Integrating Understandability in the Evaluation of Consumer Health Search Engines

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

Bevan Koopman
Australian e-Health Research Centre, CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

ABSTRACT

In this paper we propose a method that integrates the notion of understandability, as a factor of document relevance, into the evaluation of information retrieval systems for consumer health search. We consider the gain-discount evaluation framework (RBP, nDCG, ERR) and propose two understandability-based variants (uRBP) of rank biased precision, characterised by an estimation of understandability based on document readability and by different models of how readability influences user understanding of document content. The proposed uRBP measures are empirically contrasted to RBP by comparing system rankings obtained with each measure. The findings suggest that considering understandability along with topicality in the evaluation of information retrieval systems lead to different claims about systems effectiveness than considering topicality alone.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval
General Terms: Evaluation.

1. INTRODUCTION

Searching for health advice on the Web is an increasingly common practice. A recent research has found that in 2012 about 58% of US adults (72% of all US Internet users – 66% in 2011) have consulted the Internet for health advice [5]; of these, 77% have used search engines like Google, Bing, or Yahoo! to gather health information, while only 13% have started their health information seeking activities from specialised sites such as WebMD. It is, therefore, crucial to create and evaluate information retrieval (IR) systems that specifically support consumers searching for health advise on the Web. In this paper, we focus on the evaluation of IR systems for consumer health search.

Previous studies within health informatics have investigated online consumer health information beyond topicality to specific health topics; in particular, with respect to the understandability and reliability of such information. For example, Wiener and Wiener-Pla [11] have investigated the readability (measured by the SMOG reading index [7]) of Web pages concerning pregnancy and the periodontium as retrieved by Google, Bing and Yahoo!. Walsh and Volsko [10] have shown that most online information sampled from five US consumer health organisations and related to the top 5 medical related causes of death in US is presented at a readability level (measured by the SMOG, FOG and Flesch-Kincaid reading indexes [7]) that exceeds that of the average US citizen (7th grade level). Ahmed et al. [1] have highlighted the variability in readability (measured by the Flesch Reading Ease and the Flesch-Kincaid reading index [7]) and quality of concussion information accessed through Google searches. The understandability and reliability of online health information has been considered as a critical issue for supporting online consumer health search because (1) consumers may not benefit from health information that is not provided in an understandable way; and (2) the provision of unreliable, misleading or false information on a health topic, e.g., a medical condition or treatment, may led to negative health outcomes. This previous research suggests that topicality should not be considered as the only relevance factor for assessing the effectiveness of IR systems for consumer health search: other factors, such as understandability and reliability, should also be included in the evaluation framework.

Research on the user perception of document relevance has shown that users' relevance assessments are affected by a number of factors beyond topicality, although topicality has been found to be the essential relevance criteria. For example, Xu and Chen proposed and validated a five-factor model of relevance which consists of novelty, reliability, understandability, scope, along with topicality [12]. Their empirical findings highlight the importance of understandability, reliability and novelty along with topicality in the relevance judgements they collected. Nevertheless, typical evaluation of IR systems commonly considers only relevance assessments in terms of topicality[1]; this is also the case when evaluating systems for consumer health search, for example, within CLEF eHealth 2013 [6]. In this paper, we aim to close this gap in the evaluation of IR systems and focus on integrating understandability along with topicality for the evaluation of consumer health search engines. The integration of other factors influencing relevance, such as reliability, are left for future work.

The integration of understandability within the evaluation methodology is achieved by extending the general gain-discount framework synthesised by Carterette [3]; this framework encompasses the widely-used nDCG, RBP and ERR. The result is a series of understandability-biased evaluation measures. Specifically, we examine one such measure, the understandability-based rank biased precision (uRBP) – a variant of rank biased precision (RBP) [8]; variants of nDCG

MedIR July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

[1]With the recent exception of novelty and diversity, e.g., [4].

and ERR may also be derived within our framework.

The proposed evaluation measure is further instantiated by considering specific estimations of understandability based on readability measures computed for each retrieved document. While understandability encompasses other aspects in addition to text readability (e.g., prior knowledge), the use of readability measures is a good first approximation for understandability. This choice is also supported by prior work in health informatics regarding understandability of consumer health information (e.g., see [11, 10, 1]).

The impact of the proposed framework and the specific resultant measures on the evaluation of IR systems is investigated in the context of the consumer health search task of CLEF eHealth 2013 [6]; empirical findings show that systems that are most effective according to uRBP are not necessarily as effective when considering topicality alone (i.e. RBP).

2. UNDERSTANDABILITY-BASED EVALUATION

2.1 The gain-discount framework

We tackle the problem of jointly evaluating topicality and understandability for measuring IR system effectiveness within the gain-discount framework synthesised by Carterette [3]. Within this framework, the effectiveness of a system, conveyed by a ranked list of documents, is measured by the evaluation measure M, defined as:

$$M = \frac{1}{\mathcal{N}} \sum_{k=1}^{K} g(k)d(k) \tag{1}$$

where $g(k)$ and $d(k)$ are respectively the gain and discount function computed for the document at rank k,[2] K is the depth of assessment at which the measure is evaluated, and $1/\mathcal{N}$ is a (optional) normalisation factor, which serves to bound the value of the sum into the range [0,1] (see [9]).

Different measures developed within the gain-discount framework are characterised by different instantiations of its components. For example, the discount function in RBP is modelled by $d(k) = \beta^{k-1}$, where $\beta \in [0, 1]$ reflects user behaviour (high values representing persistent users, low values representing impatient users); while in nDCG the discount function is given by $d(k) = 1/(\log_2(1 + k))$ and in ERR by $d(k) = 1/k$. Similarly, instantiations of gain functions differ depending upon the considered measure. In RBP, the gain function is binary-valued (i.e., $g(k) = 1$ if the document at rank k is relevant, $g(k) = 0$ otherwise); while for nDCG $g(k) = 2^{r(k)} - 1$ and for ERR $g(k) = (2^{r(k)} - 1)/2^{r_{max}}$ (with $r(k)$ being the relevance grade of the document at rank k).

Without loss of generality, we can express the gain provided by a document at rank k as a function of its probability of relevance; for simplicity we shall write $g(k) = f(P(R|k))$, where $P(R|k)$ is the probability of relevance given the document at rank k. Note that a similar form has been used for the definition of the gain function for time-biased evaluation measures [9]. The specific instantiations of $g(k)$ in measures like RBP, nDCG and ERR can be seen as the application of different functions $f(.)$ to estimations of $P(R|k)$.

Traditional TREC-style relevance assessors are instructed to consider topicality as the only (explicit) factor influencing relevance, thus $P(R|k) = P(T|k)$, i.e., the probability that the document at k is topically relevant (to a query).

2.2 Integrating understandability

As discussed by previous work, e.g. [12], relevance is influenced by many factors; topicality being only one of them – although the most important. To integrate understandability into the gain-discount framework, we model $P(R|k)$ as the joint $P(T, U|k)$, i.e. the probability of relevance of a document (at rank k) is estimated using the joint probability of the document being topical and understandable.

To compute the joint probability we assume that topicality and understandability are compositional events and their probabilities independent, i.e., $P(T, U|k) = P(T|k)P(U|k)$. This is a strong assumption and its limitations are briefly discussed in Section 4. Following this assumption, the gain function in the gain-discount framework is expressed as:

$$g(k) = f(P(R|k)) = f\big(P(T|k)P(U|k)\big) \tag{2}$$

Different evaluation measures that may be developed within this framework would instantiate $f\big(P(T|k)P(U|k)\big)$ in different ways. In the following we will propose two RBP-based instantiations; other instantiations are left for future work.

2.3 Estimating understandability

In the traditional TREC settings, assessments about the topicality of a document to a query are collected through manual annotation of query-document pairs from assessors (i.e., binary or graded relevance assessments[3]); these are then turned into estimations of $P(T|k)$. This process may be mimicked to collect understandability assessments; in this paper however we do not explore this possibility. Instead, we explore the possibility of computing understandability as a property of a document and integrate this in the evaluation process, along with standard relevance assessments. To this aim, readability is used as a proxy for understandability. (The limitations of this choice are briefly noted in Section 4.) Its use is however justifiable because readability is one of the aspects that influence the understanding of text.

To estimate readability (and thus understandability), we employ established general readability measures as those used in [1, 10, 11], e.g., SMOG, FOG and Flesch-Kincaid reading indexes. These measures consider the surface level of the text contained in Web pages, that is, wording and syntax of sentences. In this framework, the presence of long sentences, words containing many syllables and unpopular words, are all indicators of difficult text to read [7]. In this paper, we use the FOG measure to estimate the readability of a text; the FOG reading level is computed as

$$FOG(d) = 0.4 * (avgslen(d) + phw(d)) \tag{3}$$

where $avgslen(d)$ is the average length of sentences in a document d and $phw(d)$ is the percentage of hard words (i.e., words with more than two syllables) in d.

The use of such general readability measures to assess the readability of documents concerning health information has been questioned [13] as these do not seem to adequately correlate with human judgments for documents in this domain [13]. Nevertheless, the adoption of standard readabil-

[2] For simplicity of notation, in the following we override k to represent the *rank position* k, or the *document* at rank k: the context of use will determine the meaning of k.

[3] Recall that although called "relevance assessments", in TREC-style assessments, annotators are usually instructed to consider only the topicality of a document to a query, isolating this factor from others influencing relevance in real settings.

ity measures in this paper is a first step towards demonstrating the use of the proposed understandability biased measures and analyse how system rankings would change accordingly. In addition, their usage is partially supported by previous work within health informatics on assessing the readability of online health advice [1, 10, 11].

2.4 Modelling P(U|k)

Given the readability score for a document at rank k, $P(U|k)$ needs to be estimated; this is achieved by considering user models that encode different ways in which a user is affected by document readability.

We first consider a user model $P_1(U|k)$ where a user is characterised by a readability threshold th and every document that has a readability score below th is considered certainly understandable, i.e., $P_1(U|k) = 1$; while documents with readability above th are considered not understandable, i.e. $P_1(U|k) = 0$. This is a (Heaviside) step function centred in th; this function is depicted in Figure 1 ($P_1(U|k)$) with $th = 20$, along with the FOG readability score distribution for documents from CLEF e-Health 2013 [6]. The use of a step function to model $P(U|k)$ is akin to the gain function in RBP (also a step function). The understandability-based RBP for user model one is then given by:

$$uRBP_1 = (1-\beta)\sum_{k=1}^{K} \beta^{k-1} r(k) u_1(k) \tag{4}$$

where, for simplicity of notation, $u_1(k)$ indicates the value of $P_1(U|k)$ and $r(k)$ is the (topical) relevance assessment of document k (alternatively, the value of P(T|k)); thus $g(k) = f(P(T|k)P(U|k)) = P(T|k)P(U|k) = r(k)u_1(k)$.

A second user model ($P_2(U|k)$) is proposed, where the probability estimation is similar to a step function, but smoothed in the surroundings of the threshed value; this provides a more realistic transition between readable and not-readable content:

$$P_2(U|k) \propto \frac{1}{2} - \frac{\arctan\Big(FOG(k) - th\Big)}{\pi} \tag{5}$$

where arctan is the arctangent trigonometric function and $FOG(k)$ is the FOG readability score of document at rank k; other readability scores could be used instead of FOG. The distribution of $P_2(U|k)$ values is shown in Figure 1. Equation 5 is not a proper probability distribution, but this can be obtained by normalising Equation 5 by its integral between $[\min\big(FOG(k)\big), \max\big(FOG(k)\big)]$; however Equation 5 is rank equivalent to such distribution, not changing the effect on the uRBP variant. These settings lead to the formulation of a second understandability-based RBP, $uRBP_2$, based on the second user model, by simply substituting $u_2(k) = P_2(U|k)$ to $u_1(k)$ in Equation 4.

Note that in both understandability-based measures (as well as in the original RBP) the contribution of an irrelevant document is zero, irrespective of its $P(U|k)$. The contribution (to the gain) of a relevant document with readability score above th is 1 for RBP, 0 for $uRBP_1$ and less than 0.5 for $uRBP_2$ (for $uRBP_2$ the score will quickly tend to 0 the more the readability score is above the threshold value).

Finally, note that it is possible to design other user models representing how readability influences document understandability; the challenge is to determine which model better represents the relationship between readability and document understanding.

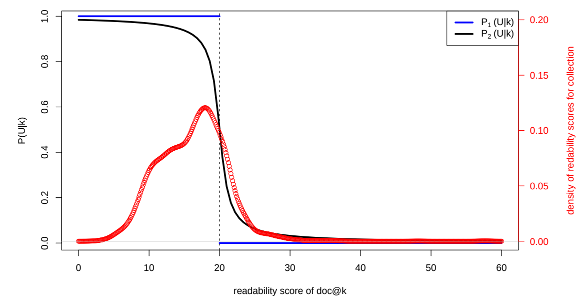

Figure 1: Distributions for $\mathbf{P_1(U|k)}$ and $\mathbf{P_2(U|k)}$ with respect to threshold $\mathbf{th = 20}$, along with the density distribution of readability scores (computed using FOG) for the documents in the CLEF eHealth 2013 qrels.

3. EMPIRICAL ANALYSIS

3.1 Experiment design and settings

To understand how accounting for understandability influences the evaluation of IR systems tailored to searching health advice on the Web, we consider the runs submitted to the CLEF eHealth 2013 [6], which specifically aimed at evaluating systems for this task. Our empirical experiments and subsequent analysis specifically focus on the changes in system rankings obtained when evaluating with standard measures (RBP) and understandability-based measures ($uRBP_1$ and $uRBP_2$). System rankings are compared using Kendall rank correlation (τ) and AP correlation [14] (τ_{AP}), which weights higher rank changes that affect top systems. We do not experiment with different values of β in RBP, and set $\beta = .95$ across RBP and uRBP.

The document collection used in CLEF eHealth 2013 has been retired due to removal of duplicates and copyrighted documents; we thus use the CLEF eHealth 2014 collection (which is a subset of the CLEF eHealth 2013 collection) to allow reproducibility of the reported results and the 2013 qrels for relevance assessment. For each document in the collection, the FOG readability scores (Equation 3) were computed – the score distribution for all documents in the CLEF eHealth 2013 qrels is shown in Figure 1. Three thresholds on the FOG readability values were explored for the computation of the two alternative formulations of uRBP: $th = 10, 15, 20$; documents with a FOG score below 10 should be near-universally understandable, while documents with FOG scores above 15 and 20 increasingly restrict the audience able to understand the text.

3.2 Results and analysis

Figure 2 reports RBP vs. uRBP of IR systems participating to CLEF eHealth 2013 for the two user models proposed in Section 2.4 and for the three readability thresholds considered in the experiments. Similarly, Table 1 reports the values of Kendall rank correlation (τ) and AP correlation (τ_{AP}) between system rankings obtained with RBP and the two versions of uRBP.

Higher correlation between systems rankings obtained with RBP and uRBP is observed for higher values of th, irrespectively of uRBP version (see Table 1). This is expected as the higher the threshold, the more documents will be characterised by a $P(U|k) = 1$ (or ≈ 1 for $uRBP_2$), thus reducing uRBP to RBP. The fact that in general $uRBP_2$ is correlated with RBP more than $uRBP_1$ is to RBP highlights the effect of smoothing obtained by the arctan function; specifically, the increase of readability scores for which $P(U|k)$ is not zero

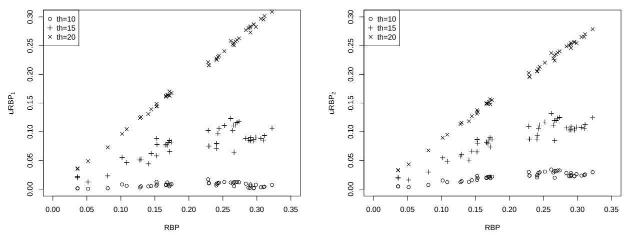

Figure 2: RBP vs. uRBP of CLEF eHealth 2013 systems (left: $\mathbf{uRBP_1}$; right: $\mathbf{uRBP_2}$) for varying values of threshold on the readability scores ($\mathbf{th = 10, 15, 20}$).

	$th = 10$	$th = 15$	$th = 20$
RBP vs.	$\tau = .1277$	$\tau = .5603$	$\tau = .9574$
$uRBP_1$	$\tau_{AP} = -.0255$	$\tau_{AP} = .2746$	$\tau_{AP} = .9261$
RBP vs.	$\tau = .5887$	$\tau = .6791$	$\tau = .9574$
$uRBP_2$	$\tau_{AP} = .2877$	$\tau_{AP} = .4102$	$\tau_{AP} = .9407$

Table 1: Correlation coefficients (τ and $\tau_{\mathbf{A}}$P) between system rankings obtained with RBP and $\mathbf{uRBP_1}$ or $\mathbf{uRBP_2}$ for different values of the readability threshold.

beyond th narrows the scope for ranking differences between systems effectiveness. These observations are confirmed in Figure 2, where only few changes in the rank of systems are shown for $th = 20$ ($\times$ in Figure 2), while more changes are found for $th = 10$ ($\circ$) and $th = 15$ ($+$). Note that the small differences in the absolute values of effectiveness recorded by uRBP with $th = 10$ should not be interpreted as a lack of discriminative power. When $th = 10$ only 1.4% of the documents in the CLEF eHealth 2013 qrels are relevant and readable, thus contributing to uRBP.

Figure 2 demonstrates the importance of considering understandability along with topicality in the evaluation of systems for the considered task. The system ranked highest according to RBP (`MEDINFO.1.3.noadd`) is second to a number of systems according to uRBP if user understandability of up to FOG level 15 is wanted. Specifically, the highest $uRBP_1$ for $th = 10$ is achieved by `UTHealth_CCB.1.3.noadd`, which is ranked 28th according to RBP, and for $th = 15$ by `teamAEHRC.6.3`, which is ranked 19th according to RBP and achieves the highest $uRBP_2$ for $th = 10, 15$.

4. LIMITATIONS AND CONCLUSIONS

In this paper, we have investigated how understandability can be integrated in the gain-discount framework for evaluating IR systems. The approach studied here is general and can be adopted to other factors of relevance, such as reliability. Information reliability plays an important role in consumer health advice search; its integration will be studied in future work.

In the proposed approach, the relevance (P(R|k)) was modelled as the joint probability $P(T, U|k)$. This joint probability was assumed to be independent and the two events to be compositional, thus allowing to derive $P(T, U|k) = P(T|k)P(U|k)$ and to treat topicality and understandability separately. This is a strong assumption and it is not necessarily true; alternatives are under investigation, e.g. [2].

The approach was demonstrated by deriving understandability-based variants of RBP; other measures can also be extended, e.g., nDCG and ERR. Note, however, that nDCG-style versions would require normalising the gain function by the ideal gain, which in turns requires finding the optimal ranking based on two criteria, relevance score and understandability, instead of one as in the standard nDCG.

Xu and Chen [12] have noted that factors of relevance influence relevance assessments in different proportions, e.g., in their study, topicality was found to be more influential than understandability. The specific uRBP measures studied here did not consider this aspect; however weighting of different factors could be accomplished through a different $f(.)$ function for converting $P(T, U|k)$ into gain values.

In this paper, we have used readability as a proxy for understandability, but this is only one aspect that influences understandability [12]; future work may explore other factors, e.g., users' prior knowledge, as well as the presence of images that further explain the textual information. Furthermore, readability was estimated using general, surface level readability measures. Previous work has shown that these measures are often not suitable to evaluate the readability of health information. For example, Yan et al. [13] claim that people experience the highest readability difficulties at word level rather than at sentence level; they further propose a new metric based on concept-based readability, specifically instantiated in the health domain. A number of alternative approaches that measure text readability beyond the surface characteristics of text have been proposed. Future work will investigate their use to estimate P(U|k), along with actual readability assessments collected from users.

5. REFERENCES

[1] O. H. Ahmed, S. J. Sullivan, A. G. Schneiders, and P. R. McCrory. Concussion information online: evaluation of information quality, content and readability of concussion-related websites. *British journal of sports medicine*, 46(9):675–683, 2012.

[2] P. D. Bruza, G. Zuccon, and L. Sitbon. Modelling the information seeking user by the decision they make. In *MUBE 2013*, pages 5–6. ACM, 2013.

[3] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *SIGIR'11*, pages 903–912, 2011.

[4] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *TREC'09*, 2009.

[5] S. Fox and M. Duggan. Health online 2013. Tech. Rep., Pew Research Center's Internet & American Life Project, 2013.

[6] L. Goeuriot, G. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, 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.

[7] D. R. McCallum and J. L. Peterson. Computer-based readability indexes. In *ACM'82 Conf.*, pages 44–48, 1982.

[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):2, 2008.

[9] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *SIGIR'12*, pages 95–104, 2012.

[10] T. M. Walsh and T. A. Volsko. Readability assessment of internet-based consumer health information. *Respiratory care*, 53(10):1310–1315, 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*, pages 1–6, 2013.

[12] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *JASIST*, 57(7):961–973, 2006.

[13] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *CIKM'06*, pages 540–549, 2006.

[14] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR'08*, pages 587–594, 2008.