Natural Language Query Processing Framework for Biomedical Literature

Carmen De Maio Giuseppe Fenza Vincenzo Loia Mimmo Parente

Department of Computer Science, University of Salerno, Italy

Abstract

The availability of huge amount of biomedical literature over the Web offers a big opportunity to carry out useful information about published research results. Nevertheless, these information are often enclosed in unstructured documents stressing the need to define suitable framework to support execution of analytics services and richer information discovery tasks. This work introduces a general framework to support natural language user's query over facet-based data model. It relies on components for knowledge extraction and ontology matching to categorize input biomedical resources with respect to existing biomedical ontologies' concepts. The framework has been instantiated implementing Fuzzy Formal Concept Analysis algorithm.

Keywords: Biomedicine, Fuzzy Formal Concept Analysis, Information Retrieval

1. Introduction

The growing amount of biomedical data available over the Web reflects the increasing interest in biomedicine which research results are mainly collected by journal repositories of unstructured biomedical information, such as: PubMed[1] [1] a free full-text archive of biomedical literature, BioMed Central[2] that publishes a range of journals across all biomedical fields, from basic life sciences to clinical medicine, WikiGenes[3] [2] a global collaborative knowledge base for the life sciences data, and so on. Furthermore, several domain ontologies in the field of bio informatics have been developed (e.g., Protein Ontology, GeneOnto[3], and so on).

Since the biomedical investigation and research results are often enclosed in unstructured textual documents, it is difficult to address analytics and natural language queries, such as: "how much is supported a specific biomedical thesis in the research studies?", or "are there any recent evaluation study about the benefits of a specific nanomaterial?", and so on. So, it is necessary to build a useful biomedical knowledge base from unstructured and heterogeneous content in order to support natural language query processing and richer information discovery tasks exploiting intelligent aggregation of different facets.

This paper relies on the definition of a general framework to categorize unstructured biomedical content with respect to available biomedical ontologies, that has been extensively described in [4]. In this work, the resulting facet-based data model will be extended to support natural language query processing feature. The framework relies on two main components that are aimed to associate each biomedical documents to different biomedical ontologies' concepts, these components performs *Knowledge Extraction* and *Biomedical Ontologies Matching*, respectively.

The former extracts an *Unsupervised Ontology* that specifies the results of unsupervised categorization process performed on unstructured data. It will be instantiated in the proposed implementation exploiting wikification services, that is the practice of representing a content with a set of Wikipedia entities (e.g., articles) [5], and consequently applying algorithms of Fuzzy Formal Concept Analysis (FFCA) to extract ontology concepts.

The latter defines a matching algorithm to find relations between concepts of *Unsupervised Ontology* and biomedical ontologies ones. The results of this matching enable a faceted search providing information about the quantities of documents belonging to each biomedical ontologies' concept (e.g., genome of bacteria, plasma membrane), that is a value of specific facet itself.

Finally, the framework introduces a *Natural Language Query Processing* component that is defined to address user's query. As highlighted in [6], we emphasize that the extracted facet-based data model enables to carry out more than just knowing the quantities of documents belonging to each facet. In fact, faceted search engine can be extended to ask for correlated facets to grouping documents across dependent multiple facets. Specifically, when natural language query happens the system will find corresponding facet-values to retrieve relevant available biomedical documents.

The proposed framework has been evaluated taking into account data resources from BioMed, WikiGenes and PubMed repositories. The evaluation has been performed submitting natural language query and measuring the Precision and Recall on the retrieved results.

The paper is organized as follows. Section 2 describes some related works. Section 3 illustrates

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.biomedcentral.com/
[3] https://www.wikigenes.org/

the general framework specifying main phases of the overall workflow and introducing the role of *Unsupervised Ontology* and *Biomedical Ontologies*. Sections 4 and 5 detail background components aimed to build knowledge base taking into account unstructured textual documents and biomedical ontologies. Section 6 describes how the system processes natural language query through facet-based data model. Finally experimental results will be shown in Section 7.

2. Related Works

The exponential increase of biomedical literature makes it mandatory to support exploration of the huge available amount of unstructured information through friendly interfaces, like as: multifaceted search, and so on. In particular, multifaceted search is a technique of semantic search that simplifies the user experience presenting information categorized according to different perspectives. With faceted navigation, users can narrow down search results by applying multiple filters called facets [7] tailored on specific application domain (e.g., e-commerce, biomedicine). Browsing, faceted-search, and query-building capabilities for more powerful Linked Data exploration, similar to OLAP, have been combined in [8]. Furthermore, in [9] a mapping between OLAP and SPARQL queries has been proposed. Although semi-structured corpora with rich metadata enables advanced querying interface enabling multifaceted search, mapping unstructured documents to hierarchical facets is still an open challenge.

In [10, 11] authors describe algorithms for extraction of hierarchical facets from a corpus based on lexical subsumption, and assignment of the documents to those facets. In [12] synsets and hypernym relations are used to accomplish a similar task. In the area of bio informatics, recent works adopt ontologies to provide an advanced access to biomedical data sources. In [13] the authors define an ontological clustering approach to conceptualize abstracts of the journals available on PubMed repository. [14] defines a Big Linked Cancer Data to support the discovery of biomedical hypotheses querying a semantic data source.

The proposed work involves a natural language query processing module that addresses user's query exploiting facet-based data model resulting from ontology extraction on biomedical literature, and ontology matching procedures between resulting *Unsupervised Ontology*'s concepts and biomedical ontologies ones. An ad-hoc algorithm will be introduced to process natural language query.

From the knowledge extraction point of view, in the last decade several methods aimed to support ontology learning from domain data have been defined. Some of them use text-mining and machine learning techniques in order to conceptualize unstructured content, such as: OntoGen [15], OntoLT [16], OntoLearn [17] and OntoEdit [18]. Others methodologies exploit Conceptual Data Analysis [19] theory for knowledge structuring and ontology building [20]. Most of them are language dependent and doesn't exploit commonsense knowledge base (e.g., Wikipedia etc.) to enrich meaning of the analyzed content. In [21] the Fuzzy extension of Formal Concept Analysis (FFCA) theory has been exploited to build hierarchical classification of the collected resources. Unlike the existing approaches, this work defines methodology for knowledge extraction according to the theoretical model of FFCA exploiting wikification services to capture semantic features characterizing input textual documents.

3. General Framework

This paper introduces a general framework to support natural language query processing on biomedical content. In Fig. 1 the overall workflow and the categorical role of the ontologies (i.e., *Unsupervised Ontology*, and *Biomedical Ontologies*) that will be detailed in the following subsections are shown.

3.1. Unsupervised Ontology

Unsupervised Ontology is the output of knowledge extraction activity performed on unstructured biomedical documents. Its name underlies that it is essentially the result of unsupervised categorization process.

The proposed general framework has been instantiated implementing FFCA algorithm on unstructured biomedical resources. However, the framework could be instantiated implementing other suitable techniques for ontology extraction (e.g., LDA [22], hierarchical clustering [13]). According to our previous works about ontology extraction, the knowledge structure resulting by applying FFCA, including concepts and their relationships, will be formalized in a machine understandable manner exploiting technologies of Semantic Web (i.e., OWL, RDFS, and so on), as detailed in [23].

3.2. Biomedical Ontologies

Biomedical Ontologies will play a crucial role to provide a domain specific interpretation of unsupervised categorization defined in *Unsupervised Ontology* resulting from Knowledge Extraction activity. Both *Biomedical Ontologies* and *Unsupervised Ontology* will be the inputs to the Biomedical Ontologies Matching component that implements a matching strategy between their concepts in order to carry out facet-based data model of input biomedical documents. This component will be detailed in Section 5.

Specifically, the framework has been implemented taking into account the following *Biomedical Ontologies* available in BioPortal[4]:

[4] http://bioportal.bioontology.org/

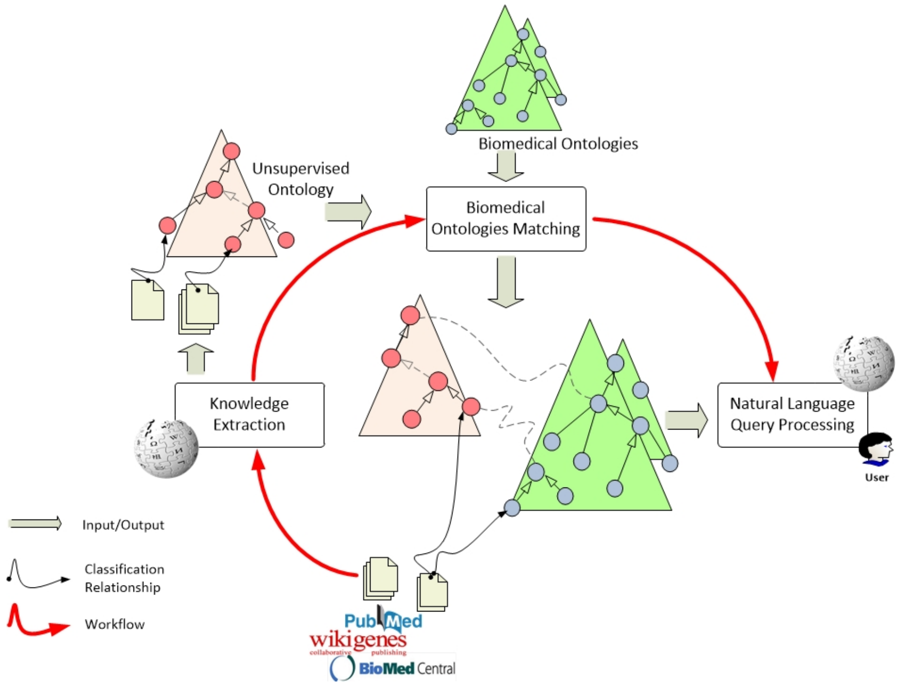


Figure 1: The workflow of the proposed general framework.

- *Ontology of Genes and Genomes - OGG*[5]: The OGG is a formal ontology of genes and genomes of biological organisms. OGG uses the Basic Formal Ontology (BFO) as its upper level ontology. This OGG document contains the genes and genomes of a list of selected organisms, including human, two viruses (HIV and influenza virus), and bacteria (B. melitensis strain 16M, E. coli strain K-12 substrain MG1655, M. tuberculosis strain H37Rv, and P. aeruginosa strain PAO1).
- *PRotein Ontology - PRO*[6]: PRO provides an ontological representation of protein-related entities by explicitly defining them and showing the relationships between them. Each PRO term represents a distinct class of entities (including specific modified forms, orthologous isoforms, and protein complexes) ranging from the taxon-neutral to the taxon-specific;
- *Gene Ontology - GO*[7]: The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases. In particular it provides an ontology of defined terms representing gene product properties in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

3.3. Workflow

The overall workflow depicted in Fig.1 is composed of the following main phases:

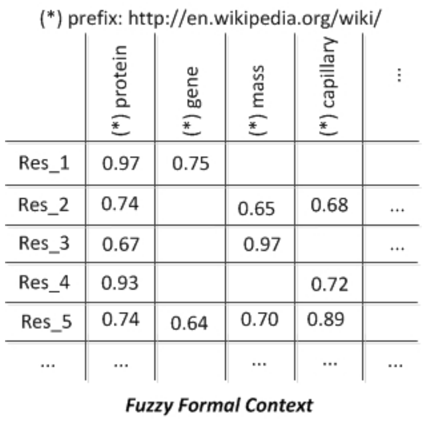
- *Knowledge Extraction.* It analyzes unstructured and heterogeneous biomedical resources to extract a common knowledge layer that is included in the resulting *Unsupervised Ontology*;
- *Biomedical Ontologies Matching.* This module performs concept matching between concepts belonging to *Biomedical Ontologies* and the ones from the *Unsupervised Ontology* ;
- *Natural Language Query Processing.* When natural language query happens, corresponding *Biomedical Ontologies*' concepts will be fired performing an ad-hoc matching algorithm. Then, biomedical resources categorized in the fired concepts will be retrieved and ranked as the results of the incoming user's query.

The following sections detail modules involved in the workflow described above.

4. Knowledge Extraction

As introduced in the previous sections, Knowledge Extraction is aimed to extract an unsupervised categorization taking into account unstructured input biomedical resources. In particular, the proposed general framework requires that the resulting categorization is represented as ontology artifact formalized by means of standard of Semantic Web (e.g., OWL, RDFS). The framework has been instantiated implementing algorithm of Fuzzy Formal Concept Analysis (FFCA) [21] to achieve the aims of the Knowledge Extraction component. FFCA is a fuzzy extension of FCA baseline introduced in [19] to face with uncertain and vague information occur-

[5] http://bioportal.bioontology.org/ontologies/OGG
[6] http://bioportal.bioontology.org/ontologies/PR
[7] http://bioportal.bioontology.org/ontologies/GO



(*) prefix: http://en.wikipedia.org/wiki/

	(*) protein	(*) gene	(*) mass	(*) capillary	...
Res_1	0.97	0.75			
Res_2	0.74		0.65	0.68	...
Res_3	0.67		0.97		...
Res_4	0.93			0.72	
Res_5	0.74	0.64	0.70	0.89	
...	...		...	...	...

Fuzzy Formal Context

Figure 2: An example of a portion of Fuzzy Formal Context with confidence threshold $T = 0.6$

ring in the representation of the domain, as well as in the unstructured textual content.

Following subsections describe how FFCA theory has been used in the proposed implementation of the proposed framework.

4.1. Fuzzy Formal Context Definition

This section details the extraction of a mathematical model to represent the text embedded in the biomedical resources.

Definition 1: *A* **Fuzzy Formal Context** *is a triple $K = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I = ((G \times M), \mu)$ is a fuzzy set.*

Recall that, being I a fuzzy set, each pair $(g, m) \in I$ has a membership value $\mu(g, m)$ in [0,1]. In the following the fuzzy set function μ will be denoted by μ_I.

Unlike FCA that uses binary relation to represent formal context, Fuzzy Formal Context enables the representation of the fuzzy relation between objects and attributes in a given domain, that in our implementation are biomedical resources and corresponding Wikipedia entities respectively. So, fuzziness enables to model relation among object and attribute in a more smoothed way ensuring more precise representation and uncertainty management. Specifically, wikification services [5] is exploited to determine the meaning of the collected unstructured text and its main concepts. We exploit Wikipedia knowledge base and wikification service to extract a set of pairs $\langle topic, rd \rangle$, where *topic* is a Wikipedia article representing the meaning of the content along with corresponding relevance degree, called *rd*, ranging in the $[0, 1]$ interval.

Thus, generalizing, the content of i-th biomedical resource will be represented via sentence wikification as:

$$Res_i = \{\langle topic_1^i, rd_1^i \rangle, \ldots, \langle topic_n^i, rd_n^i \rangle\}$$

where n is the number of topics detected by wikification of Res_i.

This representation for each biomedical resources is cross table of Fuzzy Formal Context as shown in Fig. 2. Let us note that each cell of the table contains a membership value in [0, 1], that is relevance of Wikipedia entity (rd) with respect to specific biomedical resource. Specifically, Fuzzy Formal Context shown in Fig. 2 has a confidence threshold $T = 0.6$, that means all the relationship with membership values less than 0.6 are not considered.

4.2. Fuzzy Lattice Definition

Taking into account Fuzzy Formal Context, FFCA algorithm is able to identify Fuzzy Formal Concepts and subsumption relations among them. More formally, the definition of Fuzzy Formal Concept and order relation among them are given following.

Given a fuzzy formal context $K = (G, M, I)$ and a confidence threshold T, for $G' \subseteq G$ and $M' \subseteq M$, we define $G^* = \{m \in M \mid \forall g \in G', \ \mu_I(g, m) \geq T\}$ and $M^* = \{g \in G \mid \forall m \in M', \ \mu_I(g, m) \geq T\}$.

Definition 2: Fuzzy Formal Concept. *A fuzzy formal concept (or fuzzy concept) C of a fuzzy formal context K with a confidence threshold T, is $C = (I_{G'}, M')$, where, for $G' \subseteq G$, $I_{G'} = (G', \mu)$, $M' \subseteq M, G^* = M'$ and $M^* = G'$. Each object g has a membership $\mu_{I_{G'}}$ defined as*

$$\mu_{I_{G'}}(g) = min_{m \in M'} (\mu_I(g, m))$$

where μ_I is the fuzzy function of I.

Note that if $M' = \emptyset$ then $\mu_I(g) = 1$ for every g. G' and M' are the extent and intent of the formal concept $(I_{G'}, M')$ respectively.

For instance, let us consider $c2$ in Fig. 3, it is composed of biomedical resources (objects) $I_{G'} = \{Res\_2, \ Res\_3, \ Res\_5\}$ and Wikipedia entities (attributes) $M' = \{mass, \ protein\}$ with μ_{Res_2}= 0.68, μ_{Res_3}= 0.64 and μ_{Res_5}= 0.74. So, the implementation of FFCA extracts a set of Fuzzy Formal Concepts that induces a categorization of biomedical resources.

Definition 3: *Let $(I_{G'}, M')$ and $(I_{G''}, M'')$ be two fuzzy concepts of a Fuzzy Formal Context (G, M, I). $(I_{G'}, M')$ is the* ***subconcept*** *of $(I_{G''}, M'')$, denoted as $(I_{G'}, M') \leq (I_{G''}, M'')$, if and only if $I_{G'} \sqsubseteq I_{G''} (\Leftrightarrow M'' \subseteq M')$. Equivalently, $(I_{G''}, M'')$ is the* ***superconcept*** *of $(I_{G'}, M')$.*

Let us observe in Fig. 3, the concept $c5$ is *subconcept* of the concepts $c2$ and $c3$. Equivalently the concepts $c2$ and $c3$ are *superconcept* of the concept $c5$.

Furthermore, given a Fuzzy Formal Concepts of Fuzzy Formal Context, it is easy to see that the subconcept relation $\leq$ induces a *Fuzzy Lattice* of Fuzzy Formal Concepts. As a matter of fact the lowest concept contains all attributes (Wikipedia entities) and the uppermost concept contains all object (biomedical resources) of Fuzzy Formal Context.

Fig. 3 shows an example of Fuzzy Concept Lattice. In the figure, each node can be colored in different way, according to its characteristics: a half-blue colored node represents a concept with *own* attributes; a half-black colored node instead, outlines

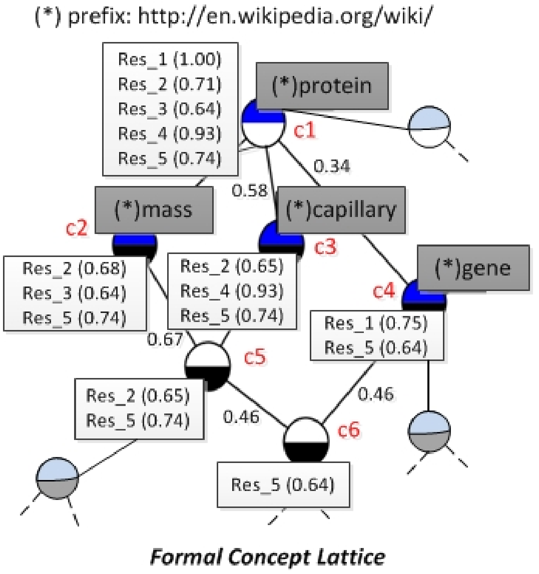


Figure 3: An example of a portion of Fuzzy Lattice with confidence threshold $T = 0.6$

the presence of *own* objects in the concept; finally, a half-white colored node can represent a concept with no *own* objects (if the white colored portion is the half below of the circle) or attributes (if the white half is up on the circle).

Furthermore, FFCA introduces also the definition of Fuzzy Formal Concept Similarity that provides degree of truth corresponding to each subsumption relation (i.e., an approximate subsumption).

Definition 4: ***Fuzzy Formal Concept Similarity*** *between concept $C' = (I_{G'}, M')$ and its subconcept $C'' = (I_{G''}, M'')$ is defined as*

$$E(C', C'') = \frac{|I_{G'} \sqcap I_{G''}|}{|I_{G'} \sqcup I_{G''}|}$$

where $\sqcap$ and $\sqcup$ refer to intersection and union operators[8] *on fuzzy sets, respectively.*

So, FFCA extracts a taxonomic arrangement of concepts and subsumption relationships (often known as a "hyponym-hypernym or is-a relationship") among them. In other words, the resulting fuzzy lattice is a hierarchical categorization of biomedical resources. The next step is aimed to translate the resulting Fuzzy Concept Lattice into a formal ontology.

4.3. Ontology Modeling

This step translates the resulting fuzzy lattice into a formal ontology according to OWL syntax. Let us underline an aspect of our framework: we use the fuzzy lattice since it enables us to deal with unstructured data in a more precise and granular manner than a crisp approach. At this point we had the choice to keep on using the fuzziness for the representation of the lattice but then, since the Biomedical Ontologies are encoded in OWL/RDFS, we chose to use OWL/RDFS. Let us stress that encode Fuzzy Formal Concept Lattice as Fuzzy OWL 2 ontology (see e.g. [24]), could be surely a subject of further investigation to compare this approach with ours.

Fuzzy Formal Concept according to the theory of FFCA is represented with both intentional and extensional information. Thus, Ontology Modeling achieves a translation of both intentional and extensional information into the corresponding classes and relations of the OWL ontology [23].

Formally, let $C = (I_{G'}, M')$ be a fuzzy concept of the Fuzzy Formal Context (G, M, I). The resulting ontology will include definition of a *owl:class* $Class_C$, where the objects of the extension $G' \subseteq G$, that in our case are biomedical resources, become its individuals (or instances), while the attributes of the intention $M' \subseteq M$, that are links to the Wikipedia entity, become its properties. The subsumption relation between couple of concepts $C' \leq C''$ produces a correspondent subclass relation, i.e., $Class_{C'}$ is subclass of $Class_{C''}$.

5. Biomedical Ontology Matching

This component implements concept matching between *Biomedical Ontologies* and *Unsupervised Ontology*. Specifically, it evaluates matching degree between each concept in the selected *Biomedical Ontologies* and the concepts in *Unsupervised Ontology*, and extracts one to many weighted relationships. The general framework has been instantiated implementing the following sub tasks:

- *Biomedical Ontology Concept Wikification.* This step performs for each biomedical ontologies' concept BOC_i wikification using its name and description. Analogously to biomedical content wikification described in Section 4.1, the wikification process returns a set of pairs $\langle topic^i_j, rd^i_j \rangle$ for $j = 1, \ldots, k$. The process filters out topics resulting by wikification service whose relevance degree rd is lower than a fixed threshold ρ (in our case $\rho = 0.6$). Then, the resulting representation of i-th biomedical ontologies' concept is:
  $$BOC_i = \left\{ topic^i_j \mid rd^i_j \geq \rho, j \leq k \right\}$$
- *Matching Evaluation.* This task performs the matching evaluation. Specifically, given biomedical ontology's concept:
  $$BOC_i = \{topic^i_1, topic^i_2, \ldots, topic^i_m\}$$
  and the unsupervised ontology one:
  $$UOC_j = \{topic^j_1, topic^j_2, \ldots, topic^j_s\}$$
  the matching degree is evaluated by performing revisited well-known measures of the Precision P and Recall R[25], as follows:
  $$P_{i,j} = \frac{|BOC_i \bigcap UOC_j|}{|UOC_j|} \tag{1}$$

[8] The fuzzy intersection and union are calculated using t-norm and t-conorm, respectively. The most commonly adopted t-norm is the minimum, while the most common t-conorm is the maximum. That is, given two fuzzy sets A and B with membership functions $\mu_A(x)$ *and* $\mu_B(x)$, $\mu_{A \sqcap B}(x) = min(\mu_A(x), \mu_B(x))$ *and* $\mu_{A \sqcup B}(x) = max(\mu_A(x), \mu_B(x))$.

$$R_{i,j} = \frac{|BOC_i \bigcap UOC_j|}{|BOC_i|} \quad (2)$$

and F-measure becomes the resulting matching degree between UOC_j and BOC_i:

$$F(BOC_i, UOC_j) = 2 \times \frac{P_{i,j} \times R_{i,j}}{P_{i,j} + R_{i,j}} \quad (3)$$

The evaluation of the intersection between BOC_i and UOC_j is computed as the maximum cardinality bipartite matching considering the graph $G = \langle V, E \rangle$:

$$V = \left\{ BOC_i \bigcup UOC_j \right\} \quad (4)$$

$$E = \{(x, y) \mid x \in BOC_i \ and\ y \in UOC_j \\ if\ WLM(x, y) \geq \tau\} \quad (5)$$

where WLM is the Wikipedia Linked Measure [26], and τ is fixed to 0.7 established during the evaluation to include in the matching the concepts that have significantly close meaning.

At the end of the execution, for each BOC_i the corresponding set of UOC_j is composed of concepts whose resulting matching degree (i.e., Eq. 3) is greater than a fixed threshold χ (in our case $\chi = 0.8$). The resulting matching induces a categorization of biomedical documents with respect to biomedical ontologies' concepts, that is facet-based data model exploited in the *Natural Language Query Processing* phase.

6. Natural Language Query Processing

The results of the matching enable faceted search providing information about the quantities of documents belonging to each biomedical ontologies' concept (e.g., THRB, THRA, Genome of Bacteria, etc.), that is a value of specific facet itself. Indeed, facet-based data model allows to carry out more than just knowing the quantities of documents belonging to each facet. For instance, let us suppose that the user asks for availability of research studies about effects of zinc with cancer, and let us assume that the selected Biomedical Ontologies define classification values corresponding to this query (e.g., materials, cancers, etc.), if the intersection of corresponding facet values is not empty it enables to conclude that it is supported by a specific number of biomedical resources. This could be extended in new researches to compute more complex analytics services on biomedical literature.

This work emphasizes that natural language query could be addressed matching the user's query with respect to extracted facets and correlating their values. Specifically, this component computes natural language input query performing the following steps:

- *Query Wikification.* Analogously to biomedical content and to biomedical ontologies' concepts wikification (described in sections 4.1 and 5, respectively), this step performs wikification of input query in order to capture meaning of user's request. In particular, given a query Q, it will be represented with topics retrieved by wikification process whose relevance degree is greater than a fixed threshold ρ (in our case $\rho = 0.6$), more formally:
$$Q = \left\{ topic_i^Q \mid rd_i^Q \geq \rho \right\}$$
- *Query and Facet Matching.* Given Q and the overall set of biomedical ontologies' concepts $BOC = \{BOC_1, \ldots, BOC_N\}$ this step computes matching degrees $F(Q, BOC_i)$ as defined in Eq. 3. The resulting set $BOC^Q \subseteq BOC$, whose $F(Q, BOC_i) \geq \chi$ is the set of facet constraints used to filter search space of the available biomedical literature. Next step ranks the results considering more relevant biomedical documents that satisfy most number of constraints in BOC^Q.
- *Results Rank.* This module ranks the results of query processing performing the scoring algorithm that carries out a degree of relevance of each retrieved biomedical resource. Let us consider the biomedical ontologies' concepts $BOC_i^Q \in BOC^Q$ resulting from the previous phase and the set of corresponding unsupervised ontology's concepts UOC_j for $j \in \{1, 2, \ldots, s\}$ whose matching degree $F(UOC_j, BOC_i^Q)$ (see Eq. 3) is greater than the threshold χ. Now, for each biomedical resource categorized as individual $I \in UOC_j$ (for $j \in \{1, 2, \ldots, s\}$) we calculate its own belonging $score_i^Q$ to the concept BOC_i^Q as follows:
$$score_i^Q(I) = \sum_{j=1}^{s} \left(\mu_j(I) \times F(UOC_j, BOC_i^Q) \right) \quad (6)$$
where $\mu_j(I)$ represents the membership degree of the resource belonging to the concept UOC_j (see Definition 2) and $F(UOC_j, BOC_i^Q)$ is the matching result between BOC_i^Q and UOC_j (see Eq. 3). Finally, the score in Eq. 6 is used to computes the final $score^*$ corresponding to the query Q as follows:
$score^*(I) =$
$$\sum_{BOC_i^Q \in BOC^Q} (score_i^Q(I) \times F(Q, BOC_i^Q)) \quad (7)$$
This value represents how much the retrieved biomedical resource (i.e., individual) is relevant with respect to the overall constraints in the input query. Finally, the result set is ranked according to the evaluated $score^*(I)$.

7. Experimental Results

The experimentation has been performed to evaluate the effectiveness of the overall process in terms of information retrieval performances. Given a natural language query, the proposed framework evaluates its matching degree with respect to *Biomedical Ontology*'s concepts and finally ranking the results according to Eq.7, as detailed in Section 6. The framework has been instantiated on a collection of biomedical resources composed of: 200 articles of PubMed; 200 of Wikigenes; and 200 of BioMed Central, and 5 natural language queries have been submitted to evaluate the performances in terms of micro-average of each precision-recall curves [25].

In particular, the submitted queries are listed following:

- Q_1 : *Zinc oxide nanoparticles*;
- Q_2 : *Nanoparticles used in cancer therapy*;
- Q_3 : *Drug Transporters in the Central Nervous System*;
- Q_4 : *Effect of gastrointestinal proteins*;
- Q_5 : *Epigenetic effects of cadmium in cancer.*

Table 1 summarizes the relevant sets, defined by expert users on the collected resources, corresponding to the queries defined above.

Table 1: Queries and corresponding relevant set distribution over the collected dataset.

Query	**Relevant Set**
Q_1	85
Q_2	50
Q_3	183
Q_4	60
Q_5	41

Formally, let be $Q = \{Q_1, Q_2, \dots, Q_n\}$ a set of queries, $RS = \{RS_{Q_1}, RS_{Q_2}, \dots, RS_{Q_n}\}$ the corresponding relevant sets, and $RR = \{RR_{Q_1}, RR_{Q_2}, \dots, RR_{Q_n}\}$ the results retrieved by the framework. For each query Q_i, we consider $\lambda = 20$ steps up to its maximum recall value and measure the number of relevant documents retrieved at each step λ. The micro-averaging of recall and precision (at the generic step λ), is formally defined as follows:

$$Rec_\lambda = \sum_{Q_i} \frac{|RS_{Q_i} \bigcap RR_{\lambda,Q_i}|}{|RS|} \tag{8}$$

$$Pre_\lambda = \sum_{Q_i} \frac{|RS_{Q_i} \bigcap RR_{\lambda,Q_i}|}{|RR_\lambda|} \tag{9}$$

where RR_λ is the set of retrieved resources at step λ and RR_{λ,Q_i} is the set of all relevant resources, retrieved at step λ, for the query Q_i.

Fig. 4 shows the tendency of the micro-average of recall/precision curve evaluated on the selected input dataset.

Specifically, it shows that increasing the recall until 0.8 the performance in terms of precision are quite constant, after that the precision reaches the minimum value of ~ 0.6.

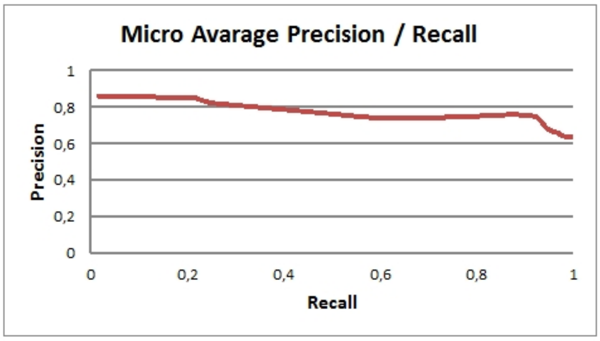


Figure 4: Micro-averaging precision/recall

8. Conclusion

This work introduces a general framework to support natural language user's query processing to retrieve relevant biomedical literature. It relies on components for knowledge extraction and ontology matching to categorize input biomedical content with respect to existing biomedical ontologies' concepts. The former extracts an *Unsupervised Ontology* that specifies the results of unsupervised categorization process performed on unstructured data implementing Fuzzy Formal Concept Analysis algorithm. The latter defines a matching algorithm between concepts of *Unsupervised Ontology* and *Biomedical Ontologies* ones. The results of the matching algorithm induce a facet-based data modelused to retrieve biomedical resources according to constraints in the input natural language query.

Further researches will follow the direction of using the facet-based data model to address analytics information request about biomedical researches. Another interesting future direction is to apply the verification techniques described in [27, 28] to check the satisfiability of input constraints with respect to extracted biomedical knowledge base.

References

[1] Maloney C, Sequeira E, Kelly C, and et al. Pubmed central. *The NCBI Handbook [Internet]. 2nd edition.*, 2013. In: Bethesda (MD): National Center for Biotechnology Information (US); 2013.

[2] Robert Hoffmann. A wiki for the life sciences where authorship matters. *Nature genetics*, 40(9):1047–1051, 2008.

[3] Tim Beißbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[4] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Domenico Parente. Biomedical data integration and ontology-driven multi-facets visualization. In *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015.

[5] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.

[6] Ori Ben-Yitzhak, Nadav Golbandi, Nadav Har'El, Ronny Lempel, Andreas Neumann, Shila Ofek-Koifman, Dafna Sheinwald, Eugene Shekita, Benjamin Sznajder, and Sivan Yogev. Beyond basic faceted search. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 33–44. ACM, 2008.

[7] Daniel Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.

[8] Georgi Kobilarov and Ian Dickinson. Humboldt: Exploring linked data. *context*, 6:7, 2008.

[9] Benedikt Kämpgen, Sean O'Riain, and Andreas Harth. Interacting with statistical linked data via olap operations. In *C. Unger, P. Cimiano, editor, Proceedings of Interacting with Linked Data (ILD 2012), workshop co-located with the 9th Extended Semantic Web Conference*, pages 36–49. Citeseer, 2012.

[10] Wisam Dakka, Panagiotis G Ipeirotis, and Kenneth R Wood. Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 768–775. ACM, 2005.

[11] Ronald Fagin, R Guha, Ravi Kumar, Jasmine Novak, D Sivakumar, and Andrew Tomkins. Multi-structural databases. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 184–195. ACM, 2005.

[12] Emilia Stoica, Marti A Hearst, and Megan Richardson. Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, pages 244–251, 2007.

[13] Hai-Tao Zheng, Charles Borchert, and Hong-Gee Kim. Goclonto: An ontological clustering approach for conceptualizing pubmed abstracts. *Journal of biomedical informatics*, 43(1):31–40, 2010.

[14] Muhammad Saleem, Maulik R Kamdar, Aftab Iqbal, Shanmukha Sampath, Helena F Deus, and Axel-Cyrille Ngonga Ngomo. Big linked cancer data: Integrating linked tcga and pubmed. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:34–41, 2014.

[15] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. *OntoGen: semi-automatic ontology editor*. Springer, 2007.

[16] Paul Buitelaar, Daniel Olejnik, and Michael Sintek. A protégé plug-in for ontology extraction from text based on linguistic analysis. In *The Semantic Web: Research and Applications*, pages 31–44. Springer, 2004.

[17] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, 18(1):22–31, 2003.

[18] York Sure, Michael Erdmann, Jürgen Angele, Steffen Staab, Rudi Studer, and Dirk Wenke. *OntoEdit: Collaborative ontology development for the semantic web*. Springer, 2002.

[19] Bernhard Ganter, Rudolf Wille, and Rudolf Wille. *Formal concept analysis*, volume 284. Springer Berlin, 1999.

[20] Gu Tao. Using formal concept analysis (fca) for ontology building and structuring. Master's thesis, Nanyang Technological University, 2003.

[21] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Sabrina Senatore. Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. *Information Processing & Management*, 48(3):399 – 418, 2012. Soft Approaches to IA on the Web.

[22] Di Wang, M. Thint, and A. Al-Rubaie. Semi-supervised latent dirichlet allocation and its application for document classification. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 3, pages 306–310, Dec 2012.

[23] Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Sabrina Senatore. Towards an automatic fuzzy ontology generation. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, pages 1044–1049. IEEE, 2009.

[24] Fernando Bobillo and Umberto Straccia. Fuzzy ontology representation using owl 2. *International Journal of Approximate Reasoning*, 52(7):1073–1094, 2011.

[25] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[26] I Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[27] Salvatore La Torre, Margherita Napoli, Mimmo Parente, and Gennaro Parlato. Verification of scope-dependent hierarchical state machines. *Inf. Comput.*, 206(9-10):1161–1177, 2008.

[28] Alessandro Ferrante, Aniello Murano, and Mimmo Parente. Enriched μ-calculi module checking. *Logical Methods in Computer Science*, 4(3), 2008.