreviations to full forms in biomedical articles," *Journal of the American Medical Informatics Association*, vol. 9, no. 3, pp. 262–272, 2002. 2.2.1

[34] Y. Kim, J. Hurdle, and S. M. Meystre, "Using umls lexical resources to disambiguate abbreviations in clinical text," in *AMIA Annual Symposium Proceedings*, vol. 2011, p. 715, American Medical Informatics Association, 2011. 2.2.1

[35] M. S. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," *Pacific Symposium of Biocomputing 2003*, 2003. 2.2.1

[36] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006. 2.2.1

[37] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001. 2.2.1

[38] L. A. Riveros Cruz *et al.*, *A knowledge-based approach to information retrieval in collections of textual documents of the biomedical domain.* PhD thesis, Universidad Nacional de Colombia, 2015. 2.2.1, 3.1.1

[39] M. Marrero, J. Urbano, S. Sanchez-Cuadrado, J. Morato, and J. M. Gomez-Berbas, "Named entity recognition: Fallacies, challenges and opportunities," *Computer Standards Interfaces*, vol. 35, no. 5, pp. 482–489, 2013. 2.2.1

[40] A. Boldyrev, "Dictionary-based named entity recognition," master's thesis in computer science, Universitat des Saarlandes Max-Planck-Institut fur Informatik Databases and Information Systems, December 2013. 2.2.1

[41] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan, "Domain adaptation of rule-based annotators for named-entity recognition tasks," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (Stroudsburg, PA, USA), pp. 1002–1012, Association for Computational Linguistics, 2010. 2.2.1

[42] L. He, Z. Yang, H. Lin, and Y. Li, "Drug name recognition in biomedical texts: a machine-learning-based method," *Drug Discovery Today*, vol. 19, no. 5, pp. 610 – 617, 2014. 2.2.1

[43] I. Segura-Bedmar, P. Martínez, and M. Segura-Bedmar, "Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems," *Drug Discovery Today*, vol. 13, no. 17–18, pp. 816 – 823, 2008. 2.2.1

[44] S. Morwal, N. Jahan, and D. Chopra, "Named entity recognition using hidden markov model (hmm)," *Int. J. Nat. Lang. Comput.(IJNLC)*, vol. 4, no. 1, pp. 15–23, 2012. 2.2.1

[45] J. R. Curran and S. Clark, "Language independent ner using a maximum entropy tagger," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, vol. 4, pp. 164–167, Association for Computational Linguistics, 2003. 2.2.1

[46] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, (Stroudsburg, PA, USA), pp. 188–191, Association for Computational Linguistics, 2003. 2.2.1

[47] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, (Stroudsburg, PA, USA), pp. 104–107, Association for Computational Linguistics, 2004. 2.2.1

[48] W. Waegeman, B. D. Baets, and L. Boullart, "Learning layered ranking functions with structured support vector machines," *Neural Networks*, vol. 21, no. 10, pp. 1511 – 1523, 2008. {ICONIP} 2007. 2.2.1

[49] T. Botsis and R. Ball, "Automating case definitions using literature-based reasoning," *Applied clinical informatics*, vol. 4, no. 4, p. 515, 2013. 2.2.1

[50] L. Cheng, H. Lin, F. Zhou, Z. Yang, and J. Wang, "Enhancing the accuracy of knowledge discovery: a supervised learning method," *BMC Bioinformatics*, vol. 15, no. S-12, p. S9, 2014. 2.2.1

[51] J. Lambek, "Categorial and categorical grammars," in *Categorial grammars and natural language structures*, pp. 297–317, Springer, 1988. 2.2.1

[52] D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. P. Sheth, and T. C. Rindflesch, "A graph-based recovery and decomposition of swanson's hypothesis using semantic predications," *J. of Biomedical Informatics*, vol. 46, pp. 238–251, Apr. 2013. 2.2.1

[53] A. Coulet, S. Nigam, G. Yael, M. Mark, and A. R. B., "Using text to build semantic networks for pharmacogenomics," *Journal of Biomedical Informatics*, vol. 43, pp. 1009–1019, 2010. 2.2.1

[54] Y.-T. Huang, H.-Y. Yeh, S.-W. Cheng, C.-C. Tu, C.-L. Kuo, and V.-W. Soo, "Automatic extraction of information about the molecular interactions in biological pathways from texts based on ontology and semantic processing," in *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, vol. 5, pp. 3679–3684, Oct 2006. 2.2.1

[55] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo, "Extracting human protein interactions from medline using a full-sentence parser," *Bioinformatics*, vol. 20, no. 5, pp. 604–611, 2004. 2.2.1

[56] D. M. McDonald, H. Chen, H. Su, and B. B. Marshall, "Extracting gene pathway relations using a hybrid grammar: the arizona relation parser," *Bioinformatics*, vol. 20, no. 18, pp. 3370–3378, 2004. 2.2.1

[57] S. Chan, C. Leung, and A. Milani, "Knowledge extraction and mining in biomedical research using rule network model," in *Brain and Health Informatics*, pp. 506–515, Springer, 2013. 2.2.1

[58] C. J. V. Rijsbergen, *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 1979. 2.2.1

[59] B.-H. Cho, C. Lee, and G. G. Lee, "Exploring term dependences in probabilistic information retrieval model," *Information Processing Management*, vol. 39, no. 4, pp. 505 – 519, 2003. 2.2.1

[60] A. Trotman, "Learning to rank," *Information Retrieval*, vol. 8, no. 3, pp. 359–381, 2005. 2.2.1, 2.2.1

[61] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM (JACM)*, vol. 7, no. 3, pp. 216–244, 1960. 2.2.1

[62] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010. 2.2.2

[63] February 2016. 3.2

[64] K. Maiti, A. Gantait, M. Kakali, B. Saha, and P. K. Mukherjee, "Therapeutic potentials of andrographolide from andrographis paniculata: a review," *Journal of Natural Remedies*, vol. 6, no. 1, pp. 1–13, 2006. 3.3.1

Comments and Signature

Fabio González

Samier Said Barguil Giraldo

Firma

Firma

Fecha

Fecha