Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 29 – 36

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Predicting the comprehension of health web documents using characteristics of documents and users

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

[a]*DEI, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal*
[b]*INESC TEC, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal*

Abstract

The Web is frequently used as a way to access health information. In the health domain, the terminology can be very specific, frequently assuming a medico-scientific character. This can be a barrier to users who may be unable to understand the retrieved documents. Therefore, it would be useful to automatically assess how well a certain document will be understood by a certain user. In the present work, we analyse whether it is possible to predict the comprehension of documents using document features together with user features, and how well this can be achieved. We use an existing dataset, composed by health documents on the Web and their assessment in terms of comprehension by users, to build two multivariate prediction models for comprehension. Our best model showed very good results, with 96.51% accuracy. Our findings suggest features that can be considered by search engines to estimate comprehension. We found that user characteristics related to web and health search habits, such as the success of the users with Web search and the frequency of the users' health search, are some of the most influential user variables. The promising results obtained with this dataset with manual comprehension assessment will lead us to explore the automatic assessment of document and user characteristics.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Consumer Health Information; Information Retrieval; World Wide Web; Comprehension

* Corresponding author. Tel.: +351 22 508 14 00.
E-mail address: melinda@fe.up.pt

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.120

1. Introduction

The number of Internet users and the amount of information on the Web, including consumer-oriented health information, increased rapidly in the past decades. It has been observed that people prefer Internet as a source when searching for health-related information[18,14]. A recent survey in the USA reported that, in 2012, 72% of all adults looked online for health information[10].

Among health consumers, the Web is frequently used as a way to access health information. In the health domain, the terminology can be very specific, frequently assuming a medico-scientific character. This can be a barrier to users who may be unable to understand the retrieved documents. Therefore, it would be very useful to automatically assess how well a certain document will be understood by a certain user.

Previously, we showed that user, task and document characteristics are good descriptors and possible predictors of comprehension[3]. The goal of our current work is to investigate whether it is possible to predict comprehension using document and user characteristics. In the present study we use an existing dataset composed by a set of annotated web pages, users' characteristics and comprehension judgements[2]. Using this collection, we built a multivariate prediction model for the comprehension of the document. In the next two sections we review the related literature, and describe the dataset and statistical analysis used in the study. The following sections present the logistic regression models and the multivariate analyses. Finally, we discuss and summarize our main findings.

2. Literature Review

Since the introduction of a guide to the Internet by Pallen in the 90's[17], information on medical and health topics started to rise on the Web. As the availability of health information increased, users' health search on the Web started to have an impact on their health care routines[11,8,9,7].

The role and representation of comprehension have been studied in detail by Kintsch[15]. He defines "the mental representation of the text and actions based on this construction" as the product of the comprehension process. He names several elementary units which enter into this process (e.g.: perceptions, ideas, concepts, images, emotions) and studies the relations among them. According to Kintsch, a "comprehender", who is, in our context, a user, has specific goals, a given perceptual situation (e.g.: the words on a page of text), and a background of knowledge and experience. All these affect the information processing at the biological level. A more coherent combination of the above factors might imply that the reader can achieve a higher comprehension level in the context of online searching[22].

Search engines are mostly restricted to topical relevance. In health information retrieval, expert and average users will satisfy their information needs with very different texts[19] and comprehension can become a typical problem to average users[19], affecting decision making[4,16,12] and diminish the value of the document[21]. Several authors have shown that many users have difficulties understanding online resources for patients[6,20].

Studies about users' comprehension of health-related information available on the Web conclude that a high reading level is necessary to increase the comprehension of web-based health information[4,20]. Other authors concluded that users do not uniformly prefer simple texts, and that the text comprehensibility level should match the user's level of preparedness[5,19]. Collins-Thompson et al. showed that it works when addressed through user's reading difficulty[5]. Often, assessing the readability of medical documents is the first step to ensure that they are readable and are thus comprehensible when shared with patients and families[21]. Readability has also been shown to be useful in the evaluation of information retrieval systems for consumer health search and to contribute to system effectiveness compared to considering topicality alone[23].

3. Dataset

Our study is based on an existing dataset composed by an annotated sample of health web documents. These documents were initially collected for a user study[1] and were later automatically and manually annotated[2]. The dataset contains the Google top-30 ranked documents for 8 information situations using 4 different queries for each, with different language and medical terminology (lay or medico-scientific). The documents were assessed by a researcher, who also defined the metadata scheme. Part of the documents (10%) was also assessed by an external

health professional[1]. The agreement rate between both assessments was very good (93%). Information about the users has been collected through questionnaires and other instruments. The users' understanding of the documents is described by the *Comprehension* variable, which has 3 assessment levels (0 - did not understand, 1 - partially understood, 2 - understood). Table 1 describes specific characteristics of web documents and users. The document characteristics were categorized as related to the content and to the website.

Table 1. Document characteristics, related to its content and website, and user characteristics used for predicting comprehension.

Content Characteristics	**Assessment Levels (in ascending order)**	**Description**
Rel	1-2	Judgements of the users about the usefulness (relevance) of the content
Specific.vocabulary	0-1	Existence of expressions concerning technical and scientific terms of health
Scientific.publication	0-1	Documents originated in scientific publications
Website Characteristics		
Domain		Name used to locate and identify sets of computers on the Internet, e.g.: generic (.com, .net) or regional (.uk, .br)
User Characteristics		
English.proficiency		English proficiency of the users
Num.medicalconcepts		Number of medical concepts included in the query, that the user knows
Age		Age of the users
Websearch.suc	1-5	Success of the users with Web search
Healthsearch.freq	1-5	Frequency of the users' health search
Healthsearch.suc	1-5	Success of the users with health search
Healthsearch.terminology	1-5	Usage of medico-scientific terminology during web searches about health subjects
Healthsearch.webpages	1-5	Level of satisfaction of the users' health information need on web pages
Healthsearch.forums	1-5	Level of satisfaction of the users' health information need on forums
Healthsearch.socialnet	1-5	Level of satisfaction of the users' health information need on social networks
Healthsearch.rss	1-5	Level of satisfaction of the users' health information need on RSS feeds

4. Statistical Analysis

In *Section 5*, we build two models to predict comprehension that include variables describing the user and the document. We select the variables which build up a model that best fits our data, using the least absolute shrinkage and selection operator (lasso). Lasso picks the best subset of predictors by shrinking the regression coefficients towards zero. It also estimates the coefficients based on logistic regression[13]. Although our models had, initially, a multinomial distribution with $Y \in \{0,1,2\}$, we merged comprehension levels 1 and 2, inducing a binomial distribution of the model with $Y \in \{0,1\}$. These models have multiple predictors, $X = (X_1, \ldots, X_p)$. We use the following logistic function:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \,, \tag{1}$$

that can be written, by definition, as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \,, \tag{2}$$

with the logit, or log-odds, transformation. The lasso regression coefficient estimates are the values minimizing the quantity composed by the sum of squared residuals (RSS) and the shrinkage penalty as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \,. \tag{3}$$

It is crucial to choose an appropriate value of the tuning parameter $\lambda \geq 0$, since it controls the relative impact of the above two terms on the regression coefficient estimates. In the shrinkage penalty, the ℓ_1 norm of a coefficient vector β is given by $||\beta||_1 = \sum |\beta_j|$. Lasso performs variable selection based on this ℓ_1 penalty, forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large. We use 10-fold cross-validation for choosing the best λ, which is feasible, as it fits the learning procedure only ten times.

After the lasso variable selection, we add the proposed features to the multiple logistic regression model. Using leave-one-out cross-validation (LOOCV), we estimate the accuracy of the model. The LOOCV error rate is estimated by averaging the *n* misclassified observations as follows

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \ . \quad (4)$$

The LOOCV approach splits the set of observations into a single observation used for the validation set, and the remaining observations used as the training set, where the prediction $\hat{y}_i$ is made for the former observation. We also build a second model, which we call reduced model, containing only the variables that were significant in the full model. Finally, we compare the LOOCV estimates of prediction (or test) errors for the two models.

5. Multivariate Analysis

We want to predict whether a document's content is comprehensible to the user or not, based on document and user characteristics, following the statistical strategy defined in the previous section.

5.1. Full model

To observe the evolution of the model coefficients, we start our analysis by fitting a lasso model on the training set for an automatically selected range of λ values, visualizing the path by plotting the coefficients against the deviance explained (*Section 5.1.1*). Next, we choose the "best" tuning parameter λ using cross-validation (*Section 5.1.2*), and use it to fit the lasso model on the full dataset (*Section 5.1.3*). Lastly, we fit a multiple logistic regression model with the variables selected by the lasso model (*Section 5.1.4*), and evaluate the results (*Section 5.1.5*).

5.1.1. Coefficient plot

Applying lasso to our data, we first generated a coefficient plot (Fig. 1.), which displays the path of the variables' coefficients against the deviance explained. Each curve on the figure corresponds to a variable. We can observe that some of the coefficients will be equal to zero, due to the forcing effect of the ℓ_1 penalty when the tuning parameter λ is sufficiently large. The numbers attached to the curves are simply the numeric codes of the variables, given by lasso. For example, the black curve annotated with label 149 refers to partially relevant documents ('Rel.1'), and the red one with label 137 refers to the variable representing an intermediate use of social networks to access health information ('Healthsearch_socialnet.3' in a 5-value scale). The vertical axis indicates the number of non-zero coefficients at the current percent deviance explained, which is expected to change sufficiently from one lambda to the next. The predictor variables enter the model as we move from left to right in the figure. The lasso can produce a model involving any number of variables depending on the value of λ.

5.1.2. Choosing the tuning parameter λ

Applying 10-fold cross-validation, based on the misclassification error, visualized in Fig. 2., we chose a value for the tuning parameter. The red dotted line indicates the cross-validation-curve, and the error bars represent the standard deviation along the λ sequence. The two vertical dotted lines indicate the value of λ (0.004) that gives minimum mean cross-validated error, and the value of λ (0.006) that gives the most regularized model, i.e., the simplest model with comparable error to the best model, given the uncertainty in the cross-validation error estimate of the latter.

5.1.3. Lasso model

Using the above information, and the list of 149 potential predictor variables, we built a model predicting the comprehension of web documents. The lasso, with the minimal λ chosen by cross-validation, yielded a prediction model containing 16 variables. The resulting coefficient estimates are summarized in Table 2. The first column contains the name of the variables and the second column the corresponding coefficients estimated by the lasso. The description of these variables is in Table 1 *(Section 3)*.

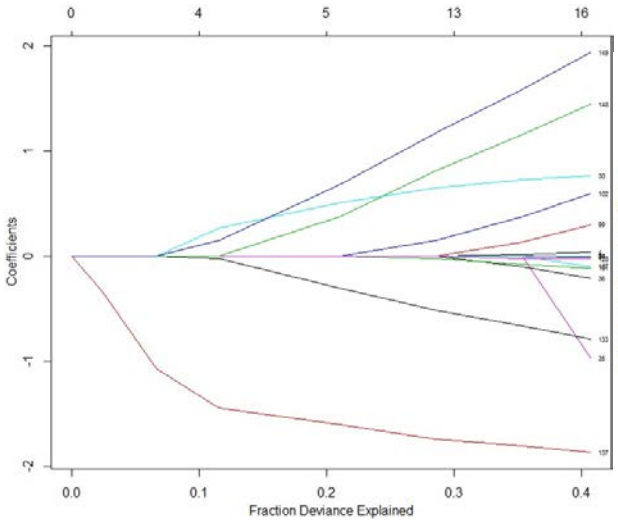

Fig. 1. Path of the variables' coefficients from the lasso model against the deviance explained.

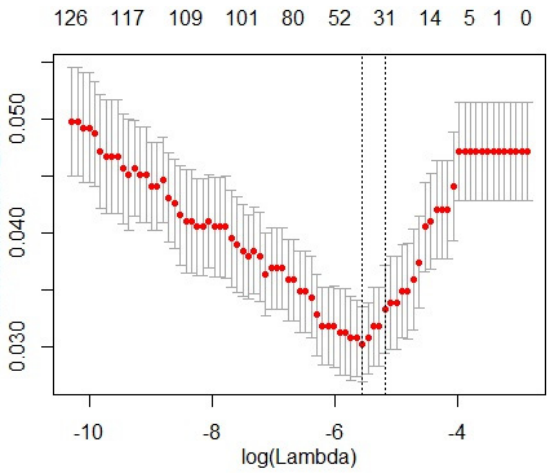

Fig. 2. Visualization of the cross-validation curve (red dotted line) with standard deviation along the λ sequence (error bars), where the vertical dotted lines indicate the optimal values of λ.

We observed that the variables regarding the partially ('Rel.1') and completely relevant documents ('Rel.2') have the largest coefficient estimates.

Table 2. Variables contained in the lasso model.

Variable	Coefficient
Rel.2	1.943
Rel.1	1.363
Specific.vocabulary.3	0.475
Healthsearch.freq.3	0.279
Healthsearchrss.3	0.159
Websearch.suc.4	0.093
Age	0.033
Num.medicalconcepts	0.016
English.proficiency	0.003
Healthsearch.freq.2	-0.005
Domain.br	-0.128
Healthsearch.webpages.5	-0.253
Healthsearch.suc.5	-0.286
Scientific.publication	-0.354
Healthsearch.forums.3	-0.574
Healthsearch.socialnet.4	-1.586

This highlights the relation between relevance and comprehension, confirming a previous finding of a significant positive association between the comprehension and relevance of the information in the documents[3]. The negative coefficients indicate that documents with the corresponding dichotomous variables are less likely to be comprehended than the documents without these characteristics, for fixed values of the remaining variables.

5.1.4. Logistic regression model

The lasso helped in variable selection, and we continue the analysis with model selection using logistic regression. The variables from Table 2 were added to the multiple logistic regression model which is summarized in Table 3.

Table 3. Summary of the coefficient estimates in the full logistic regression model.

Cat.	Variable	Estimate	Std. Error	z-value	Pr(>|z|)
D	Rel.2	4.766	0.606	7.862	**3.79E-15**
D	Rel.1	2.869	0.280	10.243	**1.27E-24**
D	Specific.vocabulary.3	0.880	0.214	4.114	**3.88E-05**
D	Scientific.publication	-0.765	0.244	-3.140	**1.69E-03**
D	Domain.br	-0.921	0.368	-2.502	**0.012**
U	Healthsearch.freq.3	1.414	0.374	3.781	**1.56E-04**
U	Websearch.suc.4	0.722	0.246	2.934	**3.34E-03**
U	Healthsearch.rss.3	0.623	0.359	1.733	0.083
U	Healthsearch.webpages.5	0.137	0.485	0.283	0.777
U	Num.medicalconcepts	0.122	0.077	1.591	0.112
U	Age	0.104	0.018	5.781	**7.41E-09**
U	Healthsearch.freq.2	0.097	0.285	0.340	0.734
U	English.proficiency	0.015	0.007	2.253	**0.024**
U	Healthsearch.forums.3	-1.422	0.289	-4.929	**8.25E-07**
U	Healthsearch.suc.5	-1.686	0.612	-2.757	**5.83E-03**
U	Healthsearch.socialnet.4	-1.785	0.662	-2.696	**7.02E-03**
LOOCV estimate of prediction error					**0.0271**

The first column contains the category of the feature, where D and U refer to document and user, respectively. The second column contains the names of the variables and the third column their corresponding estimated coefficients. The fourth column lists 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. The ones in boldface relate to the variables that are statistically significant at $\alpha = 0.05$. Our regression model was further verified by leave-one-out cross-validation (LOOCV), and its results are reported in the last row of Table 3.

5.1.5. Evaluation

We fitted a logistic regression model in order to predict comprehension, using the above 16 variables. To better assess the accuracy of this model, the model was fitted using half of the data (training dataset) and we examined how well it predicts the held out data (test data)[13]. We calculated the probabilities of the document being understood by the users, using the test dataset. Given these predictions, we determined how many observations were correctly or incorrectly classified. Our logistic regression has an accuracy of 96.56%, a specificity (true negative rate) of 75.00% and sensitivity (true positive rate) of 97.01%. The LOOCV estimate of prediction error shown in Table 3 is very low (2.71%), meaning that the regression model is of high accuracy.

5.2. Reduced model

We built a second model, only containing the statistically significant variables from the full model, to see how well it is possible to predict comprehension using less variables. In this second model, all variables remained significant at $\alpha = 0.05$. Table 4 shows the coefficient estimates for the reduced logistic regression model.

Table 4. Summary of the coefficient estimates in the reduced multiple logistic regression model.

Cat.	Variable	Estimate	Std. Error	z value	Pr(>|z|)
D	Rel.2	4.770	0.605	7.884	**3.18E-15**
D	Rel.1	2.890	0.277	10.421	**1.98E-25**
D	Specific.vocabulary.3	0.887	0.212	4.184	**2.87E-05**
D	Scientific.publication	-0.778	0.242	-3.216	**1.30E-03**
D	Domain.br	-0.896	0.363	-2.467	**0.014**
U	Healthsearch.freq.3	1.517	0.322	4.710	**2.48E-06**
U	Websearch.suc.4	0.915	0.228	4.020	**5.81E-05**

Cat.	Variable	Estimate	Std. Error	z value	Pr(>|z|)
U	Age	0.102	0.017	6.134	**8.56E-10**
U	English.proficiency	0.015	0.006	2.363	**0.018**
U	Healthsearch.forums.3	-1.128	0.241	-4.677	**2.91E-06**
U	Healthsearch.suc.5	-1.747	0.335	-5.215	**1.84E-07**
U	Healthsearch.socialnet.4	-2.021	0.378	-5.348	**8.92E-08**
LOOCV estimate of prediction error					**0.0270**

We assessed the model's accuracy using leave-one-out cross-validation, with estimated prediction error of 0.0270. Using the same training and test set we found that our logistic regression has an accuracy of 96.51%, a specificity of 73.17% and sensitivity of 97.01%. As expected, the LOOCV estimate of prediction error for this model is slightly lower than the one for the full regression model in Table 3.

6. Discussion

As expected, the best model to predict documents' comprehension is the one that contains all variables suggested by lasso. Even so, the reduced model was very close in terms of error rates and has the advantage of not requiring so much information. The variables pertaining the average use of RSS ('Healthsearch.rss.3'), frequent health searches on the Web ('Healthsearch.webpages.5'), the number of medical concepts known before the search task ('Num. medicalconcepts'), and less frequent health searches ('Healthsearch.freq.2'), were excluded from the reduced model. All these variables describe characteristics of the users.

In Table 5 we summarize the evaluation metrics of the two logistic regression models. The first row contains the number of variables included in the models. In the second row we can observe that the reduced model has the lowest prediction error estimate (LOOCV error). However, these values are almost equal in the two models, as well as the value of sensitivity. The slightly higher value of accuracy and specificity in the full model implies that the full model is best at excluding the non-comprehensible documents, which may be preferable in some retrieval systems.

Table 5. Comparison of the full and reduced logistic regression models in terms of number of variables and evaluation rates.

	Full model	Reduced model
Number of variables	16	12
LOOCV error	2.71%	**2.70%**
Accuracy	**96.56%**	96.51%
Specificity	**75.00%**	73.17%
Sensitivity	**97.01%**	**97.01%**

The analysis of these models allows us to identify important features to estimate comprehension. Documents in which the medico-scientific terms are easily understandable ('Specific.vocabulary.3' in a 3-value scale), were found to be useful to predict comprehension. Although we found relevance as a good predictor of comprehension, we don't expect it to be useful since comprehension is probably a cause for relevance and not the other way around. On the other hand, scientific publications and Brazilian documents (from the domain '.br') were associated with negative estimators, indicating that the relation is opposite. All the above characteristics have the property that they can be easily detected in documents.

We found that users' age and proficiency in English are useful to comprehension prediction. The presence of English proficiency in the model indicates that this proficiency gives users the ability to understand better the English documents. Similarly, users who feel successful in web search, and who frequently conduct health searches, were shown to comprehend documents better. This may be a consequence of their positive attitude towards search. However, users who feel highly successful on their health searches on the Web, and those who moderately use forums and social networks for health search, comprehend documents worse since these characteristics contribute negatively to comprehension.

These results are aligned with previous findings, where we reported significant positive association between comprehension and: the success of users with web search; the frequency of web search; user proficiency in English language; relevance; the use of specific medical vocabulary[3].

4. Conclusions

We conducted a multivariate analysis focused on whether the characteristics of tasks, users and documents are useful to predict document comprehension, and how. For this purpose, we built two regression models. Our best model, the full model, had a LOOCV estimate of prediction error of 2.71%, a sensitivity of 97.01%, an accuracy of 96.56%, and a specificity of 75.00%.

Among the features which were identified to predict comprehension, we found that user characteristics mainly relate to the users' web and health search habits. 'Websearch_suc.4' and 'Healthsearch_freq.3' (a success of the users with web search at level "high" and a frequency of the users' health search at level "medium") are some of the most influencing user variables. This might show that users with more computer literacy might comprehend web documents more easily. Document features possess the property of being easy to detect in documents.

In future work, we aim to use these findings as a way to improve information retrieval in the health domain. Since some of the features included in these models can be automatically assessed, they can be used to improve the estimation of comprehension by search engines, particularly on health documents on the Web.

Acknowledgements

This work was partially supported by Project "NORTE-01-0145-FEDER-000016", 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. Lopes, C. T., Ribeiro, C., 2010. Context effect on query formulation and subjective relevance in health searches. *IIiX '10 Proceedings of the third symposium on Information interaction in context,* pp. 205-214.
2. Sousa, M. I., 2011. Characterization of health web documents, Master in Information Science, *University of Porto.*
3. Oroszlányová, M., Lopes, C. T., Nunes, S., Ribeiro, C., 2015. The Influence of Documents, Users and Tasks on the Relevance and Comprehension of Health Web Documents. *In Procedia Computer Science*, pp. 771-778.
4. Berland, G. K. et al., 2001. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. *JAMA,* 285(20), pp. 2612-2621.
5. Collins-Thompson, K. et al., 2011. Personalizing web search results by reading level. *In Proc. of CIKM.*
6. Edmunds Matthew, R. et al., 2013. Patient Information in Graves' Disease and Thyroid-Associated Ophthalmopathy: Readability Assessment of Online Resources. *Thyroid,* 24(1), pp. 67-72.
7. Espanha, R. & Villanueva, F. L., 2008. Health and the Internet: Autonomy of the User. *Lisbon Internet and Networks.*
8. Fox, S., 2006. Online health search. *Pew Internet & American Life Project.*
9. Fox, S., 2011. The Social Life of Health Information. *Pew Internet & American Life Project.*
10. Fox, S. D. M., 2013. Health Online 2013. Technical report. *Pew Internet & American Life Project.*
11. Fox, S. R. L., 2002. Vital decisions: A Pew Internet Health Report. *Pew Internet & American Life Project.*
12. Hersh, W., 2008. *Information Retrieval: A Health and Biomedical Perspective (Health Informatics).* 3rd ed. Springer.
13. James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R.* Springer Texts in Statistics.
14. Kim, J., 2009. Describing and Predicting Information-Seeking Behavior on the Web. *JASIST,* 60(4), pp. 679-693.
15. Kintsch, W., 1998. *Comprehension: A Paradigm for Cognition.* Springer.
16. Lalmas, M. et al., 2006. *Advances in Information Retrieval.* Springer.
17. Pallen, M., 1995. Guide to the Internet. The world wide web.. *BMJ,* Volume 311, pp. 1552-6.
18. Savolainen, R., 2008. Source preferences in the context of seeking problem-specific information. *Information Processing and Management,* Volume 44, pp. 274-293.
19. Tan, C., Gabrilovich, E. & Pang, B., 2012. To each his own: personalized content selection based on text comprehensibility. *In Proc. of WSDM.* Seattle, Washington, USA, *WSDM.*
20. Vargas, C. R., Chuang, D. J., Ganor, O. & Lee, B. T., 2014. Readability of online patient resources for the operative treatment of breast cancer. *Surgery,* 156(2).
21. Wu, D. et al., 2013. Applying multiple methods to assess the readability of a large corpus of medical documents. *In Proc. of MEDINFO.* Copenhagen, Denmark, *MEDINFO.*
22. Yan, X., Song, D. & Li, X., 2006. Concept-based document readability in domain specific information retrieval. *In Proc. of CIKM.* Arlington, Virginia, USA, *CIKM.*
23. Zuccon, G. & Koopman, B., 2014. Integrating Understandability in the Evaluation of Consumer Health Search Engines. Gold Coast, Australia, *MedIR.*