Exploiting Query Logs and Field-Based Models to Address Term Mismatch in an HIV/AIDS FAQ Retrieval System

Edwin Thuma, Simon Rogers, and Iadh Ounis

School of Computing Science,
University of Glasgow,Glasgow, G12 8QQ, UK
thumae@dcs.gla.ac.uk,{simon.rogers,iadh.ounis}@glasgow.ac.uk

Abstract. One of the main challenges in the retrieval of Frequently Asked Questions (FAQ) is that the terms used by information seekers to express their information need are often different from those used in the relevant FAQ documents. This lexical disagreement (aka term mismatch) can result in a less effective ranking of the relevant FAQ documents by retrieval systems that rely on keyword matching in their weighting models. In this paper, we tackle such a lexical gap in an SMS-Based HIV/AIDS FAQ retrieval system by enriching the traditional FAQ document representation using terms from a query log, which are added as a separate field in a field-based model. We evaluate our approach using a collection of FAQ documents produced by a national health service and a corresponding query log collected over a period of 3 months. Our results suggest that by enriching the FAQ documents with additional terms from the SMS queries for which the true relevant FAQ documents are known and combining term frequencies from the different fields, the lexical mismatch problem in our system is markedly alleviated, leading to an overall improvement in the retrieval performance in terms of Mean Reciprocal Rank (MRR) and recall.

Keywords: Frequently Asked Question, Term Mismatch, Query Logs, Field-Based Model.

1 Introduction

We have developed an Automated SMS-Based HIV/AIDS FAQ retrieval system that can be queried by users to provide answers on HIV/AIDS related questions. The system uses, as its information source, the full HIV/AIDS FAQ question-answer booklet provided by the Ministry of Health (MOH) in Botswana for its IPOLETSE[1] call centre. This FAQ question-answer booklet is made up of 205 question-answer pairs organised into eleven chapters of varying sizes. For example, there is a chapter on "Nutrition, Vitamins and HIV/AIDS" and a chapter on "Men and HIV/AIDS". Below is an example of a question-answer pair entry that can be found in Chapter Eight, "Introduction to ARV Therapy":

[1] http://www.hiv.gov.bw/content/ipoletse

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 77–89, 2013.

Question : What is the importance of taking ARV therapy if there is no cure for AIDS?

Answer : Although ARV therapy is not a cure for AIDS, it enables you to live a longer and more productive life if you take it the right way. ARV therapy is just like treatment for chronic illnesses such as diabetes or high blood pressure.

For the remainder of this paper, we will refer to a question-answer pair as the FAQ document and the set of all 205 FAQ documents as the FAQ document collection. The users' SMS messages will be referred to as queries.

One key problem in this domain is that there will often be term mismatch between the queries from the users and the relevant FAQ documents [18,19]. For example, the user's query: *"Is HIV/AIDS gender based to some extent?"* and the FAQ document: *"Does HIV/AIDS affect women differently from men? No, the virus affects both men and women in exactly the same way i.e. by making the immune system weak, so that it cannot fight off other illnesses"* are semantically similar but lexically different. This term mismatch between the user's query and the relevant FAQ document may result in a less effective ranking by a retrieval system that relies on keywords matching in its weighting model [3].

To solve this term mismatch problem between the users' queries and the relevant FAQ documents in the FAQ document collection, query log clustering is often used [6]. Earlier work by Kim et al. [6,7] suggests that a good clustering of query logs can markedly reduce the term mismatch problem that arises in an FAQ retrieval system, thus improving the overall retrieval performance. Another approach that is often used in the Information Retrieval (IR) community to alleviate the term mismatch problem is query expansion. Various authors have reported mixed results [3,20]. For example, Voorhees [20] did not show any significant improvement if queries are expanded with terms from WordNet. On the other-hand, Fang [3] has shown significant performance improvement when hand-crafted lexical resources are used for query expansion.

In this paper, we aim to tackle this term mismatch problem in an SMS-based HIV/AIDS FAQ retrieval system by enriching the traditional FAQ document representation (Question and Answer) using terms from a query log, which are added as a separate field in a field-based model [10,16]. Our main contribution is to demonstrate that enriching the FAQ documents (Question and Answer Fields only) with additional terms from potential users of the FAQ system can alleviate the term mismatch problem that arises in our FAQ retrieval system. This will be measured by an increase in recall. Recall is the fraction of relevant documents to the query that are retrieved. We thoroughly evaluate our approach using the aforementioned HIV/AIDS question-answer booklet provided by the Ministry of Health in Botswana as our information source and a corresponding query log collected in Botswana over a period of 3 months.

The rest of this paper is organised as follows: In Section 2 we survey related work, followed by a description of our enrichment strategies in Section 3. In Section 4 we describe how the SMS queries were collected and analysed. Then we describe our experimental setting in Section 5, followed by the experimental results in Section 6 and the conclusions in Section 7.

2 Related Work

Earlier FAQ retrieval systems [4,18,21] relied on knowledge bases to alleviate term mismatch between the query and the relevant FAQ documents. For example, in the system proposed by Sneiders [18], each FAQ is analysed and annotated with three keywords types: required keywords, optional keywords and irrelevant keywords. For each user query, the system retrieves and ranks the relevant FAQs according to the three keyword types. The system rejects the match between the user's query and an FAQ document in the collection if there is at least one required keyword missing in the user's query. It is worth noting that these early representative systems rely on knowledge bases that require a lot of time to construct whenever new FAQs are added to the collection or the application domain changes.

Jeon et al. [5] and Xue et al. [22] proposed a translation based retrieval model that uses the similarity between answers of lexically different but semantically similar questions in community based question-answer archives to learn translation probabilities. They used the learned translation probabilities to search semantically similar questions and their results suggest that their approach outperforms other baseline retrieval models: the vector space model with cosine similarity, the Okapi BM25 model and the query-likelihood language model. The approach proposed by Jeon et al. and Xue et al. shows promising results for a large collection of question-answer archives. However, their approach may not work in our HIV/AIDS FAQ retrieval system because it uses a small fixed dataset of question-answer pairs (205). Learning good translation probabilities might be difficult for such a small dataset.

Kim et al. [6] on the other-hand proposed a more adaptable approach that uses query logs as knowledge sources to solve the term mismatch problem in an FAQ retrieval system. Their system called FRACT is made up of two subsystems, a query log clustering system and a cluster based retrieval system. The query log clustering system considers each FAQ as an independent category and it periodically collects and refines the users' query logs that are then classified into each FAQ category by using a vector similarity in the latent semantic space. FRACT uses the clustered query logs to associate every users' question to the relevant cluster of FAQs and ranks and return a list of FAQs based on the similarity with the cluster.

More recently Moreo et al. [11], introduced a new method called Minimal Differentiator Expression (MDE). In their approach, they solve the term mismatch problem by using linguistic classifiers that they trained using expressions that totally differentiate each FAQ. They enhance the performance of their system during the life of its operation by continuously training the classifier with new evidence from the users' queries. Their approach although different from our proposed approach also relies on query logs to resolve the term mismatch problem. In their evaluation, they reported that their approach outperformed the cluster based retrieval proposed in [6]. Other approaches that closely resemble our work

are the document expansion approach proposed in [2,17] and the query expansion approach in [1]. The document expansion approach proposed by Billerbeck and Zobel [2] yielded unpromising results and this might be partly due to the fact that the expansion terms were selected automatically without using the actual query relevance judgements. Hence this might have resulted in the wrong terms being used to expand irrelevant documents. In this work, we will rely on the query relevance judgements to avoid linking query terms to irrelevant FAQ documents.

3 FAQ Documents Enrichment Strategies

In Web IR, there is the notion of document fields and this provides a way to incorporate the structure of a document in the retrieval process [16]. For example, the contents of different HTML tags (e.g anchor text,title, body) are often used to represent different document fields [13,16]. Earlier work by [10] has shown that combining evidence from different fields in Web retrieval improves retrieval performance. In this paper, we represent the FAQ document made up of question-answer pairs into a $QUESTION$ and an $ANSWER$ field. We then introduce a third field, $FAQLog$, that we use to add additional terms from queries for which the true relevant FAQ documents are known. We aim to solve the term mismatch problem in our FAQ retrieval system by combining evidence from these three fields.

We will evaluate the proposed approach using two different enrichment strategies. First, we enrich the FAQ documents using all the terms from a query log. In this approach, all the queries from the training set for which the true relevant FAQ documents are known will be added into the new introduced $FAQLog$ field as shown in Table 1. In other words, if an FAQ document is known to be relevant to a query, then this query is added to its $FAQLog$ field. For the remainder of this paper we will refer to this approach as the Term Frequency approach. In the second approach, we will enrich the FAQ documents using term occurrences from a query log. Here, all the unique terms from the training set for which the true relevant FAQ documents are known will be added to the $FAQLog$ field as shown in Table 2. In other words, only new query terms that do not appear in the $FAQLog$ field will be added to that field. For the remainder of this paper we will refer to this approach as the Term Occurrence approach. We will apply field-based weighting models on the enriched FAQ documents using PL2F [10] and BM25F [16].

The main difference between the two enrichment approaches is that the frequencies with which users use some rare terms in specific FAQ documents can be captured if the term frequency enrichment approach is used. For example, under the term frequency approach (Table 1), the term frequencies of the terms *gender* and $infected$ in the FAQLog field are: $gender = 2$ and $infected = 2$. Under the term occurrence approach (Table 2) the term frequencies of these terms are 1 because the query terms under this approach can only be added to this field once even if they appear in many queries. Since, both BM25F and PL2F rely on term frequencies to calculate the final retrieval score of a relevant document given a

Table 1. Enrichment Using Query Term Frequencies

FIELDS	CONTENTS of FIELDS
QUESTION	Does HIV / AIDS affect women differently from men?
ANSWER	No, the virus affects both men and women in exactly the same way i.e. by making the immune system weak, so that it cannot fight off other illnesses.
FAQLog	Is HIV/AIDS gender based to some extent? Between men and women, who are most infected by HIV/AIDS? who are mainly infected male or female? which gender is mostly affected by the disease?

Table 2. Enrichment Using Query Term Occurrence

FIELDS	CONTENTS of FIELDS
QUESTION	Does HIV / AIDS affect women differently from men?
ANSWER	No, the virus affects both men and women in exactly the same way i.e. by making the immune system weak, so that it cannot fight off other illnesses.
FAQLog	is, hiv, aids, gender, based, to, some, extent, between, men, and, women, who, are, most, infected, by, mainly, male, or, female, which, mostly, affected, the, disease

query, our two enrichment strategies will always give different retrieval scores. We will investigate the usefulness of each enrichment approach in Section 5.

4 Collecting and Analysing SMS Queries

85 participants were recruited in Botswana and asked to provide SMS queries on the general topic of HIV/AIDS. Having provided SMS queries, they then used a web-based interface to find the relevant FAQ documents from the FAQ document collection using the SMS queries. This provided us with SMS queries linked to the appropriate FAQ documents in the collection. In total, 957 SMS queries were collected of which 750 could be matched to an FAQ document in the collection. The remaining 207 did not match anything in the collection and investigating how to detect such orphan queries in a real system is a subject for future work. The 750 SMS queries that could be matched spanned 131 of the 205 FAQ documents, leaving 74 FAQ documents with no SMS queries.

We analysed these SMS queries, counting the number of queries that matched each FAQ document. Our analysis shows that the distribution of queries per FAQ document was not spread evenly. There were some FAQ documents that matched more than 20 users' queries. This was more evident on a topic related to the prevention and transmission of HIV and AIDS. Similar findings were also reported by Sneiders [19] who concluded that people who share the same interest tend to ask the same question over and over again. In this paper, we exploit this repetitive nature of the query log by proposing to enrich the FAQ documents with SMS queries for which the true relevant FAQ document is known thus reducing the term mismatch problem in our FAQ retrieval system.

5 Experimental Description

We begin Section 5.1 by describing our experimental settings followed by a description of our experimental investigations and our baseline systems in Section 5.2. We then describe how we created the new enriched FAQ document representation with the query logs followed by a description of how the field weights for the field-based weighting models were optimised in Section 5.3.

5.1 Experimental Setting

For all our experimental evaluation, we used the Terrier-3.5[2] [12], an open source Information Retrieval (IR) platform. All the FAQ documents used in this study were first pre-processed before indexing and this involved tokenising the text and stemming each token using the full Porter [14] stemming algorithm. To filter out terms that appear in a lot of FAQ documents, we did not use a stopword list during the indexing and the retrieval process. Instead, we ignored the terms that had low Inverse Document Frequency (IDF) when scoring the documents. Indeed, all the terms with term frequency higher than the number of the FAQ documents (205) were considered to be low IDF terms. Earlier work in [9] has shown that stopword removal using a stopword list from various IR platforms like Terrier-3.5 can affect retrieval performance in SMS-Based FAQ retrieval. The normalisation parameter for BM25 was set to its default value of $b = 0.75$. For BM25F, the normalization parameter of each field was also set to 0.75 and these were $(b.0 = 0.75, b.1 = 0.75, b.2 = 0.75)$, representing the normalisation parameters for the $QUESTION$, $ANSWER$ and $FAQLog$ fields respectively. For PL2, the normalisation parameter was set to its default value of $c = 1$. For PL2F, the normalisation parameter for each field was set to $(c.0 = 1.0, c.1 = 1.0, c.2 = 1.0)$, representing the $QUESTION$, $ANSWER$ and $FAQLog$ fields respectively.

5.2 Experimental Investigation and Our Baseline Systems

In this study, we will investigate the following experiments:
EXV1: In this experiment, we are testing our proposed enrichment strategies. This was achieved by comparing the retrieval performance in terms of MRR and recall on the enriched collections of FAQ documents and a collection of non-enriched FAQ documents. We describe how the FAQ documents were enriched using the training set in the next section. A description of how we split the SMS query log into training and testing sets is also provided in the next section. To carry out this investigation, we used the retrieval settings described in Section 5.1. We built an index for each enriched collection of FAQ documents separately using the three fields ($QUESTION$, $ANSWER$ and $FAQLog$) so that we can use field-based weighting models such as BM25F [16] and PL2F [10] for retrieval (60 indices in total). As a baseline, we created two different indices of the original FAQ documents (non-enriched FAQ documents) using the two fields ($QUESTION$ and $ANSWER$). In the first index, we indexed the questions (Q) only and in the second index, we indexed both the question and answer (Q and A). For each index of the enriched FAQ documents, we used the associated testing set to make two runs using BM25F and PL2F as our weighing models. For each index of the non-enriched FAQ documents, we also used the 10 testing sets to make 2 runs using BM25F and PL2F as our weighting models. For this investigation, all the field weights parameters were intentionally set to 1 $(w.0 = 1, w.1 = 1, w.2 = 1)$, where $(w.0, w.1$ and $w.2)$ represent the

[2] `http://terrier.org/`

$QUESTION$, $ANSWER$ and $FAQLog$ field weights respectively. The field-based weighting models $BM25F$ and $PL2F$ are known to yield the same retrieval scores as their non field-based counterpart (BM25 and PL2 respectively) when all field weights are set to 1. To illustrate this, we also made two runs on the indexed collections with each testing set using BM25 and then PL2 as our weighting models.

EXV2: In this experiment, we investigate whether we can do better by optimising the field weights for the enriched FAQ documents collections. It is well known that significant gain in relevance can be obtained if the field weight parameters are properly optimised [15,16]. In our investigation, we use $EXV1$ as our baseline systems. We then optimise the field weights for all the enriched collections. A description of how the field weights were optimised can be found in the next section. We then perform retrieval on these enriched FAQ document collections using the associated testing set with the field weights for BM25F and PL25F set to their new optimal values.

EXV3: In experiments $EXV1$ and $EXV2$ we also investigated the effect of changing the size of the training set. In carrying out these experiments, three different collections that were enriched with queries of varying sizes were used for each testing set. A description of how these collections were created is detailed in the next section.

EXV4: To compare our approach with traditional approaches (e.g query expansion) normally used to resolve the term mismatch problem, we used the collection enrichment approach first introduced by Kwok et al. [8]. Collection enrichment is a form of query expansion where a high quality external collection is used to expand the original query terms and then retrieves from the local collection using the expanded query [8]. A local collection refers to the collection from which the final retrieved documents are retrieved. In the collection enrichment approach, we first performed retrieval on an external collection of HIV/AIDS documents crawled from the web. We crawled web pages that have a strong focus on HIV/AIDS frequently asked questions. Each web page crawled was indexed as a single document. In total, we had 3648 web page documents. For example, from *www.avert.org*, we were able to crawl 259 web documents. We provide examples of some of the domains and pages crawled in Table 3. In our collection enrichment approach, we used the Terrier Divergence From Randomness (DFR) Bo1 (Bose-Einstein 1) model to select the 10 most informative terms from the top 3 returned documents as expansion terms. These 10 new

Table 3. Examples of some of the web pages that were crawled from the web to use as an external collection for query expansion using collection enrichment approach

Web Page	Uniform Resource Locator (URL)
Avert : AVERTing HIV and AIDS	$http://www.avert.org$
FAQ |AIDS Foundation of South Africa	$http://www.aids.org.za$
What everyone should know about HIV	$http://www.hivaware.org.uk$
$AIDS_{\star}gov$	$http://www.aids.gov$

terms together with the original query terms were used for retrieval on the non enriched FAQ documents collection.

5.3 FAQ Documents Enrichment and Field Weights Optimisation

Our main contribution in this work as described in Section 1 is to demonstrate that using a field-based model to enrich the FAQ documents with additional terms from potential users of our FAQ retrieval system can alleviate the term mismatch problem that arises in our FAQ retrieval system. In order to achieve the above goals, we identified the following research hypotheses:

HP1: Enriching the FAQ documents with additional terms from queries for which the true relevant question-answer pair is known would increase the Mean Reciprocal Rank (MRR) and the overall recall in our FAQ retrieval system. Our intuition is that, additional terms introduced would help to reduce the term mismatch between the queries and the FAQ documents.

HP2: Increasing the number of queries used in enriching the FAQ documents would increase the (MRR) and the overall recall because additional terms introduced in the collection would help to alleviate the term mismatch problem.

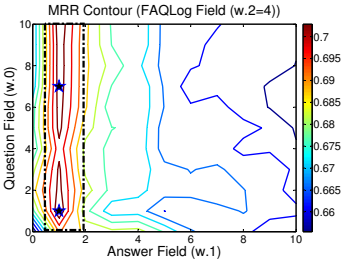
To test hypotheses *HP1* and *HP2*, we produced 10 random splits of the 750 matched SMS queries into a training set of 600 queries and a test set of 150 queries. These SMS queries were first corrected for spelling errors, so that such a confounding variable does not influence the outcome of our experiments. We plan to incorporate a spelling correction approach to our system in the future.

To test HP2, we additionally split the 600 training queries into three sets of 200 and incrementally combined them to create training sets of size 200, 400 and 600 queries (hereafter referred to as 200SMSes, 400SMSes and 600SMSes). 400SMSes is therefore a superset of 200SMSes and 600SMSes is a superset of 400SMSes. This process was chosen as it emulates the temporal nature of query collection in a real system. For each train/test split, we created 6 (3 for term frequencies and the other 3 for term occurrences) enriched collections (corresponding to 200SMSes, 400SMSes and 600SMSes) using the two enrichment approaches described in Section. 3. In total, we created 60 different enriched FAQ documents collections.

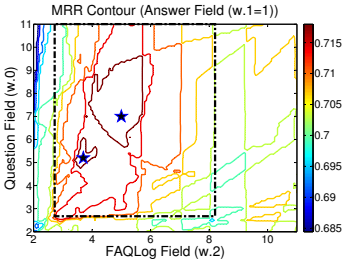
In order to infer whether using field-based weighting models does indeed help in the overall retrieval performance in terms of MRR and recall, the weights for each field were optimised. Optimisation of these field weights is vital as significant gains in relevance can be obtained if the parameters are properly optimised [15,16]. We used the 10 random splits of the 600 SMS queries of training data for optimising the field weights. The test queries for each train/test split were naturally not used for optimisation of the field weights in order to avoid over-fitting. For each training set, we randomly selected 450 SMS queries and used these to enrich the FAQ documents using our two enrichment strategies proposed in Section 3, thus giving us 2 different enriched FAQ document collections for each training set. The remaining 150 SMS queries were left for optimising the field weights.

Table 4. The mean and standard deviation for the field weights ($w.1 = 1$)

Weighting Model	Enrichment Strategy	Mean Field Weights	Standard Deviation
PL2F	Term Occurrence	$w.0 = 6.68, w.2 = 5.74$	$stdv.0 = \pm 3.18, stdv.2 = \pm 2.53$
	Term Frequency	$w.0 = 5.53, w.2 = 7.04$	$stdv.0 = \pm 3.33, stdv.2 = \pm 2.97$
BM25F	Term Occurrence	$w.0 = 5.98, w.2 = 5.94$	$stdv.0 = \pm 3.68, stdv.2 = \pm 3.06$
	Term Frequency	$w.0 = 4.02, w.2 = 6.98$	$stdv.0 = \pm 2.50, stdv.2 = \pm 3.41$



(a) Question, Answer and FAQLog Fields (w.2=4).



(b) Question, Answer and FAQLog Fields (w.1=1).

Fig. 1. The $\star$ denotes the region of highest MRR in relation to field weights $w.0$,$w.1$ and $w.2$ in this particular contour plots that were chosen randomly from our results. The higher MRR values for all the other random splits are inside the dotted rectangles.

In optimising the field weights, we used the Terrier-3.5 Information Retrieval (IR) platform. First we indexed the enriched collections separately without stopword removal and using the full Porter stemming algorithm. We then performed our optimisation using the Robust Line Search (RLS) strategy as described in [15]. For both BM25F and PL2F, we performed an initial scan of the field weights parameters $w.0, w.1$ and $w.2$ ($QUESTION$, $ANSWER$ and $FAQLog$ fields respectively) to determine the optimal values of these field weights with respect to a higher Mean Reciprocal Rank (MRR). In our initial scan, the field weights were varied linearly from 0.0 to 10.0 in steps of 1. Higher MRR values for the first scan were obtained when the $ANSWER$ field was set to 1 for most of the collections as shown in Figure 1.(a) (the $\star$ denotes the region of the highest MRR). For the $QUESTION$ and $FAQLog$ fields, higher MRR values were obtained when these fields were set to 2 or higher (Figure 1.(b)).

We then set a second starting point for each field weight to ($w.0 = 2.0, w.1 = 1.0, w.2 = 2.0$). Because the optimal value of the $ANSWER$ field was 1, this field was fixed while the others were varied linearly from 2.0 to 11.0 in steps of 0.1 for the second RLS. We increased the search space by varying parameters in steps of 0.1 instead of 1 so that we do not lose the global maximum. The above procedure was repeated for all the 10 random splits of training data. The optimal values of the field weights for these 10 random splits of training data were averaged to arrive at the final values of the field weights to use in testing our hypotheses *HP1* and *HP2*. Table 4 shows the mean and standard deviation of the field weights that we will use in our experimental investigation. It is worth pointing

out that these values were averaged taking into consideration that small changes in the parameter values of these models are known to produce small changes in the accuracy of relevance [15]. Our analysis of the various contour plots also show that the mean field weights in Table 4 are also within the region of higher MRR values that is bounded by the dotted rectangle in Figure 1.(b) for all the training samples.

6 Experimental Results and Evaluation

Table 5 summarises our experimental evaluation for research hypotheses *HP1* and *HP2*. As highlighted in [16], we can see that when setting the field weights to one (not optimised, $EXV1$), there in no improvement in retrieval performance in terms of MRR and recall for the field-based weighting models over the non field-based weighting models counterpart(BM25 and BM25F as well as PL2 and PL2F). Similar findings were also observed for the new enriched FAQ documents. However, there is a significant improvement in the retrieval performance (t-test, $p < 0.05$ for MRR) when the FAQ documents are enriched ($EXV1$).

There was a statistically significant (t-test, $p < 0.05$) increase in recall from around 0.2400 for non enriched FAQ documents to more than 0.4900 for the enriched FAQ documents. An increase in recall implies a reduction in term mismatch because previously un-retrieved documents have been retrieved. The benefit of using field-based weighting models is only realised after the field weights have been optimised ($EXV2$) as highlighted in Table 5. Higher recall values ranging from 0.68 to 0.77 and MRR values ranging from 0.67 to 0.73 were recorded, depending on the enrichment strategy. One plausible explanation for an increase in retrieval performance after optimising weights is that the fields of high importance (Question and FAQLog fields) have been assigned field weights of more than one, thus increasing the importance of term frequencies within those fields. As shown in Table 5, using the question field only without the answer field yielded better retrieval performance, suggesting that this field is more important than the answer field. Similar findings were also reported in [9].

Moreover, higher MRR values were obtained when enriching the FAQ documents using the query term frequencies rather than the query term occurrence (t-test p value for MRR ($p < 0.05$)). This is consistent with the above findings because the term frequencies approach just increases the term frequencies of repeating queries within the FAQLog field (similar to increasing the field weights). Finally, an increase in the size of the collection used to enrich the FAQ documents resulted in a slight increase in the average MRR (averaged across the 10 train/test partitions) for both PL2F and BM25F ($EXV3$). However, only the increase from 200 to 400 and 200 to 600 training SMS queries was statistically significant (t-test, $p < 0.05$), suggesting that adding more training SMS queries in the new field does indeed help to alleviate the term mismatch problem. Our approach performs better compared to query expansion ($EXV4$) using collection enrichment (t-test, $p < 0.05$) . This is because, the expansion terms were selected automatically without relevance judgement of the source documents.

Table 5. The mean retrieval performance for each Collection. Significant improvement in MRR and Recall if the FAQ documents are enriched with queries over non enriched FAQ documents, as denoted by $*$ (t-test, $p < 0.05$). Also, the was significant improvement in MRR and recall if field weights were optimised compared to non optimised field weights, as denoted by $**$ (t-test, $p < 0.05$).

					Test Evaluation Measure		
Evaluation	Collection	Enrichment Strategy	Weighting Model	Field Weights ($w.1 = 1$)	MRR	MAP	Recall
EXV1	Q(Only)	No Enrichment	BM25F/BM25	$w.0 = 1$	0.4312	0.2197	0.2495
	Q and A	No Enrichment	BM25F/BM25	$w.0 = 1$	0.4106	0.2302	0.2380
EXV4	Q(Only) and QE	Query Expansion	BM25F/BM25	$w.0 = 1$	0.4162	0.2022	0.2528
	Q ,A and QE	Query Expansion	BM25F/BM25	$w.0 = 1$	0.4317	0.2692	0.2974
EXV1 and *EXV3*	Q,A and 200SMS	Term Occurrence	BM25F/BM25	$w.0 = 1, w.2 = 1$	0.6120	0.4878	0.4951$*$
	Q,A and 400SMS				0.6614	0.4913	0.5466$*$
	Q,A and 600SMS				0.6608	0.5039	0.5924$*$
EXV2 and *EXV3*	Q,A and 200SMS	Term Occurrence	BM25F	$w.0 = 5.98, w.2 = 5.94$	**0.6774**	**0.5741**	**0.6772**$**$
	Q,A and 400SMS				**0.6692**	**0.5867**	**0.7089**$**$
	Q,A and 600SMS				**0.6666**	**0.5935**	**0.7009**$**$
EXV1 and *EXV3*	Q,A and 200SMS	Term Frequency	BM25F/BM25	$w.0 = 1, w.2 = 1$	0.6492	0.5146	0.5327$*$
	Q,A and 400SMS				0.6833	0.5491	0.5765$*$
	Q,A and 600SMS				0.6921	0.5435	0.6043$*$
EXV2 and *EXV3*	Q,A and 200SMS	Term Frequency	BM25F	$w.0 = 4.02, w.2 = 6.98$	**0.6847**	**0.6035**	**0.6902**$**$
	Q,A and 400SMS				**0.7179**	**0.6455**	**0.7546**$**$
	Q,A and 600SMS				**0.7315**	**0.6747**	**0.7484**$**$
EXV1	Q(Only)	No Enrichment	PL2F/PL2	$w.0 = 1$	0.4526	0.2720	0.2545
	Q and A	No Enrichment	PL2F/PL2	$w.0 = 1$	0.4106	0.2438	0.2711
EXV4	Q(Only) and QE	Query Expansion	PL2F/PL2	$w.0 = 1$	0.4297	0.2552	0.2815
	Q ,A and QE	Query Expansion	PL2F/PL2	$w.0 = 1$	0.4430	0.2627	0.2764
*EXV1*and *EXV3*	Q,A and 200SMS	Term Occurrence	PL2F/PL2	$w.0 = 1, w.2 = 1$	0.6068	0.5074	0.5841$*$
	Q,A and 400SMS				0.6310	0.5272	0.6168$*$
	Q,A and 600SMS				0.6831	0.5413	0.6340$*$
EXV2 and *EXV3*	Q,A and 200SMS	Term Occurrence	PL2F	$w.0 = 6.68, w.2 = 5.74$	**0.6766**	**0.5866**	**0.6950**$**$
	Q,A and 400SMS				**0.6938**	**0.6093**	**0.7188**$**$
	Q,A and 600SMS				**0.7004**	**0.6187**	**0.7465**$**$
EXV1 and *EXV3*	Q,A and 200SMS	Term Frequency	PL2F/PL2	$w.0 = 1, w.2 = 1$	0.6213	0.5432	0.5941$*$
	Q,A and 400SMS				0.6580	0.5535	0.6268$*$
	Q,A and 600SMS				0.6990	0.5848	0.6484$*$
EXV2 and *EXV3*	Q,A and 200SMS	Term Frequency	PL2F	$w.0 = 5.53, w.2 = 7.04$	**0.6701**	**0.6134**	**0.7246**$**$
	Q,A and 400SMS				**0.7112**	**0.6515**	**0.7585**$**$
	Q,A and 600SMS				**0.7254**	**0.6892**	**0.7713**$**$

This has some disadvantages as some queries might be expanded with irrelevant terms. Despite some of the disadvantages, a slight gain in MRR and recall was observed when the question and answer field were used and query expansion applied. However, there was a decrease in retrieval performance when only the question field was used, suggesting that the terms from the external collection might be adding noise to the original query.

7 Conclusions

In this paper we described a field-based approach to reduce the term mismatch problem in our SMS-Based FAQ retrieval system dealing with questions related to HIV and AIDS. Our experiments show that the inclusion of a field derived from logs of SMS queries for which the true relevant question-answer pair is known substantially improves the recall compared to query expansion using the collection enrichment approach. An increase in recall verified that the term mismatch did indeed significantly decrease (according to t-test) with the proposed approach.

In addition, we investigated how the number of queries used to enrich the FAQ documents affected performance. We saw a statistically significant increase in both recall and the average MRR when the number of queries used to enrich the FAQ documents were increased from 200 to 400 and 200 to 600. This results validates our second hypothesis $HP2$. An increase of training queries from 400 to 600 did not result in statistically significant improvement in MRR and recall. We plan to carry out further investigation with more queries to determine the point where there is no gain in retrieval performance even when the number of training queries is increased.

References

1. Billerbeck, B., Scholer, F., Williams, H.E., Zobel, J.: Query Expansion using Associated Queries. In: Proc. of CIKM (2003)
2. Billerbeck, B., Zobel, J.: Document Expansion Versus Query Expansion For Ad-hoc Retrieval. In: Proc. of ADCS (2005)
3. Fang, H.: A Re-examination of Query Expansion Using Lexical Resources. In: Proc. ACL:HLT (2008)
4. Hammond, K., Burke, R., Martin, C., Lytinen, S.: FAQ Finder: A Case-Based Approach to Knowledge Navigation. In: Proc. of CAIA (1995)
5. Jeon, J., Croft, W.B., Lee, J.H.: Finding Similar Questions in Large Question and Answer Archives. In: Proc. of CIKM (2005)
6. Kim, H., Lee, H., Seo, J.: A Reliable FAQ Retrieval System Using a Query Log Classification Technique Based on Latent Semantic Analysis. Info. Process. and Manage. 43(2), 420–430 (2007)
7. Kim, H., Seo, J.: High-Performance FAQ Retrieval Using an Automatic Clustering Method of Query Logs. Info. Process. and Manage. 42(3), 650–661 (2006)
8. Kwok, K.L., Chan, M.: Improving Two-Stage Ad-hoc Retrieval for Short Queries. In: Proc. of SIGIR (1998)
9. Leveling, J.: On the Effect of Stopword Removal for SMS-Based FAQ Retrieval. In: Bouma, G., Ittoo, A., Métais, E., Wortmann, H. (eds.) NLDB 2012. LNCS, vol. 7337, pp. 128–139. Springer, Heidelberg (2012)
10. Macdonald, C., Plachouras, V., He, B., Lioma, C., Ounis, I.: University of Glasgow at WebCLEF 2005: Experiments in Per-Field Normalisation and Language Specific Stemming. In: Proc. of CLEF (2006)
11. Moreo, A., Navarro, M., Castro, J.L., Zurita, J.M.: A High-Performance FAQ Retrieval Method Using Minimal Differentiator Expressions. Know. Based Syst. 36, 9–20 (2012)
12. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proc. of OSIR at SIGIR (2006)
13. Plachouras, V., Ounis, I.: Multinomial Randomness Models for Retrieval with Document Fields. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 28–39. Springer, Heidelberg (2007)
14. Porter, M.F.: An Algorithm for Suffix Stripping. Elec. Lib. Info. Syst. 14(3), 130–137 (2008)
15. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Info. Retr. 3(4), 333–389 (2009)

16. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields. In: Proc. of CIKM (2004)
17. Singhal, A., Pereira, F.: Document Expansion for Speech Retrieval. In: Proc. of SIGIR (1999)
18. Sneiders, E.: Automated FAQ Answering: Continued Experience with Shallow Language Understanding. Question Answering Systems. In: Proc. of AAAI Fall Symp. (1999)
19. Sneiders, E.: Automated FAQ Answering with Question-Specific Knowledge Representation for Web Self-Service. In: Proc. of HSI (2009)
20. Voorhees, E.M.: Query Expansion Using Lexical-Semantic Relations. In: Proc. of SIGIR, pp. 61–69 (1994)
21. Whitehead, S.D.: Auto-FAQ: an Experiment in Cyberspace Leveraging. Comp. Net. and ISDN Syst. 28(1-2), 137–146 (1995)
22. Xue, X., Jeon, J., Croft, W.B.: Retrieval Models for Question and Answer Archives. In: Proc. of SIGIR (2008)