UBML participation to CLEF eHealth IR challenge 2015: Task 2

Edwin Thuma, George Anderson, and Gontlafetse Mosweunyane

Department of Computer Science, University of Botswana
{thumae,andersong,mosweuny}@mopipi.ub.bw

Abstract. This paper describes the participation of UBML, a team composed with members of the Department of Computer Science, University of Botswana, to the biomedical information retrieval challenge proposed in the framework of CLEF eHealth 2015 Task 2. For this first participation, we are evaluating the effectiveness of two different query expansion strategies when searching for health related content on the web. In particular, we deploy pseudo relevance feedback, where the original query is expanded with additional terms selected from the local collection (collection being searched). In another approach, we deploy the collection enrichment approach, where the original query is expanded with additional terms from an external collection (collection not being searched). We test the generality of our results by using two different methods for selecting the expansion terms. In particular, we used the Kullback-Liebler Divergence and the Bose-Einstein 1 (Bo1) model to select the expansion terms. Our result show that we can improve the retrieval effectiveness of our system by expanding the original query with additional terms from a local collection. Furthermore, our results suggest that, when using an external collection to expand the original query, it is important to select the expansion terms from a health related external collection when searching for health related content on the web.

Keywords: Query expansion, Learning to Rank, Pseudo relevance feedback, Collection enrichment

1 Introduction

In this paper, we describe the methods used for our (University of Botswana Machine Learning and Information Retrieval Group) participation of the CLEF (Conference and Labs of the Evaluation Forum) eHealth 2015 Task 2: User-centred health information retrieval. For detailed task description, please see the overview paper of Task 2 [13]. Task 2 is part of the broader CLEF eHealth initiative, which includes Task 1A (Clinical Speech Recognition) and Task 1B (Clinical Named Entity Recognition) [5].

There is wide spread use of search engines for medical self-diagnosis [19, 17]. Task 2 focuses on solving the problem of information retrieval in this context. This is difficult particularly because search engine users who try to self-diagnose typically construct circumlocutory queries, using colloquial language instead of medical terms, making it difficult to retrieve relevant documents, which are more readily retrieved using medical terms. For example, if "baldness in multiple spots" is used as a query instead of "alopecia," it is likely few relevant documents will be retrieved [19]. We attempt to tackle this problem using query expansion techniques, making use of the local document collection, as well as external document collections.

This paper is structured as follows. Section 2 contains a background on algorithms used and related work. Section 3 describes the 10 runs submitted by UBML. In Section 4, we describe the dataset used in our experimental investigation and evaluation. Section 5 describes the experimental environment. Section 6 reports our results and discusses the results. Section 7 has our conclusion.

2 Background and Related Work

In this section, we begin by presenting a brief but essential related work and background on the different algorithms used in our experimental investigation and evaluation. We start by reviewing related work in Section 2.1. This is followed by a description of the BM25 term weighting model in Section 2.2 and learning to rank in Section 2.3. In Section 2.4, we describe the Bose-Einstein 1 (Bo1) model for query expansion, followed by a description of the Kullback-Liebler Divergence for query expansion in Section 2.5.

2.1 Related Work

CLEF 2015 eHealth Task 2 was motivated by the problem of users of information retrieval systems formulating *circumlocutory queries*, as studied by Zuccon et al. [19] and Stanton et al. [17]. Previous CLEF eHealth tasks (Task 3 in 2013 and 2014) focused on use of queries containing medical terms [3, 4]. Zuccon et al. discuss the cause of circumlocutory (colloquial) queries, which stems from users attempting to diagnose ailments, but without sufficient knowledge of medical terminology that could be relevant, therefore using layman's terms such as "hives all over body" instead of "urticaria." In their study, they found that modern search engines (such as Google and Bing) are ill-equipped to handle such queries; only 3 out of the top 10 results were highly useful for self-diagnosis. Stanton et al. also studied generating circumlocutory queries in order to train machine learning models, which would then be capable of matching queries with symptoms. Our experiments focus mainly on query expansion as an approach to tackling this problem.

Zuccon and Koopman [18] discuss the importance of understandability as an alternative desirable outcome to topicality, which is frequently used. It is not sufficient to suggest documents to the user which actually answer his queries, but which s/he does not understand. This is particularly true for medical-related queries. We therefore consider rank biased precision (RBP), understandability-based rank biased precision (uRBP), and graded understandability-based rank biased precision (uRBPgr) when evaluating our models. These reduce the number of documents that have to be evaluated for readability, by incorporating uncertainty into relevance judgements. The RBP metrics used a user-persistence parameter of 0.8, almost the same as was obtained from the work of Park and Zhang [14].

2.2 BM25 Term Weighting Model

For our baseline system and all our experimental investigation and evaluation, we used the BM25 term weighting model to score and rank medical documents. For a given query q, the relevance score of a document d based on the BM25 term weighting model is expressed as [16]:

$$score_{BM25}(d,Q) = \sum_{t \in Q} w^{(1)} \cdot \frac{(k_1+1)tfn}{k_1+tfn} \cdot \frac{(k_3+1)qtf}{k_3+qtf}. \tag{1}$$

where qtf is the number of occurrences of a given term t in the query Q. k_1 and k_3 are parameters of the model. tfn is the normalised within document term frequency. $w^{(1)}$ denotes the Robertson-Spark Jones (RSJ) weights, which is an inverse document frequency (IDF) factor and is given by:

$$w^{(1)} = \log \frac{N - dft + 0.5}{dft + 0.5} \tag{2}$$

Where N is the number of documents in the collection and dft is the number of documents in the collection that have a term t.

2.3 Learning to Rank Approach

Learning to rank techniques are algorithms that use machine learning techniques to learn an appropriate combination of features into an effective ranking model [7]. The main advantage of using learning to rank is that we can re-rank a sample of the top-ranked documents for a given query using the learned model before returning the results to the user. In general, the steps for learning an effective ranking model are as follows [8, 9]:

1. Top K retrieval: Using a set of training queries that have relevance assessment, retrieve a sample of k documents using an initial weighting model such as BM25.
2. Feature extraction: For each document in the retrieved sample, extract a set of features. These features can either be query-dependent (term weighting models, term dependence models) or query-independent (click count, fraction of stopwords). The feature vector for each document is labelled according to the already existing relevance judgements.
3. Learning: Learn an effective ranking model by deploying an effective leaning to rank technique on the feature vectors of the top k documents.

This learned model can be deployed in a retrieval setting as follows:

4. Top K retrieval: For each unseen query, the top k documents are retrieved using the same retrieval strategy as in step (1)
5. Feature extraction: A set of features are extracted for each document in the sample of k documents. These features should be the same as those extracted in step (2).
6. Re-rank the documents: Re-rank the documents for the query by applying a learned model on every feature vector of the documents in the sample. The final ranking of the documents are obtained by sorting the predicted scores in descending order.

In this work, we deploy Coordinate Ascent [11], which is a linear-based learner. A linear-based learner yields a model that linearly combines the feature values [11, 2, 9]. The final score of a document d for any given query Q, for a linear leaner is given by:

$$score_(d, Q) = \sum_{f} \alpha_i \cdot f_{i,d} \tag{3}$$

where α_i is the weight of the i_{th} feature and $f_{i,d}$ is the value/score of the i_{th} feature for the document d.

2.4 Bose-Einstein 1 (Bo1) Model for Query Expansion

In our experimental investiagtion and evaluation, we used the Terrier-4.0 Divergence from Randomness (DFR) Bose-Einstein 1 (Bo1) model to select the most informative terms from the topmost documents after a first pass document ranking. The DFR Bo1 model calculates the information content of a term t in the top-ranked documents as follows [1]:

$$w(t) = tfx \cdot \log_2 \frac{1 + P_n(t)}{P_n(t)} + \log_2(1 + P_n(t)) \tag{4}$$

$$P_n(t) = \frac{tfc}{N} \tag{5}$$

where $P_n(t)$ is the probability of t in the whole collection, tfx is the frequency of the query term in the top x ranked documents, tfc is the frequency of the term t in the collection, and N is the number of documents in the collection.

2.5 Kullback-Liebler Divergence for Query Expansion

In another approach, we used the kullback-Liebler divergence to select the most informative term from the topmost documents after a first pass document ranking. This model computes the information content of a term t in the top-ranked documents as follows [1]:

$$w(t) = (P_x(t)) \log_2 \frac{P_x(t)}{P_n(t)} \tag{6}$$

$$P_x(t) = \frac{tfx}{x} \tag{7}$$

$$P_n(t) = \frac{tfc}{N} \tag{8}$$

where $P_x(t)$ is the probability of t estimated from the top x ranked documents, tfx is the frequency of the query term in the top x ranked documents, tfc is the frequency of the term t in the collection, and N is the number of documents in the collection.

3 Description of the Different Runs

UBML_EN_Run.1: This is the baseline system. We used BM25 term weighting model in Terrier-4.0 IR platform to score and rank the documents in a document collection of around one million documents (web pages from medical web sites).

UBML_EN_Run.2: We used the baseline system (UBML_EN_Run.1). As improvement, we proposed a simple pseudo-relevance feedback method using the local collection to perform query expansion. We used the Terrier-4.0 Kullback-Leibler divergence for query expansion method to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the local collection). We then performed a second pass retrieval on this local collection with the new expanded query.

UBML_EN_Run.3: We used the baseline system (UBML_EN_Run.1). As improvement, we proposed a simple pseudo-relevance feedback method using the local collection to perform query expansion. We used the Terrier-4.0 Divergence from Randomness (DRF) Bose - Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the local collection). We then performed a second pass retrieval on this local collection with the new expanded query.

UBML_EN_Run.4: We used the baseline system (UBML_EN_Run.1). As improvement, we used the collection enrichment approach [6], where we selected the expansion terms from an external collection, which was made up of the 2004_TREC_MEDLINE_1 & 2004_TREC_MEDLINE_2 abstracts. We used the Terrier-4.0 Kullback-Leibler divergence for query expansion method to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the external collection). We then performed a second pass retrieval on the local collection with the new expanded query.

UBML_EN_Run.5: We used the baseline system (UBML_EN_Run.1). As improvement, we used the collection enrichment approach [6], where we selected the expansion terms from an external collection, which was made up of the 2004_TREC_MEDLINE_1 & 2004_TREC_MEDLINE_2 abstracts. We used the Terrier-4.0 Divergence from Randomness (DRF) Bose - Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the external collection). We then performed a second pass retrieval on the local collection with the new expanded query.

UBML_EN_Run.6 & UBML_EN_Run.7: We deployed a learning to rank approach, using Coordinate Ascent as provided in RankLib-v2.1[1]. We used the 5 training queries to train our learning to rank model. For each query in the training set, we created an initial sample of the top 1000 ranked documents using our second (UBML_EN_Run.2) and third (UBML_EN_Run.3) runs for the following runs, **UBML_EN_Run.6 & UBML_EN_Run.7** respectively. After generating this sample of documents, we generated several query dependent features (23) using Terrier-4.0 FAT Framework (for learning to rank). Later in Section 5, we provide a list of these query dependent features.

UBML_EN_Run.8: We used the baseline system (UBML_EN_Run.1). As improvement, we used the collection enrichment approach [6], where we selected the expansion terms from an external collection, which was made up of a collection of documents from Wikipedia2008. We used the Terrier-4.0 Kullback-Leibler divergence for query expansion method to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the external collection). We then performed a second pass retrieval on the local collection with the new expanded query.

UBML_EN_Run.9: We used the baseline system (UBML_EN_Run.1). As improvement, we used the collection enrichment approach [6], where we selected the expansion terms from an external collection, which was made up of a collection of documents from Wikipedia2008. We used the Terrier-4.0 Divergence from Randomness (DRF) Bose - Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the external collection). We then performed a second pass retrieval on the local collection with the new expanded query.

UBML_EN_Run.10: We used the baseline system (UBML_EN_Run1). As improvement, we used Markov Random Fields for Term Dependencies. We used the sequential dependence variant of the model, which models dependencies between adjacent query terms. In this work, we explore a window size of 2, to see what impact it has of the retrieval effectiveness. In particular, we re-ranked the documents if 2 query terms are in close proximity in ranked documents.

[1] http://people.cs.umass.edu/ vdang/ranklib.html

Table 1. Selection of Training and Test Queries used in this Task.

Training Queries
loss of hair on scalp in an inch width round
sores around mouth
puffy sore calf
eye balls coming out
white part of eye turned green

Test Queries
many red marks on legs after travelling from us
lump with blood spots on nose
dry red and scaly feet in children
whistling noise and cough during sleeping + children
child make hissing sound when breathing

4 Dataset

4.1 Document Collection

A web crawl of about one million documents is used for this task. This document collection was made available to CLEF eHealth participants through the Khresmoi project[2]. This document collection consists of web pages covering a broad range of health topics, targeted at both the general public and healthcare professionals. Web pages in the document collection are predominantly medical and health-related websites that have been certified by the Health on the Net (HON) Foundation[3] as adhering to the HONcode principles[4] (approximately 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank, Diagnosia and Trip Answers. The crawled documents were provided in their raw HTML (Hyper Text Markup Language) format along with their uniform resource locators (URL).

4.2 Queries

A total of 66 circumlocutory queries that users may pose when faced with signs and symptoms of a medical condition were provided for testing our different systems. In addition, 5 circumlocutory queries, together with their query relevance judgements were also provided for training our different systems. In Table 1, we provide a selection of training and test queries used in this task.

5 Experimental Setting

FAQ Retrieval Platform: For all our experimental evaluation, we used Terrier-4.0[5] [12], an open source Information Retrieval (IR) platform. All the documents

[2] http://khresmoi.eu/

[3] http://www.healthonnet.org

[4] http://www.hon.ch/HONcode/Patients-Conduct.html

[5] http://terrier.org/

Table 2. All query-dependent (QD) features used in this work.

Features	Type	Total
BM25 weighting model	QD	1
BB2 weighting model	QD	1
DFIC - Divergence From Independence model based on Chi-square statistics	QD	1
DFIZ - Divergence From Independence model based on Standardization	QD	1
DFR_BM25 weighting model	QD	1
DFRee weighting model	QD	1
DFReeKLIM weighting model	QD	1
Dl - A simple document length weighting model.	QD	1
DLH weighting model	QD	1
DLH13 weighting model	QD	1
DPH hypergeometric weighting model	QD	1
Hiemstra_LM - Hiemstra LM weighting model	QD	1
IFB2 weighting model	QD	1
In_expB2 - Inverse Expected Document Frequency model with Bernoulli after-effect and normalisation 2	QD	1
In_expC2 weighting mode	QD	1
InB2 - Inverse Document Frequency model with Bernoulli after-effect and normalisation 2	QD	1
InL2 weighting model	QD	1
Js_KLs - Is the product of two measures: the Jefrreys' divergence with the Kullback Leibler's divergence.	QD	1
LemurTF_IDF - The TF_IDF weighting model as it is implemented in Lemur	QD	1
LGD weighting model	QD	1
PL2 weighting model	QD	1
Tf weighting model	QD	1
TF_IDF weighting model	QD	1
XSqrA_M - The inner product of Pearson's X^2 with the information growth computed with the multinomial M	QD	1
Total		24

used in this study were first pre-processed before indexing and this involved tokenising the text and stemming each token using the full Porter stemming algorithm [15]. Stopword removal was enabled and we used Terrier stopword list. The normalisation parameter for BM25 was set to its default value of b = 0.75.

Training Learning to Rank Techniques: For our learning to rank approach, we used RankLib, a library of learning to rank algorithms. To train and test Coordinate Ascent, we used the default RankLib parameter values of the algorithms. In all our experiments, we used MAP as the objective function [8]. In Table 2, we provide a list of query dependent features used in our experimental investigation.

6 Results and Discussion

Table 3 and Figure 1 presents the retrieval results of the 10 different runs submitted to the CLEF-ehealth Task 2. From this table, we see an improvement in the retrieval performance of our baseline system when the original query is expanded with terms from a local collection (UBML_EN_Run.2 and UBML_EN_Run.3). For example, we see an increase in P@10 (from 0.3106 to 0.3197) and nDCG@10 (from 0.2897 to 0.2909). Moreover, we see an improvement in the understandability or readability of information (see Table 4) when the original query is expanded with terms from a local collection (UBML_EN_Run.2 and UBML_EN_Run.3). However, this increase in the retrieval performance and understandability or readability of information leads to a decrease in recall (see Figure 2 and Table 3). In particular, the number of relevant documents retrieved for the 66 queries decreases from 1333 (for UBML_EN_Run.1) to 1244 UBML_EN_Run.2. When we expand the original queries with additional terms from an external collection (2004 TREC MEDLINE abstracts and Wikepedia2008), we see a decrease in the retrieval performance in terms of P@10 and nDCG@10 across the different methods used for selecting the expansion terms (Table 3 and Figure 1). In addition, this leads to a decrease in the understandability and readability of information (see Table 4). Interestingly, expanding the original queries with terms from TREC 2004 MEDLINE abstracts was observed to be more effective than Wikepedia2008. These findings suggest that when using an external collection to expand the original query, it is important to select the expansion terms from a health related external collection when searching for health related content on the web.

Table 3. Retrieval Results for all 10 Runs.

Run ID	External Collection	P@5	P@10	nDCG@5	nDCG@10	rel_ret
UBML_EN_Run.1	-	0.3455	0.3106	0.3001	0.2897	1333
UBML_EN_Run.2	-	0.3455	**0.3197**	0.2920	0.2909	1244
UBML_EN_Run.3	-	0.3576	0.3182	0.2995	**0.2919**	1262
UBML_EN_Run.4	TREC_MEDLINE	0.3121	0.2742	0.2624	0.2460	1359
UBML_EN_Run.5	TREC_MEDLINE	0.3121	0.2773	0.2641	0.2500	1355
UBML_EN_Run.6	-	0.3030	0.2621	0.2336	0.2265	1244
UBML_EN_Run.7	-	0.3424	0.3091	0.2932	0.2887	1262
UBML_EN_Run.8	Wikipedia2008	0.3091	0.2652	0.2716	0.2533	1366
UBML_EN_Run.9	Wikipedia2008	0.3182	0.2697	0.2693	0.2538	1368
UBML_EN_Run.10	-	0.2818	0.2485	0.2349	0.2294	1333

In our investigation, we also deployed a learning to rank approach in order to improve the retrieval performance after expanding the original queries with terms from a local collection (UBML_EN_Run.6 and UBML_EN_Run.7). Surprisingly, this affected the retrieval performance of our system. A possible explanation for these results may be the lack of adequate training data. For example, we had 5 training queries and only 2 had relevant documents. The rest did not have relevant documents in the collection.

As an additional finding, we also investigated whether we can improve the retrieval effectiveness of our baseline by considering term dependence when ranking the documents. This was motivated by previous work in [10]. In their work, Metzler and Croft [10] reported that the sequential dependence model using term and ordered features was more effective on smaller, homogeneous collections with longer queries. Since the queries in this task were long, we deployed this sequential dependence model (UBML_EN_Run.10). Surprisingly, this sequential dependence model significantly degraded the retrieval performance and the understandability OR readability of information (see Table 3, Table 4 and Figure 1). A possible explanation for this might be that the document collection being searched was very large. This sequential dependence model has been found to be effective only in small collections [10].

Table 4. Understandability of Information

Run ID	RBP(0.8)	uRBP(0.8)	uRBPgr(0.8)
UBML_EN_Run.1	0.3294	0.2745	0.2771
UBML_EN_Run.2	0.3305	0.2709	0.2735
UBML_EN_Run.3	**0.3358**	**0.2757**	**0.2789**
UBML_EN_Run.4	0.2953	0.2255	0.2300
UBML_EN_Run.5	0.2960	0.2220	0.2279
UBML_EN_Run.6	0.2766	0.2348	0.2310
UBML_EN_Run.7	0.3339	0.2795	0.2772
UBML_EN_Run.8	0.2978	0.2352	0.2368
UBML_EN_Run.9	0.2993	0.2332	0.2362
UBML_EN_Run.10	0.2658	0.2125	0.2159

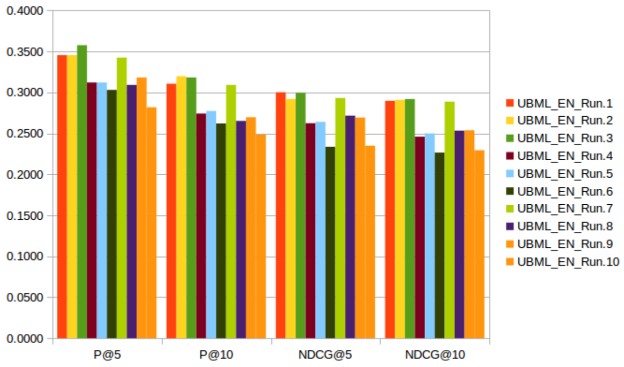

Fig. 1. Performance of the 10 Runs.

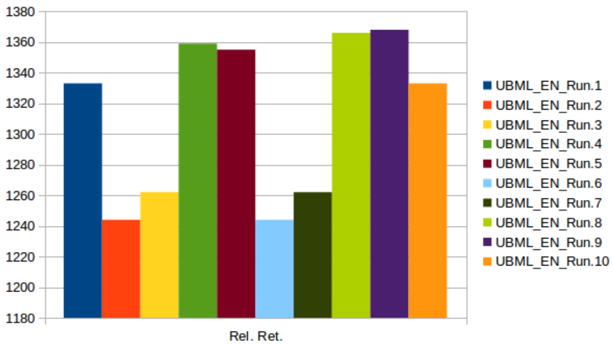

Fig. 2. Relevant Returns of the 10 Runs.

7 Conclusion

In this investigation, the aim was to assess the retrieval effectiveness of two different query expansion strategies when searching for health related content on the web. In particular, pseudo relevance feedback and collection enrichment approach. Our result show that we can improve the retrieval effectiveness of our system by expanding the original query with additional terms from a local collection (pseudo relevance feedback). Furthermore, our results suggest that, when using an external collection (collection enrichment) to expand the original query, it is important to select the expansion terms from a health related external collection when searching for health related content on the web. Overall, our results show that using an external collection to expand the original queries, degrades the retrieval performance. These results generalised well on the two different methods (Kullback-Liebler Divergence and the Bose-Einstein 1 (Bo1)) used for selecting the expansion terms. Further work needs to be done with different health related external collection to establish whether expanding the original queries with additional terms from an external collection will degrade the retrieval performance.

References

1. G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. *University of Glasgow,UK, PhD Thesis*, pages 1 – 198, June 2003.
2. C.J.C. Burges, R. Ragno, and Q.V. Le. Learning to rank with non-smooth cost functions. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, January 2007.
3. L. Goeuriot, G.J.F Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions when Reading Clinical Reports. In *CLEF 2013 Online Working Notes*, volume 8138. CEUR-WS, 2013.
4. L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G.J.F Jones, and H. Mueller. Share/clef ehealth Evaluation Lab 2014, Task 3: User-Centred Health Information Retrieval. In *CLEF 2014 Online Working Notes*. CEUR-WS, 2014.
5. L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. In *CLEF 2015 - 6th Conference and Labs of the Evaluation Forum*. Lecture Notes in Computer Science (LNCS), Springer, September 2015.
6. K.L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–256, New York, NY, USA, 1998. ACM.
7. T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, June 2009.
8. C. Macdonald, R.L. Santos, and I. Ounis. The whens and hows of learning to rank for web search. *Information Retrieval*, 16(5):584–628, October 2013.

9. C. Macdonald, R.L.T. Santos, I. Ounis, and B. He. About learning models with multiple query-dependent features. *ACM Transactions on Information Systems (TOIS)*, 31(3):11:1–11:39, August 2013.
10. D. Metzler and W.B Croft. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.
11. D. Metzler and W.B Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, June 2007.
12. I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519, Berlin, Heidelberg, 2005. Springer-Verlag.
13. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G.J.F. Jones, M. Lupu, and P. Pecina. CLEF eHealth Evaluation Lab 2015 task 2: Retrieving Information about Medical Symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
14. L.A.F Park and Y. Zhang. On the Distribution of User Persistence for Rank-Biased Precision. In *Proceedings of the 12th Australasian document computing symposium*, pages 17–24, 2007.
15. M.F. Porter. An Algorithm for Suffix Stripping. *Readings in Information Retrieval*, 14(3):313–316, 1997.
16. S.E. Robertson, S. Walker, S. Jones, M.M Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, pages 1–18, Gaithersburg, Md., USA., 1996. Text REtrieval Conference (TREC).
17. I. Stanton, S. Ieong, and N. Mishra. Circumlocution in Diagnostic Medical Queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142. ACM, 2014.
18. G. Zuccon and B. Koopman. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *Medical Information Retrieval Workshop at SIGIR 2014*, page 32, 2014.
19. G. Zuccon, B. Koopman, and J. Palotti. Diagnose This If You Can: On the Effectiveness of Search Engines in Finding Medical Self-Diagnosis Information. In *Advances in Information Retrieval (ECIR 2015)*, pages 562–567. Springer, 2015.