# إيجاد الأسئلة المماثلة أو المتشابهة في أنظمة تجميع أو (محتوى) الأسئلة والإجابات باللغة العربية

د . خليفة عبدالرؤوف نصرات ـ كلية التربية الزاوية ـ جامعة الزاوية

د . رفعت أوزكان ـ كلية الهندسة الإلكترونية ـ جامعة تورغت أوزال ـ تركيا

**ملخص الورقة:**

نظام الرد على الأسئلة من مكان تجميع أو (محتوى) الأسئلة والإجابات أصبح اليوم شائعا جدا حيث يمكن للناس الحصول على إجابات لأسئلتهم مباشرة من هذا المحتوى بدلا من تصفح عشرات المستندات نتيجة البحث على الويب ، والعثور على أسئلة مماثلة في أنظمة الرد على الأسئلة من محتوى الأسئلة والإجابات أحد أهم المشكلات نظرا ؛ لأنه من الممكن أن تكون الأسئلة المتشابهة قد تمت الإجابة عليها بالفعل. ومعظم الدراسات السابقة في هذا الموضوع كانت مخصصة للغة الإنجليزية فقط. وهناك دراسات محدودة للغاية في مجال اللغة العربية (محتوى سؤال وجواب باللغة العربية) ، وفي هذه الدراسة قمنا بفحص الأداء والإنجاز باستخدام مجموعة من النماذج ( Models ) لمشكلة إيجاد أسئلة مماثلة في نظام الرد عن الأسئلة من مكان تجميع الأسئلة والإجابات باللغة العربية .

# Finding Similar Questions in Arabic Community Question Answering Systems

Dr . Khalifa Abdullrauof Nusrat ، Dr. Rifat Özcan

**Abstract**

Community Question Answering (CQA) systems have become very popular recently. People can obtain direct answers to their questions instead of browsing tens of documents as a result of searching on the Web. Finding similar questions in CQA systems is one of the most important problems since it is possible that very similar questions might have been already answered. Even though there is an extensive literature focusing on this problem,

most of the studies are for only English. There are very limited studies concerning Arabic CQA systems. In this study, we investigate the performance of various retrieval models on finding similar questions problem for an Arabic CQA system.

**Keywords**

Community question answering; finding similar questions; Retrieval models.

## 1. INTRODUCTION

Community Question Answering (CQA) services have become very popular in recent years. CQA services such as Yahoo! Answers [1] , BaiduKnows[2] , StackOverflow[3] are social collaborative applications and became important information resources on the Web. They have millions of users who seek for an answer to their questions and/or provide answers to diff erent questions in diverse subjects. CQA services have several advantages over using Web Search engines. One benefit of these services is that users

[1]https://answers.yahoo.com/

[2]http://zhidao.baidu.com/

[3]http://stackoverflow.com/

can directly obtain answers rather than a list of potentially relevant documents, which is the case in web search. However, the user may have to wait for some time so that his/her question gets answered by the community. The waiting time to obtain the answers may be reduced if there is a large number of questions-

answers and a similar question may have already been answered previously [5] . One of the problems in CQA services is to find similar questions for a submitted query so that the user can retrieve the answers for very similar questions. This retrieval task requires matching of existing questions /answers that are semantically similar to the user's question. The major challenge for this problem is the word mismatch between the user's question and existing question- answer pairs in the CQA sites. The Arabic language is one of the major languages used On the Web and has many users which is increasing daily.

Arabic Information Retrieval (IR) has gained significant attention in the last decade due to the increasing volume of the Arabic text on the Web and the Arabic language is ranked

As the seventh top language on the Web. The number of Arab Internet users grew from 2.5 million in 2000 to 65mil-lionin 2011 [3]. As of November 2015, this number reached

to168. 1million and ranked fourth (following English, Chinese, and Spanish)[4] .Arabic is an important language Islam Religion since the Quran, one of the four Holy books, was revealed in Arabic. There are more than 1.2 billion Muslims in the world and they pray five times aday using the Arabic language. Geographically, Arabic is an official language in 25 countries including the members of the Arab league.

These countries are populated with more than 400 million People that making Arabic the fifth most commonly spoken Language in the world. Due to the increase in the number of Arabic speaking users on the Web, there are a lot of CQA services used by Arabic people but there are only a few studies on this area. In 2014,

Darwish and Magdy, in their Arabic IR survey [3], report ranking similar questions /answers for a new question as an open research area in Arabic IR .Towards this goal, Semantic Evaluation Workshops in 2015[5] and 2016[6] reserved one of their tasks to problems related to Arabic CQA systems.

Therefore, finding similar questions in Arabic CQA systems is a hot research topic. In this study, we compare the effectiveness of three different retrieval models, vector space model and two probabilistic approaches (BM25 and language model), for finding similar

---

[4]http://www.internetworldstats.com/stats7.htm

[5]http://alt.qcri.org/semeval2015/task3/

[6]alt.qcri.org/semeval2016/task3/

Questions in an Arabic CQA system .We first construct our dataset including 5,000 questions and report some characteristics of these questions and answers. We compare three retrieval models using NDCG and average relevance metrics and also investigate the effects of two different stemmers on this problem. This paper is organized as follows: Section 2 presents related work on finding similar questions in CQA services. We describe our dataset in Section 3.We mention the basics of the three retrieval model in Section 4. Section 5 presents our experimental results. We conclude the paper in Section 6.

## 2. RELATEDWORK

There are various studies on finding similar questions in CQA systems but most of them focus on English question-answer

pairs. There are three main directions for question similarity, namely, lexical, syntactic, and semantic approaches. In an early study, Burkeetal. [1] Propose a hybrid approach consisting of a lexical approach based on vector space model and a semantic similarity approach (using WORDNET) for finding similar questions from frequently asked questions (FAQs). Jijkoun and deRijke [6] also use vector space model for ranking questions in FAQs. Finding similar questions problem is also called "question search" and language model based approaches are also proposed in [4,2]. In later studies, translation based models are proposed for question retrieval in order to bridge the lexical gap between new question and historical questions in CQA systems. Translation based approaches requires a data set to Learn the translation probabilities between different terms. This normally requires a parallel corpus where sentences in one language are mapped to sentences in another language (for statistical machine translation). Zhou et al. [13] propose A translation model for question search for CQA systems. Questions and their answers are considered as a parallel corpus in order to learn translation probabilities. They used about 1million question-answer pairs from Yahoo! Answers. Wang et al. [12] propose a composite kernel approach that captures both lexical semantics and syntactic information in a question sentence by focusing on word sequence, POS tag Sequence and syntactic tree (using parse tree). They found that composite kernel achieves better P@10 and MAP results for finding similar questions compared to methods relying on vector space model and language models. All of the studies mentioned so far are for English language. To the best of our knowledge, there was not any study related to question

retrieval in Arabic CQA systems until SemEval2015 [8] and SemEval2016 [9] workshops. In SemEval2015 workshop, the only task using Arabic CQA Data set is to classify answers for a question as "definitely relevant", "potentially useful", and "irrelevant" [8]. The Arabic data set includes question-answer pairs collected from Fatwa website [7], which contains religious questions about Islam. We also construct our data set using the same website but in SemEval2015 the dataset is biased towards shortest questions and answers as it is noted in [8]. Our dataset is larger (5,000 questions compared to 1,730) and more representative as we did not put any constraint on question and answer lengths. In the next year's challenge in the same workshop, medical related question-answer pairs are used as the Arabic dataset. The task was similar to previous year such that for

---

[7]http://fatwa.islamweb.net/

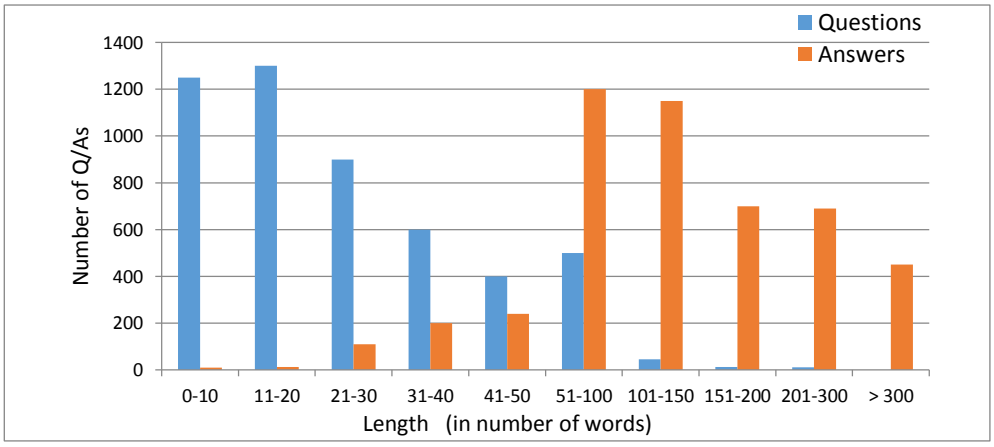



Figure 1: Distribution of questions and answers by their Length in number of words.

a given new question, the task is to re-rank first 30 related Questions retrieved by a search engine. Our study is similar to this task but we compare effectiveness of three different baseline retrieval models (Vector space model, BM25, and Language model). Furthermore, we investigate the effect of two different stemmers for the Arabic language for similar Question retrieval task.

## 3. DATASET

Our dataset consists of 5,000 questions constructed from the Fatwa website. Task 3 in SemEval2015 [8] workshop also use 1,730 questions from the same website as the dataset. This website contains questions related to Islamic religion from regular users and these questions are answered by a group of scholars in this field. Therefore, there is exactly one answer for each question. We choose 50 questions randomly as our test set and used (random) 30 of them in our experimental results. Figure 1 shows the distribution of questions and answers by length. Average number of words in a question is 27 and average number of words in an answer is 167. As it can be seen from the figure, answers are very long compared to questions. There are153 answers with more than 500 words.

## 4. RETRIEVAL MODELS

### 4.1 Vector Space Model

Vector space model is a well known retrieval model proposed in one of the pioneering work by Salton et al. [11]. Documents and queries are represented as vectors in a multidimensional space and cosine similarity is computed. Document and queries are

generally represented as a bag of Words and different weighting approaches are used as term frequency-inverse document frequency (TF-IDF) being the most popular.

## 4.2 Probabilistic Retrieval Models

### 4.2.1 BM25

BM25 is a probabilistic ranking function that relies on inverse document frequency of query terms, term frequency and document length normalization. BM25 score of a document *d* for a query *q* is computed as follows: [8]

$$BM25(d,q) = \sum_{i=1}^{n} IDF(qi) \frac{f(qi,d)(k1+1)}{f(qi,d)+k1(1-b+b\frac{|d|}{avgd1})}$$

(1)

[8]https://en.wikipedia.org/wiki/Okapi_BM25

In this equation, *f (qi, d)* is the frequency of query term *q* In document, *d*, |*d*| denotes the document length and avgdl is the average document length. IDF (*qi*) is inverse document frequency of query term *qi* .Parameters, *k1* and *b*, control the effect of term frequency and document length normalization, respectively.

### 4.2.2 Language Model

Language models are based on probability distributions for words. Query likelihood language model is used for ranking documents for a given query. A language model is constructed

for each document in the collection. The logic behind this model is to compute the probability of producing /generating the query terms under each document language model. For a given document d, its document language model $Ld$ is used to compute the probability of generating the query $q$. Under unigram language model, this probability $P\ (q \mid Ld\ )$ is computed by the following equation.

$$\prod_i P\ (q \mid Ld\ ) = (1-\lambda)Pml\ (qi \mid Ld\ ) + \lambda P(qi \mid LC\ ) \quad (2)$$

In this equation, $Pml$ is the maximum likelihood of qi given the document language model $Ld$ . $P(qi \mid LC)$ is the probability of term $qi$ given the collection language model $LC$ and $\lambda$ Controls the interpolation between document and collection Language models in order to avoid zero probabilities.

## 5. EXPERIMENTS

We use Lucene library[9] as the implementation of vector Space model and BM25 ranking function. Lucene's default Similarity function uses a customized version of vector space model. We use BM25 similarity for ranking with default parameters $k1$=1.2, b=0.75.We applied the Arabic Analyzer available in this library to perform stopword elimination (using120 stopwords).This analyzer also applies Light stemming for Arabic [7]. We implement our query Likelihood language model with collection smoothing with $\lambda$=0.05. Stopword elimination and stemming is also performed in order to be comparable with vector space model and BM25 ranking functions. As an alternative stemming

algorithm, we also apply Khoja Stemmer [10] which uses 168 stopwords (we also experiment with Lucene stopwords here but it causes very minor/negligible changes in results). This stemmer removes diacritics, prefixes and suffixes in order to extract the root word. Note that we use each question in our test set as the query. Each question-answer pair in our dataset is represented as a single document in our experimental setting. We compute top-20 results for each question in our test set using three different retrieval models with light stemming. We repeat this experiment using the Khoja Stemmer [10]. Stopword elimination and stemming are also applied for each question in our test set. All results are evaluated using 3-scale relevance ratings given in Table 1 by a native Arabic speaker. We report NDCG and average relevance (AvgRel) metrics for ranks@5, @10, and@20. We combine top-20 results of a question for all cases and construct an ideal ranking for top-20 among these results to compute NDCG. Figure 2 shows comparison of three different retrieval models for NDCG@5 and

[9]https://lucene.apache.org

Table 1: Relevance Levels

| Relevance Score | Label |
|---|---|
| 2 | Highly relevant |
| 1 | Partially relevant |
| 0 | Irrelevant |

AvgRel@5. LM stands for language model and Lucene represents vector space model. It is seen that BM25 ranking Function achieves the highest NDCG@5 (0.726) and the highest

AvgRel@5 (1.38) with Light stemming. Vector space Model is better than unigram language model in both metrics. We see that Khoja Stemmer has inferior results compared to light stemming in all three retrieval models. We run significance tests to see whether differences are statistically significant or not. According to paired t-test results, the difference between BM25 and two other retrieval models in NDCG@5 are statistically significant (with p-values 0.049 and 0.006 for Lucene and LM, respectively.) Similarly, Vector space model achieves statistically significant improvement over LM with p-value 0.01. On the other hand, the difference between BM25 and vector space model inAvgRel@5 is found to be not statistically significant, while we have significant improvements for BM25 over LM with p-value 0.01 and for vector space model over LM with p-value 0.02. We also test the difference between Light and Khoja stemmers.

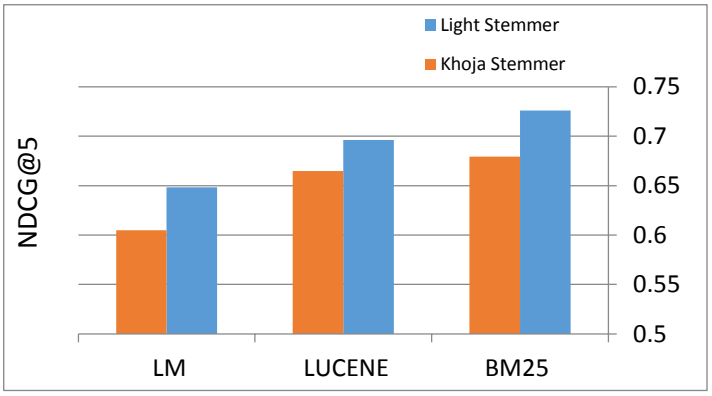
Even though Light stemmer achieves superior results for all retrieval models, the difference in NDCG@5 and AvgRel@5 With Khoja stemmer is not found to be statistically significant. Figure 3 shows NDCG@10 and AvgRel@10 results for three different retrieval models, again with Light and Khoja stemmers. The results are generally slightly lower compared to the results at rank 5 in Figure 2. However, trends are the same such that BM25 achieves the top NDCG and average relevance and LM has the lowest results. The results at rank 20 is similar with the same trends and are not shown here for space limitations.

## 6. CONCLUSION

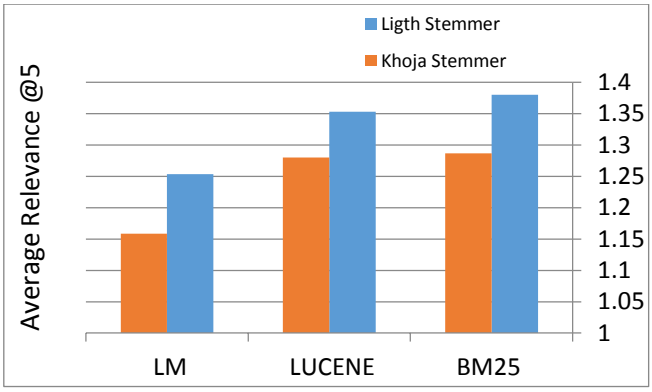
Community Question Answering services are valuable resources for people seeking answers to their natural language questions.

These social services may compensate for long and natural language queries in which web search engines have difficulties. Even though Arabic language and the number of Arab Internet users constitute a major portion of the Web population, there are limited studies about Arabic CQA systems. Recently, SemEval workshops in 2015 and 2016, put initial efforts towards this goal.

In this study, we compare effectiveness of three different Retrieval models for finding similar questions in Arabic CQA systems. We also investigate the effect of two different stemmers on this task. Our results indicate that BM25 ranking function with Light stemmer achieves the highest NDCG and Average Relevance at ranks 5, 10, and 20. Therefore, we suggest that this retrieval model should be used as the baseline when comparing a new method for this task. As a future work, we plan to work on finding similar questions problem using syntactic and semantic features in addition to lexical features.

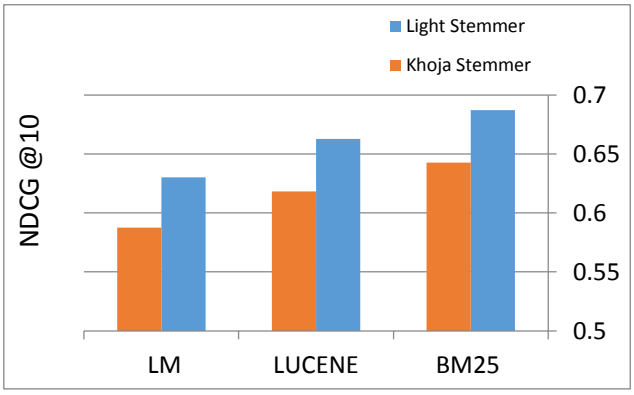



(a)

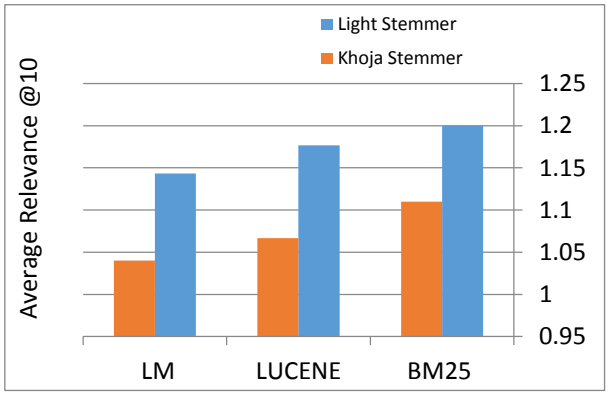



(b)

Figure2: Comparison of retrieval models for a)NDCG@5, b) Average .Relevance.at rank5.




(a)




(b)

Figure3: Comparison of retrieval models for a)NDCG@10, b) Average .Relevance.at rank10.

## 7. REFERENCES

[1]. R.D.Burke,K.J.Hammond,V.A.Kulyukin,S.L. Lytinen,N.Tomuro , and S.Schoenberg.Question

Answering from frequently asked questionfiles: Experiences with the FAQ FINDER system.AI

Magazine, 18(2):57–66,1997.

[2] X.Cao,G.Cong,B.Cui,C.S.Jensen,and C.Zhang.The use of categorization information in language

Models for question retrieval.In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM'09, pages 265–274, New York, NY,USA,2009.ACM.

[3] K.DarwishandW.Magdy.Arabicinformation retrieval. *Found. Trends Inf. Retr*., 7(4):239–342,Feb.2014.

[4] H.Duan, Y. Cao, C.Lin,and Y. Yu. Searching questions by identifying question topic and question

focus. *In ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational*

*Linguistics, June 15-20,2008,Columbus,Ohio,USA*, pages 156–164, 2008.

[5] J. Jeon , W .B. Croft, and J. H. Lee. Finding similar

Questions in large question and answer archives. *In Proceedings of the14th ACM International Conference On Information and Knowledge Management,* CIKM '05, pages 84–90, New York, NY, USA, 2005.ACM.

[6] V. Jijkoun and M. deRijke . Retrieving answers from frequently asked questions pages on the web.*In Proceedings of the 14th ACM International Conference On Information and Knowledge Management*, CIKM '05, pages 76–83, NewYork, NY, USA, 2005. ACM.

[7] L. S. Larkey, L. Ballesteros, and M.E. Connell. *Arabic Computational Morphology :Knowledge -based and Empirical Methods*, chapter Light Stemming for Arabic Information Retrieval, pages 221–243. Springer Netherlands, Dordrecht, 2007.

[8] P. Nakov, L.M`arquez, W. Magdy, A. Moschitti, J.Glass, and B.Randeree.Semeval-2015 task3: Answer selection in community question answering. In *Proceedings of the 9th International Work shop on Semantic Evaluation (*SemEval2015), pages269–281, Denver, Colorado, June2015. Association for Computational Linguistics.

[9] P. Nakov, L. M`arquez, W. Magdy, A. Moschitti, J.Glass, and B. Randeree. SemEval-2016 task3: Community question answering. In *Proceedings of the 10th International Work shop on Semantic Evaluation*, SemEval'16, SanDiego, California, June2016. Association for Computational Linguistics.

[10] R. G. S. Khoja. Stemming Arabic text. Technical report, Computing Department, Lancaster University, 1999.

[11] G. Salton, A. Wong, and C.S. Yang. A vector space Model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov.1975.

[12] J. Wang, Z. Li, X. Hu, and B. Hu. A novel composite Kernel for finding similar questions in CQA services. In *Web-Age Information Management, 11th International Conference, WAIM 2010, Jiuzhaigou, China, July 15-17, 2010. Proceedings*, pages 608–619, 2010.

[13] G. Zhou, L. Cai, J. Zhao, and K. Liu. Phrase-based Translation model for question retrieval in community Question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* –Volume 1, HLT'11, pages 653–662, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.