Verifying Hadith Correctness in Islamic Web Pages using Information Retrieval Techniques

Mohammed Q. Shatnawi
Faculty of Computer and Information Technology
Jordan University of Science and Technology

Qusai Q. Abuein
Faculty of Computer and Information Technology
Jordan University of Science and Technology

Omar Darwish
Faculty of Computer and Information Technology
Jordan University of Science and Technology

ABSTRACT
Certain types of web information need to be verified for correctness once being posted on the web (i.e. the information has a reference that cannot be tampered with). Such these types of information need systems in order to verify and assure their correctness. "The Hadith"; which is the sayings of Prophet Mohammad is an example of these types of information. Hadith in Islam is a very important tool to understand Qur'an, and in matter of Islamic law.
This paper presents a reliable method to extract Hadith text from web pages, in order to find out the degree of correctness for each one; by looking it up in Sheikh Al- Albani Hadith Database (correct series and weak series) which contains 17000+Hadeeth texts and their degrees of correctness according to Sheikh Al-Albani studies. The proposed approach uses positional index built over the Sheikh Al-Albani Hadith Database (SAHDB) and extracts Hadith terms as queries from the passed web page, and execute these queries over SAHDB index.

General Terms
Information retrieval.

Keywords
Hadith Database, web information, SAHDB, Al-Albani, Weak Hadith, correct Hadith.

1. INTRODUCTION
Huge amount of information of different types is being posted on the web on daily basis. There are tools available to enable users to search for this vast amount of information. Different types of information needs different processing approaches. Figure 1 provides the different types of web information. For example, processing structured-based text differs from processing unstructured-based text.
Hadith has a special importance in Islamic creed, because it is the second source of legislation after the Holy Qur'an. Once Muslims need to verify and evaluate their actions, they refer first to Qur'an, and if they could not resolve the issue, they refer to the Hadith as the second source of Islamic Law.
As a result, the science of Hadith has become one of the most important fields in Islamic studies. The core interest of science of Hadith is verifying the correctness of the speeches, deeds, acts or characteristics that have been adopted by prophet Mohammad (PBUH) .
Thousands of Islamic contents websites have been launched in the last few years. Unsurprisingly, most of them have their content written using the Arabic language. In addition to the availability of other websites that offer Islamic services like Fatwa, Islamic e-book download and Qur'an recitations. Hadith text can be found in other web sites. For example, Hadith text can be found, almost, in most of Arabic discussion boards and forums or even in thousands of Islamic e- mails that being sent daily by individuals. As a matter of fact, the Hadith is available widely on the web. Therefore, there should be an available tool to be used in validating the Hadith correctness, which, in fact there isn't a known tool to help the Muslim or an interested communities verify the correctness of Ahadith (i.e. plural of Hadith).

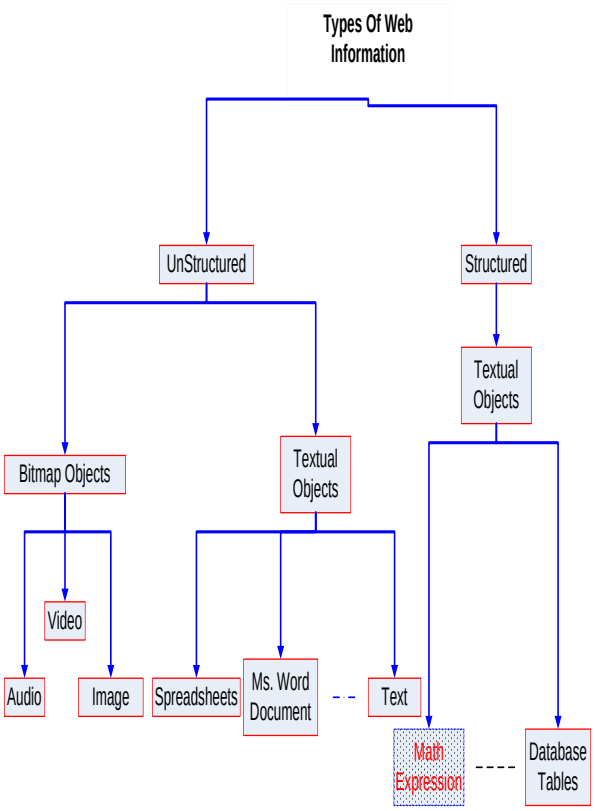

Fig 1: Types of Web Information

2. MOTIVATION
Because of the availability of this very large number of Islamic contents web pages, and the nature of the Internet, it is almost impossible to track the degree of correctness for every Hadith mentioned in every web page and try to look its degree up. If the author of that web page did not verify the correctness of the posted Hadith, it would be extremely difficult for the reader to verify the correctness of those Ahadith. Especially, with the existence of false speeches that

were never found to have any relation to prophet Mohammad (PBUH). In addition, web users do not have the potential to start looking for the Hadith correctness over the web; instead, they need a tool that does this important and critical task on behalf of them. As a solution, this paper proposes a method to automatically retrieve Hadith text and find its degree of correctness according to SAHDB so it would be considered as trusted Islamic source of legislation enabling web surfers to trust or distrust such those Ahadith.

3. SHEIKH AL-ALBANI HADITH DATABASE (SAHDB)

Sheikh Mohammad Nasser Al-Deen Al-Albani [1] (1914-1999) is the most famous modern Muslim scholar in the area of Hadith. He is the author of a many books in various areas in Islamic sciences. This paper focuses on two of his famous work.

- Al-Selselah AlSahihah (The Correct Series), which contains thousands of Hadith texts that were believed to be Sahih. Hadith Sahih is the one that is "conveyed by a trustworthy, completely competent person, either in his ability to memorize or to preserve what he wrote" [1] [2] or at least Hasan (Good).
- Al-Selselah ALDa'eefah (The Weak Series) in which he listed thousands of Hadith texts that were found to have lower degree of correctness like Da'eef (Weak) or Mawdo'u (Discarded) [1].

This database is provided by Sheikh Al-Albani's official website [1] and presented in the form of Microsoft Access database. It contains more than 17000 Hadith texts from both correct series and weak series, along with their degrees (i.e. the level in which can be used to trust the Hadith). Each Hadith is enlisted in one of the categories (degrees) shown in Table 1.

Table1: Hadith Degrees

Degree (Arabic)	Count	Degree (Arabic)	Count
أثر صحيح	9	لا أصل له	99
أثر ضعيف	8	لا يصح	15
باطل	78	لم تتم دراسته	119
حسن	1542	متفق عليه	12
شاذ	45	مرسل	1
صحيح	11083	منكر	288
ضعيف	2878	موضوع	848
ضعيف جدا	537	موقوف	2
كذب	10	**Total**	**17545**

One of the disadvantages of this database is that it may contain more than the original text of Hadith like some notes and advices along with the original Hadith text. However, it remains the most appropriate Hadith database since it contains texts for known weak/ Sahih Hadith (i.e. weak/ correct Hadith).

4. RELATED WORKS

The concept of Auto-indexing refers to the operation of automatic selection of the main/key words in a text document. This issue differs in difficulty according to the used language. Every language is described by its domain, logical structure, and its syntax [3]. In fact, languages that have complicated grammatical rules, such as Arabic, need complicated indexing methods.

Many of those methods have been discussed in the literature.

N-gram frequency statistic technique is used in classification algorithms for Arabic language [4]. An Arabic part-of-speech tagger that uses statistical and rule-based methodologies has been proposed [5]. These parts-of-speech are further portioned into particles, verbs, and nouns. Stemmer to eliminate all of a word's affixes has been used to generate the stem. The stemmer problems occur when particular letters that appear to be affixes are part of the word itself, and when certain letters change to other letters when an affix is added.

Gawrysiak, Gancarz, and Okoniewski [6] express the unigram and N-gram document representation methodologies that are used commonly in text mining and disuses their shortcomings.

Another technique that looks at the word's positions in the document, was presented. In Larkey and Connell [7], the authors employ some standard techniques to handle co-occurrences. Several monolinguals were presented and cross-language runs and shows that stemming enhances the proposed system.

Billhardt, Borrajo, and Maojo in [8] use term co-occurrence data for document indexing.

Authors in [9] provide a multilingual Internet search engine that includes Arabic word stemming. Stemming is performed by eliminating postfix, prefix, infix and parts of words, and also by reducing plural, singular and adjective forms into a canonical form. Light stemmers have been developed by Larkey, Ballesteros, and Connell in [10] where they depend on heuristics and a statistical stemmer based on co-occurrence for Arabic retrieval. Because of that, rarely limited work has been done on finding indices for Arabic text and especially in Hadith approach.

Aljaly, M. and O. Frieder in [11] found that the main technique in decreasing the size of index term and gaining a higher degree of relevancy for the retrieved document is using stemming methods. They also found out that stemming methods would shrink the size of the document representation by 20-50% regarding to the full words representation.

5. POSITIONAL INDEX AND HADITH EXTRACTION

In order to effectively extract Hadith text from a web page, a positional index [2] has been built for Hadith database; each Hadith has been treated as a single document, and hence a unique document ID was assigned for each document in the index. The following subsections illustrate the main steps for index building and Hadith text extraction.

5.1 Tokenization and Normalization

In the tokenization step, terms have been simply split up into tokens based on the space delimiter. Apparently, this will make some common Hadith terms ""عن ابي هريرة"" and "صلى الله عليه و سلم" be handled incorrectly because of using the space as a delimiter.

In the information retrieval approaches, the frequently appear terms can be removed, because they cannot be considered as words that distinguish between documents. To overcome this problem stop word list and positional index have been used as will be explained in the following subsection. Each token is then normalized to a simple form. Normalization involves two steps:

1. Remove any character other than the known 28 Arabic alphabet letters. This ensures that any Arabic vowel marks (Tashkeel or Harakat) and commonly used word stretching character "_" like in "محمـد" are removed.
2. Replace all forms of Alef "إ، أ، آ، ا" with simple "ا" without any signs. This ensures that even mistyped occurrences of Alef are handled.
   Once the token is normalized, it is added to the positional index.

5.2 Stop Word List and Index Building

It is important to have stop word and stop list since we focus on the precision when extracting Hadith text from web pages. Therefore, having very common phrases in Hadith text may degrade the precision. For instance, consider the phrase "عن ابي هريرة رضي الله عنه عن النبي صلى الله عليه و سلم" which is common in more than 1300 Hadith. It is totally inappropriate to extract all these 1300 Hadith because one of them is mentioned in the web page. Experimentally, we have found that terms with occurrence for more than 750 times in all of 17000 Hadith should not be indexed, thus it is added to a stop word list, which contains eventually 122 terms.
After excluding stop words, a positional index is built from the remaining terms. We have used the format specified in [12] to build up the positional index for all terms in Hadith database, which counts 56087 terms.

6. EXTRACTING HADITH FROM WEB PAGE

This is the core step in Hadith verification. Extracting Hadith involves several steps illustrated below:

1. Clean the HTML code from the web page. This step was performed using HTML cleaner Java package [12].
2. Generate a list of terms that occur in Hadith positional index and discard other terms.
3. Set a value K so that K is the minimum number of contiguous words that should occur in both positional index and terms list in the same order. Practically, K values of 3 and 4 were found to give more accurate results.

When all documents (Hadith texts) are retrieved, we look up each one by its document ID to find out the degree of correctness for each one.

7. EVALUATION

In the experiment, five web pages that contain Hadith text have been randomly chosen. Total number of Hadith text in those web pages was 63, the number of the retrieved text was 83 with 32 correctly extracted (i.e. relevant).
Based on the precision and recall measures and the experiment outcomes, the following results have been provided:

Precision = 32 / 83 = 38.6%
Recall = 32 / 63 = 50.7%

The reason which led to this result is the existence of common phrases inside Hadith text. For example, the phrase "لا اله الا الله و حده لا شريك له له الملك و له الحمد" occurs in document 1. This phrase exists in about 36 Hadith texts. Such phrases lead to at least 35 irrelevant documents to be returned. However, if the precision and recall are calculated without document 1, the result would be:

Precision = 16 / 21 = 76.1%
Recall = 16 / 38 = 42.1%

Notice that the precision is much higher in this case. One suggested solution to solve this problem is to make stop word determination policy more strict by decreasing the maximum number of term occurrences before considering it as a stop word, but this would lead to another problem which is the inability to extract short Hadith texts that consist of four or less terms. Another suggested solution is to apply this method only on a larger database that contains more Hadith text. This process will guarantee that all Hadith texts in a given web page are already known and indexed. In addition, there is a need to have cleaner database that gives more accurate results as it has been suggested in the conclusion.

8. CONCLUSION AND FUTURE DIRECTIONS

The suggested method of extracting Hadith gives relatively a good result as appears in the previous section. It is clear that the accuracy of this method needs to be enhanced. The following suggestion shows the problems along with the proposed method and suggestions to enhance the accuracy for future work:

1. The used Hadith database needs to be cleansed because Hadith text currently contains additional terms that are not part of the Hadith itself. For example, a Weak Hadith text may contain another suggested correct Hadith terms; such texts need to be cleansed so the database contains only the original text without any extra terms that usually occur in almost most of the Hadith texts [13].
2. Redundant Hadith texts need to be removed. The same Hadith may contain two or more alternating texts recorded as completely two different texts whereas they are the same. Removal of such redundancies and keeping only one alternate for each Hadith text tends to give higher precision.
3. The Hadith database needs to be larger than the used one. Currently, the used Hadith database still misses a lot of known Hadith texts whether it is correct or weak. Using large enough database allows more Hadith texts to be recognized and extracted, which gives higher recall.

9. REFERENCES

[1] The website of Sheikh Al- Albani, http://www.alalbany.net/, [Accessed in October 2011]

[2] Mohammad Bin Isma'eil AlBukhari, "Sahih AlBukhari", Dar Al-Islam Publisher, Alreyad

[3] Harter, S. (1986). Online information retrieval: Concepts, principles, and techniques. Orlando, FL: Academic Press.

[4] Khreisat, L. (2006). Arabic text classification using N-gram frequency statistics: A comparative study. In Proceedings of the 2006 international conference on data mining (pp. 78–82).

[5] Khoja, S. (2001). APT: Arabic part-of-speech tagger. In Proceedings of student workshop at the second meeting of the North American chapter of the association for computational linguistics.

[6] Gawrysiak, P., Gancarz, L., & Okoniewski, M. (2002).Recording word position information for improved document categorization. In Proceedings of the PAKDD text mining workshop.

[7] Larkey, L., & Connell, M. (2001).Arabic information retrieval at UMASS. In Proceedings of the 10th text retrieval conference (pp. 562–570).

[8] Billhardt, H., Borrajo, D., & Maojo, V. (2000).Using term co-occurrence data for document indexing and retrieval. In Proceedings of BCSIRSG 22nd annual colloquium on information retrieval research (pp. 105–117).

[9] Rachidi, T., Iraqi, O., Bouzoubaa, M., Ben Al Khattab, A., El Kourdi, M., Zahi, et al. (2003). Barq: Distributed multilingual internet search engine with focus on Arabic language. In Proceedings of the IEEE conference on systems, man and cybernetics.

[10] Larkey, L., Ballesteros, L., & Connell, M. E. (2002). Arabic information retrieval: Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis .In Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval.

[11] Aljaly, M. and O. Frieder, 2002. On Arabic Search: Improving the Retrieval Effectiveness via light Stemming Approach.

[12] Christopher D. Manning, Prabhakar Raghavan , Hinrich Schütze (2008). An Introduction to Information Retrieval, Campridge University Press.

[13] HTML Cleaner Project Home Page, http://htmlcleaner.-sourceforge.net, [Accessed in October 2011]