ISSN 2277-3061
Volume 16 Number 2
INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

A review on the Detection of Missing Content Queries in FAQ Retrieval Systems

Edwin Thuma[1], Moemedi Lefoane[2], Gontlafetse Mosweunyane[3]
[1]Computer Science Department, University of Botswana, Gaborone,Botswana
thumae@mopipi.ub.bw
[2]Computer Science Department, University of Botswana, Gaborone,Botswana
moemedi.lefoane@mopipi.ub.bw
[3]Computer Science Department, University of Botswana, Gaborone,Botswana
mosweuny@mopipi.ub.bw

ABSTRACT

When developing an automated FAQ retrieval system, the information supplier constructs question candidates in advance using their own knowledge. Then they answer these question candidates to create question-answer pairs to use in the FAQ retrieval system. However, these question-answer pairs will not always satisfy the users' information needs. When there is no relevant question–answer pair to a users' query, such a user may submit various query reformulations browsing over the long results list and may abandon the search before their information need has been satisfied. Such users many never return to use the system again because of the inability of the system to return relevant question-answer pairs to their query. In order to alleviate this, modern automated FAQ retrieval systems use a Missing Content Query (MCQ) detection subsystem to detect those queries that do not have the relevant question–answer pair. In this article we conduct a review of the different approaches proposed in the literature for detecting these MCQs. In particular, we provide a comprehensive review of the different systems that deployed the binary classification approach, the thresholding approach and the hybrid approach in the detection of MCQs. Moreover, we describe the strength and weaknesses of each approach.

Indexing terms/Keywords

Frequently Asked Questions, Missing Content Queries.

Academic Discipline And Sub-Disciplines

Computer Science

SUBJECT CLASSIFICATION

Information Retrieval

TYPE (METHOD/APPROACH)

Survey/Review

1. INTRODUCTION

Missing Content Queries (MCQs) [16], also known as Out-of-Domain (OOD) queries [1], are those queries for which there are no relevant FAQ documents (question – answer pairs) in the FAQ document collection (collection being searched) [1,16]. According to Sneiders [13,14], MCQs together with other factors such as vocabulary mismatch problem, where the terms used in the users' query do not match the terms used in the FAQ document collection, can force users to abandon their search before their information need has been satisfied because of the inability of the system to retrieve relevant FAQ documents. Moreover, Sneiders [13] suggests that when developing FAQ retrieval systems, detecting MCQs should be taken into consideration in order to avoid users iterating with a system longer when there is no relevant FAQ document in the FAQ document collection. More recent attention has focused on the detection of these MCQs. In particular, the Forum for Information Retrieval Evaluation (FIRE) organized the SMS Based FAQ retrieval task from 2011 to 2013 to advance research in the FAQ retrieval domain and the detection of MCQs in FAQ retrieval systems [1]. Several approaches have been proposed in the literature for detecting these MCQs. In particular, Yom-Tov et al. [16], Thuma et al. [15] and Leveling [10] proposed a binary classifier to detect these queries that do not have the relevant FAQ documents in the FAQ document collection. In their approaches, they deployed binary classifiers using different feature set to detect these MCQs. These classifiers yielded varying performance in terms of classification accuracy and ROC area (AUC). The ROC area signifies the overall ability of the classifier to identify *MCQs* and *non −MCQs*. The best classifier has an area of 1.0 and a classifier with an area of 0.5 or lower is considered ineffective. In addition, several teams that participated in the Forum for Information Retrieval Evaluation (FIRE) SMS-Based FAQ retrieval task deployed the thresholding technique to detect theses MCQs [5, 11, 12]. The participating teams devised different heuristics to come up with a threshold value to determine whether a user query is a missing content query or a non-missing content query (non-MCQ). Other teams that participated in the FIRE SMS-Based FAQ retrieval task deployed a hybrid approach, where a combination of approaches was deployed to detect these MCQs. The main aim and objective of this article is to provide a comprehensive review of the approaches proposed in the literature for detecting MCQs. In our review, we describe the strengths and weaknesses of each approach.

The rest of this article is organized as follows: In Section 2, we provide a comprehensive review of the different approaches for detecting MCQs using a binary classifier. In Section 3, we provide a review of the different approaches for

detecting MCQs using the thresholding technique, this is followed by the approaches that use the hybrid technique in Section 4. We provide concluding remarks in Section 5.

2. Binary Classification Approach

The study of the detection of missing content queries (MCQs) was first carried out by Yom-Tov et al. [16] in their investigation of the applications of query difficulty predictors. The intuition behind this is that queries that are predicted to be difficult are likely to be MCQs and those that are predicted to be easy are likely to be non-MCQs. In their investigation, they artificially created 166 MCQs by deleting the relevant documents for the 166 queries form the TREC-8 collection that had 200 description-part queries and 200 title-part queries. They then trained a tree-based estimator to classify MCQs and non-MCQs using the complete set of 400 queries. In their experiment, they used a query difficulty predictor trained by analyzing the overlap between the results of the full query and the results of its sub-queries to pre-filter easy queries before identifying MCQs with a tree-based classifier. Their results suggest that identifying MCQs can be improved by combining the MCQs classifier with a query difficulty estimator. The main strength with this approach is that the query difficulty predictor first pre-filter easy queries so that they are not grouped with MCQs.

Similarly, Leveling [10] viewed the detection of MCQs as a classification problem. In their approach, they trained an IB1 classifier as implemented in TiMBL [3] using query difficulty predictors which were essentially numeric features generated during the retrieval phase on the training data (FIRE2011 SMS-Based FAQ retrieval monolingual English data) to distinguish between MCQs and non−MCQs. The aforementioned features were comprised of the result set size for each query, the raw BM25 FAQ document scores for the top five documents (5 features), the percentage difference of the BM25 FAQ document scores between the consecutive top 5 documents (4 features), the normalised BM25 FAQ document scores for the top five retrieved documents (5 features) and the term overlap scores for the SMS query and the top 5 retrieved FAQ documents (5 features). Their approach essentially yielded a binary classifier that can determine whether a query is MCQ or non−MCQ.

This approach is much simpler compared to the hybrid approach that we discuss later in Section 4.0, which was proposed by Hogan et al. [9] because it relies on a single classifier instead of relying on several classifiers. Leveling evaluated this approach using a leave-one-out validation approach and reported a marked improvement in the detection of MCQs when the binary classifier is trained on features collected during retrieval with no stopword removal. In their empirical evaluation, they reported a highest classification accuracy of 85.6 for MCQs and 75.1% for non-MCQs. A possible explanation to this high classification accuracy is that the retrieval scores of the top 5 retrieved FAQ documents are a good indicator of their relevance to the query. Essentially, if the retrieval scores are very low, it is highly likely that the query is a MCQ. Similarly, if the retrieval scores are high, it is likely the query is a non-MCQ. Also, the percentage difference of the BM25 FAQ document scores between the consecutive top 5 document can act as a good discriminator of MCQs and non-MCQs. Intuitively, a small percentage difference in the retrieval scores may indicate a cohesive retrieved result set, which may indicate that all the retrieved documents are relevant to the query, hence a query is a non-MCQ. Similarly, a bigger percentage difference in the retrieval scores may indicate a very diverse retrieved result set, which may indicate that all the retrieved documents cover different aspect of the query, hence none of the retrieved FAQ documents is relevant to the query and the query is a MCQ.

One study by Thuma et al. [15] examined different feature sets in order to determine the best combination of features that can be used to build a model that would yield the highest classification accuracy when classifying MCQs and non-MCQs. In order to be able to make a general conclusion, they used two different datasets in their empirical evaluation. The first dataset used was a collection of HIV/AIDS FAQ documents and a query log of HIV/AIDS related MCQs and non-MCQs collected in Botswana over a period of three months. The other dataset was a collection of FAQ documents from the Forum for Information Retrieval Evaluation (FIRE2012) English monolingual SMS-Based FAQ retrieval training data and the associated query log (SMS queries). They empirically evaluated different binary classifiers, which were built using three different feature sets. The first feature set were represented by a vector of attributes representing word count information from the text contained in the query string. The second feature set were created using the approach we saw earlier, which was proposed by Leveling [10]. The third feature set were represented by query difficulty predictors, some of which were introduced by Hogan et al [9]. We provide a list of these query difficultly predictors later in Section 4.0.

In their empirical evaluation, they reported that the best set of features for building a binary classifier for detecting MCQs is a vector of attributes representing word count information from the text contained in the query string. Furthermore, they reported that the classification accuracy of such a classifier can be improved by combining this feature set with the other feature sets proposed by Leveling [10] and Hogan et al. [9]. Their results generalized well across the two different datasets and different classifier.

3.0 Thresholding Approach

In addition to the binary classification approach, several researchers deployed the thresholding technique in order to detect MCQs. In particular, Shivhre [12] used the total sum of the retrieval scores of the top 5 retrieved documents as a threshold to decide whether queries are MCQs or non-MCQs. The retrieval scores for the top 5 retrieved FAQ documents were scaled between 0 and 1. In their approach, if all the matching question parts of the top 5 retrieved FAQ documents had a total score below that set threshold, the SMS query was considered a MCQ. The approach proposed by Shivhre [12] yielded fairly reasonable results as they were able to accurately detect 72.5% MCQs and 54% of the non-MCQs using the FIRE2011 SMS-Based FAQ retrieval task dataset [1]. Since the threshold values was set based on the score of the question part only, this method is likely to perform well for the MCQs because fewer FAQ documents are less likely to contain the words used in the query in their question part of the FAQ document. However, the approach may perform

poorly in the non-MCQs as evidenced by the 54% detection accuracy because there may be a term mismatch problem between the user query and the relevant FAQ document in the FAQ document collection.

In the same vein, Gupta [5] experimentally set a threshold to determine whether to mark an SMS query as a MCQ or a non-MCQ. Any user query with a retrieval score of the top ranked FAQ document, which was less than the *No. of Tokens (tokens in the SMS query)*C* was considered a MCQ. The threshold value of *C* was obtained experimentally and was set at 1.15. The simple heuristic devised by Gupta for identifying MCQs and the non-MCQs also performed fairly well on the FIRE2011 SMS SMS-Based FAQ retrieval task as they reported that 56.5% MCQs and 59.3% non-MCQs were accurately detected. The main weakness with the approach proposed by Gupta is that it places more emphasis on the query length. With this approach, if a very long MCQ is made up of terms in the vocabulary of the corpus being searched, the top ranked FAQ document is likely to have retrieval score that is more than the *No. of Tokens (tokens in the SMS query)*C.* This will result in this long MCQ query being wrongly classified as a non-MCQ.

Shaikh et al. [11] also set a threshold value for the retrieval scores of the question part in order to determine if a query is a MCQ or a non-MCQ. In their work, they did not describe how they determined this threshold. However, they reported a reasonably high detection accuracy of 74% for the non-MCQs and 84.7% for the MCQs. The main strength with this approach proposed by Shaikh et al. is that it places more emphasis on the vocabulary used in the query. Hence, it is likely to work well for on-topic MCQs (MCQs related to the FAQ document collection) and perform poorly for off-topic MCQs (MCQs not related to the FAQ document collection). If the query contains a lot of terms not in the vocabulary of the FAQ document collection being searched, that query is likely to be identified as a MCQ. On the other hand, if a query contains a lot of terms that are in the vocabulary of the FAQ document collection being searched, that query is likely to be identified as a non-MCQ.

4.0 Hybird Approach

Hogan et al. [9] deployed a hybrid approach in the detection of MCQs. In particular, they deployed the binary classification approach and the thresholding approach to identify MCQs. This was achieved by combining 3 different lists of MCQs generated through three different approaches and then applied a simple majority voting approach to identify MCQs. The first list of candidate MCQs was generated using an approach proposed by Ferguson et al. [4] for determining the number of relevant documents to use for query expansion. In this approach, a score for each query was produced based on the inverse document frequency (IDF) component of the BM25 score for each query without taking into consideration the term frequency and the document length. First, the maximum score possible for any document was calculated as the sum of the IDF scores for all the query terms. Following this approach, documents without all the query terms will have a score less than the maximum score. A threshold was then used to determine if a query should be added to the list of candidate *MCQs*. They added queries that had all their document scores below 70 % of the maximum score to this list.

The second list of candidate *MCQs* was generated by training a k-nearest neighbour classifier to identify *MCQs* and *non −MCQs*. The features used to train this classifier included query performance estimators (Average Inverse Collection Term Frequency (AvICTF) [8], Simplified Clarity Score (SCS)) [7], the derivates of the similarity score between collection and query (SumSCQ, AvSCQ, MaxSCQ) [17], result set size and the un-normalised BM25 document scores for the top five documents. The third list of candidate *MCQs* was generated by simply counting the number of term overlaps for each incoming query and the highest ranked document (For example, if the query consists of more than one term and had only one term in common with the document, that query was marked as *MCQ*). Hogan et al. used the held-out training data to evaluate their approach and they concluded that combining the three lists of candidate *MCQs* through a simple majority voting yielded better results. In their empirical evaluation, they reported a classification accuracy of 85.6 for MCQs and 70.2% for non-MCQs with an overall classification accuracy of 78 % on the FAQ SMS training data using a leave-one-out validation.

One major drawback with the hybrid approach proposed by Hogan et al. is that two of the approaches proposed for detecting MCQs rely heavily on the ranking function used by the FAQ retrieval system. The classification accuracy may degrade in performance when the user query is semantically similar to the relevant FAQ documents in the collection but lexically different to that FAQ documents. The retrieval scores for the top 5 FAQ documents ranked after submitting such a query are likely to be below 70% of the maximum score possible resulting in the query being added to the list of MCQs. Similarly, the lexical difference between the relevant FAQ documents and the user query can result in some queries being added to the MCQs list when only one term in the top relevant FAQ document overlaps with one term in the query.

5.0 Discussion and Conclusion

In this article, we surveyed three different approaches for the detection of MCQs in FAQ retrieval systems. These approaches are the binary classification approach, the thresholding approach and the hybrid approach. In common, binary classification approaches produce the best results compared to the thresholding techniques. One plausible explanation for this is that the binary classification approaches are leveraging several feature sets to help in identifying MCQs. Indeed, seeking to achieve the highest classification accuracy – the hybrid approach has been shown to outperform all the other approaches at the FIRE SMS-Based FAQ retrieval task since it leveraged information from multiple sources to detect MCQs. In contrast, using the thresholding techniques has shown some inconsistent results with some approaches producing above average classification accuracy while other approaches produced just above average classification accuracy.

REFERENCES

1. Contractor, D., Subramaniam, L., Deepak, P., and Mittal, A. (2013). Text Retrieval Using SMS Queries: Datasets and Overview of FIRE 2011 Track on SMS-Based FAQ Retrieval. In Multilingual Information Access in South Asian Languages, volume 7536 of Lecture Notes in Computer Science, pages 86–99, Berlin, Heidelberg. Springer-Verlag.
2. Cronen-Townsend, S., Zhou, Y., and Croft, W. (2002). Predicting query performance. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02, pages 299–306, New York, NY, USA. ACM.
3. Daelemans,W., Zavrel, J., Sloot, K., and Bosch, A. (2002). TiMBL: Tilburg Memory-Based

   Learner - version 4.3 - Reference Guide.
4. Ferguson, P., O'Hare, N., Lanagan, J., Smeaton, A., McCarthy, K., Phelan, O., and Smyth, B. (2011). CALRITY at the TREC 2011 Microblog Track. In Proceedings of the 20th TREC Conference, pages 1–6, Gaithersburg, Md., USA. Text REtrieval Conference (TREC).
5. Gupta, A. (2013). Mapping SMSes to Plain Text FAQs. In Multilingual Information Access in South Asian Languages, volume 7536 of Lecture Notes in Computer Science, pages 157–162, Berlin, Heidelberg. Springer-Verlag.
6. Hauff, C., Murdock, V., and Baeza-Yates, R. (2008). Improved query difficulty prediction for the web. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, pages 439–448, New York, NY, USA. ACM.
7. He, B. and Ounis, I. (2004). Inferring Query Performance Using Pre-retrieval Predictors. In Proceedings of the String Processing and Information Retrieval, pages 43–54, Berlin, Heidelberg. Springer-Verlag.
8. He, B. and Ounis, I. (2006). Query Performance Prediction. Information Systems, 31(7):585–594.
9. Hogan, D., Leveling, J., Wang, H., Ferguson, P., and Gurrin, C. (2011). DCU@FIRE 2011: SMS-based FAQ retrieval. In FIRE 2011, 3rd Workshop of the Forum for Information Retrieval Evaluation, 2-4 December, IIT Bombay, pages 34–42.
10. Leveling, J. (2012). On the Effect of Stopword Removal for SMS-Based FAQ Retrieval. In Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems, pages 128–139, Berlin, Heidelberg. Springer-Verlag.
11. Shaikh, A., Jain, M., Rawat, M., Shah, R., and Kumar, M. (2013). Improving accuracy of sms based faq retrieval system. In Multilingual Information Access in South Asian Languages, volume 7536 of Lecture Notes in Computer Science, pages 142–156, Berlin, Heidelberg. Springer-Verlag.
12. Shivhre, N. (2013). SMS Based FAQ Retrieval. In Multilingual Information Access in South Asian Languages, volume 7536 of Lecture Notes in Computer Science, pages 131–141, Berlin, Heidelberg. Springer-Verlag.
13. Sneiders, E. (1999). Automated FAQ Answering: Continued Experience with Shallow Language Understanding. Question Answering Systems. In Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium, pages 97–107, California, USA. AAAI Press.
14. Sneiders, E. (2009). Automated FAQ Answering with Question-specific Knowledge Representation for Web Self-service. In Proceedings of the 2Nd Conference on Human System Interactions, pages 295–302, Piscataway, NJ, USA. IEEE Press.
15. Thuma, E. Rogers, S. and Ounis, I (2014). Detecting Missing Content Queries in an SMS-Based HIV/AIDS FAQ Retrieval System, In Advances in Information Retrieval, volume 8416 of Lecture Notes in Computer Science, pages 247-259, Berlin, Heidelberg. Springer-Verlag.
16. Yom-Tov, E., Fine, S., Carmel, D., and Darlow, A. (2005). Learning to Estimate Query Difficulty: Including Applications to Missing Content Detection and Distributed Information Retrieval. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA. ACM.
17. Zhao, Y., Scholer, F., and Tsegay, Y. (2008). Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, pages 52–64, Berlin, Heidelberg. Springer-Verlag.