SINGIDUNUM UNIVERSITY

To the scientific council of the Department of Postgraduate Studies

Based on the decision of the Department of Postgraduate Studies of Singidunum University No. 4-110-1 /2015 at the meeting held on 30.03.2015, we appointed to the Committee for the assessment of the doctoral thesis of candidate Mr. Hesham Naser Elzentani, heading entitled "Arabic XML Documents: Summarizing, Managing, and Securing". Hereby, we submit the following report on evaluation of the Ph.D. thesis:

1. About the Candidate

Ph.D. candidate Hesham Naser Elzentani, was born on 4th of Jun 1978 in Obari, Libya, where he finished primary and secondary school. Mr. Hesham Naser Elzentani has attained the diploma of the Bachelor of science in computer engineering at Engineering Academy, Tajoura, Libya, as well as Master diploma in computing and control engineering at Faculty of technical sciences, University of Novi Sad. Success at Bachelor with the average grade of 72.04%, and in the Master he obtained 60 ECTS credits with the average grade of 9.67.

Mr. Hesham Naser Elzentani is interested in Programming and computer networks. He has advanced skills in object-oriented programmings such as Java and Delphi, as well as C++, C and Databases, such as XML DBMS, SQL DB, MDB, and Paradox.

Date of application of Ph.D. proposal was on 21.01.2015. Ph.D. Candidate is working successfully on his doctorate. He explores the systematization of Arabic XML documents summarization knowledge and securing Arabic XML documents and their presentation to find some mechanisms to satisfy the needs of XML summarization and to protect the summarized document.

2. Thesis's abstract

W3C proposed the XML as standard that used in web applications, transactions, documentations, database management systems and to exchange information between systems over the Internet. XML allows storing different data regardless of how it will be displayed. XML has been used to create, update and query databases. Create and write clear human-readable XML documents as well as machine-readable are easy, so it's easy to create applications that process these XML documents. Generally, all kinds of information can be expressed as XML documents.

In recent years, the growth of Arabic content and numbers of users on the Internet has greatly increased. Arabic is a widely spoken language with more than 375 million speakers and over 155 million, or over forty percent of these Arabic-speaking people use the Internet. This represents nearly five percent of all the Internet users in the world. The number of Arabian speaking Internet users has grown by a factor of sixty in the last fifteen years (2000-2015). This growth in usage has outpaced the growth in information retrieval systems, summarization of Arabic text (such as documents and web pages), query processes and natural language processors.

The Arabic word has different forms of syntax and morphologies with different meanings. Grammatically, documents contain different forms of words including derivations. This causes problems in text processing, document summarization, and information retrieval systems. Furthermore, there is a high level of information loss during the processes of querying, a document summarizing and information retrieval, especially with large documents, as information loss is directly proportional to the size of documents during these processes.

The thesis describes an RAX System designed for ranking Arabic documents in information retrieval processes. The proposed solution basically depends on the similarity of textual content. The model we have designed can be used for documents stored in the different formats and written in the Arabic language. Due the complex lingual semantics of this language, the proposed solution uses a purely statistical approach. The design and implementation are based on existing text processing frameworks and referent Arabic grammar. The main focus of our research has been the evaluation of different similarity measures used for classifying Arabic documents from different domains and different document categories based on query criteria provided by the user.

Further, the thesis studied the security issues of the RAX system. These issues combined between XML security (XML digital signature and XML encryption) and the SOAP message to create a secret environment between an end user and the RAX system model as well as study the security attacks and countermeasures.

Keywords: Text similarity measures, Text classification, Processing Arabic documents, XML Documents, Securing XML Documents.

3. Objective, Hypothesis, and Methodology of the Thesis

The research objective focused on improvements of existing summarizing and securing evaluations approaches to reflect the requirements of XML summarization process as well as querying, ranking and securing. However, the research objectives are:

1) Summarizing Arabic text and/or Arabic XML documents.
2) Ranking the summarized text and/or documents based on queries and similarity measures.
3) Applying security issues on XML documents according to XML digital signature and XML document encryption.
4) Design and build a system model, which will establish these objectives.

General hypothesis is particularly in the domain of XML data representation, information retrieval systems, summarization of Arabic XML documents, XML queries, ranking Arabic XML documents and securing Arabic XML documents.

Objectives and hypotheses of the research led to developing a new methodology that will be able theoretically to evaluate various kinds of Arabic XML documents from different domains and different categories. Theoretical evaluation presents a powerful methodology to analyze the influence of XML structures and its contents over the designed model. This thesis describes a model system, which's designed for ranking Arabic documents stored in the different formats in information retrieval processes and security model to protect user queries and documents. The research results should help engineers, network administrators, database designers and information retrieval systems to deal with Arabic XML documents. Furthermore, methodologies in Bibliography guided this research to achieve the objectives.

4. Bibliography

The bibliography consists of 99 citations, mostly international and recent.

5. Conclusion and Future Work

This thesis proposed the RAX System which designed for ranking Arabic documents based on content similarity. RAX model was applicable to documents stored in different formats and written in the Arabic language. The design and implementation were based on existing text processing frameworks and referent Arabic grammar. The main focus of the research was on evaluating different similarity measures used for classifying Arabic documents from different domains and different categories.

In the preparation stage, the RAX system was used to process Arabic text taking into account the character encoding for the Arabic language (UTF-8, Windows-1256 etc). The preparation stage of the processing of Arabic text was established in 4 steps: extraction of full text from documents; normalization (remove diacritics, remove non-letters and remove punctuation marks); removal of stopwords from the normalized text and stemming (remove prefixes, remove suffixes and finally extract roots or stems words). The well-formed Arabic XML document was created from the stemmed text and loaded into XDBMS which manages end user queries over a collection of XML documents.

In the implementation stage, the RAX system managed XML documents via an XML database management system using XPath and XQuery languages. The Arabic text in queries was processed in 3 steps: normalization, removal of stopwords and stemming (preparation stage). The RAX system uses cosine similarity to measure the similarity metric in n-dimensional space. This is based on the finding that when two vectors are similar in rate and direction from the origin to their end points, they will be close to each other in the vector space, with a small angular separation, and vice versa. The cosine value lies between 1 and -1. Therefore, the cosines of small angles are close to 1, which means high similarity, while the cosines of large angles are close to -1, which means low similarity.

We conclude that the Arabic text was fully represented in the processing of Arabic documents.

Furthermore, the total of the term frequencies of the documents and the weights of queries were equal to the totals of the whole collection. There was a proportional relationship between the number of terms of a query and its result. The RAX system excludes terms which are not matched. System performance could be improved by changing the type of stemmer.

There are two main advantages of the RAX system. Firstly, the query results are more comprehensive and wider when using the roots of words or stems. Secondly, the similarity measures are calculated after the completion of the query process i.e. comparing the collection of terms extracted from the collection of XML Arabic documents and the query terms. So, the ranking is calculated according to this comparison.

The thesis proposed a survey, which's studied the security issues of the RAX system. These issues combined between XML security (XML digital signature and XML encryption) and the SOAP message to create a secret environment between an end user and the RAX system model, as well as study the security attacks and countermeasures.

Following are the verified hypothesis, which's presented in the hypothesis section:

1) Different forms of a word have caused problems in text processing, document summarization, and information retrieval systems.
2) In every summarization process, there was information loss that directly proportional to the size of the document.
3) The well-summarized document contains the whole important information, but with a big document, it's impossible to get well-summarized document without losing important information.
4) The summarized document always has a smaller size than the original. However, the parsers can process it in the small amount of memory.
5) As we can see from the results of the RAX system that the documents are ordered according to the similarity measures, and this ranking is helpful in information retrieval systems.
6) The similarity measures between XML documents and their summarized documents are close.
7) The similarity between a query and its result depending on the terms of query and content of summarized XML document.
8) The security issues will be powerful if the security attacks and their countermeasures are taking into account.
9) The XML digital signature and the encryption do not affect the summarized XML document.

As regards future work, the RAX system could be improved in various ways. We plan to work on making it more efficient. This will mean that the stemmer will need to be improved and enhanced in capabilities and effectiveness to deal with the huge volume of Arabic roots in large data sets (stopword list, compatibility between prefixes and suffixes in stemming process, etc). We also aim to use DTD and XML schema to create XML documents as well as to enhance their summarization. Finally, we plan to upgrade the RAX system to find and replace any query term which has a zero term frequency.

6. Conferences and Papers

The published conferences and papers are illustrated below:

1) Hesham Elzentani, "An Open-Source Based Application for Creating and Verifying Digital Signatures for XML Documents", International Conference on Computing, Communication System and Informatics Management (ICCCSIM), Dubai, UAE, 29–30 July, 2012, International Journal of Information Technology and Computer Science (IJITCS), 4(2): 1–10, 2012.
2) Hesham Elzentani, Mladen Veinović, "Summarization of XML Documents", Konferencije Elektronika, Telekomunikacije, Računarstvo, Automatika I Nuklearna tehnika (ETRAN), Zlatibor, Serbia, pp. VI2.2.1 – 2.2.6, 3–6 June, 2013.
3) Hesham Elzentani, "Managing XML Trees Using XPath, Xquery, Clustering and Tree Tuples over Sedna XML database", 1st Singidunum University International Scientific Conference SINTEZA, Belgrade, Serbia, ISBN: 978-86-7912-539-2, pp. 878–881, 25–26 April, 2014, DOI: 10.15308/SinteZa-2014-878-881.
4) Hesham Elzentani, Mladen Veinović and Goran Šimić, "Managing Semi-Structured Data Using XPath, XML Trees and Tree Tuples over a Wireless Network", Special issue of International Journal of Information Technology and Computer Science (IJITCS), 13(1): 45–55, 2014.

The following papers are submitted for publication:

1) Hesham Elzentani, Mladen Veinović and Goran Šimić, "RAX system to rank Arabic XML documents", Submitted to Kuwait Journal of Science (KJSE, ISSN: 1024-8684) on 21/3/2016.
2) Hesham Elzentani and Mladen Veinović, "Arabic Text: Summarizing and Querying", Submitted to The International Arab Journal of Information Technology (IAJIT, ISSN: 2309-4524) on 04/05/2016.

7. Opinion of the Committee

PhD. dissertation of Mr. Hesham Naser Elzentani was performed according to the previously approved application. The candidate has achieved the goal of research with the application of relevant scientific and professional knowledge in the subject area. The dissertation is an autonomous research work of the candidate. The candidate has shown his ability for an original approach to the analysis of observed phenomena that are the subject of the dissertation. The structure of the doctoral dissertation is correct which enabled the candidate to accomplish the research purposes, to establish a quality basic hypotheses which were starting point to work and to provide relevant answers to questions.

The doctoral dissertation has presented RAX system to summarize, manage and rank Arabic documents based on XML syntax over XML database management system, as well as it has proposed the security issues to secure the RAX system.

Taking into consideration all these facts, we propose to the Council of Department for Postgraduate Studies and International Cooperation of the Singidunum University to accept doctoral dissertation of Mr. Hesham Naser Elzentani, entitled "Arabic XML Documents: Summarizing, Managing, and Securing" and approve its public defense.

Belgrade, 09 May 2015

Members of the Committee:

Prof. Mladen Veinovic, PhD,
University Singidunum

Prof. Aleksandar Jevremovic, PhD,
University Singidunum

Prof. Goran Šimić, PhD,
Military Academy
of the Ministry of Defense