Ranking Health Web Pages with Relevance and Understandability

Joao Palotti
Vienna University of Technology, Austria
palotti@ifs.tuwien.ac.at

Lorraine Goeuriot
Université Grenoble Alpes, France
lorraine.goeuriot@imag.fr

Guido Zuccon
Queensland University of Technology, Australia
g.zuccon@qut.edu.au

Allan Hanbury
Vienna University of Technology, Austria
hanbury@ifs.tuwien.ac.at

ABSTRACT

We propose a method that integrates relevance and understandability to rank health web documents. We use a learning to rank approach with standard retrieval features to determine topical relevance and additional features based on readability measures and medical lexical aspects to determine understandability. Our experiments measured the effectiveness of the learning to rank approach integrating understandability on a consumer health benchmark. The findings suggest that this approach promotes documents that are at the same time topically relevant and understandable.

1. INTRODUCTION

An increasing number of people rely on online health information to understand and manage their health; this information is commonly accessed through search engines [4]. The retrieval of incorrect or unclear health information poses potential risks as people may dismiss serious symptoms, use inappropriate treatments or unfoundedly escalate their health concerns about common symptomatology [1, 10]. However, an extensive number of studies has shown that the average user experiences difficulty in understanding the content of a large portion of the results retrieved by current search engine technology, e.g., see [11].

In the context of consumer health information seeking, search engines should not only retrieve relevant information, but they should also promote information that is understandable by the user and that is reliable and verified [10]. This paper tackles one aspect of this problem by investigating the effectiveness of a learning to rank approach aimed at retrieving documents that are at the same time topically relevant and understandable by the user. Specifically, we employ a range of standard *retrieval features* to capture information about relevance, and exploit a number of *readability features* comprising of readability measures and medical lexical aspects to determine document understandability.

Through experiments on a consumer health search collection, we show that our approach improves health search results, demonstrating that the combination of retrieval features and readability features within a learning to rank approach best promotes search results that are relevant and understandable for the user.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914741

2. RELATED WORK

Our work tackles the problem of retrieving health information in answer to queries issued by laypeople. This problem has been largely investigated in the context of the CLEF eHealth Evaluation Lab[1], from which we take the data to evaluate the proposed approach. Specifically, the 2015 task provides a test collection to evaluate the effectiveness of search engines in answering self-diagnosing queries [6]. The evaluation framework explicitly accounts for both the topical relevance of the search results and their understandability, interpreted as how easy it is for a layperson to understand the content of a specific search result. This is done using understandability-biased evaluation measures, where gains obtained from relevant information are weighted by how hard it is for a layperson to understand that information [15, 16]. In this paper we use the CLEF eHealth 2015 test collection along with the explicit understandability assessments distributed and the understandability-biased RBP measure (see [6, 15, 16]). In addition, we further expand the understandability-biased evaluation framework by modifying the Bpref measure in the same spirit of understandability-biased RBP (see Section 4.4).

Our approach exploits a number of readability measures as features for the learning to rank approach. The measures we employ are based on surface-level characteristics of text, such as characters, syllables and word counts [3]. For example, the Dale-Chall readability formula is based on a corpus of words that can be understood by fourth-grade students; the Flesch-Kincaid measure instead computes a readability score based on a weighted combination of the number of words and the number of syllables in a sentence. The Gunning-Fog index combines the intuitions of these two approaches using sentence length and frequency of "complex" words. Previous work that has explored the under-

[1] http://sites.google.com/site/clefehealth/

standability of the health content retrieved by search engines extensively relied on these measures (along with other surface-level measures we also use to compute readability features) [12]. Readability measures specific to the health domain have been proposed, e.g. [13]. These however rely on the mapping of the text content of documents to health terminologies and the assessment of readability based on hierarchy and relationships encoded in it: a computationally intensive and error prone process. We defer the use of these techniques to compute readability features in the context of learning to rank to future work. We do however use clinical terminologies (Mesh, ICD) and dictionaries (Drugbank, CHV) to compute readability by testing whether a word is present in such resources.

We are not the first to explore the use of readability features to improve search engine results. Collins-Thompson et al. have shown the benefits of personalising search results to the reading levels of individual users [2]. Similarly, Tan et al. have modelled both the comprehensibility of texts and the users reading proficiency to improve content ranking [9]. The understandability issue is crucial in consumer health search, and the application of readability measures has not been well explored yet. Zuccon et al. have encoded readability measures as language modelling priors but have found no improvements in search results [17]. In this paper, we investigate the application of new approaches to effectively include understandability in consumer health search.

3. FEATURES FOR LEARNING TO RANK

In this work we study the effectiveness of a learning to rank approach that exploits retrieval features and readability features. The hypothesis is that the combined use of these feature sets not only improves results in terms of topical relevance, but promotes search engine results that are more understandable by the general public. Here, readability measures (and other features) are used as a proxy for document understandability. To validate this hypothesis, we investigated a number of retrieval and readability features, which we then alternated and combined to verify the contribution of each feature type to the improvement of search engine results.

The features used in our investigation are summarised in Table 1. As retrieval features, we used the matching scores provided by a number of common retrieval models, along with query independent features and document score modifiers. To compute these features, we indexed two fields: the document titles only; and the entire document bodies.

As readability features, we used a large number of existing readability measures, as well as lexical and morphological measures. Readability features fall into four main categories:

Traditional formulas: these are the existing well known readability formulas for general text. A thorough description of these formulas can be found in [3].

Surface measures: these are basic syntactic and lexical features, based on document statistics. Examples of this type of feature are the number of characters, syllables, words, and sentences present in a document. This category also includes the word length distribution in documents, e.g., #(|Word|>6) is the number of words in a document with more than 6 characters.

General vocabulary related measures: these are common lexical features used to assess text difficulty, e.g. proportion of numbers, stopwords, and common words in documents.

Medical vocabulary related measures: these are lexical and morphological features specifically adapted to the scientific domain, such as acronyms, or greco-latin affixes; or adapted to the medical domain, such as the number of terms present in lexicons such as Drugbank, Mesh or the ICD. We also compute the number of terms from the Consumer Health Vocabulary (CHV) contained in documents, as well as the sum of the terms' difficulty scores from CHV and the document mean score.

For features in the last three categories, we computed their raw values, as well as their average values per sentence and per word, i.e., we divided the value of the feature by the number of words and sentences in the document.

Feature Type	Feature Category	Feature Name
IR Features (18)	Common IR Models (14)	BM25 $\star$ PL2 $\star$ DirichletLM $\star$ LemurTF_IDF $\star$ TF_IDF $\star$ DFRee $\star$ Hiemstra_LM $\star$
	Query Independ. (2)	Document Length $\star$
	Doc. Score Modifier (2)	Divergence from Randomness Markov Random Field
Readability Features (72)	Traditional Formulas (8)	ARI Index Coleman Liau Index Dale-Chall Score Flesch Kincaid Grade Flesch Reading Ease Gunning Fog Index LIX Index SMOG Index
	Surface Measures (25)	# Characters $\diamondsuit\dagger$ # Sentences $\diamondsuit$ # Syllables $\diamondsuit\dagger$ # Words $\dagger$ # (| Syllables(Word) | > 3) $\diamondsuit\dagger$ # (| Word | > 4) $\diamondsuit\dagger$ # (| Word | > 6) $\diamondsuit\dagger$ # (| Word | > 10) $\diamondsuit\dagger$ # (| Word | > 13) $\diamondsuit\dagger$
	General Vocabulary Related Features (12)	Numbers $\diamondsuit\dagger$ English Dictionary $\diamondsuit\dagger$ Dale-Chall List $\diamondsuit\dagger$ stopwords $\diamondsuit\dagger$
	Medical Vocabulary Related Features (27)	Acronyms $\diamondsuit\dagger$ Mesh $\diamondsuit\dagger$ DrugBank $\diamondsuit\dagger$ ICD10 (International classification of Diseases) $\diamondsuit\dagger$ Medical Prefixes $\diamondsuit\dagger$ Medical Suffixes $\diamondsuit\dagger$ Consumer Health Vocabulary $\diamondsuit\dagger$ Sum(chv Score) $\diamondsuit\dagger$ Mean(chv Score) $\diamondsuit\dagger$

Table 1: Features used in the learning to rank process; the number of features for each group is reported in parenthesis. $\star$: scores from both titles and document bodies are used (thus doubling the number of features). $\diamondsuit$: raw feature values and values normalised by number of words in a documents are used. †: raw feature values and values normalised by number of sentences in a document are used.

4. EXPERIMENTS AND ANALYSIS

4.1 Dataset

To investigate methods that provide users with search results that are both topical and understandable in answer to health queries, we use the CLEF 2015 eHealth Evaluation Lab collection [6]. This collection contains approximately 1 million web pages and 66 queries from people seeking self-diagnosing information. A key aspect of this collection is that it contains explicit graded assessments of both the topical relevance of documents to queries and the understandability of documents, which indicates whether a document

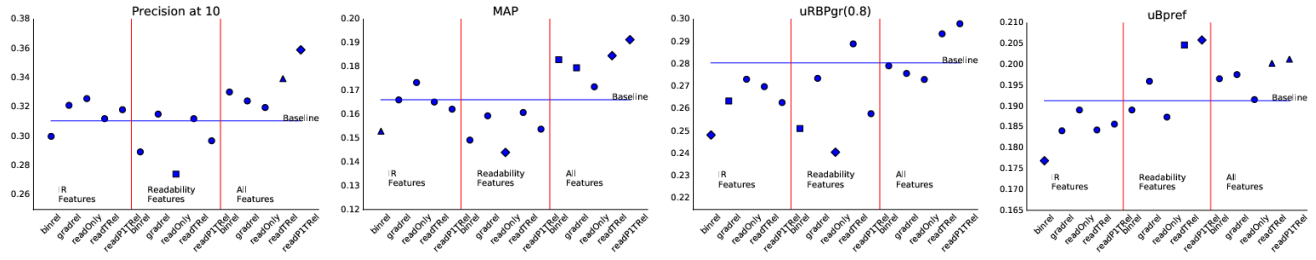

Figure 1: Comparison of different combinations of feature sets and relevance functions (x-axis). The score of the baseline system (BM25) for each metric is shown using the horizontal line. Diamond (square/triangle) markers are for experiments when $p < 0.05$ ($p < 0.1$ / $p < 0.15$) for a paired t-test w.r.t the baseline system.

is hard to read (label 0), somewhat hard to read (1), somewhat easy to read (2), and easy to read (3). Almost 70% of the documents are judged as easy or somewhat easy to read, and just 2% are both highly relevant and easy to read.

4.2 Experiment Settings

We used the learning to rank framework provided in Terrier 4.0 with Jforest and LambdaMART [5]. To extract readability features, we preprocessed the documents using *boilerplate* following the methods by Palotti et al. [7]. To learn and evaluate the learning to rank models, we used a leave-one-out approach where we train models using 60 topics, validate them using 5 topics to optimise the number of iterations and finally test with the learned model using a single topic; this process is repeated to test on all 66 topics. We set nDCG as the metric to optimize as it considers different grades of relevance, and we explored five functions $f(d)$ for the relevance label assigned to each document d in the training and validation steps (listed in Table 2). *ReadTRel* is the direct product combination of topical relevance and understandability, which would assign label 0 (i.e., not relevant) to documents that have understandability of 0 (i.e., very hard to read). As this renders irrelevant those relevant documents with an understandability score of 0, we designed the function *readP1TRel* in which the relevance of very hard to understand documents has a not null contribution to the score of the documents.

4.3 Retrieval evaluation

We first tested retrieval and readability features[2] separately; we then evaluated the effectiveness of their combination. The two left-most plots of Figure 1 report the values of P@10 and MAP obtained by the learning to rank approaches. A baseline system based on BM25 with parameters set to their default values in Terrier is also shown as a horizontal line. Note that P@10 was the primary measure used in CLEF 2015. As would be expected, the results show that retrieval features contributed more than readability ones to increase the relevance of document ranking. However, the best P@10 and MAP were obtained when both feature types were combined, suggesting that (1) learning to rank using both feature types improves the general result ranking, and (2) readability features improve relevance-based ranking.

[2]In these experiments, we included in the readability features also the raw retrieval score of the baseline (BM25), but not all other retrieval features.

Function Name	Description
binrel	binary relevance only
gradrel	graded relevance only
readOnly	graded readability only
readTRel	graded readability × graded relevance
readP1TRel	(graded readability + 1) × graded relevance

Table 2: The five variants used for document labels for the learning to rank approach.

4.4 Understandability-biased evaluation

In this section, we study effectiveness according to the understandability-biased evaluation. Figure 1 reports the value of uRBPgr, a graded version of RBP where the gain of a document is a joint function of the relevance label and the understandability label [15, 16]. The persistency parameter of RBP (ρ) was set to 0.8 (as in previous work [15, 16]). The results show that retrieval features alone did not improve uRBPgr; neither did the readability features alone. Their combination however improved uRBPgr, but not consistently across different label functions.

Further analysis of the results revealed that rankings obtained by the learning to rank models trained with readability only features contained many unassessed documents (on average only $\sim$ 79% of top 10 documents were assessed); while most of the documents obtained with retrieval features only or combination were assessed ones ($\sim$ 94% of the top 10 documents were assessed). Thus, results for readability only features may be affected by the lower coverage of the assessments. To overcome this limitation, we adopted a version of Bpref modified in the spirit of the understandability-biased evaluation framework (uBpref). Bpref considers only documents that have been explicitly assessed with respect to their relevance. uBpref also considers only assessed documents (for relevance and for readability); the (binary) gain from the relevance status of an assessed document (0: irrelevant, 1: relevant) is multiplied by the graded gain from the understandability assessment (with weights as in uRBPgr).

Figure 1 reports the results evaluated with uBpref. The results suggest that the effectiveness of learning to rank with readability only features was underestimated because of the many unassessed documents. Readability only features, in fact, led to increased uBpref over the baseline; often higher than the combination. While using all features did not lead to the highest uBpref, they provided consistent gains over the baseline across different label functions.

4.5 Feature analysis

We performed a feature ablation study to analyse the impact of features on the effectiveness of systems. We experimented with the best two models from our previous experiment (Figure 1), using all features (combining retrieval and readability features) and using as document labels the prod-

Funct.	System	P@10	uRBPgr	uBPref
-	**Baseline**	0.3106	0.2805	0.1913
readTRel	**All Features**	0.3394 △	0.2935	0.2003 △
	All - Formulas	0.3409 ◇	0.2915	0.1970
	All - Surface	**0.3606** ◇	**0.2999**	**0.2009** △
	All - General Voc.	0.3348	0.2948	0.1974
	All - Medical Voc.	0.3379 △	0.2879	0.1983
readP1TRel	**All Features**	**0.3591** ◇	**0.2980**	**0.2013** △
	All - Formulas	0.3258	0.2820	0.1951
	All - Surface	0.3485 ◇	0.2797	0.1951
	All - General Voc.	0.3439 □	0.2773	0.1916
	All - Medical Voc.	0.3303	0.2899	0.1994

Table 3: Feature ablation study based on the two best methods in Figure 1. The best results are in bold. Diamonds, squares, triangles markers indicate statistical significance (paired t-test w.r.t. baseline) with $p < 0.05$, $p < 0.1$, $p < 0.15$, respectively.

System	BPREF	τ_{AP}
BM25 w.o. learning	0.4432	0.4628
Readability	0.4682	0.4994
ReadTRel	0.4451	0.4459
ReadP1TRel	0.4443	0.4483

Table 4: Effectiveness comparison of different system variations with respect to the rank obtained using understandability assessments.

uct of topical relevance and readability scores ($readTRel$ and $readP1TRel$). At each ablation step, we removed a feature group from the readability features, and then learned and evaluated a new model. Retrieval results are shown in Table 3. When document labels were assigned using $readTRel$ as a function, the removal of some feature groups improved results over the model that used all features. For example, removing surface features improved the results of P@10 from 0.3394 to 0.3606.

4.6 Is Understandability Learnt?

We analysed whether learning to rank did learn to prefer more understandable documents. We trained a model using only readability features, not including the raw retrieval score of the baseline, and experimented with a different function for document labels. We report Bpref for each system variation with respect to understandability assessments only, as we are only interested to know if the assessed documents were correctly ranked. Additionally, we compared the ranking generated by each system with the perfect order of documents according to their understandability; for that, we used τ_{AP} [14], as it is based on average precision and assigns more weight to differences in the top rankings[3].

The learning to rank model learned how to better rank documents according to understandability when it was exclusively trained with functions based on understandability labels (see Table 4, second row). In this case, baseline effectiveness was improved by 5% for Bpref (where understandability assessments are used in place of relevance ones) and 7% for τ_{AP}. On the other hand, when the model was trained combining readability and retrieval features, document understandability did not appear to be learnt.

5. CONCLUSIONS

We describe a method that integrates understandability in the ranking of search results for consumer health search. This method is based on a learning to rank approach that combines features capturing topical relevance and features measuring the readability of health documents. We found that the combination of retrieval features and readability features indeed did improve search engine results, both for relevance and understandability retrieval measures. The provision of documents that are both relevant and understandable in answer to health related queries is an important requirement for next generation search engines [8]. Source code, all data analysed in this paper, and results (including other baseline models and evaluation measures) are available online at http://github.com/ielab/sigir2016-ranking-relevance-understandability.

As future work, we want to perform a wider feature analysis and evaluation. The results in Section 4.5, in fact, indicate that feature selection may improve system effectiveness.

[3]Usually τ_{AP} is used to compare systems rankings; here we used it to compare documents rankings.

Acknowledgement

JP and AH were supported by Horizon 2020 program (H2020-ICT-2014-1) n°644753 (KCONNECT). JP has also been supported by the ESF project ELIAS.

6. REFERENCES

[1] M. Benigeri and P. Pluye. Shortcomings of health information on the internet. *Health Prom. Inter.*, 2003.

[2] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proc. of CIKM*, 2011.

[3] W. H. Dubay. The principles of readability. *Costa Mesa, CA: Impact Information*, 2004.

[4] S. Fox and M. Duggan. Health online. Technical report, Pew Research, 2013.

[5] C. Macdonald, R. L. Santos, I. Ounis, and B. He. About learning models with multiple query-dependent features. *ACM Trans. Inf. Syst.*, 2013.

[6] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. Jones, M. Lupu, and P. Pecina. CLEF eHealth Evaluation Lab 2015, Task2: Retrieving Information About Medical Symptoms. In *Proc. of CLEF*, 2015.

[7] J. Palotti, G. Zuccon, and A. Hanbury. The Influence of Pre-processing on the Estimation of Readability of Web Documents. In *Proc. of CIKM*, 2015.

[8] N. Pletneva, A. Vargas, and C. Boyer. D8.1.1. requirements for the general public health search. Technical report, Khresmoi Project, 2011.

[9] C. Tan, E. Gabrilovich, and B. Pang. To each his own: personalized content selection based on text comprehensibility. In *Proc. of WSDM*, 2012.

[10] R. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. In *ACM Trans. on Inf. Sys*, 2008.

[11] R. C. Wiener and R. Wiener-Pla. Literacy, Pregnancy and Potential Oral Health Changes: The Internet and Readability Levels. *Maternal and child health journal*, 2013.

[12] D. Wu, D. Hanauer, Q. Mei, P. Clark, L. An, J. Lei, J. Proulx, Q. Zheng-Treitler, and K. Zheng. Applying multiple methods to assess the readability of a large corpus of medical documents. In *Proc. of MEDINFO*, 2013.

[13] X. Yan, D. Song, and X. Li. Concept-based Document Readability in Domain Specific Information Retrieval. In *Proc. of CIKM*, 2006.

[14] E. Yilmaz, J. A. Aslam, and S. Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *Proc. of SIGIR*, 2008.

[15] G. Zuccon. Understandability biased evaluation for information retrieval. In *Proc. of ECIR*, 2016.

[16] G. Zuccon and B. Koopman. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *MedIR Workshop at SIGIR*, 2014.

[17] G. Zuccon, B. Koopman, and A. Nguyen. Retrieval of Health Advice on the Web: AEHRC at ShARe/CLEF eHealth Evaluation Lab Task 3. In *CLEF eHealth*, 2013.