A Framework for Semantic Text Clustering

Soukaina Fatimi[1], Chama EL Saili[2], Larbi Alaoui[3]
TIC Lab, International University of Rabat
Sala Al Jadida
Morocco

***Abstract*—Existing approaches for text clustering are either agglomerative, divisive or based on frequent itemsets. However, most of the suggested solutions do not take the semantic associations between words into account and documents are only regarded as bags of unrelated words. Indeed, traditional text clustering methods usually focus on the frequency of terms in documents to create connected homogenous clusters without considering associated semantic which will of course lead to inaccurate clustering results. Accordingly, this research aims to understand the meanings of text phrases in the process of clustering to make maximum usage and use of documents. The semantic web framework is filled with useful techniques enabling database use to be substantial. The goal is to exploit these techniques to the full usage of the Resource Description Framework (RDF) to represent textual data as triplets. To come up a more effective clustering method, we provide a semantic representation of the data in texts on which the clustering process would be based. On the other hand, this study opts to implement other techniques within the clustering process such as ontology representation to manipulate and extract meaningful information using RDF, RDF Schemas (RDFS), and Web Ontology Language (OWL). Since Text clustering is an indispensable task for better exploitation of documents, the use of documents may be more intelligently conducted while considering semantics in the process of text clustering to efficiently identify the more related groups in a document collection. To this end, the proposed framework combines multiple techniques to come up with an efficient approach combining machine learning tools with semantic web principles. The framework allows documents RDF representation, clustering, topic modeling, clusters summarizing, information retrieval based on RDF querying and Reasoning tools. It also highlights the advantages of using semantic web techniques in clustering, subject modeling and knowledge extraction based on processes of questioning, reasoning and inferencing.**

Keywords*—Text clustering; similarity measure; ontology; semantic web; RDF; RDFS; OWL; reasoning; inferencing rules; SPARQL; topic modeling; summarization

I. Introduction

It has been a while since the web has changed from the web of documents to the web of data. Before knowing this upgrade, the information on the web was designed to be human-understandable only. Therefore a device or a robot could not access information in the same manner as humans, and artificial intelligence cannot evolve under these circumstances. This particular issue is considered as the motivation behind the evolution of information representation and the launch of the Semantic web as a web of connected data. The concept base is to transform the web of unstructured data to a network of interconnected chunks of information. Hence, both humans and machines can navigate between bits of data to explore it and retrieve more information from it. This collection of interrelated data is referred to as Linked Data [2]. The Linked Data is guided with a set of principles to allow easy sharing of structured data planet-wide. To represent and enable the use of this linked data and to allow the navigation between pieces of information, special representation should be used. The Resource Description Framework is at the core of the linked data paradigm. The RDF model, in which the data is represented as triples of interconnected subject and object with the intermediary of a predicate, is the mainstay of the interconnection of the information in the semantic web. This model enables the navigation between pieces of information following RDF links. It is indeed true that unstructured representation of information is still used, and that studies have provided significantly valuable tools for the manipulation tasks of textual documents, such as text clustering, information retrieval topic identification, etc. Still, the advantages of the linked data are captivating. Therefore, providing semantic data manipulation based on a semantic web model needs to be explored and strongly highlighted.

Text clustering has been widely explored for textual document manipulation. Yet proposed methodologies have lacked in the use of semantic relationships between words. Generally, documents are considered a bag of unrelated words and semantics are not explored in the process of text clustering.

Nevertheless, this work aims to use the semantic web approach for a semantic text clustering using graph-based representation model RDF with the respect of the linked data principles. We propose a system that is an integrated set of techniques in which the textual documents are transformed into an RDF graphs representation and divided into homogenous clusters based on a semantic clustering approach. These documents are further explored using semantic web techniques such as querying and information retrieval using inferencing and reasoning tools.

Text clustering is an indispensable task for better exploitation of documents to retrieve information, identify topics in more efficient ways. The provided system is a holistic approach allowing better understanding and use of textual documents with the mean of a semantic framework based on the RDF model. The purpose of working with RDF is due to its countless advantages, the self-explanatory or semantic characteristics of RDF data and is very beneficial for better semantic similarity computing and more efficient clustering.

We present an overall framework, and show how to apply machine learning techniques to mine textual documents using

This work is within the framework of the research project "Big Data Analytics - Methods and Applications (BDA-MA)". Authors S. Fatimi and C. El Saili are financially supported by a PhD grant of International University of Rabat.

Linked Data principles and highlight the importance of text clustering and the use of semantics in text clustering based on the RDF model.

The rest of the paper is organized as fellows. The next section introduces a review and presents the general context of our work, such as text clustering, semantic web, semantic similarity measurement, and topic identification. In the third section, the overall framework is presented and the steps of the system are discussed in the subsections emphasizing the clustering process. Finally, a conclusion and perspective work are given in the last section.

II. Related Work

The semantic oriented clustering approach that we are presenting in this paper is a combination of interesting concepts and techniques, from text clustering to similarity measurement and also to semantic web concepts and frameworks. In this section, we will give an overview of all the above notions and mention some of the related studies and works on these fields.

A. Semantic Web

The Semantic Web concept was introduced by Tim Berners-Lee as a novel form of web content that is understandable by humans [2]. The main goal of this concept is to interconnect and structure data in the World Wide Web to create an environment where programs can ramble between different pages to understand, process, and question existing information. Semantic web has caught the attention of many researchers ([3], [4], [5], [6]). Tim Berners-Lee introduced several principles for semantic web concept, he defined Resource Description Framework (RDF) a graph model to present data on the web, RDF interconnects data as triplets of a subject, predicate and object where subject and object are nodes and property is an arc. These RDF elements may be a textual value, or a blank and may be represented as Universal Resources Identifiers (URI) to distinct notions and relations that can connect them [2]. The use of RDF allows machines to understand the meaning of these notions and their linkage. This type of data is stored in special repositories called Triple Stores [7]. One other fundamental component of Semantic Web is ontology creation. Researchers have intensely studied this concept and its application in many domains like biomedical network security [8], smart cities [9] and robotic application [10]. Ontology is defined as a collection of information that describes a concept and provides its vocabulary. Ontologies are understandable by both humans and machines and allow semantics and syntactic exchange. Definite web ontology languages have been unified due to research in the Semantic Web that allows to efficiently describe a domain with the use of the semantic web languages RDF Schemas (RDFS), and Web Ontology Language (OWL). Ontologies are engineered based on the domain concepts referred to by "classes" and the relationships between these concepts which can be hierarchical as subclass relationships or predefined as properties. The models can also include constraints on the expressed information.

B. Text Clustering

Document clustering is an unsupervised learning process that separates documents into significant groups. It's one of the main techniques of text mining [11]. Document clustering is to designate a corpus of content documents into distinctive bunches so that documents within the same gather depict the same subject. Clustering of documents contains three categories: partitioning methods, agglomerative and divisive clustering. Researchers proposed several document clustering algorithms like K-Means, Hierarchical Agglomerative clustering and Frequent Itemsets based clustering and more algorithms having been utilized in this learning process.

Text clustering is a dynamic field that caught the researcher's consideration. The enormous textual data shared on the net is considered as a bag of information and can be labeled as the crude fabric of information. Diverse methods are actualized to move forward the extraction of profitable data from this information. Text clustering consists of indexing, crawling and filtering the information. We distinguish four steps within the process of text clustering: the collection of data, the preprocessing at that point, the clustering and the post processing of the clusters. Initially, documents are collected and put away, and it is basic to preprocess all these documents to dispose of the commotion [11] before clustering these documents.

The Internet is nowadays advancing from a Web of documents to a Web of Information, employing a graph-based representation and a set of basic standards, known Linked Data Principles [12]. In any case, statistics on the LOD Cloud (April 2017) reports that over 149 billion realities are as of now put away as RDF triples in 9960 information sources. Hence, the number of RDF datasets distributed on the Net is continuously and rapidly expanding. A client willing to use these datasets will begin to have to investigate them in order to decide which data is pertinent to his particular needs. Therefore, to encourage this interaction, a topical see of an RDF dataset can be given by applying the clustering instrument which can be characterized as making a set of homogeneous clusters with expansive intra-cluster similarity and expansive inter-cluster disparity [12].

C. Text Documents to RDF Triples

In order to handle documents of unstructured text data with semantic web techniques, it is obvious that the conversion of text documents to RDF triples is the major step to be done. The objective of this transformation is to change plain text into data units understandable by machines. Authors in [33] proposed another approach that converts a given content into RDF triples based on the semantic and syntactic structure of sentences. Based on this approach they built a system called T2R that creates important triples with all fundamental linguistic relations and semantic parts of the text. This approach can be used for any plain text. T2R inputs a text document into a syntactic parser using the Stanford tool and semantic parser utilizing the Senna tool.

LODifier [34] is one of the inspiring approaches in the Knowledge graph construction process to provide a tool for the conversion of texts to RDF. It is based on both deep semantic analysis and named entity recognition systems. Based on

LODifier, authors in [13] proposed a conversion of tweets into RDF triple where tweets are assembled topic-wise, by utilizing topic identification methods and shaping homogeneous clusters using the K-Means algorithm. Only the tweets containing named substances in DBpedia datasets are used. Each topic corpus is summarized then transformed into an RDF chart utilizing the LODifier tool. In the work [14], another model that collects resumes data from the internet and classifies them based on the cosine similarity measure was proposed. In this model, the data is represented using the semantic web like RDF based on Protégé tool and SPARQL. Another methodology is proposed by authors of [15] as a combination of the techniques of similarity computing, visualization procedures and RDF query language (SPARQL) to manipulate academic contents. They utilized also the ontological model to form syllabus information justifiable by both people and computers. Another framework for automatic knowledge graph (KG) extraction from unstructured text was proposed by authors of [16] and extended in their second work [17]. They underlined two RDF extraction steps. Firstly, candidate generation by focusing on the importance of mapping predicates to a referential KG for more searchability increase, and secondly, candidate selection process using pre-defined ontologies. Following the same path, [18] proposed an open-source platform for KG construction that includes graph management and downstream application support and is based on tools such as Stanford CoreNLP, Neo4j and Apache Solr.

D. Similarity Measures

The literature has many methods for computing the semantic similarity between terms. Semantic similarity measures can be classified into four categories: Edge Counting Measures, Information Content Measures, Feature Based Measure and Hybrid Methods. In the followings we present the overall idea behind similarity measurement and highlight the antecedent works about semantic similarity measure and its uses.

1) Similarity measure concept: A similarity measure is a function that assigns a non-negative real number to each pair of patterns, defining a notion of resemblance and having the target range between [0,1]. Similarity measures form the basis for many patterns matching algorithms. Besides that, similarity measures compare vectors which should be symmetrical and assign a value to them becoming larger when they are similar and getting the largest value when they are identical. Usually measured as the cosine of the angle between vectors, that is, the so-called cosine similitude, the Cosine similarity is one of the foremost well-known closeness measures in various data recovery applications and clustering as well. A Jaccard degree was introduced in [19] and is in some cases alluded to as the Tanimoto coefficient measures closeness between limited test sets and is characterized as the estimate of the intersection isolated by the estimate of the union of the test sets. For this measure, the Jaccard coefficient compares individuals for two sets to see which individuals are shared and which are unmistakable. The foremost AHC strategies do a calculation on this similarity matrix and develop a progressive structure to indicated connections or proximities among the data.

2) Semantic similarity measure: Research on the semantic similarity measures based on RDF data has mainly been done for the similarity measurement of RDF graphs for the query matching. The goal is to extract the best matching result. A similarity measure (gSemSim) was proposed to progress ordinary similarity measures to decrease their impediments. The notable feature of this semantic similarity measure is its capacity to display more reasonable similarity between concepts in the viewpoint of space information. Reference [20] demonstrates pairwise word interactions and displays a new similarity center instrument to recognize vital correspondences for superior similitude estimation. These thoughts are executed in a neural network design that illustrates state-of-the-art precision on three SemEval assignments and two reply determination tasks.

E. Semantic Text Clustering

Document clustering is one of the main techniques of text mining that is considered as an unsupervised learning process that separates documents into significant groups. It is to designate a corpus of content documents into distinctive bunches so that documents within the same gather depict the same subject. Researchers proposed several document clustering algorithms like Hierarchical Agglomerative clustering and Frequent Itemsets based clustering and others that are used in this learning process [21]. Traditional text clustering methods usually focus on the frequency of terms in documents to create connected homogenous clusters, thus, documents can be semantically related so these approaches will conduct inaccurate clustering results. The complexity of natural language results in the complexity of having accurate and efficient text clustering. Researchers have made use of semantic web technologies such as ontologies to take advantage of the semantic relationship between words in clustering. Walaa K. Gad and Mohamed S. Kamel [34] proposed a semantic similarity-based model (SSBM) to handle the semantic in documents. They incorporated the use of ontology in their case WordNet to obtain the semantic similarities between words, such as synonyms and hypernyms, and the documents vector is constructed based on a refined terms weight that includes term frequency (TF) and Inverse document frequency (IDF) and the semantic weight based on terms semantic relationships. Following the same path, authors in [35] applied text clustering Based on the semantic body for Chinese spam mail Filtering, the proposed methodology is based on lexical chains and HowNet semantic similarity to handle the words' synonyms, this technique helps to overcome defiance related to synonyms and near-synonyms by merging them. Thus, the results of the experiment were good, but the use of HowNet resulted in some limitations since it doesn't cover all possible similarities between words. [22] also presents an approach using lexical chains combined with WordNet; A WordNet-based semantic similarity measure for solving the problem of Polysemy and synonymy, and lexical chains to extricate a little subset of the semantic features which not as it denoted the topic of documents but moreover are advantageous to clustering. In [15] a combination of the techniques, methods, and algorithms such as cosine similarity, visualization procedures have been used for semantic text clustering, moreover, the ontological model to form syllabus information intelligible for both people and computers. In [12], using Candidate Description (CD) as a set of predicates, a form of RDF clustering algorithm has been developed, it used similarity matrix which contains the pairwise similarities between CDs clusters and utilized Cosine Similarity, Jaccard similarity and Sorensen Dice To measure the similarity between CDs.

F. *Topic Modeling*

Topic models are unsupervised machine learning techniques used to thematically describe a set of documents, it intends to detect the group of words that characterize and describe the collection of documents. Topic modeling is among important techniques used for the measurement of document similarity for classification [23], the clustering and cluster labeling, summarizing documents, and more [24]. Topic modeling was firstly introduced for textual documents. Yet, its use for unstructured types of data such as images has been explored. In multiple researches, topic modeling has been combined with semantic web [25] to improve the topic modeling results. However, few are the techniques that have been provided in order to apply topic modeling over unstructured topics. [26] Proposed a framework for applying topic modeling to RDF graph data based on LDA, they highlighted some of the major challenges in using topic modeling over RDF data. These challenges are related to the sparseness and the unnatural language of the RDF graphs and gave some methods to tackle it. In [27] a method to profile RDF datasets on Knowledge-based modeling techniques is given with the goal to describe the content of the datasets. The extracted representative topics for the RDF dataset are annotated with Wikipedia categories. Knowledge-based topic modeling has been earlier used for entity summarization in [36] using a probabilistic model called ES-LDA that uses a modified version of the LDA algorithm was used to handle the challenges of working with the RDF model.. The model uses prior knowledge for statistical learning techniques to create representative summaries for the large semantic web documents in order to facilitate the use of semantic web entities.

The whole approaches presented have not provided significantly valuable tools for the manipulation tasks of textual documents, such as text clustering, information retrieval topic identification, etc. Still, the disadvantages of the linked data are captivating. Therefore, providing semantic data manipulation based on a semantic web model needs to be explored and strongly highlighted. This work aims to take advantage of most of the semantic web techniques' benefits and present an overall framework for semantic text clustering based on RDF data more efficient than these approaches.

III. METHODOLOGY

Text is considered the essential and mostly utilized representation of data, numerous investigations and strategies have been examined to move forward the information disclosure based on textual information. The aim behind transforming textual data into an RDF model is to make it understandable by both humans and machines. The transformation should take into consideration syntactic and semantic relations between terms. The goal is to analyze, summarize, and extract information from this data. All these errands require a profound understanding of the basic structures and semantics of the documents. Exploring large amounts of data in order to retrieve relevant information can be a frustrating task. Based on the RDF framework and using associated techniques such as SPARQL for querying the data, RDF Schema (RDFS) and Web Ontology Language (OWL) to apply reasoning and inference support on the data.

Furthermore, clustering methods result in improving these interactions in order to provide better results and is considered as the pillar for other knowledge discovering tasks such as summarization and visualization. Classic clustering methods ignore the semantics between the words, generally, documents are considered as a bag of words, and do not make use of the relations that may exist between the words. Words can have multiple meanings depending on the context there are used in. Therefore, separating words from their context can lead to a misunderstanding of the words. The use of RDF based models for textual documents clustering is a step toward preserving semantics in documents and providing efficient clustering with better accuracy.

In this sense, and as previously mentioned, this work aims to take advantage of most of the semantic web techniques' benefits. Therefore, it proposes a graph data model for clustering and mining text documents. The model is based on the use of semantic web technology RDF to represent the information, SPARQL, RDFS and OWL to use reasoning engine in order to retrieve information from it.

The proposed methodology starts with the extraction of RDF representation from textual documents based on the semantic and syntactic nature of sentences, and then several mining techniques are introduced. This methodology focuses on clustering based on a proposed similarity measure of the RDF graphs, in order to group related documents in homogenous clusters, and topic modeling process to extract the underlying topics presented in the documents. Finally, an inferencing model is introduced based on RDFS and OWL language in order to extract more facts from the data.

Fig. 1 summarizes the proposed framework whose main components are explained in the subsequent.

A. *Extract RDF from Data*

The first step toward our semantic-based clustering system and documents querying is to transform the textual unstructured documents into RDF triples representation. As previously discussed in the above section, there have been many studies tackling the transaction from text to RDF triples. The main goal of this transformation is to switch from textual sentences that are understandable by humans only to interconnected information and intelligible by both humans and machines.

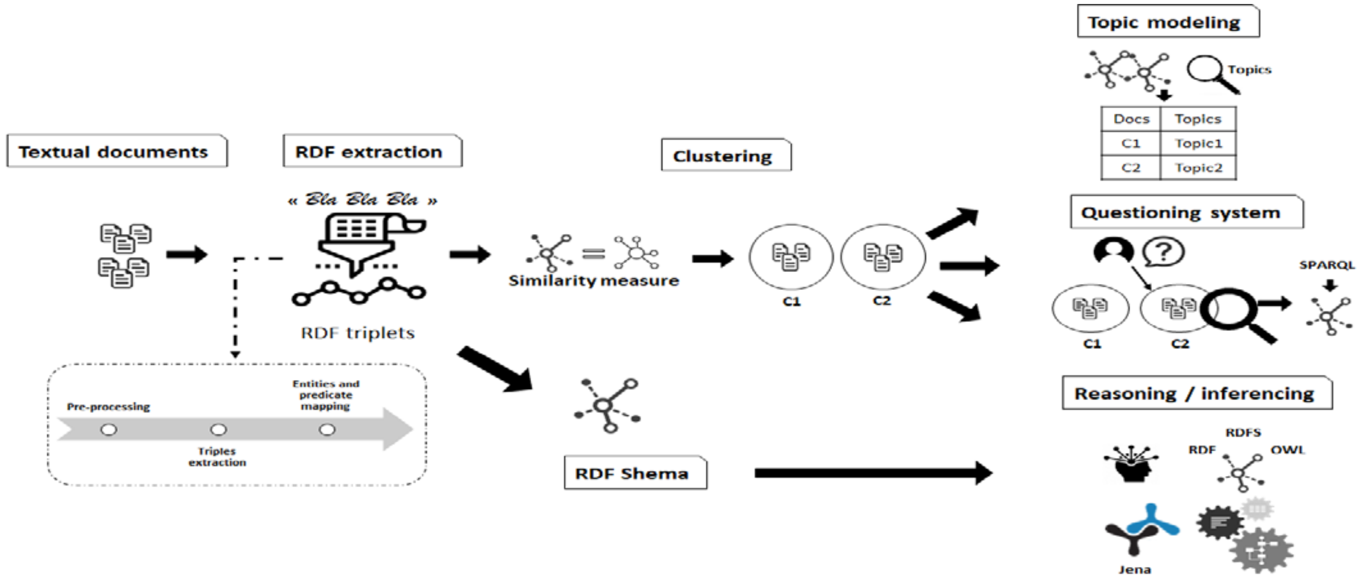

Fig. 1. Semantic-based Text Clustering Framework.

An RDF graph G is defined as a collection of statements. A statement is a triple (t) representing the relationship - named predicate (p) which is generally presented as a URI - between a subject (s) and an object (o) in the form of t=(s,p,o). A subject can be either a resource (URI or IRI) or a basic string (Literal), while an object can be denoted as a resource, a literal or an abstract identifier (Blank).

Overall, the task of text transformation to RDF is an iterative process that consists of converting each sentence into RDF triples under its semantic and syntactic form. This process can be represented by the schema given in Fig. 2.

The preprocessing phase is vital before addressing the triples extraction. Usually, sentence parsers that can be used for the triples extraction cannot handle some special case words, such as capital names, multiple word names which are considered as independent words and also the ambiguities in distinguishing named entities. These issues can be handled during the preprocessing phase to prepare the sentences for the RDF extraction. Another cause for concern is the multi-clause sentences, the parsing of these sentences will lead to shortage or false representation of the real meaning of the sentence. Isn this case, multi-clause sentences should be split into single-clause ones with the maintenance of the semantics in the original text.

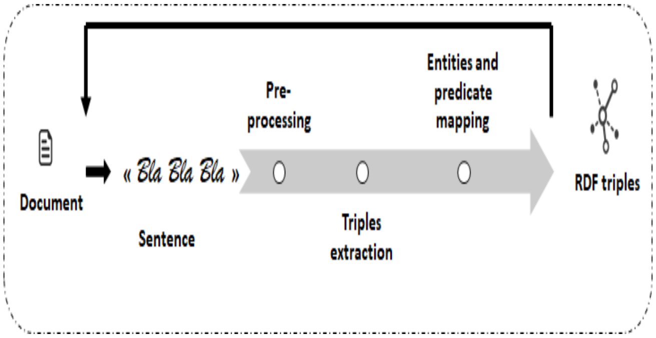

Fig. 2. RDF Extraction Process.

For each extracted single-clause sentence, the Stanford parser [28] can be used to analyze the grammatical structure of sentences, and in particular to identify the subject or object of a verb, in order to represent sentences in the form of the triple (subject, verb, object). Senna parser [29] can also be used to enrich the RDF extraction. Senna provides useful information, by allowing entity name recognition. It allows labeling of the named entities with given categories such as organizations, monetary value, person, etc. Its semantic role labeling can as well be used to enhance the discovery of the semantic sense of words in a given sentence.

Now that the triplets have been discovered, it is time to map to each extracted entity and predicate its Unified Resource Identifier (URI). DBpedia is a data set powered by Wikipedia articles that relates an entity to a Wikipedia article and provides a URI to identify it. In the mapping of RDF triples, using URIs provided by DBpedia. The use of DBpedia is due to the richness of the subject's details and the Multilanguage's description provided as well as the continuity of updates of the data sets. The identification of the most relevant meaning of words can be done based on the synsets provided by WordNet, which is a large-scale lexical database for English. After identifying the most relevant meaning of an entity based on its context and the syntactic and semantic role, it can be associated with its DBpedia URIs or WordNet URIs if it existed.

The named entity recognition Wikifier [30] can be used to obtain links to Wikipedia of the associated articles to the named entity. The following example illustrates the transformation from an unstructured text to an RDF graph.

Considering the sentence: "The WHO declared Covid-19 a pandemic". The Stanford parser will enable the tagging of this sentence (Table I).

Using WordNet to discover the appropriate sense of the words and select the named entity that corresponds to it in order to assign its DBpedia URIs/IRIs, Fig. 3 represents an RDF graphic representation of the sentence.

TABLE I. PART OF WORDS TAGGING USING STANFORD PARSER

	Word	POS		POS	Word
1	The	DT	4	COVID-19	VBN
2	WHO	NN	5	A	DT
3	declared	VBN	6	Pandemic	NN

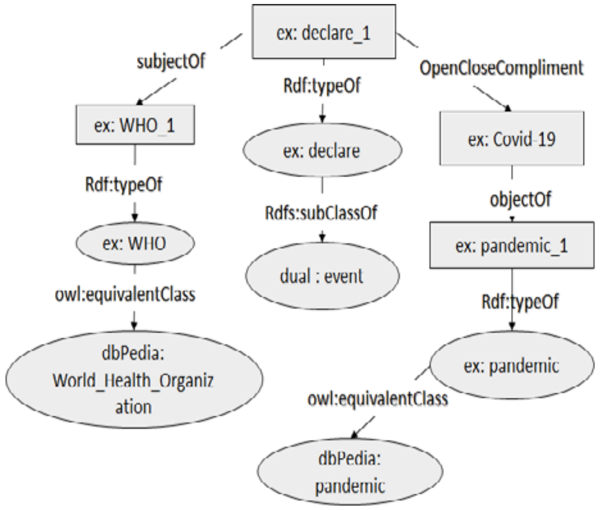

Fig. 3. Example of RDF Schema.

The extracted sets of triplets of each document not only allows the comprehension of documents for machines, but it will also be used in more sophisticated tasks such as document clustering, query answering and text summarization.

B. RDF based Clustering

After retrieving RDF triples for each text document we can proceed to the clustering of these documents. The clustering process consists of grouping documents in related clusters based on the similarity between them. In this case, the documents are represented using RDF graphs. The RDF triples in the graphs correspond to the document sentences, where each subject, object, or predicate are identified using a URI from the DBpedia datasets. In traditional clustering, the similarity between documents is calculated based on the text words considered as independent items. The use of RDF representation allows the incorporation of the semantic relationship of terms. However, the key to an efficient clustering is the use of a similarity measure that results in better matching between documents, not only based on relevant words with the highest strength or occurrence frequency in the documents as feature words for clustering but taking into consideration the semantic relationship between words and between documents. The extracted RDF graphs are loaded with semantic and syntactic information about the texts.

Our goal is to put forward a semantic similarity measure based on the RDF model with the exploitation of the semantic web tools. To tackle this issue, a similarity function based on RDF graph matching is going to be set up to compute the similarities between documents.

As earlier discussed, textual unstructured data is transformed into RDF graphs, the matching of two RDF graphs consists of the matching of their unitary elements which are the RDF triples and precisely it's about the calculation of the similarity between triples' subject, predicate, and object. Graph matching algorithms for RDF have been used for the matching of RDF queries and RDF graphs in order to implement searching processes or to put in place Linked Data recommendation processes. This study introduce the Graph matching algorithm to calculate the similarity between RDF graphs corresponding to a documents' dataset in order to perform clustering over these documents. In the following, we introduce and discuss the RDF graph distance computing in the clustering process.

1) Similarity computing between documents and clustering: We assume in the following that the previous step of RDF extraction is completed and that each document is represented as an RDF graph, where every RDF graph consists of a list of RDF statements.

The matching of two graphs can be translated to an assignment problem that would be solved using the Hungarian matching algorithm over a bipartite graph. The bipartite graph whose vertices are the set of triples of the two RDF graphs, each RDF's triples group is considered as an independent vertex of the bipartite graph, whereas the weight of the edges of the graph is the similarity measure between the triples.

Considering two documents D1 and D2, we represent these documents by two RDF graphs G1 and G2 respectively, and we define the bipartite graph BG as BG:=(U,V,E), U and V being the BG partition's parts such that U is the set of nodes related to the triples of the document D1, and V represents the triples of the second document D2. Moreover, E is stating the edges of the graph and the weight of these edges is the computed similarity measure between the nodes connecting the edges. The Hungarian matching algorithm is used over the BG to find the maximum similarity matching between the pairs of triples represented by U and V. Based on the matching result, the overall similarity measure can be computed between the two RDF graphs. This ability of measuring the similarity between a pair of documents yields to a similarity matrix based on the computed results. The computed similarity matrices can be used in multiple clustering techniques to provide homogenous clusters. This proposed framework allows therefore introducing an agglomerative hierarchical clustering to extract the clusters.

The following subsection discusses how we can obtain the similarity measure between a pair of triples.

2) Similarity computing between triples: As a result of the previous section, computing the similarity between triples is most crucial tasks in the clustering process that is to put in place since it can impact the clustering efficiency. As already mentioned, several methods consider the text as a dissociated bag of words, ignoring the semantics in texts. This is what we aim to tackle by handling unstructured textual data within the context of an RDF model in order to preserve the semantic relationships in the text, linking it to the important knowledge base in our case DBpedia and furthermore include a semantic similarity measure to compute the distance between RDF triples.

It is trivial that in order to compute the similarities between documents we need to measure the unitary similarity of the triples pairs. One of the major advantages of RDF representation of the unstructured textual documents is to be able to utilize the data values of resources to calculate the resources' similarity scores.

Considering two triples t1=(s1, p1, o1) and t2=(s2, p2 , o2) the similarity between t1 and t2, Sim_t (t1, t2), is related to the similarities between subjects Sim_s(s1,s2), between predicates Sim_p(p1,p2) and between objects Sim_o(o1,o2). Firstly, to compare words, in this case, it is essential to use a linguistic similarity measure based on a reliable source such as WordNet. Several researchers have tackled the use of WordNet in the similarity computing based on the provided synsets, and one of the most used formulae is Lin's similarity. Secondly, and in the case of a string value or not being able to find the word in WordNet database we can proceed to a string similarity measurement such as the normalized compression distance and the Levenshtein Distance [31]. Finally, for the URI form of data, if the corresponding value can be matched to a WordNet word then we could use the linguistic computing method, and if not the string similarity measurement could be used instead. The triple's object similarity will be handled in the same way as the subject, as for the predicate, we can consider the fact that is two triple's subjects (s1, s2) and objects (o1, o2) are similar then it is very likely that the predicates (p1, p2) are also similar, otherwise linguistic and string similarity computing methods can be used.

C. Topic Modeling

Clustering goal is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic. One of the major challenges related to clustering is cluster labeling and how to provide a clear description of the clusters. Therefore, we can note that for identifying and describing the constructed clusters Topic models can be used. In this case, clustering allows the inference of more coherent topics. A topic is a group of words that resume and refers to the content of the cluster. The identification of the cluster's topic allows a previous view over its content and eases the searching process. Topic models were firstly introduced for text documents and are easily adaptable to the case of RDF graphs. In [26] an approach to use topic modeling with RDF data was proposed using Latent Dirichlet Allocation (LDA) which is a commonly used model to identify the topics of documents. LDA aims to extract thematic information from documents' collections and it is based on the bag of words as vocabulary extracted from these documents. In our use case, the documents are the RDF Graph and the words are the extracted words from the graphs' triples. [26] introduced several limitations and challenges of using topic modeling for RDF graphs; Firstly, the sparseness of RDF data which means that even when having large datasets, the preprocessing of this data could result in a restricted set of words that could be used as a bag of words. Secondly, the lack of context is encountered since used words can have several meanings. In the case of RDF data, the context is hard to be determined due to unnatural language of data and/or to the sparseness of RDF graphs. The unnatural language is related to the graph representation, unlike sentence representation that enhances the understanding of the words, and finally, the short text problem which can be handled by either text supplementing or providing a modified version of the LDA algorithm. However, the strengths of our RDF graph representation process help overcoming these challenges. since textual documents were converted to RDF graphs, the use of semantic and syntactic parsing tools and the introduction of Dbpedia and WordNet synsets for efficient entity recognition based on the context of the sentence helps to tackle the issues related to unnatural language nature of RDF science the identified entities are based on the context of the documents, in order to identify the most relevant meaning of an entity. On the other hand, the RDF graph is enhanced with semantic role relations and Dbpedia classes allow overcoming the likely sparseness nature of RDF and text shortness problems.

D. Summarizing Clusters and Questioning System based on RDF Clustered Data

In order to enhance the exploration of RDF data and due to the big amount of RDF data and its complexity, RDF summarization was introduced to assist the understanding and use of this type of data. Summarization aim is to provide brief, concise, and significant information. Our framework goal is to make better use of the textual documents through a semantic text clustering system. Therefore the use of the RDF summarization techniques in the proposed framework is guided with the attention to improve the information extraction from the handle datasets. Hence, it is important to assist the queuing system based on RDF data since the extracted RDF graphs clusters can be significantly large resulting in a querying process that is extremely expansive with regards to resource and time.

Summarization can be used for various reasons or applications such as ontology extraction from RDF graphs, assisting users by providing graph visualization, and improving the querying process. In our case, we are basically interested in these applications related to the advancement of the querying task in many ways. In particular, indexing is when summary graphs are seen as an index for the larger RDF graph. In this case, a query is initially matched with the summary graph for finding equaling index nodes, and then the original graph is explored after detecting the matching nodes. Thus this process reduces the computation time and improves the querying task. To be noticed is that a summary will also help identify the best matching data partition to apply the querying when working with distributed systems.

There are multiple RDF summarization approaches, some include ontologies to handle the summarization of an RDF graph, and others can ignore their use and only work on the bare RDF graph. Based on some recent reviews such as [32], a RDF summary can either be compact information that contains the major meanings of the graph or can be a graph that is exploitable rather than the massive original graph. In [32] the existing summarization approaches can be classified based on multiple criteria. Among others we cite the input and output type, the purpose and the methods which can be structural, statistical, pattern-mining, or hybrid. In order to reach our goal structural quotient summaries for its wide applicability in the indexing and query answering tasks. Quotient summarization graphs are summary graphs where each summary node is connected to the IDs of the original graph nodes. These graphs can be obtained based on equivalence relations such as bi-similarity.

E. Reasoning using Jena Inference Engine

RDF schema (RDFS) allows defining and organizing RDF data vocabularies. In RDFS, the relationships between properties and resources are defined using RDF which offers a typing arrangement for RDF models. These relations are hierarchical like the notion of classes, subclasses, and properties, a huge amount of links between elements can be identified by specifying properties of classes and inheritance between classes, therefore RDF objects are considered as an instance of one or many classes and are specified with the class properties and parent class specifications. Many projects have incorporated the use of RDF(S) representation format such as Protégé, and Mozilla. Web Ontology Language enhances the describing properties and classes by providing an extended vocabulary. It allows for example expressing the cardinality of relations between classes, offers other assets such as equality and symmetry of properties, and so on [1]. These OWL characteristics result in more detailed ontologies allowing high performance in documents reasoning tasks. As we already mentioned semantic web concept is all about allowing both humans and machines to understand data, therefore data should be presented in a well-structured form and rules should be provided in a well-defined language in order to implement reasoning process and allows data to be shared onto the web [2]. Logically, notions are inferred from ontologies if they conform to their associated models; this process is referred to as reasoning. The clustering semantic-based framework proposed in this work can be enhanced by including a reasoning layer that allows deriving additional truth from the RDF graphs. Tools such as the Jena framework have been widely used to extract data from RDF graphs and OWL ontologies.

IV. Conclusion

This paper showed how semantic web techniques can be used for the textual documents clustering and exploration. It underlined some of the existing works of manipulating RDF data and got inspired from these to present a connected pipeline of semantic processes for the semantic text clustering based on RDF. The main contribution consists on presenting an overall framework for semantic text clustering based on RDF data modeling. This framework combines multiple techniques in order to get an efficient and accurate system, allowing exploring textual documents using machine learning techniques combined with semantic web principles. The system allows documents RDF representation, clustering, topic modeling and clusters summarizing as well as information retrieval using both RDF querying and reasoning tools. The aim is to take advantage of the semantic web in order to enhance the exploration of documents and enhance the use of semantics along the whole process. In future work we intend to validate our framework and improve it by choosing most relevant tools and techniques in each of the framework's steps. An experiment of the proposed approach on a real dataset will be further attacked.

References

[1] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 web ontology language primer," https://www.w3.org/TR/owl-primer/, 2012.

[2] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 1st ed.; Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

[3] M. Noura, A. Gyrard, S. Heil, and M. Gaedke, "Automatic knowledge extraction to build semantic web of things applications," IEEE Internet Things J., vol. 6, no. 5, pp. 8447–8454, Oct. 2019.

[4] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," J. of Web Semantics, vol. 36. Elsevier, pp. 1–22, Jan. 01, 2016.

[5] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì, "Semantic web machine reading with FRED," Semant. Web, vol. 8, no. 6, pp. 873–893, Aug. 2017.

[6] P. Pauwels, S. Zhang, and Y. C. Lee, "Semantic web technologies in AEC industry: A literature overview," Automation in Construction, vol. 73. Elsevier B.V., pp. 145–165, Jan. 01, 2017.

[7] M. Trianes Torres, A. Sánchez Sánchez, M. Blanca Mena, and J. García, "Competencia social en alumnos con necesidades educativas especiales: nivel de inteligencia, edad y género," Rev. Psicol. Gen. y Apl. Rev. la Fed. Española Asoc. Psicol., vol. 56, no. 3, pp. 325–338, 2003.

[8] G. Xu, Y. Cao, Y. Ren, X. Li, and Z. Feng, "Network security situation awareness based on semantic ontology and user-defined rules for Internet of Things," IEEE Access, vol. 5, pp. 21046–21056, Aug. 2017.

[9] N. Komninos, C. Bratsas, C. Kakderi, and P. Tsarchopoulos, "Smart city ontologies: improving the effectiveness of smart city applications," J. Smart Cities, vol. 1, no. 1, pp. 31–46, Nov. 2016.

[10] G. Sarthou, R. Alami, and A. Clodic, "Semantic Spatial Representation: a unique representation of an environment based on an ontology for robotic applications," in Proc. Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP), pp. 50-60, Association for Computational Linguistics, 2019.

[11] G. S. Reddy, T. V. Rajinikanth, and A. A. Rao, "A frequent term based text clustering approach using novel similarity measure," Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014, pp. 495–499, 2014.

[12] S. Eddamiri, E. M. Zemmouri, and A. Benghabrit, "An improved RDF data Clustering Algorithm," Procedia Comput. Sci., vol. 148, pp. 208–217, 2019.

[13] M. Achichi, Z. Bellahsene, D. Ienco, and K. Todorov, "Towards Linked Data Extraction From Tweets," https://hal.archives-ouvertes.fr/hal-01411403/document, 2016.

[14] A. M. Abirami, A. Askarunisa, R. Sangeetha, C. Padmavathi, and M. Priya, "Ontology based ranking of documents using Graph Databases: a Big Data Approach,": https://www.amrita.edu/icdcn/sangeetha-r.pdf, 2014.

[15] V. Saquicela, F. Baculima, G. Orellana, N. Piedra, M. Orellana, and M. Espinoza, "Similarity detection among academic contents through semantic technologies and text mining," in IWSW, pp. 1-12, 2018.

[16] N. Kertkeidkachorn and R. Ichise, "T2KG: An end-to-end system for creating knowledge graph from unstructured text," AAAI Work. - Tech. Rep., vol. WS-17-01-, pp. 743–749, 2017.

[17] N. Kertkeidkachorn and R. Ichise, "An automatic knowledge graph creation framework from natural language text," IEICE Trans. Inf. Syst., vol. E101D, no. 1, pp. 90–98, 2018.

[18] R. Clancy, I. F. Ilyas, J. Lin, and D. R. Cheriton, "Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond," in Proc. Second Workshop on Fact Extraction and VERification (FEVER), pages 39–46 Hong Kong, November 3, 2019. Association for Computational Linguistic.

[19] T. Tran, H. Wang, and P. Haase, "Hermes: Data Web search on a pay-as-you-go integration infrastructure," J. Web Semant., vol. 7, no. 3, pp. 189–203, 2009.

[20] H. He and J. Lin, "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement," 2016 Conf. North Am.

Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf., pp. 937–948, 2016.

[21] H. Patil and R. S. Thakur, "Frequent Term-Based Text Clustering Using Hidden Support," Proc. Int. Conf. on Recent Advancement on Computer and Communication, B. Tiwari et al. (eds.), Lecture Notes in Networks and Systems 34, vol. 34. Springer Singapore, 2018.

[22] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, “A semantic approach for text clustering using WordNet and lexical chains,” Expert Syst. Appl., vol. 42, no. 4, pp. 2264–2275, 2015.

[23] V. S. Anoop, S. Asharaf, and P. Deepak, “Topic modeling for unsupervised concept extraction and document ranking,” Adv. Intell. Syst. Comput., vol. 683, pp. 123–135, 2018.

[24] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” Springerplus, vol. 5, no. 1, 2016.

[25] L. Yao et al., “Incorporating knowledge graph embeddings into topic modeling,” 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 3119–3126, 2017.

[26] J. Sleeman, T. Finin, and A. Joshi, “Topic modeling for RDF graphs,” CEUR Workshop Proc., vol. 1467, pp. 48–62, 2015.

[27] S. Pouriyeh, M. Allahyaril, G. Cheng, H. R. Arabnia, K. Kochut, and M. Atzori, “R-LDA: Profiling RDF Datasets using knowledge-based topic modeling,” Proc. - 13th IEEE Int. Conf. Semant. Comput. ICSC 2019, pp. 146–149, 2019.

[28] “The Stanford Natural Language Processing Group.” https://nlp.stanford.edu/software/lex-parser.shtml.

[29] “SENNA,” https://ronan.collobert.com/senna/.

[30] J. Brank, G. Leban, and M. Grobelnik, “Annotating documents with relevant {Wikipedia} concepts,” Proc. Slov. Conf. Data Min. Data Wareh., p. 4 pages, 2017.

[31] K. Al-Khamaiseh, "A Survey of String Matching Algorithms," Int. J. of Engineering Research and Applications, vol. 4, Issue 7 (Version 2), pp.144-156, 2014.

[32] Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika, “Summarizing semantic graphs: a survey,” The VLDB J., vol. 28, no. 3, pp. 295–327, 2019.

[33] K. Hassanzadeh, M. Reformat, W. Pedrycz, I. Jamal, and J. Berezowski, "T2R: System for Converting Textual Documents into RDF Triples;" in Proc. IEEE/WIC/ACM Int. Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013.

[34] Augenstein, I., S. Padó, and S. Rudolph, "Lodifier: Generating linked data from unstructured text," In ESWC, pp. 210–224, 2012.

[35] Q.-Y Zhang, P. Wang, and H.-J Yang, "Applications of Text Clustering Based on Semantic Body for Chinese Spam Filtering," J. of Computers, vol. 7, no. 11, pp. 2612-2616, 2012.

[36] Seyedamin Pouriyeh, Mehdi Allahyari, Krzysztof Kochut, Gong Cheng, and Hamid Reza Arabnia. Es-lda: Entity summarization using knowledge-based topic modeling. in Proc. of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, pages 316–325, 2017.