Predicting the Situational Relevance of Health Web Documents

Melinda Oroszlányová[a], Carla Teixeira Lopes[a,b], Sérgio Nunes[a,b], Cristina Ribeiro[a,b]

[a]Department of Informatics Engineering (DEI), Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal

[b]INESC TEC, Porto, Portugal

Abstract — **Relevance is usually estimated by search engines using document content, disregarding the user behind the search and the characteristics of the task. In this work, we look at relevance as framed in a situational context, calling it situational relevance, and analyze if it is possible to predict it using documents, users and tasks characteristics. Using an existing dataset composed of health web documents, relevance judgments for information needs, user and task characteristics, we build a multivariate prediction model for situational relevance. Our model has an accuracy of 77.17%. Our findings provide insights into features that could improve the estimation of relevance by search engines, helping to conciliate the systemic and situational views of relevance. In a near future we will work on the automatic assessment of document, user and task characteristics.**

Keywords - health information retrieval; Web; situational relevance.

I. Introduction

It is estimated that 3.5 billion individuals (47.3% of the population) were Internet users in 2016 worldwide [26]. The number of users increased and so did the amount of information that has become available to the users in the past decades, including consumer-oriented health information. Consequently, the number of people being affected by such information has also increased. Studies have shown that people consider the Internet to be a credible source when seeking health-related information [22, 12]. The latest national survey reported that in 2012, among all adults in the U.S., 72% looked online for health information [6]. Several user studies have been conducted with the aim to learn how people use online resources for their health concerns [5, 3], and how internet users search for health information on the Web [4, 6]. The goal of current research is to assess and improve relevance estimation of consumer-oriented health information on the Web. Search engines typically estimate relevance using document characteristics [20], leaving out features from users and tasks that can be useful for relevance estimation. The objective of the present study is to analyze which characteristics influence the relevance of health web documents, with the help of an existing dataset, composed by annotated web pages, characteristics, users, tasks and relevance judgments. We aim to find good descriptors and potential predictors of situational relevance.

II. Background

A. The Concept of Relevance in Information Retrieval

The notion of relevance has been studied for decades. Several information retrieval (IR) models have been developed to predict documents' relevance (e.g., the classical Boolean model, vector space model and probabilistic model). Generally, they consist of a framework including representations of documents, queries, relationships among them and, in some cases, a ranking function. IR models rely on evaluations which consider traditional user and task models. Such models are though inadequate, as for example they do not capture all types of information-seeking tasks, activities, and situations [11]. These models do not seem to be sufficient to approximate the relevance judgments [1]. It is important to consider the associated context, other than just document properties. A range of relevance models have been introduced and discussed, from Saracevic's stratified model of interaction levels [20] till Borlund's model [1]. The stratified model is based on theoretical concepts of human-computer interaction (HCI), and the stratificational theory developed in linguistics. It considers the contemporary reality of IR, and the nature of relevance in information science, and optimizes the strengths, and minimizes the weaknesses of both the systems-centered and user-centered approaches to IR [20]. Borlund's model is based on an analytic approach, considering the temporal dimension [1]. Besides explicit relevance models [25], multidimensional relevance modeling has been as well studied [27].

Every search engine has to estimate the usefulness of the information accessed via web pages, referred to as the relevance of a document to a user [20]. As the retrieval of relevant information is the main concern of any IR system [15], there are several types of user-based relevance, depending on the context and on the user. In this study we consider situational relevance (i.e., utility), expressed by the usefulness of the documents to the user task [20]. The aim of the estimation of situational relevance is that knowing how relevance depends on the user and document characteristics can bring insights on new features. For this reason, the concept of relevance, that is, the retrieval of relevant information, is central to information retrieval in all domains [15]. In IR processes, the role of users is also an important factor in relevance assessment. Users evaluate web pages and decide about their utility for different information-seeking tasks, based on certain criteria. Such features include textual, structural and qualitative aspects, as well as non-textual items and physical properties of the web documents [24]. Research, topic, scope, data, influence, affiliation, web characteristics, and authority have been identified as key relevance criteria [2], indicating the complexity of web users' relevance judgments, and are important in the design of IR systems. Other user-defined relevance criteria such as specificity, topicality, familiarity and variety are frequently used in relevance judgments [21].

B. *Consumer Health Information Seeking*

Since the 90's, when a guide to the Internet was introduced by Pallen [18], the healthcare providers started to share information on medical and health topics with the public on the Web. More and more information had became available, and online health seekers started to look for the information not only for themselves, but also often for someone else [7, 4, 5, 3]. Thus, the health search has an impact on people's health care routines. Theoretical models of health IR are summarized in [16], where the reviewed studies suggest the usefulness of multidisciplinary approaches and of conceptual models. A wild range of literature about IR evaluation has been reviewed, providing "a baseline for the growth and maturation of the specialty" [16]. This historical overview documents the evolution of the IR evaluation methods of 40 years, analyzing 127 selected articles, which the readers can use as a baseline bibliography of the area.

Based on the results of the evaluation of user-centered health information retrieval, the development of retrieval techniques for medical queries for lay users proved difficult [9]. Related research on automatic generation of queries [8] explores new topic generation strategies, with the aim of generating queries that are representative of patients information needs. Investigation on the effectiveness of search engines in retrieving information about medical symptoms has been conducted, focusing on designing systems which improve health search [19]. It resulted in the conclusion that query expansion is an important factor in improving search effectiveness. Further development of search technologies for consumer health search considers self-diagnosis information needs and needs related to treatment and management of health conditions [28]. The relevance assessments were shown to be influenced by user, task, query and document characteristics (e.g., age, gender, health search experience, medical specialty, task clarity) [13]. A previous study showed that user and task characteristics are also good descriptors and possible predictors of relevance [17]. In the present work we want to predict the relevance of a document to a user, with the help of the available features [13].

III. Methodology

A. *Datasets*

The present study is based on an existing dataset composed by an annotated sample of 4533 health web documents. It was initially collected for a user study [14], where the participants performed 8 tasks, associated with different health information seeking situations, based on questions submitted to the health category of the Yahoo! Answers service. From the list of open questions of this category, starting with the most popular one, 8 questions about treatments to a symptom/disease were selected. For each question 4 different search queries were defined, 2 in English and 2 in the participants native language. In each language, the 2 queries were formulated by using lay and medico-scientific terminology, respectively. Queries were built concatenating the 8 symptoms or diseases (painful urination/dysuria; head itching/head pruritus; high uric acid/hyperuricaemia; mouth inflammation/stomatitis; bone infection/osteomyelitis; heartburn/pyrosis; hair loss/alopecia; joint pain/arthralgia) with the word treatment with different medical terminology (lay/medico-scientific). To reduce the risk of Google learning from the previous submitted queries, it was ensured that returned links were never clicked. Further, to prevent changes in the search engine, all queries were submitted within a very short time span. For each query, the top-30 results were collected. For these documents, a metadata scheme was defined and used for a latter annotation with manual and automatic approaches [23]. The documents were assessed by university students in terms of relevance and comprehension, using a 3-valued scale. To evaluate the quality of the annotation, 10% of the documents were also assessed by an external health professional [23]. The agreement rate between both assessments was measured through Kappa de Cohen, where 38 indicators had concordance values greater than 0.8, 3 indicators had concordance values between 0.6 and 0.8, and 1 indicator had between 0.4 and 0.6. Thus, the way the characteristics were evaluated/annotated was, in general, well defined. Information about the users has been collected through questionnaires. The metadata scheme that was used to annotate the dataset contains specific characteristics of web documents, tasks and users, listed in Table I. The document features were categorized according to its content (e.g.: is it readable? is it a scientific publication?), to its web characteristics (e.g.: articles, academic works), to the entity responsible for the website (e.g.: are there contacts of the author and web-master? is it of scientific nature?), and to the website (e.g.: its objective, domain or type). Task related characteristics include users' feedback on the tasks clarity, easiness and familiarity. User characteristics describe the user in terms of their age, English proficiency, health literacy and health search experience.

In the present work, situational relevance is assessed by a question where users were asked to evaluate the usefulness of each document in a 3-level scale (0 - non-relevant, 1 - partially relevant, 2 - totally relevant). The task characteristics contain the comprehension of the documents by the users, which has 3 assessment levels, as described in Table I.

TABLE I. Description of the Documents, Task and User Characteristics

Characteristics/Variables	**Scale**
Documents - Content	
Readability indicators (Ari, Colemanliau, Fleschkincaid, Fleschreading, Gunningfog, Smog, Smogindex)	Continuous

Characteristics/Variables	Scale
The possible impact of information on the user, e.g., the use of "positive" or "negative" expressions (character of the information)	0-Negative 1-Neutral 2-Positive
Existence of "real" cases given by specialists (clinical cases)	0-Not present 1-Present
Whether the content is divided into several pages in case of html formats (split content)	0-Not present 1-Present
Language of the content (annotated according to ISO 639-1 (e.g.: pt; en))	0-Not present 1-Present
Testimonies of the users	0-Not present 1-Present
Accreditation (HON code, URAC)	0-Not present 1-Present
Last update date, annotated according to ISO 8601 (YYYY-MM) and with "0" if it did not exist)	Nominal
Indication of sources (references)	Continuous
Parallel interests (commercial intent, advertisements)	0-Not present 1-Present
Terminology (specific vocabulary)	1-Little understandable 2-Understandable 3-Completely understandable
Type of the content (audio, image, text, video)	0-Not present 1-Present
Electronic format of the document (e.g.: html, pdf)	Nominal
Number of pages of the document	Continuous
Documents from a publication of scientific character (e.g.: scientific papers)	0-Not present 1-Present
Type of medical information contained in the document (epidemiologic data, pathologic definition, diagnosis, indication of health professionals, place of treatment, prevention, prognosis, treatment)	0-Not present 1-Present
Links to other sites/internal pages of the URLs	0-Not present 1-Present
Documents – Web Documents	
Main type of the content (Article, informative, message, questionnaire, comment, academic work)	Nominal
Rank of the documents chosen by the users	0-Not present 1-Present
Documents – Responsible entity	
Author (contacts, name)	0-Not present 1-Present
Attainment of the author	0-No attainment was mentioned 1-Attainment in the health domain 2-Attainment in another area
Webmaster (contacts, name)	0-Not present 1-Present
Reputation (scientific nature, governmental nature)	0-Not present 1-Present
Documents – Website	
Mission (objective)	0-Not present 1-Present
Domain (e.g.: .com, .gov, .edu)	Nominal
Type (collaborative, personal institutional-scientific, institutional-not scientific, electronic commerce)	Nominal
Disclosure (copyrights, privacy policy)	0-Not present 1-Present
Editorial review (team of revision, process of revision)	0-Not present 1-Present
Tasks	
Correct answers in the tasks	0-No 1-Yes
Language of the query	Nominal
Medical terms in the query	0-No 1-Yes
Previous search of the user about the given tasks	0-No 1-Yes
The user had an exact idea about the information in the tasks	1 (Disagree) to 5 (Agree)
Level of clarity, easiness and familiarity of the tasks for the users	1 (Unclear/Easy/Unfamiliar) to 5 (Clear/Complex/Familiar)
Whether the user succeeded in the task (task completion status)	1 (Unsuccessful) to 5 (Successful)
Whether the users knew the technical terms	0-No 1-Yes
Comprehension of the documents by the users	0 – Did not understand 1 – Partially understood 2 – Understood
Users	
English proficiency of the users	Continuous
Health literacy of the users	Continuous
Number of medical concepts included in the query, that the user knows	Continuous
Age of the users	Continuous
Gender of the users	Nominal
Health status of the users	1 (Not healthy) to 5 (Very healthy)
Experience of the users with Web search and with health search	0-No 1-Yes
Frequency of the users' Web search and health search	1 – Once a year 2 – Once a month 3 – Once a week 4 – Once a day 5 – More often
Success of the users with Web search and health search	1 (Never) to 5 (Always)
Health search in Portuguese, English and other language	1 (Never) to 5 (Frequently)
Usage of medico-scientific terminology during Web searches about health subjects	1 (Never) to 5 (Always)
Level of satisfaction of the users' health information need on web pages, blogs, forums, social networks, chats, newsletter and RRS feeds	1 (Never) to 5 (Frequently)

B. *Statistical Analysis*

In Section IV we analyze how multiple variables from our data collection relate with relevance. We build a prediction model with the aim to foresee the relevance of a document based on its characteristics, as well as those for users and tasks. With this goal in mind, we first select the variables that build up a model that best fits our data. To do so, we use the least absolute shrinkage and selection operator (lasso), which selects the best subset of predictors by shrinking the regression coefficients towards zero, and estimates the coefficients [10]. It is based on logistic regression, which models the probability of documents' relevance given their characteristics, as well as those for users and tasks. We can write it as *Probability(relevance = yes|characteristics*), where the probability values *p(characteristics)* range between 0 and 1. Originally, our model had a multinomial distribution with three relevance levels (0, 1 and 2). Here we merge relevance levels 1 and 2, inducing a binomial distribution of the model.

After the lasso variable selection, we include the chosen characteristics in the multiple logistic regression model, and estimate its accuracy using leave-one-out cross-validation (LOOCV). The LOOCV error rate in our classification setting is estimated by averaging the misclassified observations. The LOOCV approach splits the set of observations into a single observation, used for the validation set, and the remaining observations which form the training set, where the prediction is made for the former observation.

IV. Multivariate Analysis of Situational Relevance

In this section we describe how we build the models, the models and their evaluation. We build a second model, called reduced model, that contains only the significant variables from the full model, and compare the LOOCV estimates of prediction (or test) errors for the two models. We decided to build the reduced model to analyze whether we could reach similar results using a lower number of features, what would ease the process of relevance estimation.

A. Full Model

Our first model considers all variables. We start our analysis by fitting a lasso model on the training set. Using cross-validation we then choose the “best” tuning parameter, and use it to fit the lasso model on the full dataset (*Model definition process*). With the variables selected by the lasso model, we fit a multiple logistic regression model (*Logistic regression model*), and evaluate the results (*Evaluation*).

1) Model definition process.

Applying the lasso to our dataset, and using the potential predictor variables discussed in Section III. *A*, we built a model predicting the relevance of web documents. The lasso, with the minimal tuning parameter chosen by cross-validation, yielded a prediction model containing candidate variables to be analyzed with the multiple logistic regression model.

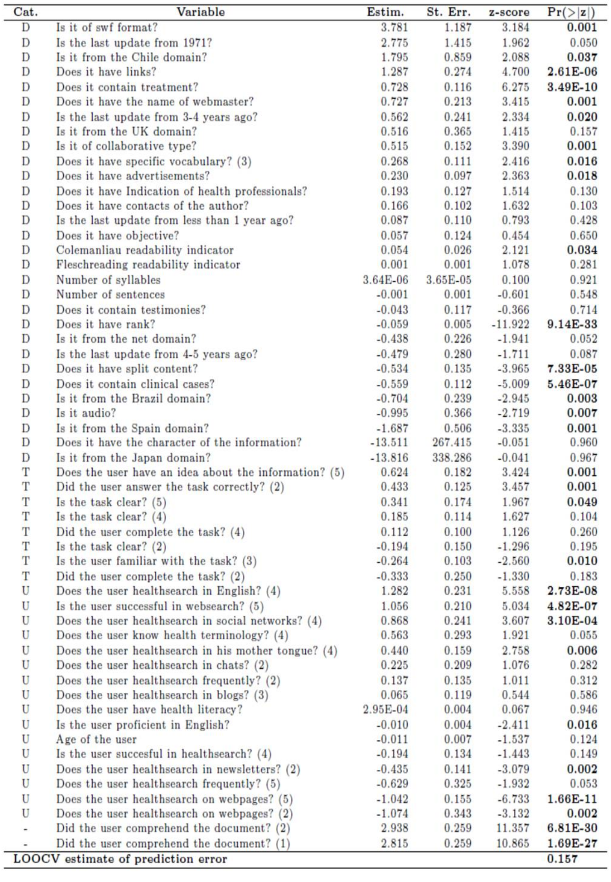
2) Logistic regression model.

The lasso helped in variable selection, and we continued the analysis with model selection using logistic regression. The resulting variables from the lasso model were added to the multiple logistic regression model which is summarized in Table II. The letters D, U and T in the first column identify the feature as pertaining to the document, user or task, respectively. In the second column we list the variables. The numbers in the parentheses indicate the levels of the variables (according to the scales defined in Table I.). The continuous variables, naturally, do not have such indications, nor the dichotomous (binary) variables. The latter are the ones scaled with 1 in Table I. The third column lists the variables' corresponding estimated coefficients. The fourth column contains the standard error when assessing the accuracy of the coefficient estimates. The fifth column contains the z-statistic, where a large (absolute) value indicates evidence against the null hypothesis of the coefficients being equal to zero. The last column lists the corresponding p-values.

3) Evaluation.

Our regression model was verified by leave-one-out cross-validation, and its results are reported in the last row of Table II. The p-values associated with the variables, marked with bold in Table II., are statistically significant at α = 0.05. The negative coefficients indicate that documents with the corresponding variables are less likely to be relevant than the documents without these characteristics, for fixed values of the remaining variables. Variables with large coefficient estimates highlight the importance of such variables (e.g. comprehension) for relevance. To assess the accuracy of the model, we have fitted the model using half of the data (training dataset), and then examined how well it predicts the held out data (test dataset) [10]. Using the test dataset we then computed the probabilities of the document being relevant, allowing us to compute the accuracy, sensitivity and specificity of the model. Given these predictions, we determined how many observations were correctly or incorrectly classified. Our logistic regression has an accuracy of 77.17%, a specificity (true negative rate) of 68.01% and sensitivity (true positive rate) of 78.98%. The LOOCV estimate of prediction error from Table II. is low (15.73%), meaning that the regression model is of high accuracy.

TABLE II. Summary of the Coeffcient Estimates in the Full Model.



Cat.	Variable	Estim.	St. Err.	z-score	Pr(>|z|)
D	Is it of swf format?	3.781	1.187	3.184	**0.001**
D	Is the last update from 1971?	2.775	1.415	1.962	0.050
D	Is it from the Chile domain?	1.795	0.859	2.088	**0.037**
D	Does it have links?	1.287	0.274	4.700	**2.61E-06**
D	Does it contain treatment?	0.728	0.116	6.275	**3.49E-10**
D	Does it have the name of webmaster?	0.727	0.213	3.415	**0.001**
D	Is the last update from 3-4 years ago?	0.562	0.241	2.334	**0.020**
D	Is it from the UK domain?	0.516	0.365	1.415	0.157
D	Is it of collaborative type?	0.515	0.152	3.390	**0.001**
D	Does it have specific vocabulary? (3)	0.268	0.111	2.416	**0.016**
D	Does it have advertisements?	0.230	0.097	2.363	**0.018**
D	Does it have Indication of health professionals?	0.193	0.127	1.514	0.130
D	Does it have contacts of the author?	0.166	0.102	1.632	0.103
D	Is the last update from less than 1 year ago?	0.087	0.110	0.793	0.428
D	Does it have objective?	0.057	0.124	0.454	0.650
D	Colemanliau readability indicator	0.054	0.026	2.121	**0.034**
D	Fleschreading readability indicator	0.001	0.001	1.078	0.281
D	Number of syllables	3.64E-06	3.65E-05	0.100	0.921
D	Number of sentences	-0.001	0.001	-0.601	0.548
D	Does it contain testimonies?	-0.043	0.117	-0.366	0.714
D	Does it have rank?	-0.059	0.005	-11.922	**9.14E-33**
D	Is it from the net domain?	-0.438	0.226	-1.941	0.052
D	Is the last update from 4-5 years ago?	-0.479	0.280	-1.711	0.087
D	Does it have split content?	-0.534	0.135	-3.965	**7.33E-05**
D	Does it contain clinical cases?	-0.559	0.112	-5.009	**5.46E-07**
D	Is it from the Brazil domain?	-0.704	0.239	-2.945	**0.003**
D	Is it audio?	-0.995	0.366	-2.719	**0.007**
D	Is it from the Spain domain?	-1.687	0.506	-3.335	**0.001**
D	Does it have the character of the information?	-13.511	267.415	-0.051	0.960
D	Is it from the Japan domain?	-13.816	338.286	-0.041	0.967
T	Does the user have an idea about the information? (5)	0.624	0.182	3.424	**0.001**
T	Did the user answer the task correctly? (2)	0.433	0.125	3.457	**0.001**
T	Is the task clear? (5)	0.341	0.174	1.967	**0.049**
T	Is the task clear? (4)	0.185	0.114	1.627	0.104
T	Did the user complete the task? (4)	0.112	0.100	1.126	0.260
T	Is the task clear? (2)	-0.194	0.150	-1.296	0.195
T	Is the user familiar with the task? (3)	-0.264	0.103	-2.560	**0.010**
T	Did the user complete the task? (2)	-0.333	0.250	-1.330	0.183
U	Does the user healthsearch in English? (4)	1.282	0.231	5.558	**2.73E-08**
U	Is the user successful in websearch? (5)	1.056	0.210	5.034	**4.82E-07**
U	Does the user healthsearch in social networks? (4)	0.868	0.241	3.607	**3.10E-04**
U	Does the user know health terminology? (4)	0.563	0.293	1.921	0.055
U	Does the user healthsearch in his mother tongue? (4)	0.440	0.159	2.758	**0.006**
U	Does the user healthsearch in chats? (2)	0.225	0.209	1.076	0.282
U	Does the user healthsearch frequently? (2)	0.137	0.135	1.011	0.312
U	Does the user healthsearch in blogs? (3)	0.065	0.119	0.544	0.586
U	Does the user have health literacy?	2.95E-04	0.004	0.067	0.946
U	Is the user proficient in English?	-0.010	0.004	-2.411	**0.016**
U	Age of the user	-0.011	0.007	-1.537	0.124
U	Is the user succesful in healthsearch? (4)	-0.194	0.134	-1.443	0.149
U	Does the user healthsearch in newsletters? (2)	-0.435	0.141	-3.079	**0.002**
U	Does the user healthsearch frequently? (5)	-0.629	0.325	-1.932	0.053
U	Does the user healthsearch on webpages? (5)	-1.042	0.155	-6.733	**1.66E-11**
U	Does the user healthsearch on webpages? (2)	-1.074	0.343	-3.132	**0.002**
-	Did the user comprehend the document? (2)	2.938	0.259	11.357	**6.81E-30**
-	Did the user comprehend the document? (1)	2.815	0.259	10.865	**1.69E-27**
LOOCV estimate of prediction error					**0.157**

B. Reduced Model

We built a second model, including only the statistically significant variables from the full model. In this second model, all variables remained significant except the one pertaining to the third level of task familiarity. Table III. shows the coefficient estimates for a logistic regression model that uses the selected 30 variables to predict the probability of a document being relevant or not relevant for the user. We assessed the model's accuracy using leave-one-out cross-validation, with an estimated prediction error of 0.1585. Our logistic regression has an accuracy of 77.53%, a specificity of 70.85% and sensitivity of 78.72%. As expected, the LOOCV estimate of prediction error for this model is slightly higher than the one for the full regression model in Table II.

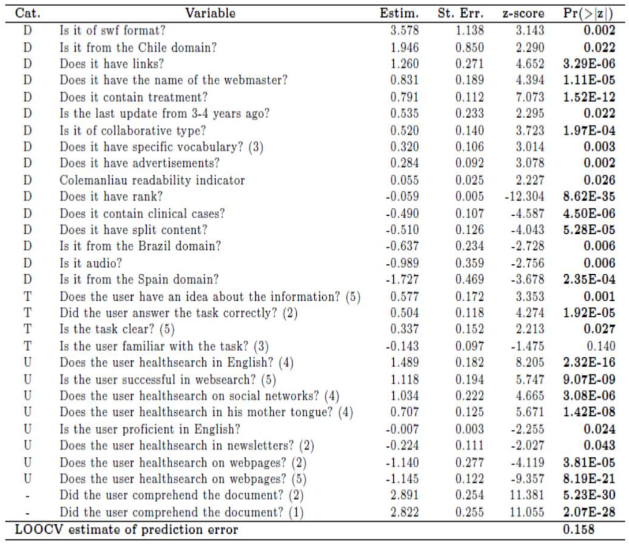
V. Discussion

As expected, the best model to predict documents' relevance is the one that contains all variables suggested by lasso. However, the reduced model was very close in terms of error rates and has the advantage of not requiring so much information. In Table IV. we summarize the evaluation metrics of the full and reduced logistic regression models. The first row contains the number of variables included in each model. In the second row we can see that the full model has the lowest prediction error estimate (LOOCV error).

The slightly higher values of sensitivity in the full model support this finding as well. However, its accuracy and specificity, indicated in the third and fourth row, are slightly lower than the one of the reduced model. This implies that the reduced model with higher accuracy and specificity is better at excluding the non-relevant documents, what may be preferable in a retrieval system. We note that the unbalanced data regarding the proportion of relevant documents in the dataset might affect accuracy and yield a very optimistic estimate, what is a common phenomenon in binary classification. Since some of the features were annotated with manual approaches, it might be more difficult to automatically predict them. On the other hand, features annotated with automatic approaches might be easier to predict automatically.

The analysis of these models allows us to identify important features to estimate relevance. Documents containing links to other sites were found to be useful to relevance prediction. On the other hand, the variables related to the rank of the document and to documents with content divided into several pages were associated with negative estimators, indicating that the relation is the other way around. The presence of information about a treatment, and medical terminology understandable by the user, also contribute to the document being relevant. However, the presence of information about clinical cases given by specialists was found to contribute negatively to relevance, as well as documents from the domain '.es' and '.br' (i.e., the Internet country code top-level domains for Spain and Brazil). Documents from collaborative websites, Chilean web domains ('.cl'), which contain the name of webmaster, and which were recently updated were, as well, shown to be useful features to predict relevance. Users seem to value the use of some media, e.g. flash documents in '.swf' format, but not content including audio files. In case of .swf format, .cl domain and search for health information in newsletters, the reason of such findings might be related to the number of documents.

TABLE III. Summary of the Coefficient Estimates in the Reduced Multiple Logistic Regression Model.



Cat.	Variable	Estim.	St. Err.	z-score	Pr(>|z|)
D	Is it of swf format?	3.578	1.138	3.143	**0.002**
D	Is it from the Chile domain?	1.946	0.850	2.290	**0.022**
D	Does it have links?	1.260	0.271	4.652	**3.29E-06**
D	Does it have the name of the webmaster?	0.831	0.189	4.394	**1.11E-05**
D	Does it contain treatment?	0.791	0.112	7.073	**1.52E-12**
D	Is the last update from 3-4 years ago?	0.535	0.233	2.295	**0.022**
D	Is it of collaborative type?	0.520	0.140	3.723	**1.97E-04**
D	Does it have specific vocabulary? (3)	0.320	0.106	3.014	**0.003**
D	Does it have advertisements?	0.284	0.092	3.078	**0.002**
D	Colemanliau readability indicator	0.055	0.025	2.227	**0.026**
D	Does it have rank?	-0.059	0.005	-12.304	**8.62E-35**
D	Does it contain clinical cases?	-0.490	0.107	-4.587	**4.50E-06**
D	Does it have split content?	-0.510	0.126	-4.043	**5.28E-05**
D	Is it from the Brazil domain?	-0.637	0.234	-2.728	**0.006**
D	Is it audio?	-0.989	0.359	-2.756	**0.006**
D	Is it from the Spain domain?	-1.727	0.469	-3.678	**2.35E-04**
T	Does the user have an idea about the information? (5)	0.577	0.172	3.353	**0.001**
T	Did the user answer the task correctly? (2)	0.504	0.118	4.274	**1.92E-05**
T	Is the task clear? (5)	0.337	0.152	2.213	**0.027**
T	Is the user familiar with the task? (3)	-0.143	0.097	-1.475	0.140
U	Does the user healthsearch in English? (4)	1.489	0.182	8.205	**2.32E-16**
U	Is the user successful in websearch? (5)	1.118	0.194	5.747	**9.07E-09**
U	Does the user healthsearch on social networks? (4)	1.034	0.222	4.665	**3.08E-06**
U	Does the user healthsearch in his mother tongue? (4)	0.707	0.125	5.671	**1.42E-08**
U	Is the user proficient in English?	-0.007	0.003	-2.255	**0.024**
U	Does the user healthsearch in newsletters? (2)	-0.224	0.111	-2.027	**0.043**
U	Does the user healthsearch on webpages? (2)	-1.140	0.277	-4.119	**3.81E-05**
U	Does the user healthsearch on webpages? (5)	-1.145	0.122	-9.357	**8.19E-21**
-	Did the user comprehend the document? (2)	2.891	0.254	11.381	**5.23E-30**
-	Did the user comprehend the document? (1)	2.822	0.255	11.055	**2.07E-28**
LOOCV estimate of prediction error					0.158

Besides the above document characteristics, we found that several user health search habits help in estimating the relevance of documents. Users who feel successful in web search, and who frequently conduct health search in English or Portuguese language (which are the languages of the queries in the dataset), were shown to assess documents as relevant more often. Users' proficiency in English language was shown to contribute negatively to relevance, as well as frequent health search on web pages and newsletters. The advanced comprehension level of the documents by the users was shown to highly influence the prediction of its relevance. The clarity of the tasks was also found to contribute positively to relevance, while the familiarity of users with the tasks showed negative contribution. More experienced users might be more demanding, what is inline with the findings in [13] and [20].

Regardless the high values of estimates, some of the features included in the model might be less useful, rather being just a reflection of the dataset (e.g.: there were only a few documents in SWF format (0.14%)). As well, in case of variables with multiple levels it is useful to consider only one level at once. For instance, we might prefer the second level to the first one for the variable Comprehension, because its estimate is higher or because we want to make predictions for

TABLE IV. Comparison of the Full and Reduced Logistic Regression Models in Terms of Number of Variables and Evaluation Rates.

	Logistic regression models	
	full	***reduced***
Nº of variables	56	30
LOOCV error	15.73%	15.85%
Accuracy	77.17%	77.53%
Specificity	68.01%	70.85%
Sensitivity	78.98%	78.72%

documents which are completely understood by the users.

VI. Conclusions

We conducted a multivariate analysis focused on whether the characteristics of tasks, users and documents are useful to predict document relevance, and how. For this purpose we built two regression models. Our best model had the following evaluation metrics: the LOOCV estimate of prediction error for the full model which considered all variables suggested by lasso (15.73%); sensitivity for the full model including all variables (78.98%). Accuracy was almost equal in the full and reduced models (77.17% vs. 77.53%); and specificity was slightly higher for the reduced model (68.01% vs. 70.85%). The model with higher accuracy and specificity is best at excluding the non-relevant documents, which may be preferable in some retrieval systems.

Among the features which were identified to predict relevance, we found several characteristics related to the user and tasks. These mainly relate to the users' health search habits, their comprehension and the tasks' clearness. Some features included in the models are easy, others are relatively or very difficult to assess automatically. They can be useful to improve the estimation of relevance by search engines, particularly of health documents on the Web. Therefore, in the future we will work on the development of methods to automatically detect these features. The application of these models to other datasets might be also interesting, allowing the generalization of our results. This might be important because there are features that are only present in a small number of documents what may be interfering with the model. Another future study might consider incorporating some of the features (e.g. considering users understandability of the documents) to improve the performance of search engines.

Acknowledgment

This work was supported by Project "NORTE-01-0145-FEDER-000016" (NanoSTIMA), financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] M. J. Bates, Understanding information retrieval systems: management, types, and standards. Auerbach Publications, eBook, 2011.

[2] A. Crystal and J. Greenberg, "Relevance criteria identified by health information users uuring web searches," JASIST 57(10), pp. 1368–382, 2006.

[3] R. Espanha and F. L. Villanueva, "Health and the internet: autonomy of the user. Technical report," Lisbon Internet and Networks, 2008.

[4] S. Fox, "Online health search. Technical report," Pew Internet & American Life Project, 2006.

[5] S. Fox, "The social life of health information. Technical report," Pew Internet & American Life Project, 2011.

[6] S. Fox and M. Duggan, "Health online 2013. Technical report," Pew Internet & American Life Project, 2013.

[7] S. Fox and L. Rainie, "Vital decisions: a Pew Internet health report. Technical report," Pew Internet & American Life Project, 2002.

[8] L. Goeuriot et al., "Building realistic potential patient queries for medical information retrieval evaluation" In: Proceedings of the LREC workshop on Building and evaluating resources for health and biomedical text processing, 2014.

[9] L. Goeuriot et al., "ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval," In CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Shefeld, UK, 2014.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to statistical learning with applications in R. Springer, eBook, 2013.

[11] D. Kelly, S. Dumais, and J. O. Pedersen, "Evaluation challenges and directions for information-seeking support systems," IEEE, pp. 44–50, 2009.

[12] J. Kim, "Describing and predicting information-seeking behavior on the web," JASIST 60(4), pp. 679–693, 2009.

[13] C. T. Lopes and C. Ribeiro, "Context effect on query formulation and subjective relevance in health searches," In: IIiX '10 Proceedings of the third symposium on Information interaction in context, pp. 205–214, New York, NY: ACM, 2010.

[14] C. T. Lopes and C. Ribeiro, "Measuring the value of health query translation: An analysis by user language proficiency," JASIST 64(5), pp. 951–963, 2013.

[15] Ch. D. Manning, P. Raghavan and H. Schütze, Introduction to information retrieval. Cambridge University Press, eBook, 2009.

[16] Ch. Marton, and Ch. W Choo, "A review of theoretical models of health information seeking on the web," Journal of Documentation, 68(3), pp. 330–352, 2011.

[17] M. Oroszlányová, C. T. Lopes, S. Nunes and C. Ribeiro, "The influence of documents, users and tasks on the relevance and comprehension of health web documents," Procedia Computer Science, Vol. 64, pp.771–778, 2015.

[18] M. Pallen, "Guide to the internet. The world wide web," BMJ 311, pp. 1552–6, 1995.

[19] J. Palotti et al., "CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms," In CLEF 2015 Online Working Notes. CEUR-WS, 2015.

[20] T. Saracevic, "Relevance reconsidered," In: Information science: Integration in perspectives, pp. 201–218. CoLIS 2, Copenhagen, 1996.

[21] R. Savolainen and J. Kari, "User-defined relevance criteria in web searching," JDOC 62(6), pp. 685–707, 2006.

[22] R. Savolainen, "Source preferences in the context of seeking problem-specific information," Information Processing and Management 44, pp. 274–293, 2008.

[23] M. I. Sousa, Characterization of health web documents. Master's thesis, Master in Information Science, University of Porto, 2011.

[24] A. Tombros, A. Ruthven and J. M. Jose, "How users assess web pages for information seeking," JASIST 56(4), pp. 327–344, 2005.

[25] S. Vargas, P. Castells and D. Vallet, "Explicit relevance models in intent-oriented information retrieval diversification," SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, NY, pp. 75–84, 2012.

[26] World Telecommunication/ICT Indicators database 2016 (20th Edition/June 2016), http://www.itu.int/en/ITU-D/statistics

[27] Y. Zhang, J. Zhang, M. Lease and J. Jacek Gwizdka, "Multidimensional relevance modeling via psychometrics and crowdsourcing," SIGIR '14 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, NY, pp. 435–444, 2014.

[28] G. Zuccon et al., "The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred health information retrieval," In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, 2016.