International Journal of Web Information Systems

CASONTO: An efficient and scalable Arabic semantic search engine based on a domain specific ontology and question answering

Awny Sayed Amal Al Muqrishi

Received 26 December 2015
Revised 3 March 2016
6 March 2016
Accepted 6 March 2016

CASONTO

An efficient and scalable Arabic semantic search engine based on a domain specific ontology and question answering

Awny Sayed
Department of Computer Science, Minia University Faculty of Science, Al Minia, Egypt, and
Amal Al Muqrishi
Ibri College of Applied Science, Ibri, Oman

Abstract

Purpose – The purpose of this paper is to present an efficient and scalable Arabic semantic search engine based on a domain-specific ontological graph for Colleges of Applied Science, Sultanate of Oman (CASOnto). It also supports the factorial question answering and uses two types of searching: the keyword-based search and the semantics-based search in both languages Arabic and English. This engine is built on variety of technologies such as resource description framework data and ontological graph. Furthermore, two experimental results are conducted; the first is a comparison among entity-search and the classical-search in the system itself. The second compares the CASOnto with well-known semantic search engines such as Kngine, Wolfram Alpha and Google to measure their performance and efficiency.

Design/methodology/approach – The design and implementation of the system comprises the following phases, namely, designing inference, storing, indexing, searching, query processing and the user's friendly interface, where it is designed based on a specific domain of the IBRI CAS (College of Applied Science) to highlight the academic and nonacademic departments. Furthermore, it is ontological inferred data stored in the tuple data base (TDB) and MySQL to handle the keyword-based search as well as entity-based search. The indexing and searching processes are built based on the Lucene for the keyword search, while TDB is used for the entity search. Query processing is a very important component in the search engines that helps to improve the user's search results and make the system efficient and scalable. CASOnto handles the Arabic issues such as spelling correction, query completion, stop words' removal and diacritics removal. It also supports the analysis of the factorial question answering.

Findings – In this paper, an efficient and scalable Arabic semantic search engine is proposed. The results show that the semantic search that built on the SPARQL is better than the classical search in both simple and complex queries. Clearly, the accuracy of semantic search equals to 100 per cent in both types of queries. On the other hand, the comparison of CASOnto with the Wolfram Alpha, Kngine and Google refers to better results by CASOnto. Consequently, it seems that our proposed engine retrieved better and efficient results than other engines. Thus, it is built according to the ontological domain-specific, highly scalable performance and handles the complex queries well by understanding the context behind the query.

Research limitations/implications – The proposed engine is built on a specific domain (CAS Ibri – Oman), and in the future vision, it will highlight the nonfactorial question answering and expand the domain of CASOnto to involve more integrated different domains.

International Journal of Web Information Systems
Vol. 12 No. 2, 2016
pp. 242-262

1744-0084
DOI 10.1108/IJWIS-12-2015-0047

This work is funded by TRC (The Research Council) Sultanate of Oman from 2012 to 2015.

Originality/value – The main contribution of this paper is to build an efficient and scalable Arabic semantic search engine. Because of the widespread use of search engines, a new dimension of challenge is created to keep up with the evolution of the semantic Web. Whereas, catering to the needs of users has become a matter of paramount importance in the light of artificial intelligence and technological development to access the accurate and the efficient information in less possible time. However, the research challenges still in its infancy due to lack of research engine that supports the Arabic language. It could be traced back to the complexity of the Arabic language morphological and grammar rules.

Keywords Web search and information extraction,
Language and representation issues of Web semantics, Metadata and ontologies

Paper type Research paper

1. Introduction

Current Web contains billions of documents that has many administrative problems and limitations. Its content is still readable only by humans. To solve these problems, Tim Berners-Lee introduced the Semantic Web as a conceptual model of Web that makes the contents to be read and used by human and intelligently by machines. Semantic Web can enhance the currently existing Web by formulating the structure that may be understandable not only for humans but also for machines. In addition, this will permit computer to understand what the Web page is about, and hence draw conclusions about it. This innovation could highlight and reveal the meaning of contextual vocabularies, which will enable to unleash a revolution of new probabilities. This is illustrated by the information which can easily be found, shared, integrated and exchanged (Karin *et al.*, 2010; Samhaa *et al.*, 2007).

Basically, there are two types of searching the *Classical Search* and *Semantic Search*. The first one is known by keyword-based search that have popularized keywords in which means the users can submit their queries to the search engine, and a ranked list of outcomes is returned back to the user (Agrawal *et al.*, 2002). Currently, the keyword-based search engines such *as Google*, *Bing* and *Yahoo* have lack in support of the concept of semantics; therefore, they give many irrelevant results to the users and without any accuracy (Antoniou and van Harmelen, 2008). Fundamentally, this concept is so far from understanding the searcher intent and the contextual meaning of the user's query. Therefore, it is regarded as a critical challenge that has been addressed and solved by the *Semantic Search*.

In the context of Web, ontology is considered as one of the corner stone's of semantic search engines. It has been used in the past years by the *artificial intelligence* and *knowledge representation* communities; however, nowadays, it is becoming as a part of the standard terminology of a much wider society including information systems modeling (Gruber, 1993). In addition to this, it is becoming quite significant for the cause of lack standards, such as shared knowledge, which is rich in semantics and represented in machine understandable form. Another related point to this, it has been proposed to resolve the problems that arise from using variety of terminologies to define the same concept or use the same term to identify different concepts (Nicola and Pierdaniele, 1995).

Ontology is also provided with the required conceptualizations and knowledge representation to face the current challenges of the classical search. First is the inability to use the abundant information resources on the Web. While the Web has tremendous set of useful information, however, getting information from the Web is quite difficult.

Search engines are restricted to a simple keyword-based technique. Interpretation of information contained in Web documents is left to the human user. Second challenge is the difficulty of the information integration, while the integration of data from different sources is a challenging task for the reason of synonyms and homonyms. Finally are the issues of knowledge management. Multi-actor scenario involved in distributed information production and management. If the people and machines do not speak a common language, they cannot share knowledge among them.

The knowledge in the ontology could be represented as a *specific domain* according to the degree of conceptualization (Ontogenesis, 2010). *Domain ontology* is provided with vocabularies about concepts in the same domain only and their relationships or about the theories which governs the domain. Furthermore, it is rich with axiomatic theories whose focus on clarifying the intended meanings of terms. It is designed to not only cater the needs of specific community but also provide terminological structure that can share among different communities.

In recent years, the combination of the exponential growth of Web and the explosive demand for getting better information has stimulated the interest in the question answering (QA) (Magdy and Shaheen, 2012). The main principle of the QA is to provide inexperienced users with a flexible access to information and allowing them for writing a query in natural language and obtaining not the disoriented documents which include the answer, conversely the concise answer itself. Thus, the multi-user queries have created wider horizons of the challenges based on the Arabic ontological graph, as the technique that used to answer the questions is entirely different from the normal user queries technique.

According to the foregoing, there are few of the *ontological search engines* that support *Arabic language* and *QA*. The fundamental reasons could be traced back to the *natural language processing* (NLP) (Giunchiglia *et al.*, 2008) and the challenges to address the syntactic search and produce synonym meaning of words. Another point is the particularities like short vowels, absence of capital letters, grammar and the morphological complexity such as the diacritics. Thus, in this paper, we present an early version of *an Arabic semantic search* based on a specific ontological domain and hold the QA which is called CASOnto. Even though, CASOnto system supports Arabic language and English as well, we put our attention to orient the Arabic search which we hope to release as soon as possible to enrich the Arabic content on the World Wide Web. It gives a generic picture of the different components of the system and also draws results of the already implemented parts.

The rest of this paper is structured as follows. Section 2 introduces the research efforts to develop some of the ontological search engines, namely, *Wolfram Alpha, Kngine and Google*. The next section discusses the difficulties of Arabic language and ontology. Section 4 highlights the QA in the ontological graph. Sections 5 and 6 represent our proposed ontological Arabic search engine in detail and with some experimental evaluations with other popular engines. Finally, Section 7 concludes the paper and gives some suggestions to improve the ontological engine in the future.

2. Related works

Ontology is considered as a portal to make the engines more intelligent and powerful. It is a respectful mission for the current generation of the Web which is known as Web 3.0 and the future mission for Web 4.0. Ontology is powerful and has correct and reliable data that store in their repositories that are called the ontological

graphs. It enables the user to get and retrieve a direct answer without any complexities.

There are several ontological graphs developed according to the developers' interest; some of them serve one domain, while others develop to involve multiple domains such as the electronic government. Our purpose focuses on developing Arabic and English CASOnto, which stands for Ibri College of Applied Sciences Engine. It is a domain specific that called a reference ontology. It is focused on the college information such as academic departments, academic staffs, students, where they live and so on. Developers already had created some reference ontologies that focus on academic community, for instance *HERO ontology* (Ghomari, 2013), *Univ-Bench ontology* (Ghomari, 2013), *university ontology* (Ghomari, 2013; LG, 2013) and *AIISO ontology* (LG, 2013; Mesaric and Dukic, 2007). Currently, there are some engines that are based on the concept of semantic such as *Kngine* (Ramachandran and Sujatha, 2011), *Wolfram Alpha* (AlphaTeam, 2013) and the most popular engine nowadays *Google*.

Kngine (Ramachandran and Sujatha, 2011) is the first multi-language QA engine which supports four languages along with English and Arabic. Kngine stands for Knowledge Engine that is Web 3.0 Knowledge Engine. It is designed to provide customized and exact meaningful search results, for instance semantic information about the keywords, user's queries, list things and finding out the relations between the keywords. The exciting characteristics of this search engine gives precise results which link different kinds of related information together to present them to the user such as *movies*, *photos*, *prices* and the *users reviews*.

Wolfram Alpha (AlphaTeam, 2013) is a computational knowledge engine or answer engine which was developed by Wolfram Research. It is an online website that answers factual queries directly by computing the answer from externally sourced "curated data" or structured data, rather than providing a list of documents or Web pages.

There are several techniques that used in semantic engines such as artificial intelligence, NLP (Guo and Ren, 2009) and machine learning. As shown in Table I, Kngine uses the efficiency of *knowledge-based approach* and the power of the *statistical approach* (KingineTeam, 2013), while Google used its own search technology which is called as *Hummingbird algorithm* (Danny, 2013). That means "precise and fast" data or query's answer are the powerful features for any search engine. On the other hand, all these engines have their own mobile application that facilitates them to be more popular and portable for the customers throughout the world. Furthermore, they have an advanced feature that is known as "voice recognition" which enables the operating system to convert spoken words into written text. Moreover, Table I indicates that most of the search engines support English language, while there are few engines that support Arabic language such as Google and Kngine; however, these engines have a wide domain that do not cover academic community. In addition, there are some weaknesses such as giving incorrect outputs, ignoring Arabic diacritics and giving results in English, while the searching process is done in Arabic. Therefore, according to the aim of this paper, our proposed CASOnto search engine tries to cover these issues.

Search engine	Kngine	Google	Wolfram Alpha
Specialty	Knowledge engine	Search engine	Computational knowledge engine
Repository	Wikipedia and other sites	Wikipedia	Curated data of other sites
Search approaches	Knowledge-based approach and the statistical approach	Hummingbird approach	Its own computational approaches
Results	Direct answer or link to web pages	Direct answer	Direct computational answer
Voice recognition	Yes	Yes	Yes
Portability	Yes	Yes	Yes
Language support	Multi-language (supports Arabic)	Multi-language (support Arabic)	Multi-language (doesn't support Arabic)

Table I. Ontological search engines

3. Arabic language and ontologyical engines

3.1 Importance of Arabic language

Arabic language is considered as integral to the vast majority of the population of the Middle East and the rituals of Muslims, whereas over 20 countries, there are over 300 million native speakers of this language because it is regarded as their mother tongue and the religious language of all Muslims of a variety of ethnicities throughout the world. However, there are minority groups of native speakers all over the world as well. Furthermore, it is an official language of the United Nations, the Arab League, the Organization of Islamic Conference and the African Union (Black *et al.*, 2006; Wekipedia, 2014; Majdi *et al.*, 2011).

Arabic language is a semitic language that has 28 alphabets. It outperforms the English language, and all the languages in the world where it consists of more than 12 million words without any frequency. Moreover, Arabic is also one of the six official languages throughout the world which is expected to be in the future and with expectation of extinction of some familiar languages currently (Saleh and Al-Khalifa, 2009).

3.2 Arabic ontological engines

The ontological engines that support Arabic language are so significant to enrich Arabic content in the Web repositories and to get the precise and concise piece of information during the searching process without suffering to many browsers. However, the research communities are still in its infancy of supporting Arabic ontological engines.

The Arabic language has a collection of specialties that may obstruct the development of semantic Web engines. The complexity in Arabic can be traced back to its complex morphological, grammatical and semantic aspects, as it is a highly inflectional and derivational language. Because of these reasons, there are few ontological search engines available in the market, and the current NLP tools cannot directly accommodate the desires of the Arabic language. Therefore, our CASOnto tries to fulfill the Arab nations based on the current approaches of developing ontological engines.

4. Question answering and Arabic ontological engines

4.1 Importance of question answering

QA is one of the important computer science disciplines within the fields of information retrieval and NLP. It is concerned with building systems that automatically answer questions posed by humans in a natural language. Thus, QA is an adequate technique that helps a wide range of users who are looking for a concise and precise answer to their questions.

QA is used in a completely different approach to evaluate the user's query and retrieve only the most specific and accurate answer from the ontological graph. By typing the question in a natural language query, some analytical methods are followed such as determining the part of speech, the synonyms or antonyms from the WordNet and others without any user intervention.

4.2 Question answering in the Arabic ontological engines

QA in the semantic search engines is not available for most popular world's languages, for instance the Arabic language. Most of the accessible systems are lacking maturity and high efficiency due to the complexity of the Arabic language morphological and grammar rules such as stop words, short vowels, absence of capital letters, diacritics, complex morphology and others.

This paper represents the CASOnto which tries to develop a full collection of Arabic system to cater the user's needs; thus, it supports the QA. In reality, there are two main significant types of questions which are the factorial questions and non-factorial questions. The factorial questions focus on answering a particular category of facts-seeking questions. Simple interrogative questions await an answer related to a named entity. Examples of such questions are "Who is the dean of Ibri CAS?" or "How many student in the college?"; some other question are presented in Section 5.6. Whereas non- factorial questions gives more narrative details about the question that focuses on handling Why and How questions. Another significant aim of this paper is to focus on the factorial questions which classifies between 4W + 1H (that includes Who is? Where? When? What is? and How many?). The type of answer and other details will be discussed in Section 5.6.

5. The proposed engine: CASOnto

Our semantic search engine (CASOnto) was designed for College of Applied Sciences (CAS), Sultanate of Oman. This system is based on two types of data sets which are the resource description framework (RDF) data set for the keyword search and the ontological graph for the semantic search. Moreover, it supports both languages Arabic and English. This paper demonstrates the design of the ontological graph more because we already mentioned the RDF on other paper (Almuqrishi *et al.*, 2015). It is also distinguished with the feature of answering the Arabic factorial questions. Furthermore, the semantic search engines are built based on a variety of structures; however, most of them follow the same fundamental phases which are *designing*, *inference*, *storing*, *indexing*, *searching*, *query processing* and *the user's friendly interface* as it is illustrated in Figure 1. Next sections will describe each phase in more details.

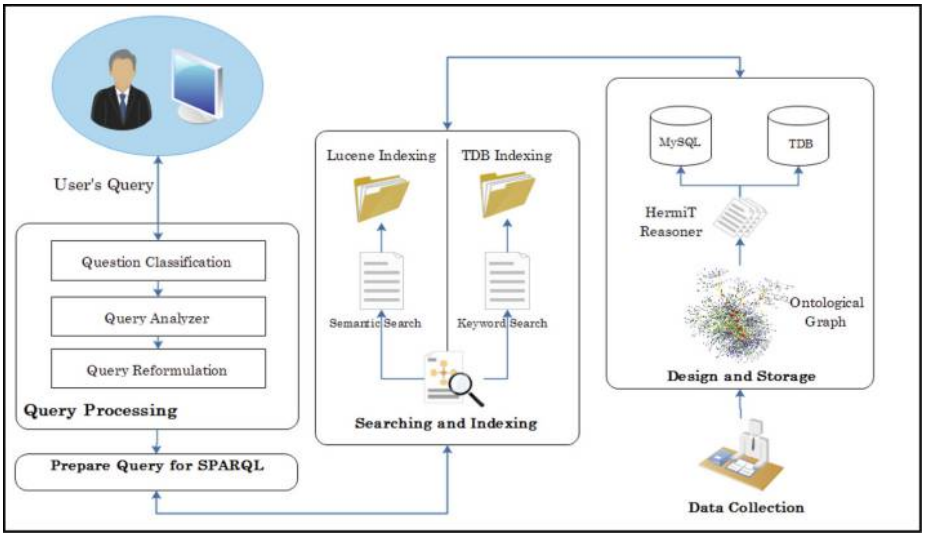

Figure 1.
CASOnto architecture

5.1 CASOnto design and storage

Our CASOnto is designed based on a specific domain of the IBRI CAS to highlight the academic and nonacademic departments. It uses the TDB and MySQL to store the ontological inferred data. The next gives more details about system's design, inference and storage.

5.1.1 CASOnto design. The Arabic ontological graph of CASOnto is shown in Figure 2. It includes three major concepts thing: *classes*, *properties* and *relationships*. Super classes and subclasses have been defined in protégé, and each new class is sub-class from the general class that is called thing. Our CASOnto have three main classes for English and Arabic (person, organization and location) (الشخص-الموقع-المؤسسة). It also defines some relationships among different classes. Some classes are equivalent to other classes. For example, the class (Dean of College) (عميد_الكلية) is equivalent to (Head of College) (رئيس_الكلية). On the other hand, individuals, who are known as instances, are considered as a member of the class. CASOnto instances reach to more than 1,000 individuals.

The last significant concept is the relationships. There are different types of relations such as the relationship between classes or among the classes and individuals. Besides, it is defined as the domain and range for each property. *Domain* means the start edge of the relation, while *range* means the end edge of the relation. For example, it defines the relation "*headOf*"."رئيس" between the class "*dean*" "عميد" and "*college*" "الكلية". It is an *inverse* relation that has a *domain dean* and the *range college*. Finally, after the *classes*, *properties* and *relationships* are created, we need to interpret some things that are not understood by ontology itself. In the Semantic Web, the inference is used to discover new relationships between the data that modeled as a set of defined relationships between the resources. It works as automatic procedures that deriving additional information by generating new relationships based on the ontology data set. Furthermore, it plays an important role to reduce self-join issues among the triples. There are variety of reasons, which can plugin inside the ontology environment such as Pallet (PelletTeam, 2013), FaCt++ (FaCt++Team, 2013), HerMiT (HerMitTeam, 2013), etc. [...]. Our CASOnto uses HerMiT that is an open source which plugins in protégé 4.3. It easily gives inferred *SubClasses*, *equivalentClasses*,

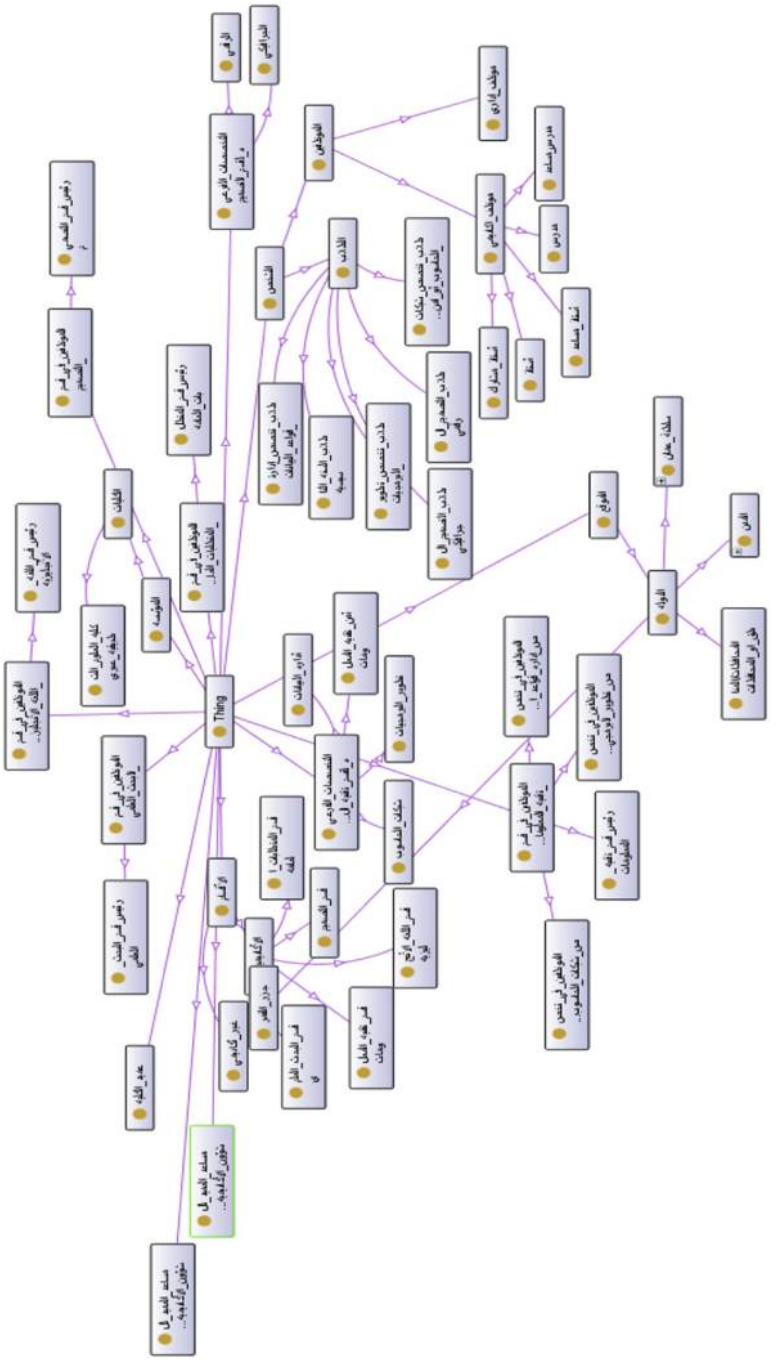

Figure 2.
Arabic CASOnto

SubObjectProperty, *equivalentObjectProperty*, *SubDataProperty*, *equivalentDataProperty*, *inverseObjectProperty*, *Class assertion (Individual)* and *Properties assertion (value)*(W3C wiki, 2010). For instance, the class "department" "قسم" has equivalent class, which is "شعبة" "*section*". Moreover, two asserted individuals "مساعد العميد للشؤون الأكاديمية والبحث العلمي" "Assistant Dean for Academic Affairs & Scientific Research" and "مساعد العميد للشؤون الأكاديمية المساندة" "Assistant Dean for Academic Support Affairs". In addition, the department individual is inferred all data and object property values.

5.1.2 CASOnto storage. Our search engine uses two directions the *Relational Database* and *Triple-store* as it is illustrated in Figure 2. Triple-store means the database management systems for data modeled using RDF. It is unlike the relational database management systems (RDBMS), which store data in relations (or tables). Moreover, the RDBMS are queried using SQL, while the triple-store stores RDF triples and are queried using SPARQL. A key feature of many triple-stores is the capability to do inference. It is essential to note that a DBMS typically presents the capacity to deal with concurrency, security, logging, recovery and updates, in addition to loading and storing data. In CASOnto, we decide to use the Jena TDB because it is a component of Jena for RDF storage and query. It also supports the full range of Jena APIs. In addition, TDB can be used as a high performance of RDF store on a single machine. It also includes automatic protection against multi-JVM usage, which prevents this under most circumstances. On the other hand, the MySQL is used as RDBMS for the keyword searching purpose.

5.2 CASOnto indexing and searching processes

Indexing is a high-level concept among the developers of the search engines, to retrieve the data from the ontology data set faster and efficient. In our search engine, we use two ways for indexing, which store in the Jena TDB and the RDBMS MySQL as it is shown in Figure 2. It uses the *TDB indexing* which is built on the Fuseki for Jena TDB data set. Many of the persistent data sets in the TDB triple-store use a custom implementation of threaded B+ Trees. The threaded nature is referred to the meaning of the long scans of indexes proceeds without needing to traverse the branches of the tree. Lucene is also used for indexing the MySQL database. Lucene consists of a chain of logical steps after gain access to the original content. The steps are that acquire the content, build document, build document, build document.

The searching process in the CASOnto is implemented via two types of searching which are the *Keyword Searching* and *Semantic Searching*. First, the *Keyword Searching* is done by the support of Apache Lucene, which provides with the access to the Lucene indexes. This type of searching gets the matched keywords as a full-text query without understanding the concept behind it. Second, semantic searching is known by entity search that is supported by Apache Jena Fuseki. It provides a SPARQL server that can use the Jena TDB for persistent storage. In addition, it provides with the SPARQL protocols for query, update and rest update over the HTTP. Moreover, the SPARQL query offers the searching over the triple-store and retrieves the needed results.

5.3 Query processing and user's interface

Query processing is a very important component in search engines that helps to improve the user's search results and makes the system efficient and scalable. The architecture of the

question processing module is shown in Figure 3. It is classified in three main components which are *Query Analyzer*, *Question Classification* and *Query Reformulation*.

5.3.1 Query analyzer. The natural language query given by the user is analyzed using various types of the NLP techniques. CASOnto handles this issue via *Spelling Correction, Query Completion, Stop Words Removal and Diacritics Removal.* It offers a way to check the syntax of the user query and raises an error if the query needs a preprocessing to enhance the search results. It may transform the representation of the query before searching by suggesting different query from the original one. The misspelled query suggestions are generated automatically based on the Lucene Spell-Checker that relies on CASOnto data set. If the user enters wrong word "Awny Sayyyed" (" عونيي سيييد "), for instance, associated and valid terms like "Awny Sayed" ("عونيي سيد ") is suggested (Figure 4). This function saves the user's effort and time, and in addition provides with additional database information from the same domain. There are some Arabic *stop words* that are unnecessary in the searching, thus based on the Khoja list of the stop words, the java application discarded them. There are several reasons that stop words should be removed from the question or query. First, they make the question look heavier for analysis. Moreover, they are not required for the analysis process to be completed successfully, and its removal do not have any effects on the result. On the other hand, although *diacritics* are quite crucial in the *Classical Arabic Language*, the dialects and slang rarely use them. This version of CASEOnto handles diacritics by removing them from both the user query as well as the ontological database.

5.3.2 Question classification. There are several type of natural language questions. CASOnto handles the factorial questions and has classified them into various sets

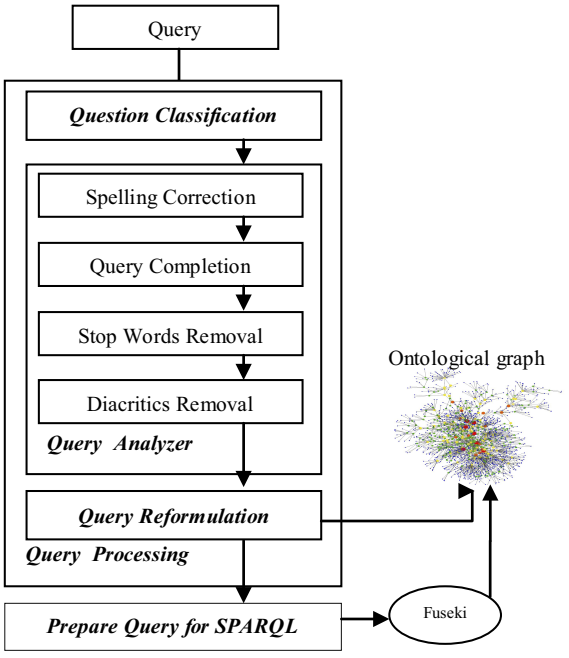

Figure 3.
CASOnto query processing

Figure 4.
CASOnto
auto-complete

(4W + 1H) for extracting more precise and accurate answers. Each question has an identify question keywords as shown in Table II. For instance, "Who is the dean of College?" "من هو عميد الكلية ؟", the identify question "Who is" "من هو" refers to a person "شخص", while the rest of keywords removed as stop words or pass on the SPARQL query as a concept (College, الكلية) or relations among them (dean, عميد).

5.3.3 Query reformulation. The user queries may be reformulated by adding the ontological information in the data set. This technique will give the user different query, but it has the same meaning. In this way, the visualization of the user will expand because the ontology has defined most of synonyms and relationships among the concepts. For example, "Who is the dean of College?" "من هو عميد الكلية ؟", it has the word dean which is a synonym to the ontology head "Who is the head of College?" "من هو رئيس الكلية" as shown in Table III.

5.3.4 CASOnto interface. Our CASOnto system provides with a usable interface that enables users to interact with the engine easily. Thus, a powerful system with a poorly designed user interface has little value that could put the system in the trap. CASOnto offers three parts of searching which are keyword searching, SPARQL expert and CAS queries. Each one of them provides with a guide that helps the user to search probably. Whereas, the CAS queries include a set of predefined queries based on our Arabic and English ontology. The next one is the SPARQL expert, which requires an expert of writing SPARQL query because it forces the user to write a manual query. The last part is the keyword searching that retrieves the results based on the full-text matching of the query.

Table II.
Types of questions

Question form		Expected answer	
Who is. . .?	من هو. . .؟	Person	شخص
Where. . .?	أين. . .؟	Place	مكان
When. . .?	متى. . .؟	Time	زمان
How many. . .?	كم عدد. . .؟	Number	عدد
What is. . .?	ما هو. . .؟ ، ما هي. . .؟	Thing	شئ

Original queries	Reformulated queries	SPARQL queries
من هو عميد الكلية؟	من هو عميد الكلية؟ من هو رئيس الكلية؟	PREFIX rdf: www.w3.org/1999/02/22-rdf-syntax-ns# PREFIX rdfs: www.w3.org/2000/01/rdf-schema# PREFIX xsd: www.w3.org/2001/XMLSchema# PREFIX owl: www.w3.org/2002/07/owl# PREFIX انتولوجي: www.ibri.cas.edu.om/انتولوجي# SELECT ?y WHERE { ?x rdf:type انتولوجي:رئيس_الكلية. ?x انتولوجي:الأسم ?y }

من هم رؤساء الأقسام الأكاديمية؟	من هم رؤساء الأقسام الأكاديمية؟
	من هم المسؤولين عن الأقسام الأكاديمية؟
		PREFIX rdf: www.w3.org/1999/02/22-rdf-syntax-ns#
		PREFIX rdfs: www.w3.org/2000/01/rdf-schema#
		PREFIX xsd: www.w3.org/2001/XMLSchema#
		PREFIX owl: www.w3.org/2002/07/owl#
		PREFIX انتولوجي: www.ibri.cas.edu.om/انتولوجي#
		SELECT DISTINCT ?y WHERE {
		{?x rdf:type انتولوجي:رئيس_قسم_تقنيه_المعلومات.
		?x انتولوجي:الأسم ?y} UNION{
		?x rdf:type انتولوجي:رئيس_قسم_المتطلبات_العامه.
		?x انتولوجي:الأسم ?y} UNION{
		?x rdf:type انتولوجي:رئيس_قسم_اللغه_الأنجليزيه.
		?x انتولوجي:الأسم ?y} UNION{
		?x rdf:type انتولوجي:رئيس_قسم_التصميم.
		?x انتولوجي:الأسم ?y} UNION{
		?x rdf:type انتولوجي:رئيس_قسم_البحث_العلمي.
		?x انتولوجي:الأسم ?y} }

أسماء الموظفين الأكاديميين الذين حققوا مؤهل الدكتوراه؟	أسماء الموظفين الأكاديميين الذين حققوا مؤهل الدكتوراه؟ سماء أعضاء هيئة التدريس الذين حققوا مؤهل الدكتوراه؟	PREFIX rdf: www.w3.org/1999/02/22-rdf-syntax-ns# PREFIX rdfs: www.w3.org/2000/01/rdf-schema# PREFIX xsd: www.w3.org/2001/XMLSchema# PREFIX owl: www.w3.org/2002/07/owl# PREFIX انتولوجي: www.ibri.cas.edu.om/انتولوجي# SELECT ?y WHERE { ?x rdf:type انتولوجي:موظف_اكاديمي. ?x انتولوجي:المؤهلات "الدكتوراه". ?x انتولوجي:الأسم ?y }

(*continued*)

Table III.
SPARQL test queries

Table III.

Original queries	Reformulated queries	SPARQL queries
من هم طلاب تخصص تطوير البرمجيات؟	من هم طلاب تخصص تطوير البرمجيان من هم تلاميذ تخصص تطوير البرمجيات؟ من هم متعلمي تخصص تطوير البرمجيات؟ من هم دارسي تخصص تطوير البرمجيات؟	PREFIX rdf: www.w3.org/1999/02/22-rdf-syntax-ns# PREFIX rdfs: www.w3.org/2000/01/rdf-schema# PREFIX xsd: www.w3.org/2001/XMLSchema# PREFIX owl: www.w3.org/2002/07/owl# PREFIX انتولوجي: www.ibri.cas.edu.om/انتولوجي# SELECT DISTINCT ?y WHERE { ?x rdf:type انتولوجي:طلاب_تخصص_تطوير_البرمجيات. ?x انتولوجي:الأسم ?y }

6. Experimental results

Our CASOnto is based on two types of searching which are the classical search (keyword-based search) and the semantic search (entity-based search). The purpose of the classical search is to measure the matching of the keywords with the RDF data set as well as the ontological graph. It is arranged based on the high score of matching. The aim of semantic search is to get the exact answer from the ontological graph. In addition, it is built to understand the context of the searching text and retrieve the coherent answers without going on a maze as the classical search.

We conduct two experiments to measure the performance of our proposed search engine. First, compare the (keyword-based search and entity-based search) of RDF and the ontology based on simple and complex queries. Second, compare our proposed engine CASOnto with other engines such as Wolfram Alpha, Kngine and Google. As mentioned above, the data set used is an ontological graph that holds information about departments, staff, faculty and students for College of Applied Sciences, Ibri, Oman. As shown in Table IV, the CAS-Ontology data set contains 31,279, which is classified into 2,159 subjects, 132 predicates and 5,575 objects, whereas the English CAS-Ontology data set contains 32,322, which is categorized into 3,035 subjects, 150 predicates and 6,507 objects.

6.1 Evaluation metrics

The analysis evaluation of search engine is measure based on different metrics to get a quality model that is presented based on ISO 9,126 standards for system quality. In this section, it distinguishes between three varieties of evaluation measurements which called *Recall*, *Precision* and *Accuracy*.

- *Recall*: It is referred to the fraction of the documents that are relevant to the query which are successfully retrieved (i.e. sum of all true positives and false negatives). It is known as a lexical recall or a correct recall (Rc):

$$\textbf{Recall} = \textbf{Number of retrieved relevant/Number of possible relevant}$$

- *Precision*: This measure (Pc) is defined as the fraction of the documents retrieved that are relevant to the user's information need. It is called lexical precision or a correct precision:

$$\textbf{Precision} = \textbf{Number of total relevant/Number of total retrieved}$$

- *Accuracy*: This metric gives a good overall view of the competency of a search engine and how accurate it is. It is computed by dividing the number of correct outputs (i.e. the sum of true positives and true negatives) by the total number of queries.

Data set	Triples	Subject	Predicate	Object
Arabic CAS_ Ontology	31,279	2,159	132	5,575
English CAS_ Ontology	32,322	3,035	150	6,507

Table IV.
Data set descriptions

6.2 Resource description framework and ontology evaluation

RDF and ontology are the main two backbones of CASOnto system. In our engine, RDF is designed to be a keyword-based search, while the ontology is considered as a classical search and a semantic search. In our experimental of RDF and ontology, we classify approximately 60 queries (which exists in the paper's Appendix) into two categories, which are simple and complex queries based on the number of self-joins. As we have seen in Table V and Figure 6, the comparison of simple and complex queries is done under two types of searching: keyword-based search and entity-based search. Clearly, in the case of comparison, we depend on first answer if it is true, and ignoring the rest of the retrieved answers. Systematically, it seems that the semantic search is better than the classical search in both simple and complex queries. Clearly, the accuracy equals to 100 per cent in both types of queries. While the classical search is better with the simple queries as illustrated in Figure 5, the accuracy is 16.6 per cent; however, it equals 0 per cent in complex. We retrieved only five relevant results because this searching is based on the full text that means all the keywords should be exist in the same triple to get the result. The total relevant of queries is efficient with the semantic search; therefore, our next experiment will take this to compare our CASOnto with other semantic search engines.

6.3 Semantic search engines comparison

> The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the RDF, which integrates a variety of applications using XML for syntax and URIs for naming.

W3C Semantic Web – The Semantic Web is a charter that allows publishing, sharing and reusing data and knowledge on the Web and across applications, enterprises and community boundaries.

	Simple query				Complex query		
	Keyword-based search		Entity-based search		Keyword-based search		Entity-based search
Query no.	RDF	Ontology	Ontology	Query no.	RDF	Ontology	Ontology
Q1	–	–	✓	*Q1*	–	–	✓
…	…	…	…	…	…	…	…
Q12	✓	✓	✓	*Q12*	–	–	✓
Q13	✓	✓	✓	*Q13*	–	–	✓
Q14	✓	✓	✓	*Q14*	–	–	✓
Q15	✓	✓	✓	*Q15*	–	–	✓
Q16	✓	✓	✓	*Q16*	–	–	✓
…	…	…	…	…	…	…	…
Q30	–	–	✓	*Q30*	–	–	✓
Total relevant	5	5	30	Total relevant	0	0	30
Total retrieved	30	30	30	Total retrieved	30	30	30
% Precision	16.6	16.6	100	% Precision	0	0	100
% Recall	100	100	100	% Recall	0	0	100
% Accuracy	16.6	16.6	100	% Accuracy	0	0	100

Table V.
Performance of CASOnto engine

Our experiment of semantic search engines is compared with four types of famous semantic engines, which are CASOnto, Wolfram Alpha, Kngine and Google. We submitted 15 different queries against the tested engines that exist in the paper's Appendix. As shown in Table VI, our engine retrieved 12 of 15 answers (that is defined by the symbol √), while the rest of engines have irrelevant answers (which is defined by the symbol -) as well as no responses (that is defined by the symbol 0). The ratios of precision are comparable between the rests of the engines where Wolfram Alpha, Kngine and Google have 53.3, 46.47 and 55.56, respectively, as it is illustrated in Figure 6. In addition, the accuracy of our engine is also high compared to other engine; it has 100 per cent, while Wolfram Alpha, Kngine and Google have 53.3, 46.47 and 33.3, respectively, as we shown in Figure 6 as well as Table V. Consequently, it seems that our engine retrieved better and efficient results than other engines. Thus, it is built according to the ontological domain-specific, highly scalable performance and handles the complex queries well by understanding the context behind the query.

7. Conclusion and future work

In conclusion, though new improved keyword-based technologies for searching the WWW are evolving constantly, the growth rate of these improvements is likely to be slight. Problems of imprecise and irrelevant results will continue to hinder Web searchers, especially with the continued expansion of the Web. Search engines based on a new concept as the semantic Web technology are effectively able to handle the above-mentioned problems. A domain-specific ontology based on semantic search engine as ours CASOnto is advantageous in several ways. First, our approach has been able to successfully eliminate the problem of irrelevant results, which is one of the main problems encountered by the users of a regular search engine. By using the mapping technique between instances and classes, the search engine effectively fetches the exact information. Second, by producing exact information as the result, the search engine eliminates the need to go through numerous results as in case of a regular search engine. Third, the number of searches and time required by the semantic search engine is less than that of a regular search engine. In the future work, we shall extend the RDF and the ontological graph to contain all information about ministry of higher education (MoHE). In addition, we try to handle the nonfactorial questions and to demonstrate a good indexing mechanism, which is suitable to deal with the large data set. With

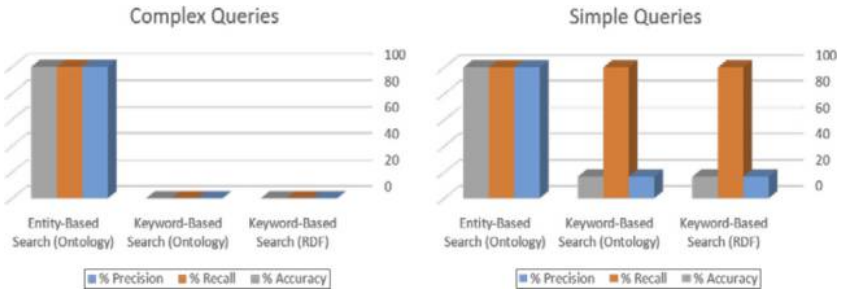

Figure 5. Classical and semantic search of CASOnto

Query no.	CASOnto	Google	Kngine	Wolfram Alpha
Q1	✓	✓	✓	✓
Q2	✓	0	✓	✓
Q3	✓	✓	✓	✓
Q4	✓	–	–	✓
Q5	✓	✓	–	✓
Q6	✓	0	✓	✓
Q7	✓	–	–	–
Q8	✓	–	✓	–
Q9	✓	0	–	–
Q10	✓	0	–	–
Q11	✓	–	–	✓
Q12	✓	0	–	–
Q13	X	✓	✓	–
Q14	X	✓	✓	✓
Q15	X	0	–	–
Retrieved relevant	12	5	7	8
Retrieved irrelevant	0	4	8	7
Total retrieved	12	9	15	15
Not retrieved relevant	0	6	0	0
Not retrieved irrelevant	3	0	0	0
% Precision	100	55.56	46.47	53.3
% Recall	100	45.45	100	100
% Accuracy	100	33.3	46.47	53.3

Table VI.
Performance of different ontological search engines

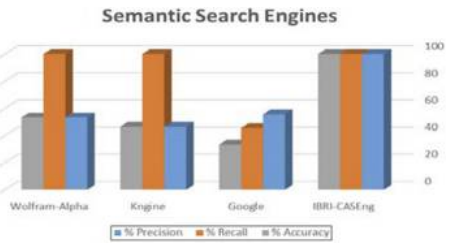

Figure 6.
Performance of semantic search engines

the consideration of the time, store and information retrieval (IR), which are important to retrieve data, fast, scalable and efficient.

References

Agrawal, S., Chaudhuri, S. and Das, G. (2002), "DBXplorer: a system for keyword-based search over relational databases", *ICDE Conference, San Jose, CA*.

Almuqrishi, A., Sayed, A. and Kayed, M. (2015), "Caseng: arabic semantic search engine", *Journal of Theoretical and Applied Information Technology*, Vol. 75 No. 2.

AlphaTeam (2013), "Alpha search engine", available at: www.wolframalpha.com/

Antoniou, G. and van Harmelen, F. (2008), *A Semantic Web Primer*, The MIT Press, Cambridge, MA, London.

Black, W., Rodriguez, S.E.H., Alkhalifa, M., Vossen, P. and Pease, P. (2006), "Introducing the Arabic WordNet project", *Proceedings of the 3rd International Global WordNet Conference, Jeju Island, South Korea, 22-26 January*, pp. 295-299.

Danny, S. (2013), "FAQ: all about the new Google 'Hummingbird' Algorithm | Why is it called Hummingbird?", available at: http://searchengineland.com/google-hummingbird-172816 (accessed 26 March 2015).

FaCt++Team (2013), "Logical reasoner Fact++", available at: http://owl.man.ac.uk/factplusplus/ (accessed 1 March 2013).

Ghomari, L.Z.G. (2013), "Process of building reference ontology for higher education", *Proceedings of the World Congress on Engineering, London*.

Giunchiglia, F., Kharkevich, U. and Zaihrayeu, I. (2008), "Concept search: semantics enabled syntactic search", *Semantic Search 2008 Workshop (SemSearch2008) at the 5th European Semantic Web Conference, ESWC, Tenerife*.

Gruber, T. (1993), "A translation approach to portable ontology specifications", *Knowledge Acquisition*, Vol. 5, pp. 199-220.

Guo and Ren (2009), "Towards the relationship between semantic web and NLP", available at: www.kngine.com/Technology.html

HerMitTeam (2013), "Logical reasoner Hermit", available at: http://hermit-reasoner.com/ (accessed 1 March 2013).

Karin, B., Casanova, M.A. and Truszkowski, W. (2010), *Semantic Web: Concepts, Technologies and Applications*, Springer, London.

KingineTeam (2013), *How Kngine Works?*, available at: www.kngine.com/Technology.html (accessed 1 March 2013).

LG. (2013), "Higher education reference ontology", available at: http://datahub.io/dataset/higher_education_reference_ontology (accessed 24 May 2015)

Magdy, A. and Shaheen, M. (2012), "A survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends", *The 13th International Arab Conference on Information Technology ACIT'2012, Balamand, North, 10 December 2013*.

Majdi, B., Abdul Rahim, A. and Roslan, I. (2011), "An Arabic language framework for semantic web", *2011 International Conference on Semantic Technology and Information Retrieval*, Putrajaya, Malaysia, 28-29 June 2011.

Mesaric, J. and Dukic, B. (2007), "An approach to creating domain ontologies for higher education in economics", *Proceedings of 29th International Conference on Information Technology Interfaces, Cavtat, Croatia*, pp. 75-80.

Nicola, G. and Pierdaniele, G. (1995), "Ontologies and knowledge-bases towards a terminalogical clarication", Mars, N.J. (Ed.), *Towards Very Large Knowledge Bases-Knowledge Building and Knowledge Sharing*, IOS Press, Amsterdam, pp. 25-32.

Ontogenesis (2010), "Referance and application ontologies", available at: http://ontogenesis.knowledgblog.org/295 (accessed 19 March 2015).

PelletTeam (2013), "Logical reasoner Pellet", available at: http://clarkparsia.com/pellet/ (accessed 1 March 2013).

Ramachandran, A. and Sujatha, R. (2011), "Semantic search engine: a survey", *IJCTA*, Vol. 2 No. 6.

Saleh, L.M.B. and Al-Khalifa, H.S. (2009), "AraTation: an Arabic semantic annotation tool", *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, ACM, Kuala Lumpur*, pp. 447-451.

Samhaa, R.E., Maryam, H. and Ahmed, R. (2007), "Ontology based annotation of text segments", *Proceedings of the 2007 ACM symposium on Applied Computing*, Seoul, 11-15 March 2007.

W3C wiki (2010), "Semantic web tools", available at: www.w3.org/2001/sw/wiki/tool (accessed 19 March 2015).

Wekipedia (2014), "Triplestore", available at: http://en.m.wikipedia.org/wiki/Triple_store (accessed 20 April 2015).

Appendix

(1) Simple test queries:

- Dean of the college
- Head of the college
- Academic departments
- Head of academic departments
- Heads of academic departments
- Head of Scientific Research Department
- Academic majors
- Head of Software Development
- Heads of academic majors
- Assistant professors name
- lecturers in the college
- Awny sayed Sayed email
- Mohamed Kayed qualification
- Munesh nationality
- Head of it department email
- Head of English department nationality
- Governorates in Oman
- Regions in Oman
- Cities in Oman
- Cities in Sultanate of Oman
- Cities of Al Dhahira
- IT majors
- Information Technology majors
- Design majors
- Foundation students
- Academic staff
- Staff emails
- Design faculty
- Design staff
- Researchers name

(2) Complex test queries:

- Academic staff who have achieved the PhD degree
- Academic staff who have achieved the master's degree
- Students who live in North Batinah and has Digital major
- Male Students who live in Ibri that located in Al Dhahirah
- Batch 2011 students from Al Dhahirah who are male
- Academic staff emails from Design Departement who have achieved the PhD degree
- Male students from Information Security
- Egyptian academic staff who has PhD
- Iraqi academic staff who has master's degree
- Non-Omanis female students from IT
- Non-Omanis female students from Information Technology
- Omani academic staff emails from Desing Departement
- Number of Omanis in the academic departments who have achieved the Phd PhD degree
- Female students of Graphic Design from Muscat
- IT staff who are lecturers and their nationality is Indian
- IT faculty who are lecturers and their nationality is Indian
- Omani students from batch 2012 studies Network
- Network students from ALBuraymi and their gender is male
- Administrator of IT department who is female and has a PhD
- Network staff and Security faculty who are Jordanian and has a PhD
- Number of Digital and Graphic student from Muscat and their nationality Comoros
- Number of IT staff who is male and has a PhD
- Academic staff who is head of major and has a PhD
- Lecturers from English department who has master's degree from UK
- IT major that has more than 2 two PhD staff from Egypt
- Number of IT faculty who are lecturers from data management
- Number of Design faculty who are Assistant Professors and their major is graphic design
- Number of IT students who are from software development and their gender is female
- Academic department that has more than 50 student who lives in Ibri from Al Dhahirah region
- Omani students from batch 2010 who are male and their major is Network

(3) Queries to compare semantic search engines:

- How many Regions in Oman
- How many Governorates in Oman
- How many Cities in Oman
- How many Cities in Sultanate of Oman
- How many Regions in Sultanate of Oman
- How many Governorates in Sultanate of Oman

- Number of Regions in Oman
- Number of Governorates in Oman
- Number of Regions in Sultanate of Oman
- Number of Governorates in Sultanate of Oman
- How many Cities in Muscat
- Number of Cities in Al Dakhiliya
- Who is the Oman President
- What are the Colleges in Oman
- Who is Squ dean

Corresponding author
Awny Sayed can be contacted at: awny.sayed@mu.edu.eg