Research Plan (Items 1-8 not to exceed 25 pages)

1. Background (maximum 2 pages)
2. Significance (maximum 1 page)
3. Preliminary data or studies.
4. Research design and methods.
5. Anticipated results and evaluation criteria.
6. Strategy for project continuation
7. Plans for disseminating research results
8. Rebuttal to the peer reviewer’s comments for resubmitted proposals (when applicable).

Relational Arabic Text Mining Framework   
based on Morphological Analysis

|  |  |
| --- | --- |
| Fadi Zaraket  Assitant Professor  Electrical and Computer Engineering Department  American University of Beirut | Rehab M. Duwairi  Associate Professor  Computer Science and Engineering Department  University of Qatar |

## Background (maximum 2 pages)

This proposal aims at enabling relational text mining in Arabic text. In the following we will introduce text mining, then we will discuss related work on Arabic text analysis. Next, we discuss the problems with the current approaches that lightly massage the Arabic text and then apply text mining techniques that work for Latin texts. Finally, we will articulate our approach of using unique features of the Arabic language as well as relational based queries and we will present evidence of the feasibility of this work.

### Text mining

Research in text mining started in the mid 1980s when Swanson realized that slicing and combining seemingly unrelated medical articles led to the discovery of new hypotheses [[1](#JNi06)]. Till present, most of the interest in text mining comes from biological sciences. Lately we saw applications of text mining in the advertisement, political campaigning, and other businesses. Text mining is the process of analyzing a large collection of documents to discover previously unknown information, which is extremely difficult to find otherwise. Text mining involves information retrieval, natural language processing, information extraction, and data mining [[2](#ASr09)]. Information retrieval (IR) tools are similar to a Google search engine where documents of interest are selected based on the relevance to a set of keywords of interest to the user. Natural language processing (NLP) targets the automation of understanding human languages. This task is the oldest and most difficult task in the artificial intelligence domain. The main difficulty behind NLP comes from the ambiguity of natural languages [[3](#Osm08)]. While NLP techniques cannot deliver their final target yet, they can deliver analysis of sentences into nouns, verbs, and adjectives. They can also narrow the meaning of a word or a phrase based on the context to one of its possible meanings. Finally they can parse a sentence into a relation between the discovered entities using abstractions such as verb-name abstractions [[3](#Osm08)].

Information extraction (IE) targets forming data structures and instances of these data structures out of a collection of documents. It usually does this by using the output of IR and NLP and fitting it to templates of interest to the user.

Data mining can be applied at the end on the output of the IE process to answer queries the user might have about the documents [[4](#JHa05)].

### Related work on Arabic text analysis

In the following we review the major works on Arabic text mining in academy and industry.

**Arabic text mining in academy.** We propose to build a native Arabic text analysis tool that takes advantage of unique features in the Arabic language to provide answers to relational queries. Prior work in Arabic text mining addressed the classification and categorization of Arabic text and made little use of the unique features of the Arabic language. Buckwalter [[5](#Bee89),[6](#Tim04)] uses morphological analysis to enable relaxed stemming of Arabic words. His work is used in several open source spell checkers as well as natural language processing frameworks [[7](#Col09)]. The work of Al-Zoghby [[8](#Ham07)] addresses the problem of entities extracted from Arabic text with similar content but slightly different names via using the derivation feature in the Arabic language as a similarity measurement. This work is limited to the names of the fields in the data structures extracted from Arabic texts. El-Halees used statistical analysis to address the problem of classifying Arabic text documents with maximum entropy [[9](#AEL07)]. In [[10](#Abd07)] Mesleh evaluated six different techniques to classify Arabic text using a support vector machine. El Dost [[11](#MEl03)] used root words in the Arabic language to speed up automated and hierarchical indexing. Most of academic research in the field is covered in the almost comprehensive book “Arabic Computational Morphology” [[12](#Sou07)]. The book summarizes recent work and presents strong evidence of shortage of historic research and an emerging interest in research in this area.

**Arabic text mining in industry.** In addition to the academic research, few companies in the industry provide primitive Arabic text analysis tools. Sakhr [[13](#Sak09)] provides a keyword extractor, a categorization tool that groups texts under a limited predefined groups (10 groups including politics, sports, …), and a summary generator tool that selects sentences from the text and displays them on top of the text. Basis technology [[14](#Bas09)] provides Rosette which is a lexical analysis tool that normalizes Arabic text as a pre-processing step for other tools such as indexers and general purpose data mining tool to analyze the Arabic text. Basis also provides REX; an entity extractor.

### Novel methodology and insight

We will differ categorically from the above mentioned work on Arabic text and data mining. Researchers in [[9](#AEL07)] [[8](#Ham07)] [[10](#Abd07)] [[11](#MEl03)] tried to massage and preprocess the Arabic text to make it presentable to known text and data mining techniques that tend to Latin text and Latin based queries. We think that this approach does not tend directly to the Arabic users of the Arabic text mining techniques who write their texts and queries in Arabic and thus think in Arabic syntax and semantics while formulating their text and queries. We hypothesize that Arabic queries use semantic rules and inhibit relations that are similar to the Arabic text under investigation, and thus techniques native to the Arabic language will do better in Arabic text analysis.

We find evidence in the work of the CADIM (قادم) project [[7](#Col09)] at the University of Columbia to support our hypothesis where they observe that Arabic texts used as training data for common NLP techniques proved ineffective and thus they build semi-automatic models to address the problem of automatic speech recognition in Arabic.

Breifly, we will augment current morphological analyzers to accept relaxed rules of the Arabic grammar that deals with word forms (الصرف). We will also augment the current digital dictionaries to resolve primitive entity semantics and targeted qualifications such as clear adjectives. We limit our NLP techniques to this extent and extract semantic graphs from each text document and then use graph analysis and statistical techniques to establish relations between these graphs. We leave the details of this approach to Section “Research Design and Methods”.

# Significance

Arabic text mining will benefit sectors where Arabic text documents are key such as security reports, governmental records, taxation files, health records, as well as the trading floors of stock exchanges. The application of text mining techniques to Arabic text documents will result in great benefits to many research fields such as Arabic literature, Islamic studies, Hadith authentication, and Arabic history and culture in general. We briefly discuss two possible case studies and show the benefits of using Arabic text mining. More detailed description and additional cases can be found in Section “Anticipated Results”.

***Authenticity of literature case:*** We have two sets of texts. The first is several books of historical accounts. A chain of narrators precedes each account. The second is several books of biographies. A biography may include an evaluation of the authenticity of the narrator therein. With Arabic text mining we can automatically exhaustively query *all* accounts related to women and travel, relate these accounts to their narrators, relate the narrators to their biographies and thus extract the authenticity of the accounts. A similar manual query takes ages to answer and is error prone.

### *Security case:* Given a huge set of investigation reports, an officer may throw a list of names of convicted people and check all the reports for possible relations between them in terms of criminal action, locality, or third party names unknown to the officer who happen to know the convicts. The query may return at the click of a button, two convicted people who had no direct relation in the reports, but who are related only through a third person unknown to the investigator. Such a query is very hard to answer manually.

### Direct benefits of Arabic text mining

In the following we list few benefits and usage examples of an Arabic text mining framework.

* Qualitatively enhance research in the cultural and literature fields.
* Allow exhaustive coverage and complex queries of finite data sets such as the hadith literature.
* Find relations and inconsistencies in security related reports that can help save lives.
* Find faults and inconsistencies in claims in the insurance sectors.
* Help investors analyze Arabic news and data from Arab markets prior to making important decisions.

### Indirect benefits of Arabic text mining

The use of Arabic text mining in the above mentioned fields will have a high impact on the prosperity and safety of the Arab world in particular and the world in general. For example, having the tools to check the authenticity of a hadith at a click of a button may save youth from the hands of extreme religious leaders. Extracting a relation between two criminals may uncover a third criminal who is ready to commit a crime. Finding inconsistencies in claims may curb fraud and corruption in the insurance sector. In addition to these benefits, building the necessary academic infrastructure for Arabic text mining in Lebanon and Qatar will open the way for investments in the field and provide work opportunities for native workers.

# Preliminary Data or Studies

Our preliminary studies are promising. With a small manually entered dictionary, few programmed grammar rules and patterns, and two hardcoded queries we were able to solve selected samples from the literature and security case studies presented in Section “Significance”.

These preliminary results serve as a good proof of concept to our hypothesis. They also confirm that our approach performs better than currently existing techniques from Sakhr [[13](#Sak09)] and Basis [[14](#Bas09)] as well as from open source tools [[7](#Col09)].

The Co-Lead PI worked and published extensively in this domain. A selected list of her recent publications follows.

* Rehab Duwairi, Mohammad Al-Refai, Natheer Khasawneh, "Feature Reduction Techniques for Arabic Text Categorization". Journal of the American Society for Information Science and Technology (JASIST), Volume 60, Issue 11, pages: 2347-2352, 2009.
* R. M. Duwairi, “Arabic Text Categorization”, International Arab Journal for Information Technology (IAJIT), Volume 4, Number 2, pages 125 – 132, 2007.
* R. M. Duwairi, "Machine Learning for Arabic Text Categorization". Journal of the American Society for Information Science and Technology (JASIST), Vol. 57, Issue 8, pages 1005-1010, 2006.

The Lead PI worked in the area of software arabization since 1996 as he worked on developing Arabization modules for Windows NT and an Arabic string manipulation library [[15](#Zar96)]. The Lead PI is also involved in working on an open source project that aims at comprehensive automation of authenticity checks against a corpus of literature heritage [[16](#Zar06)]. The Lead PI is leading an effort at AUB to introduce Computational Arabic in the curriculum of the ECE program. He is also leading a team of students in an effort to place the infrastructure needed for this project to start.

The lead PI worked and published about relational logic and logic solvers applied to structured languages as follows.

* Sequential circuits for relational analysis. F. Zaraket, A. Aziz, and S. Khurshid. *International Conference on Software Engineering*, Minneapolis MN, 2007.
* Sequential circuits for program analysis. F. Zaraket, A. Aziz, and S. Khurshid. *Automated Software Engineering*, Atlanta GA, 2007.

# Research Design and Methods:

We see more and more Arabic documents such as texts, books, publications, hospital and governmental records emerging every day. Most of the newly generated documents are produced in digital textual form while also old paper documents are ported to digital form. These documents include huge amounts of information that is qualitatively different than structured information such as that contained in database entries. Text mining is the technology that automates the discovery of information in non structured text.

Text mining concerns the partitioning of text segments into classes, the clustering of text segments, the extraction of concepts and entities from text, and the discovery and modeling of relations between text entities and classes. Text mining techniques perform natural language processing (NLP) to analyze text and perform information extraction (IE) to extract information into data structures. Fundamental to IE are the name entity (NE) and the name entity relation (NER) extraction techniques which capture features from sentences and phrases. A key challenge to NE and NER is that sentence structures and words are often ambiguous in natural languages. While research to address ambiguity in Latin languages is still lagging behind [[17](#Red08)], not much has been done to address NE and NER in the context of the Arabic language.

The application of text mining techniques to Arabic text documents will result in great benefits to many research fields such as Arabic literature, Islamic studies, Hadith authentication, and Arabic history and culture in general. Arabic text mining will also benefit sectors where Arabic text documents are key such as the security sector, the government personal records, the taxation department, the health sector, as well as the trading floor of a stock exchange.

The difficulty of applying text mining techniques to Arabic text documents lays in the absence of automated tools that understand the unique features of the Arabic language. We will make use of an example to illustrate the process. Given an Arabic text document we want a tool that can segment and isolate all names, dates, time of the day occurrences, tool names, and locations as entities. Then we want another tool to identify identities amongst these entities, and relate them in their order of appearance in the document. We want another tool to look at the verbs and actions in the sentences and try to build relations between the identified entities based on these verbs. If we were successful to do that, we now want to order all entities associated with dates based on a chronological order.

To do the above, we need tools that are able to automatically differentiate between names, verbs, adjectives and other vocabulary structures. This is not a simple task with the Arabic language. For example, Arabic grammar allows name-based sentences or sentences with no verbs (جمل اسمية) . Another unique feature is the possibility to have names in Arabic that express action such as noun verbs (اسم فعل، فاعل...) . In addition to these structural differences, there are features in the Arabic language that need special treatment such as derivatives and stems (المشتقات والجذور). We also need to build dictionary tools that associate words and phrases based on their meanings and their context. Often time, and depending on the application and the user of the tool, we need to allow the user to assign meanings and semantics for patterns in the text.

In the following we discuss our research methodology to develop the proposed Arabic text analysis tools.

***Lexical analysis library:*** We need a library of routines that enables primitive lexical analysis of Arabic text. For example, simple primitive tasks such as comparing two strings need special attention in Arabic since each Arabic letter has four different forms depending on its position in the word; isolated form, start, middle and end of a word. Arabic diacritics should also be ignored most of the time when comparing two strings. These lexical analysis routines should also pay attention to the form of a letter when used in tokenization as a letter in and end of word form denotes the end of the word even if it was not followed with a space. Such a library can benefit from existing Unicode enabled string libraries such as Qt. It also can provide a layer of abstraction that hides the details of the encoding of Arabic letters used in the text documents.

***Syntax analysis library:*** This library should allow sentence structure understanding. We will explore the best way to develop this library. Several options exist such as pattern based matching, tokenization and dictionary based lookups, or a language grammar that allows more than one matching rule. This library needs to be extendible to allow additions of new rules or patterns. We will explore Arabic grammar books such as The Principles of Arabic [[18](#Sha73)] and others [[19](#Abd001)] [[20](#Abd00)] to augment current morphological analyzers [[6](#Tim04)] and syntax analysis tools [[7](#Col09)]. We will use relaxed NLP techniques to overapproximate results. Then and after a relational query, we will apply statistical analysis techniques to rank the matching results.

***Dynamic modern Arabic dictionary:*** The dictionary will allow identities based on semantics, derivation rules, as well as user assigned rules. The dictionary will return modern grammatical categories for words classifying them as nouns, verbs, and adjectives instead of the classical Arabic categories that specify how the word is derived from an original radix. It remains to be explored how to represent these categories without exploding the size of the dictionaries with redundant words of multiple categories. This work will augment existing open source Arabic dictionaries and will allow for better morphological analysis and entity extraction.

***An entity extraction library:*** We will explore building a library that allows extraction of entities from Arabic text. These entities can be proper names, dates, locations, and qualification adjectives to list few possibilities. The extractor will use the dynamic modern Arabic dictionary as well as the syntax library to extract entities from a sentence and to establish identity between entities across sentences.

***A relation extraction library:*** We will explore building a library that allows extraction of relations from Arabic text. We will explore the use of pairing techniques to establish relations between the extracted entities. The relations will form graphs where the nodes are the entities and the previously extracted relations and the edges are the newly discovered relations. The routines in this library will take as input a template for an expected relation and will return a graph that describes whether this relation exists between the subject documents and where. These relations will be ranked using statistical analysis. The statistical criterion and the ranking methods remain to be decided.

***A logic solver:*** This is a logical solver that will solve query expressions to resolve matching text and relations. We will explore the proposal of a logical query language that is restricted enough to be computable and expressive enough to allow interesting properties. We will use of the shelf solvers and matches from the entity and relation extractors to evaluate the expression. The choice of solvers remains to be decided after the query language is formed.

***Tagger:*** We will explore the development of an automatic tagging interface that allows fast and guided inspection and correction mechanism.

We will integrate all the mentioned techniques in a scalable framework for Arabic text analysis. The framework as a whole will allow for exhausting Arabic text analysis and allow for extensions of the framework. A user should be able to iteratively and repetitively apply the abode techniques on his input text.

## Work Plan

Our work plan is detailed in the following table. In the activity type column the letter D stands for documentation, P stands for programming, L for learning, and R for research.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DELIVERABLE | Activity type | Time period | | | | | | | | | | | |
| First year | | | | Second year | | | | Third year | | | |
| Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 |
| Lexical analysis lib | P |  | | | |  |  |  |  |  |  |  |  |
| Syntax analysis lib | P,R |  |  | | | | |  |  |  |  |  |  |
| Dynamic dictionary | P,R |  |  | | | |  |  |  |  |  |  |  |
| Entity extraction lib | P,R |  |  | | | | | | | | |  |  |
| Relation extraction lib | P,R |  |  |  | | | | | | | |  |  |
| Query language and Logic solver | P,R |  |  |  |  | | | | | | | | |
| Tagging interface | P |  | | |  |  |  |  |  |  |  |  |  |
| Literature case study | P,R |  | | |  | | | | | | | | |
| Security case study | P,R |  | | |  |  |  | | | | | | |
| Health case study | P,R |  | | |  |  |  | | | | | | |
| Search engine lib | P |  |  |  | | | | | |  |  |  |  |
| Integrated framework | P |  | | | | | | | | | | | |
| Continuous training | L |  | | | | | | | | | | | |
| NLP research | R |  | | | | | | | | | | | |
| IE research | R |  | | | | | | | | | | | |
| Documentation | D |  | | | | | | | | | | | |

In the development cycle, we will use an Extreme programming methodology in the process of building the Arabic text mining framework. We will use open source tools, libraries, and frameworks to build our libraries.

Our research will concentrate on NLP techniques native to the Arabic language to analyze the Arabic text and we will avoid preprocessing that feeds common text and data mining techniques. We will build on existing open source programming and text mining tools as well as open source Arabic text analysis tools based on NLP such as MADA+TOKAN [[21](#Rot08)], NEMLAR [[22](#RAl09)], and IJAES [[23](#Int09)].

We will use a dokuwiki based website [[24](#Dok09)] to document the progress online and will use subversion [[25](#Sub09)] as a source control tool.

# Anticipated Results and Evaluation Criteria

We anticipate that the research will lead to the production of an Arabic text analysis framework that allows relational queries between Arabic text documents. The components of the framework are listed in Section “Research Design and Methods” and are repeated for convenience as follows.

* Lexical analysis library
* Syntax analysis library
* Dynamic modern Arabic dictionary
* An entity extraction library
* A relation extraction library
* A logic solver
* Tagger

We will use these tools to handle several case studies from different fields. In the following we discuss three case studies that we anticipate our proposed framework to solve.

### The literature case study

Assume we have two sets of texts. The first set is several books of historical accounts where each account is preceded by a chain of narrators who narrated the account. The second set is several books of biographies where each biography of a person may include an evaluation of the authenticity of that person. We want a tool that can answer a complex query such as relating all historical accounts in the first set that discuss women and vehicles (such as tamed animals) to the narrators of the accounts, and then relate these narrators to the second set of books of biographies with an emphasis on authenticity evaluations.

We need tools that can search and index the historical accounts to return a subset that relates to women and travel. Then we need tools that can identify names, and succession relations between names to build a chain of narrators for each historical account in the subset. We need tools that can identify the names in the biographies and detect evaluation adjectives.

Provided we have these tools we can exhaustively at the click of a button visualize the authenticity of all the historical accounts that may allow or forbid women from driving according to Islamic laws. We can go one step further to detect locations and dates and relations based on chronological orders and further question the consistency of the historical accounts. Such query takes currently huge amounts of manual effort that is also prone for error and far from being exhaustive.

We here envision that text mining applied to the authenticity of historical accounts will have a huge impact in the field.

Note that most of the Islamic history and Hadith books as well as a huge library of Arabic literature books are already digitized and available on the internet.

### The security case study

Given a huge set of investigation reports, an officer may throw a list of names of convicted people and check all the reports for possible relations between them in terms of criminal action, locality, or third party names unknown to the officer who happen to know the convicts. The query may return at the click of a button, two convicted people who had no direct relation in the reports, but who are related only through a third person unknown to the investigator. Such a query is very hard to answer manually.

### The health case study

Given a set of clinical reports gathered by the Order of Medical Doctors, the ministry of health may query to relate all reports on a specific sickness to the doctors treating it and evaluate the response to the different types of medications as classified in different areas in the country each with a different climate. Such query, while possible at a click of a button with text mining tools if the data is available in digital form, would take a huge manual effort to conduct surveys, collect and gather the data. In this case, digitizing the data may be a less expensive effort.

## Evaluation Criteria

The Lead PI will conduct biweekly meetings to track progress on the project. The team will maintain a dokuwiki based website to continuously post updates on the progress of the project. In Section “Research Design and Methods” we presented a schedule for delivering the different products of the project. The schedule can be used as a timeline for measuring the progress of the project.

The lexical and syntax analysis libraries will be compared with current open source libraries against commonly used corpora of data such as a collection of books or a collection of news reports. The dictionary module will be an augmentation over existing digitized dictionaries such as the Buckwalter [[6](#Tim04)] morphological analyzer. In the entity and relational extraction tools we will use techniques native to the Arabic language and we will compare against English entity and relational extraction tools of similar or translated texts. We will also compare techniques that pre-process Arabic and then pass it to English IE and RE techniques. The logic solver will build over the IE and RE tools and will use off-the shelf quantifier and propositional solvers such as MiniSat [[26](#Een03)]. The tagger is a GUI interface that allows fast manual inspection and refinement of the automated results. Each tool in the framework can start on its own with a prototypical implementation of the other needed tools.

Each tool in the framework stands alone and can operate on a well defined input and generate the desired output. The input of one tool, such as the entity extraction, can be automatically generated by another tool, such as the dictionary and the syntax analyzer. However, the input can also be manually generated if desired using the straightforward GUI tagger facility. Thus the tagger facility stands as a desired utility on its own and at the same time represents a contingency plan in case any module in the framework failed to deliver.

The team will publish the results of the research in the form of refereed conference and journal papers as well as tool papers addressing each of the tools. The team will also hold workshops to educate possible users on the tools once the tools are available.

# Strategy for Project Continuation

Once completed the Arabic text analysis framework will be an open source contribution and developers from the Arabic world in the open source community are expected to join efforts to keep it going. Furthermore, once completed the Arabic text analysis framework can sustain itself by charging on support or expert use. The framework or service on the framework can easily be commercialized to tend to the security, health, insurance, news, media, or literature business sectors.

The area of Arabic text analysis will be a research area for a long time as NLP is one of the most complex computational problems. The Arabic text analysis framework will be used by NLP researchers in the Arabic language field and those are expected to maintain and advance the framework.

# Plans for Disseminating Research Results

The team will publish the results of the research in the form of refereed conference and journal papers as well as tool papers addressing each of the tools. The team will also hold workshops to educate possible users on the tools once the tools are available.

The Arabic text analysis framework will be partially based on existing work in the open source domain and will thus be available as an open source tool. Interested researchers and users will be able to download the tools for free and use them for non-profit non-commercial purposes.

## References (reproduced here for convenience)

x

|  |  |
| --- | --- |
| [1] | J. Nightingale, "Digging for data that can change our world," *The Guardian*, 10 January, 2006. |
| [2] | A. Srivastavaand M. Sahami, *Text Mining: Classification, Clustering, and Applications*.: Chapman & Hall, 2009. |
| [3] | H. Al-Fawareh and S. Jusoh and W. Sheikh Osman, "Ambiguity in Text mining," *IEEE proceedings of the international conference on computer and communication engineering*, May 2008. |
| [4] | J. Han and M. Kamber, *Data mining concepts and techniques*.: University of Illinois at Urbana Champaign, 2005. |
| [5] | K Beesley, T Buckwalter, and S Newton, "Two-level finite-state analysis of Arabic morphology," in *Seminar on Bilingual Computing in Arabic and English*, Cambridge, England, 1989. |
| [6] | Tim Buckwalter, "Buckwalter Arabic morphological analyzer version 2.0," Linguistic Data Consortium (LDC) , catalog number LDC2004L02, 2004. |
| [7] | (2009) Columbia's Arabic Dialect Modeling Group. [Online]. <http://www1.ccls.columbia.edu/~cadim/> |
| [8] | A. Al-Zoghby and A. SharafEldin and N. Ismail and T. Hamza, "Mining Arabic Text Using Soft-Matching," in *Computer Engineering & Systems*, 2007. |
| [9] | A. EL-Halees, "Arabic Text Classification. Using Maximum Entropy," *Islamic University journal* , 2007. |
| [10] | Mesleh Abdelwadood, "Support vector machines based Arabic language text classification system: feature selection comparative study," in *12th WSEAS International Conference on Applied Mathematics*, 2007. |
| [11] | M Eldost, "Arabic text data mining: A root-based hierarchical indexing model," *International journal of modelling & simulation*, vol. 23, 2003. |
| [12] | Abdelhadi Soudi, Antal van den Bosch, and Neumann Guenter, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*.: Springer, 2007. |
| [13] | (2009, September) Sakhr Inc. [Online]. <http://www.sakhr.com/products/Mining> |
| [14] | (2009) Basis technology. [Online]. <http://www.basistech.com/base-linguistics/arabic/> |
| [15] | Fadi Zaraket, Rabih Zbib, and Monah Khalil, "Yaraah: Arabization module for Windows NT," AUB, technical FYP. 1996. |
| [16] | Fadi Zaraket, Marwan Zeineddine, Hussein AlAsadi, and et al. (2006) Hadithopedia: comprehensive hadith encylopedia. [Online]. <http://code.google.com/p/hadithopaedia/> |
| [17] | Judy Redfearn, "Text Mining," *Joint Information Systems Committee*, September 2008. |
| [18] | R. Shartouni, *The principles of Arabic/Mabādi al-Arabiyah*. Beirut: Le Levant, 1973. |
| [19] | Abdeh AlRajehi, *Semantical Practice/At-tatbiq Alnahawi*, 1st ed. Beirut: Renaissance (nahda), 2000. |
| [20] | Abdeh AlRajehi, *The morphological practice/At-tatbiq Assarfi*, 1st ed. Beirut: Renaissance (An-nahda), 2000. |
| [21] | Ryan, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin Roth, "Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking," in *Association for Computational Linguistics (ACL)*, Columbus, Ohio, 2008. |
| [22] | G. Kanaan, M. Yaseen, B. Al-Sarayreh, N. A. Al-Naji R. Al-Shalabi, "Arabic Query Expansion Using Interactive Word Sense Disambiguation," in *International Conference on Arabic Language Resources & Tools*, cairo, 2009. |
| [23] | (2009) International Journal of Arabic-English Studies. [Online]. <http://www.ijaes.org/> |
| [24] | (2009) DokuWiki. [Online]. <http://www.dokuwiki.org> |
| [25] | (2009) Subversion. [Online]. <http://subversion.tigris.org/> |
| [26] | Niklas Een and Niklas Sörensson, "An Extensible SAT-solver," in *SAT*, 2003. |
| [27] | QT C++ Library. [Online]. <http://www.qtsoftware.com> |

x