Semantic Web Technologies for Classification and Indexing of E-government Archive

Yousef A. Aburawi
Misurata University / Faculty of IT , Misurata,Libya
yaburawi@it.misuratau.edu.ly

Abdulbaset S. Albaour
Misurata University / Faculty of IT , Misurata,Libya
baset.albaour@umit.edu.l

Abstract— **Semantic web Meta data and natural language processing technologies can be combined to produce systems that working more efficiently for classification and retrieval of documents. The main aim of this Paper is applying semantic web technologies and tools for indexing, classifying, and retrieving Arabic content documents within E-government archive by using semantic web annotation technology and natural language processing (NLP).**

Index Terms: **Semantic web, NLP, document classification, and document retrieval.**

I. INTRODUCTION

The Information and Knowledge has been a subject of the great meaning since the beginning of the human history. The amount of knowledge grows from day to day and the requirements for the simplicity and speed of requests grow as well.

In today's world, the documents play a major role in our life. Therefore, the classification of these documents will be very sensitive, especially in huge enterprises such as big companies and E-government.

Document management and classification procedures are significant factors in the modernization process of the E- government entities. Efficient procedures result into:

- Access to historical information.
- Easy retrieval of relevant documents.

The proposed system aims to introduce a system for classification of E-government documents by utilizing semantic web technologies and natural language processing.

Received 9 March 2016; revised 16 February 2016; accepted 21 March 2016.
Available online 24 March 2016.

II. THEORETICAL BACKGROUND

A. *Current Issues*

According to current situation, the classification and archiving of E-government documents is manual and there is no software for automatic classification Of Arabic content documents. Consequently, it sometimes takes a lot of time and effort to categorize a document in correct classification.

To meet the general goal of this research there are some issues that will need to be considered. For example, the classified document has to be formed in predefined format (template) to insure that all document features are extracted correctly.

B. *Semantic Web Technologies*

Semantic Web is a group of methods and technologies to allow machines to understand the meaning - or "semantics" - of information on the World Wide Web [1] or a specific document.

The story of semantic web has started in the book weaving the Web. The Original Design and Ultimate Destiny of the World Wide Web [2] written by Sir. Tim Berners-Lee, inventor of the World Wide Web and chairman of the World Wide Web Consortium (W3C). Then, he has clarified the idea in Semantic Web article: The Semantic Web is an extension of the current web, in which information is given well-defined meaning, to enable people and computers to work in cooperation..... These developments will give important functionality such as the ability of processing and understanding the data that they merely display at present. [3]

Nowadays the web has come into view as the largest distributed information repository in the world. Human knowledge appeared on the Web in various digital forms: Web pages, news articles, blog posts, digitized books, videos, speech transcripts, etc.

The fundamental question here: how the knowledge will be extracted from this data?

Knowledge management (KM) discipline has been developed concepts and technologies to retrieve information and knowledge from heterogeneous data sources. Those technologies like semantic web and ontologies are used to fill the gap between data and knowledge.

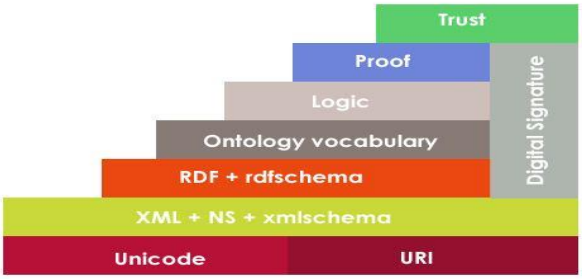

Figure 1. Semantic Web Stack. [4]

The semantic technologies working by building an extra metadata layer upon the current content and using the description language to describe the resources or pieces of information in a specific domain. In the domain every resource is identified by Uniform Resource Identifier (URI), we can think of a resource as a subject or a "thing" we want to talk about. Such identification enables the interaction of the resource using specific ways. The Triples of URIs produce Resource Description Framework (RDF) that represents basic data model of Subject, Predicate, and Object (Figure 2), which is known usually as a statement.

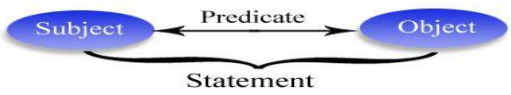

Figure 2. RDF Model

C. *Metadata*

The core Idea of the semantic web is metadata, which generally defined as:

"Metadata are data about data"

Metadata is structured data on the particular type of information in the document that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

There are three kinds of metadata:

Descriptive: describes a resource and so it is easier to identify this resource. Such as: author, keywords, etc. Structural describes how two objects fit together. Such as Order of page numbers.

Administrative: all metadata, supporting rights management (e.g. creator) or digital archiving (e.g. version).

An important reason for generating metadata is to facilitate detection of relevant information. In addition to resource detection, metadata can help arrange information resources, provide digital recognition, and support archiving and protection.

Extraction tools will automatically generate metadata from an analysis of the content of digital resource. These tools are generally dedicated to textual resources. The quality of the metadata extracted can vary significantly based on the tool's algorithms as well as the structure and content of the source document.

These tools should be considered as an aid to creating metadata. Especially if resource documents cover particular subject [5].

D. *Ontology*

The ontology term has borrowed from philosophy, a subfield of philosophy that study the nature of existence or the kinds of things that actually exist, and how to describe them. However, in latest years, it has become one of the many words are taken by computer discipline and given a specific technical meaning that is different from the original one.

Therefore, we can define the ontology as:

Ontology is a formal representation of concepts that exist in a domain and the relationships between those concepts. [6]. In general, the ontology consists of a finite number of terms and the relationships between them, these terms denote the concepts of the domain. So, the ontology is a model for a specific subject: It shows the relationships between concepts (classes), and it shows hierarchical structure to explain these relationships.

Ontology's relationships may include information such as:

- Properties (Professor teaches Students).
- Value restrictions (only Professor A may teach Class).
- Disjointness statements (Professors Disjoint Stuff),
- Specifications of logical relationships between objects (every department must include at least 10 Professors).

This formal organization of concepts will provide a shared and clear understanding of the domain, which will overcome the differences in terminology by mapping specified concept in the domain and all terms that denote it.

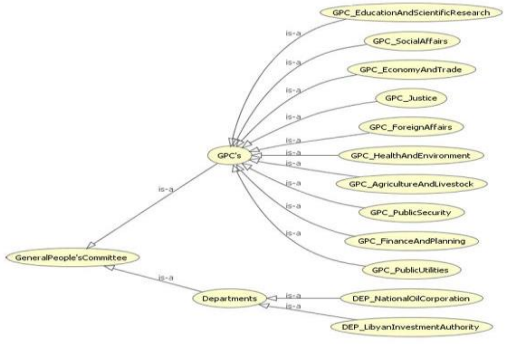

Figure 3. A Simple Ontology of Libyan Government.

E. Natural Language Processing (NLP)

Natural language has been used in the computing field for years and is still a popular form of input because it is understood by everyone. Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages [8].It contains a broad set of techniques for automated generation, manipulation, and analysis of natural human language. NLP plays an essential role in any system that interacts between computers and natural language.

F. Information Extraction

Information Extraction (IE) is a NLP technology for extracting specific types of information from text resources to be populated to the knowledge base. Its aim is to extract structured knowledge from unstructured text to enhance the use and reuse of that information.

This structured information source is then used for some other purposes, such as searching or analyzing.

E-government documents archiving requires Ontology- based Information extraction tasks to be classified depending on the content of the document.

OBIE is the process of identifying in text resources relevant concepts, properties, and relations expressed in the specific ontology. The ontology contains concepts (classes) organized hierarchically, relations between concepts, and properties. This ontology represents the domain of application. Once the information has been extracted from text resource, the ontology has to be populated with

all gathered concepts, and then the system has to decide how to classify this document.

G. knowledge Management (KM)

Knowledge is the full utilization of data and information, joined with the potential of people's skills, ~~ideas~~ and intuitions.

Knowledge is more relevant to develop business than capital, labor or land. Nevertheless, it remains the most neglected factor. It is essential for action, performance and adaption. Knowledge provides the ability to respond to new situations.

Knowledge is found in ideas, judgments, talents, relationships, perspectives and concepts. And it has to translate to organizational processes, documents, products, services, facilities and systems. For knowledge to be of value it must be focused, current, tested and shared. Both of Metadata and Ontologies are used in knowledge management discipline to create an explicit organization of information.

There is no exact definition for KM, because it contains of different varieties of strategies and practices used in an organization to identify, create, represent, distribute, analyze and compare all information and experience. [7]

With the beginning of information age, the value of knowledge rises rapidly because every collection of data can be used to gain new knowledge.

Knowledge management has three stages:

- Data collection.
- Information extraction from collected data.
- Gaining new knowledge from the information.

New knowledge will control the human activity (e.g. Business) to get the wise behavior (Wisdom).

Figure 4. Representations of the DIKW hierarchy. [7]

H. Similar efforts

Knowledge base is a collection of data representing related organized information that are related to particular domain or area. The information extraction techniques can be a very useful tool to build a knowledge base from unstructured free text like text document and tables.

At the present time, there exist many systems that generate knowledge base from extracted information. SOBA (SmartWeb Ontology-Based Annotation) [8] is one of these systems. It can automatically extract information from football match reports based on an underlying ontology. These reports may exist in a heterogeneous formats such as text, images, and image captions. It is actually used to extract information from football web pages for automatic population of a knowledge base that can be used for domain specific question answering. SOBA realizes a tight connection between the ontology, knowledge base and the information extraction component. Extracted information has stored in a knowledge base, and in turn uses the knowledge base to understand and connect newly extracted information with respect to already existing entities.

The system consists of three main parts: web crawler, linguistic annotation components and a component for the transformation of linguistic annotations into a knowledge base.

The web crawler acts as an observer on relevant websites, automatically downloads documents from them and sends these to a linguistic annotation web service. Linguistic annotation and information extraction are based on the Heart-of-Gold (HoG) architecture, which generates a uniform and flexible infrastructure for building multilingual applications that use XML-based natural language processing components. The linguistically annotated documents are processed by the semantic

transformation component, which provides a knowledge base of football-related events (matches, goals, etc.) and entities (players, teams, etc.) by mapping annotated events or entities to ontology classes and their properties. As a conclusion, an information-extraction system which relies on ontology to formalize and semantically integrate (link) extracted information from diverse resources in a knowledge base is done by SOAB.

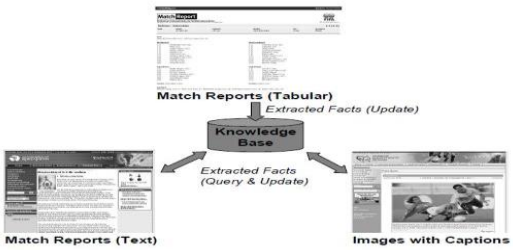

Figure 5. SOAB Architecture. [8]

The authors in Ontology-based Information Extraction for Business Intelligence [9] paper have aimed to use information extraction as an acquisition tool to provide valuable information to customers or decision makers. The combination between GATE (General Architecture for Text Engineering) and the domain ontology make possible to create robust system to extract and merge information from a variety of resources to produce reasonable knowledgw base and valuable information.

The robust and adaptable technology has been developed for the extraction of relevant semantic information to be used in variant business intelligence processes in different areas such as financial risk management, internationalisation, and IT operational risk management. The relevant applications in these areas are international company intelligence, country or region selection, risk identification and mapping. All these applications need the extraction and merging of information from a number of reliable but diverse data sources (news reports, financial reports, database, and company websites).

In conclusion, the system generates ontological annotations which are transformed into tuples for ontology population. The system already extracts information from a range of sources and for specific applications in financial risk management and internationalisation.

Applications are being created which use the priceless information in the knowledge based to provide valuable information to customers. Performance measured through quantitative evaluation in both extraction and cross- source co reference look promising.

III. RESEARCHER WORK

A. System Construction

The background research has covered the use of NLP technology will be sufficient for reading the content of unclassified document. Semantic web technology will be used for building E-government ontology that represents the hierarchy and taxonomy of different departments. In order to build the proposed system the following development skills and tools will be needed:

- Java: will programming language to build the proposed system and NetBeans IDE 6.5.1 will be used as a Java IDE.
- GATE (General Architecture for Text Engineering): a powerful tool for NLP.
- Protégé-OWL: a powerful tool for building semantic web ontology.
- SPARQL query language: Querying Linked Data.
- Gena Framework: reading and writing RDF models in Ontologies.

B. System skeleton

Classification Process Classification process for unclassified document can be divided into four stages :

1. Reading unclassified document.
2. Document annotation.
3. Data extraction and classification.
4. Ontology Population.

These stages are detailed as following:

- The unclassified document will be loaded to the system.
- ANNIE in GATE will extract information from the document in the following order:
  1. The tokeniser will split the text into simple tokens or main types, such as numbers, symbols, and words.
  2. The sentence splitter will divide the text into sentences.
  3. The entity recognition process will be done based on E-government gazetteers that list the relevant items of one E- government department or ministry.
- The software will count the number of keywords of each department and will guess the classification that documents relevant to.
- Other features of a document like date of issuing and document title will be extracted at this stage.
- If the software fails in document classification, it will ask the user to classify the document manually.
- Classification of the document and its features will be populated to the ontology that represents the knowledge base.

The documents will be able to retrieve based on their features such as: classification, title and issuing date.

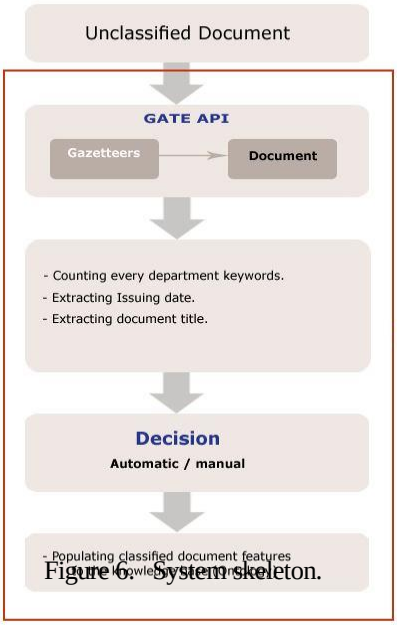

Figure 6. System skeleton.

IV. RESULTS / DISCUSSION

To find out whether the project has been successfully utilizing a variety of proposed techniques to meet the objectives. Testing is comprehensive and covers every action that might be possible under different circumstances.

A. *Classification Testing*

The group of Arabic documents was elected as inputs; this election is based on available GPC gazetteers that built during the project building. Then the example of program classifications is obtained as following:

Table 1. Classified Documents

Document	Expected Classification	Actual classification
Doc1	وزارة الخارجية	وزارة الخارجية
Doc2	وزارة الخارجية	وزارة الخارجية
Doc3	وزارة الخارجية	وزارة الخارجية
Doc4	مصلحة التسجيل العقاري	مصلحة التسجيل العقاري
Doc5	وزارة الزراعة	وزارة الزراعة

B. *Retrieval Testing*

Classified documents covered in the previous example will be populated into a knowledge base with specific features; the retrieving process will be available with a combination of document features such as classification, date and any part of the document title. The previous group of documents has populated to knowledge base with following features:

Table 2. Documents Features

Document	Date	Class.	Title
Doc1	28/09/2013	وزارة الخارجية	باعتماد اتفاق للتعاون في مجال التعليم والبحث العلمي
Doc2	02/10/2013	وزارة الخارجية	باعتماد مذكرة تفاهم بين ليبيا وحكومة مالطا
Doc3	17/10/2013	وزارة الخارجية	باعتماد محضر اجتماعات الدورة الثامنة للجنة المشتركة الليبية النمساوية
Doc4	17/10/2013	مصلحة التسجيل العقاري	الإشراف على التعويضات وتقرير حكم
Doc5	17/10/2013	وزارة الزراعة	باعتماد برنامج توزيع المزارع المستصلحة

Once the SPARQL query built in the correct format, it will be executed by SPARQL engine that is a part of GENA framework and relevant results will be retrieval successfully.

Retrieval process produced following results:

Table 3 . Retrieved Documents

Selected Feature	Result
Classification = وزارة الخارجية	Doc1 ,Doc2,Doc3
Document name = "مذكرة تفاهم"	Doc2
Document date = "09/2010"	Doc1
Classification = وزارة الخارجية + Document date = "10/2013"	Doc2,Doc3

C. User Testing

A documents classification program is available to be used by non-technical users, so it is important to test the program by users from different technical backgrounds. Users tested the program on a laptop in a room on their own. Users were given a practice task to perform before actually testing the real program. The intention of this was to get users familiar with the program. When the users had completed the practice task, they performed several tasks with the real program. When the users had completed the tests they were asked to complete a questionnaire.

Users were divided into three groups:

- Group 1: advanced users such as information technology students.
- Group 2: users with medium skills of computer using.
- Group 3: users with normal skills of computer using.

For each group, the arithmetic mean has taken for each question to represent question score.

The usability testing was a useful tool for evaluating the usability of the application and changes were made to the user interface.

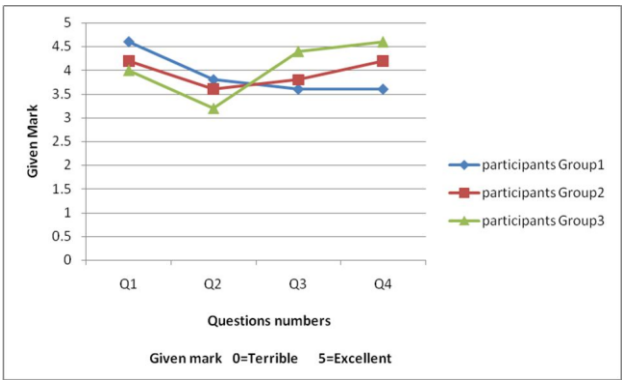

Figure 7. Questionnaire Results.

V. CONCLUSION / FUTURE WORK

A. Conclusions

The completed process of classification of Arabic content documents has met the objectives set at the beginning of the research. The program has a well designed user interface that is easy to use. Through the use of semantic annotation for the document by specific gazetteers and analyzing this layer of annotations program well guess the suitable classification for the target document.

The big challenge faced the project was using Arabic language in the treated document. This problem was overcome by using Unicode Transformation Format (utf-8) that supports Arabic language. Moreover, the program enables users to populate features of classified documents into knowledge base and retrieve those documents based on a group of search keys.

Classification process: Classification process done by many sequential stages by utilizing semantic web concepts that allow us to build metadata layer based on specific gazetteers that cover a set of terms or keywords in the project area.

By counting the number of keywords in each sub area the program will guess the document classification.

Document features such as document date and document title will be extracted at this stage.

Population process: Once the document classification is guessed successfully and the features are extracted, all this information will be populated into knowledge base (domain ontology) that represents data storage and provide the enterprise hierarchy.

If the program failed in guessing the classification, it will allow the user to enter the correct classification manually.

Document body also will be stored on specific area on the hard disk

Retrieve process: Classified documents could be searched and retrieved by building user queries based on document features. The program allows the user to brows a set of retrieved documents.

B. Future Work

There are three main areas where further developments can be made. These areas are:

Developing program package: The program can be rewrite using Java programming language to produce the whole package contains all program resources to make the program runs on any platforms and operating system, including Mac, UNIX and Linux platforms. And make them easy to be installed and removed.

Developing Programs for Other Languages: Due to the similarity in characters sets and Unicode Transformation Format (utf-8) support. It is a possibility that new programs could be created for other languages that use Arabic alphabet but are fundamentally a different language such as: Persian, Pashto and Urdu.

Comparing with other archiving ways: It will be very useful to compare the using of semantic web technologies with any existing methods in building similar programs; this comparing can conclude the best method and its efficiency

REFERENCES

[1] Semantic Web [Online]. Wikipedia. Available at: http://en.wikipedia.org/wiki/Semantic_Web [September 10, 2015].

[2] T. Berners-Lee and M. Fischetti, 1999. Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. Harper San Francisco.

[3] T. Bemers-Lee, J. Hendler, and O. Lassila, 2001 .The semantic web. Scientific American. vol. 284(5), pp. 34–43.

[4] G.Antoniou and F. Harmelen, 2008.A Semantic Web Primer. 2nd edition. The MIT Press.

[5] Understanding metadata [Online]. NISO Press.Available at: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf [September 17, 2015].

[6] G.Antoniou and F. Harmelen, 2008.A Semantic Web Primer. 2nd edition. The MIT Press.

[7] Jennifer Rowley, 2007. The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science. Vol. 33 (2), pp. 163-180.

[8] P.Buitelaar,et al . Ontology-based Information Extraction with SOBA [Online]. Available at: http://www.dfki.de/~paulb/lrec2006.SmartWeb.pdf [November 12, 2015]

[9] H.Saggion,et al . Ontology-based Information Extraction for Business Intelligence[Online]. Available at: http://gate.ac.uk/sale/iswc07/musing/musing-iswc07.pdf [November 14, 2015]