Effects of Language and Terminology of Query Suggestions on the Precision of Health Searches

Carla Teixeira Lopes(✉) and Cristina Ribeiro

DEI, Faculdade de Engenharia, Universidade do Porto and INESC TEC,
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{ctl,mcr}@fe.up.pt

Abstract. Health information is highly sought on the Web by users that naturally have different levels of expertise in the topics they search for. Assisting users with query formulation is important when users are searching for topics about which they have little knowledge or familiarity. To assist users with health query formulation, we developed a query suggestion system that provides alternative queries combining Portuguese and English language with lay and medico-scientific terminology. Here, we analyze how this system affects the precision of search sessions. Results show that a system providing these suggestions tends to perform better than a system without them. On specific groups of users, clicking on suggestions has positive effects on precision while using them as sources of new terms has the opposite effect. This suggests that a personalized suggestion system might have a good impact on precision.

Keywords: Query suggestion · Health · Language · Terminology
English proficiency · Health literacy · Topic familiarity

1 Introduction

Health information is an online popular pursuit being sought by 80% of U.S. Internet users [2]. In this type of searches, users frequently have difficulties finding the correct terms to include in their queries [7,21], lacking the knowledge of the proper medical terms [19,22] or misspelling them [6,13]. For these reasons, support in query formulation may contribute to an improved retrieval experience.

Previous findings [8,9] led us to develop a system that, based on an initial user query, suggests 4 different queries combining two languages (English and Portuguese) and two bodies of terminology (lay and medico-scientific). To the best of our knowledge, no previous works have explored cross-language query suggestions in the health domain.

The usage given to the suggestions provided by this system was studied before [10] as well as their effect on the medical accuracy of the knowledge acquired during the search session, considering different user characteristics [18].

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 101–111, 2018.
https://doi.org/10.1007/978-3-319-98932-7_9

In this work we assess the impact of presenting and using these suggestions on the precision of the search session. As search assistance should be personalized to achieve its maximal outcome [4], we have considered users' English proficiency, health literacy and topic familiarity in this analysis.

2 Related Work

Query formulation is one of the most important aspects of information seeking. Query suggestion provides alternative ways to help users formulate queries and explore less familiar topics [5]. This technique is particularly important in topics about which users have little knowledge or familiarity. In these situations, users lack of vocabulary and knowledge may hinder query formulation.

In the health domain, the terminology gap between medical experts and lay people often causes additional difficulties in searches conducted by consumers [23]. To mitigate some of these difficulties, different query modification approaches have been proposed. Most of the approaches use specialized vocabularies from the Unified Medical Language System (UMLS). This is the case of the assistant proposed by Zeng et al. [21] that compute the semantic distance between the query and suggested terms using co-occurrences in medical literature and log data as well as UMLS semantic relations. iMed [11] and MedSearch [12] are two health search engines that suggest related medical phrases to help the user refine the query. In these systems, the phrases are extracted and ranked based on MeSH (Medical Subject Headings), the collection of crawled webpages, and the query. Similarly to our two-terminology query suggestion system, Zarro and Lin [20] also use MeSH along with social tagging to provide users with medico-scientific and lay terms.

All these works assess their systems through user studies, although focusing on different issues. In three of them [11,12,21], users were randomized into 2 groups, one receiving suggestions and the other not receiving them. In the study conducted by Zarro and Lin [20], 10 subjects were lay and the other 10 were expert. All subjects used the same system. The evaluation that is most resembled with a precision assessment is conducted by Luo et al. [12]. Authors combine relevance and diversity in a metric they call usefulness and each document is either useful or not. This metric is then used to compute the NDCG metric. Zeng et al. [21] assess the rates of successful queries, i.e., queries with at least one relevant result among the top 10. In the assessment of iMed [11], a search session is considered successful if the user can list one of the correct diseases associated with the medical case's situation. Note that this evaluation does not consider the relevance of each document. Zarro and Lin [20] focused on the differences between lay and expert subjects. Zarro and Lin [20] found that both user groups preferred MeSH terms because their quality was considered superior to the quality of social tags. All the assistance approaches described here were considered successful.

Regarding multilingual query suggestion, Gao et al. [3] proposed a system providing suggestions in a language different from the original query's language.

After using query logs to translate queries, authors used word translation relations and word co-occurrence statistics to estimate the cross-lingual query similarity. They used French-English and Chinese-English tasks for the evaluation and found that these suggestions, when used in combination with pseudo-relevance feedback, improved the effectiveness of cross-language information retrieval.

Since 2014, the Conference and Labs of the Evaluation Forum (CLEF) eHealth lab began to propose a multilingual information user-centred health retrieval task, incorporating queries in several languages in its dataset. The low number of teams proposing multilingual approaches makes us conclude that this type of approaches could be more explored.

3 Suggestion Tool

We developed a search suggestion system that, given a health query, suggests queries in two languages, Portuguese (PT) and English (EN), using medico-scientific (MS) and lay terminology.

In Fig. 1, we present the architecture of the suggestion tool, which will be briefly described in the following paragraphs. More details on the suggestion system can be found in a previous publication [18].

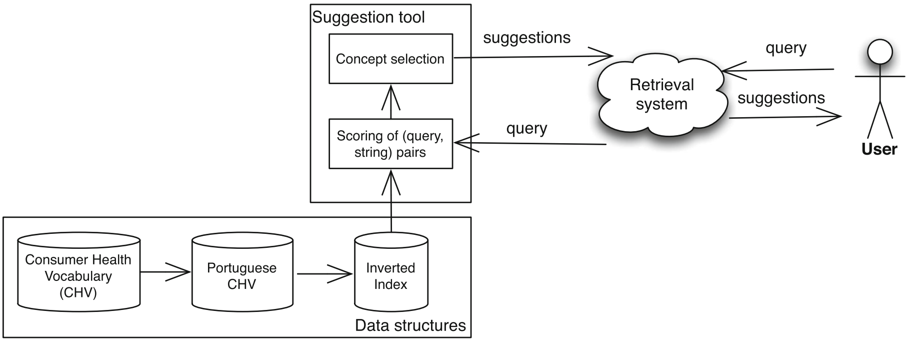

Fig. 1. Architecture of the suggestion tool [18].

Our system uses the Consumer Health Vocabulary (CHV) [14], a vocabulary that connects informal expressions about health to technical terms used by health care professionals. Each expression is associated with an Unified Medical Language System (UMLS) concept which, in turn, may be associated with several expressions or strings. Each string is associated with a CHV and an UMLS preferred names. Given that queries will probably be formulated in Portuguese, we use a Portuguese translation of the CHV.

After stemming the terms included in the CHV, we created an inverted index in which we associate each term with an inverse string frequency (isf_t) and a postings list, i.e., a list of the strings in which the stemmed term appears. The inverse string frequency is computed as $isf_t = log(N/sf_t)$, where sf_t is the number of strings in which the term appears and N is the total number of strings.

As the probability of finding multiple occurrences of the same term in a string is very small, we decided to ignore the term frequency in each string ($tf_{t,s}$). Each (query, string) pair is assigned the following $score(q, s) = \sum_{t \in (q \cap s)} isf_t$. Because the length of strings and queries has a very small variance, we decided to not normalize the above formula.

To limit the number of suggestions, for each query, we only select the string with the maximum score. For this string, we identify the associated concept and return its CHV and UMLS preferred names in English and Portuguese. This results in a maximum of 4 suggestions for each query. As an example, a set of suggestions could be: "colectomia", "remoção do cólon", "colectomy" and "colon removal".

Our retrieval system used the Bing Search API to obtain web results for users' queries. To increase the usability of the interface with regard to learning, we decided to keep the interfaces very simple and similar to those used in the most popular search engines. All the suggestions are presented in a single line.

4 Experiment

We conducted a user study with 40 participants (24 female; 16 male), with a mean age of 23.48 years (standard deviation = 7.66). English proficiency was evaluated using an instrument developed by the European Council that grades English proficiency in the Common European Framework of Reference for Languages, a widely accepted European standard for this purpose. To evaluate the users' health literacy, we have used the Medical Term Recognition Test, a brief and self-administered instrument proposed by Rawson et al. [15]. Users' familiarity with each topic was self-assessed on a five-level scale. The sample of users is heterogeneous in these characteristics.

Each user was assigned a set of 8 tasks, half of them conducted in a system presenting the suggestions (SYS+) and the other half on a system without suggestions (SYS). Each task was associated with a simulated work task situation [1]. Situations were rotated and counter-balanced across subjects and systems. Before the user study, to define these situations, we asked 20 persons with no medical expertise from 30 to 68 years old and a wide range of education levels (from high school to PhD) to state the health topic for which they had most recently searched on the Web. From these, we randomly selected 8 and created a scenario for each. Note that these persons were not participants of the study. The information situations were described to the users in Portuguese.

In each task the user had to formulate 3 queries in a language of their choice and assess the relevance of the top 10 results for each query, considering his own context in a 3-value scale. In the first query, users formulated the query without help from the system. Users had no restrictions in query formulation, being able to use their preferred language and terminology. Based on the initial query, the system presents suggestions that can, or not, be used for the formulation of the second query. The same happens when the user is moving from the second to the third iteration. The set of 3 iterations constituted a search session. After the third iteration, they were asked to evaluate the feeling of success with the iterations in a 5-level scale. More details on the user study can be found in a previous publication [18].

Our experiment was motivated by the following research questions: (1) Does a system that includes this suggestion tool lead to a higher precision? (2) How does clicking on a suggestion and using suggestions as sources of terms affect precision? (3) Does this effect differ with the language and terminology of the suggestions? (4) Does this effect differ with the English proficiency, health literacy and topic familiarity of the user?

5 Data Analysis

To evaluate the impact of suggestions provided by this system, we considered that users might use them as suggestions, clicking or not on them, or as a source of terms they can use in later queries. Considering this last scenario, we computed the proportion of suggestion's terms that were used in the subsequent query (termsUsed) and the proportion of the suggestion's terms that were used in the following query and were not used in the previous query (newTermsUsed). The former is useful to assess the quality of suggestions' terms and the latter is also useful to assess the utility of the suggestions for the users. Let Q_{it} be the set of unique stemmed terms belonging to the query of the iteration it and S_{it} the set of unique stemmed terms belonging to the suggestion presented in the iteration it, these proportions are computed as follows: $termsUsed_{it} = \frac{|Q_{it} \cap S_{it}|}{|S_{it}|}$ and $newTermsUsed_{it} = \frac{|(Q_{it} \cap S_{it}) \setminus Q_{it-1}|}{|S_{it}|}$. With these proportions, we were able to analyze three scenarios of suggestions' use as source of terms: using or not using suggestions' terms (Terms?), using or not using all the terms of a suggestion (All terms?), using or not using suggestions' terms that were not used in the previous query (NewTerms?). Note that users may use all the terms from a suggestion without clicking it, or they can change the order of the terms or even mix them with other terms.

We used Graded Average Precision (GAP), a measure proposed by Robertson et al. [16], based on a probabilistic model that generalizes average precision to the case of multi-graded relevance in which the user has a binary view of relevance even when using a non-binary scale of relevance. Based on the results presented by GAP's proponents, we used an equally balanced g_1 and g_2, i.e., $g_1 = g_2 = 0.5$, meaning that the levels 1 and 2 of our relevance scale have the same probability of being the grade from which the user starts considering the documents relevant.

The analysis was done comparing iterations where suggestions were used with iterations where suggestions were not used. We compared GAP means between groups of iterations (with and without the use of suggestions) using the Student's t-test. When the assumption of homogeneity of variances was not verified, we applied the Welch t-test. To compare groups of users, we applied the one-way ANOVA and the Tukey's test to locate the differences. Reporting our results, we use a * to mark significant results at $\alpha = 0.05$ and a ** to mark significant results at $\alpha = 0.01$.

6 Results

We found that the first iteration has a higher mean GAP than the second iteration (Tukey's adj. p = 0.009**) and the third iteration (Tukey's adj. p = 3.3e−06**). Differences may be explained by users' criteria in judging relevance. We found that documents with reoccurring contents, because they are no longer useful, are assigned lower relevance scores. Given these differences, we decided to base our analysis on the variation of GAP between iterations. For each iteration we have therefore computed a $\Delta GAP_{it} = GAP_{it} - GAP_{it-1}$. We found no significant differences between ΔGAP_2 and ΔGAP_3.

In SYS+, GAP tends to decrease less (Δ GAP mean of −0.031) than in SYS (Δ GAP mean of −0.033). In all the four use scenarios (Terms?, All terms?, NewTerms? and Click?), we did not find significant differences between using or not using suggestions. After this general analysis, we repeated it by suggestion's language and terminology. Almost all the comparisons were non-significant. The only exceptions occur when new terms from Portuguese (t(147.5) = 2.4, p = 0.01**) or lay (t(139.5) = 2.78, p = 0.003**) suggestions are used, situations in which the impact of suggestions on precision is negative.

6.1 Analysis by English Proficiency

We compared the mean Δ GAP in the four use scenarios, in each group of English proficiency (Table 1). With respect to Portuguese suggestions, although we haven't found significant differences, the general tendency is to have higher precision when users do not use the suggestions. In terms of English suggestions, we found an opposite effect when a suggestion is clicked or when all the terms of a suggestion are used. Yet, this tendency is only significant when *proficient* users click on English suggestions. Proficient users tend to benefit when they use new terms from an English suggestion and, surprisingly, the same happens with *basic* proficiency users when they use terms from English suggestions.

6.2 Analysis by Health Literacy

Comparing the precision of lay and medico-scientific queries by level of health literacy, we found few significant differences. As can be seen in Table 2, two of the three exceptions occur when *marginal* (t(89.9) = 2.3, p = 0.01*) and *functional* (t(33.7) = 2.0, p = 0.03*) health literate users employ terms from lay suggestions they have not used before. These suggestions have a negative impact on the precision of the search sessions of these users. The use of lay suggestions tends to be beneficial to precision when *low* health literate users use all the suggestion's

Table 1. Δ GAP means by language. Boldface represents the maximum in each group and scenario. Square brackets are used there are significant differences between scenarios. EP stands for English proficiency.

	Terms? [w/o — w/]	All terms? [w/o — w/]	NewTerms? [w/o — w/]	Click? [w/o — w/]
Portuguese				
Basic EP	−0.05 — **−0.03**	**−0.04** — −0.05	**−0.04** — −0.08	**−0.04** — −0.05
Independent EP	**−0.01** — −0.04	**−0.02** — −0.06	**−0.02** — −0.06	−0.02 — −0.02
Proficient EP	**−0.03** — −0.08	**−0.04** — −0.05	**−0.03** — −0.10	−0.04 — −0.04
English				
Basic EP	−0.04 — **−0.03**	−0.04 — **−0.03**	**−0.04** — −0.06	−0.04 — **−0.03**
Independent EP	**−0.02** — −0.05	−0.02 — **−0.01**	**−0.02** — −0.08	−0.02 — **−0.01**
Proficient EP	**−0.04** — −0.05	−0.05 — **0.01**	−0.04 — **−0.02**	[−0.05 — **0.03**]*

terms and when these users and the *marginal* health literate users click in the suggestions.

In medico-scientific suggestions, the use of all the suggestion' terms tends to be favorable to precision in all levels of health literacy. Moreover, the use of new terms from suggestions and suggestions' clicks are beneficial to precision in the *low* and *functional* health literacy groups. Of these, the only significant difference occurs when *low* literate users click medico-scientific suggestions (t(35.4) = −1.9, p = 0.03*), showing the positive impact of this type of suggestions.

6.3 Analysis by Topic Familiarity

Considering topic familiarity, whose results are also in Table 2, we found that *familiar* users (TF2) have significantly higher precision in iterations in which they do not use new terms from lay (t(72.9) = 2.2, p = 0.01*) or medico-scientific suggestions (t(76.4) = 2.0, p = 0.02*). We also found that *extremely familiar* users tend to have higher precision with medico-scientific suggestions or when they click or use all the terms from lay suggestions. *Non-familiar* users also seem to benefit from clicks in both lay and medico-scientific suggestions.

Comparing the mean Δ GAP of the several topic familiarity levels in each scenario and type of terminology, we found that, when using new terms from medico-scientific suggestions, *non-familiar* users have a significantly higher precision than *familiar* users (Tukey's adjusted p = 0.01*). This is simultaneously due to the increase in precision in *non-familiar* users and the significant decrease found in the *familiar* users, when using these suggestions.

7 Discussion

On Table 3 we summarize the significant findings previously reported, only found when users use new terms from suggestions or when they click on them.

Answering the first research question, the system with suggestions (SYS+) tended to demonstrate better performance in terms of precision. The positive

Table 2. ΔGAP means by terminology. Boldface represents the maximum in each group and scenario. Square brackets are used there are significant differences between scenarios. HL stands for Health Literacy.

	Terms? [w/o — w/]	All terms? [w/o — w/]	NewTerms? [w/o — w/]	Click? [w/o — w/]
Lay				
Low HL	**−0.04** — −0.06	−0.05 — **−0.03**	**−0.04** — −0.07	−0.05 — **0.00**
Marginal HL	**−0.02** — −0.04	**−0.02** — −0.04	[**−0.02** — −0.08]**	−0.03 — **−0.01**
Functional HL	**−0.04** — −0.10	**−0.05** — 0.10	[**−0.05** — −0.11]**	**−0.05** — −0.09
Not familiar	**0.01** — −0.03	**0.00** — −0.03	**0.00** — −0.06	0.00 — **0.02**
Familiar	**−0.04** — −0.06	**−0.04** — −0.05	[**−0.03** — −0.11]*	−0.04 — −0.04
Extremely familiar	**−0.04** — −0.05	−0.05 — **−0.04**	**−0.04** — −0.06	−0.05 — **−0.01**
MS				
Low HL	**−0.04** — −0.04	−0.04 — **−0.05**	−0.04 — **−0.04**	[−0.05 — **0.01**]*
Marginal HL	**−0.02** — −0.03	−0.03 — **−0.03**	**−0.02** — −0.06	**−0.03** — −0.04
Functional HL	**−0.04** — −0.08	−0.06 — **−0.03**	−0.05 — **−0.05**	−0.05 — **−0.04**
Not familiar	**0.00** — −0.01	0.00 — **0.02**	0.00 — **0.01**	0.00 — **0.03**
Familiar	**−0.04** — −0.06	**−0.04** — −0.06	[**−0.03** — −0.10]*	**−0.04** — −0.06
Extremely familiar	−0.05 — **−0.04**	−0.05 — **−0.04**	−0.05 — **−0.04**	−0.05 — **−0.04**

effects of the suggestion system has also been previously shown in the medical accuracy of the knowledge obtained in the session [18].

Pertaining the second research question, we found no significant differences between using or not using the overall set of suggestions, either as a whole or as a source of terms. Answering the third research question, negative effects of language and terminology are found when users use new terms from Portuguese and lay suggestions. No other significant effect was found.

Moving on to the fourth research question, as seen in Table 3, the effect differs in the two use scenarios, clicking on suggestions has positive effects on precision and using new terms from them has negative effects. English suggestions are advantageous to *proficient* users when they click on them. The use of lay and medico-scientific suggestions has also a good effect on precision. Surprisingly, lay suggestions increase precision not only in *non-familiar* users but also in the *extremely familiar* group.

The precision increase found when *low* health literate users click on medico-scientific suggestions is consistent with what has been found in a previous study, that is, "less subject expertise seems to lead to more lenient and relatively higher relevance ratings" [17]. This means that these users may be assessing documents regarding their relation with the topic instead of their utility to themselves. Findings in *non-familiar* users could be explained the same way but, since we have found in a previous study [18] that medico-scientific suggestions increase their answers' correct contents and tends to decrease their incorrect contents, we have reasons to believe this is not the case. Moreover, this agrees with what we found in a previous study [9] where we concluded that health literacy is more important to comprehend medico-scientific documents than topic familiarity.

Table 3. Summary of the significant findings. ↑ denotes increases and ↓ decreases in each outcome.

	NewTerms?	Click?
General		
English		Proficient EP (↑)
Portuguese	General (↓)	
Lay	General (↓)	
	Marginal HL (↓)	Non-familiar (↑)
	Functional HL (↓)	Extremely familiar (↑)
	Familiar (↓)	
Medico-scientific	Familiar (↓)	Low HL (↑)
		Non-familiar (↑)
		Extremely familiar (↑)

With the use of new terms from suggestions, precision decreases with the use of Portuguese and lay suggestions. The same happens with lay suggestions in higher levels of health literacy and with both lay and medico-scientific suggestions in users familiar with the topic.

We have also compared the performance of different groups of users in each use scenario for each type of suggestion. In these comparisons, we found that users *familiar* with the topic have lower precision than *non-familiar* users when new terms from medico-scientific suggestions are used. This may be explained by the benefits that *non-familiar* users seem to obtain from medico-scientific suggestions in every scenario.

8 Conclusion

We describe a query suggestion system for the health domain and study its impact on the precision of the search session considering several user characteristics. This analysis takes into account the utility of the suggestions as new whole queries and as sources of terms.

We found that a system with these suggestions is beneficial for the precision of the search session. In a previous work we have reached a similar conclusion regarding medical accuracy [18]. The best precision effects of the suggestion tool are achieved when users use it strictly as a suggestion tool, that is, when they click in suggestions.

Previously, we found that English suggestions are preferred to the Portuguese by the general user, both in terms of clicks and as source of new terms [10]. We have also found that clicking English suggestions is beneficial for the medical accuracy of the knowledge acquired during the search session. In this work, we have shown that the benefit of whole English suggestions is not restricted to medical accuracy but also applies to precision if the user is proficient in this language. The benefits of these suggestions allied with users preference for them show the potential of English suggestions for users with other native languages in the health domain. This corroborates the finding of a previous study [8] that suggest that non-English–speaking users having at least elementary English proficiency can benefit from a system that suggests English alternatives for their queries.

In terms of terminology, medico-scientific suggestions are preferred to lay ones by the general user, in clicks and as source of terms, a preference that increases with health literacy [10].

In terms of medical accuracy, both terminologies are favourable to the general user. In the present work we found that both terminologies are advantageous to specific groups of users but not to the general user. This suggests that personalizing the suggestion system might have good effects on precision.

As future work, we would like to explore the effects of this suggestion system on motivational relevance.

Acknowledgments. 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. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Inf. Res. **8**(3) (2003). http://informationr.net/ir/8-3/paper152.html
2. Fox, S.: Health topics. Technical report, Pew Internet & American Life Project, Washington, DC (2011)
3. Gao, W., Niu, C., Nie, J.Y., Zhou, M., Wong, K.F., Hon, H.W.: Exploiting query logs for cross-lingual query suggestions. ACM Trans. Inf. Syst. **28**(2), 1–33 (2010). https://doi.org/10.1145/1740592.1740594
4. Jansen, B.J., McNeese, M.D.: Evaluating the effectiveness of and patterns of interactions with automated searching assistance. J. Am. Soc. Inf. Sci. **56**(14), 1480–1503 (2005). https://doi.org/10.1002/asi.20242
5. Kelly, D., Gyllstrom, K., Bailey, E.W.: A comparison of query and term suggestion features for interactive searching. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 371–378. ACM, New York (2009). https://doi.org/10.1145/1571941.1572006
6. Kogan, S., Zeng, Q., Ash, N., Greenes, R.A.: Problems and challenges in patient information retrieval: a descriptive study. In: Proceedings AMIA Symposium, pp. 329–333 (2001). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243602/
7. Kriewel, S., Fuhr, N.: Evaluation of an adaptive search suggestion system. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 544–555. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_47
8. Lopes, C.T., Ribeiro, C.: Measuring the value of health query translation: an analysis by user language proficiency. J. Am. Soc. Inf. Sci. Technol. **64**(5), 951–963 (2013). https://doi.org/10.1002/asi.22812

9. Lopes, C.T., Ribeiro, C.: Effects of terminology on health queries: an analysis by user's health literacy and topic familiarity, vol. 39, chap. 10, pp. 145–184. Emerald Group Publishing Limited (2015). https://doi.org/10.1108/S0065-283020150000039013
10. Lopes, C.T., Ribeiro, C.: Effects of language and terminology on the usage of health query suggestions. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 83–95. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_7
11. Luo, G., Tang, C.: On iterative intelligent medical search. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 3–10. ACM, New York (2008). https://doi.org/10.1145/1390334.1390338
12. Luo, G., Tang, C., Yang, H., Wei, X.: MedSearch: a specialized search engine for medical information retrieval. In: Proceeding of the 17th ACM Conference on Information and Knowledge Mining, CIKM 2008, pp. 143–152. ACM, New York (2008). https://doi.org/10.1145/1458082.1458104
13. McCray, A.T., Tse, T.: Understanding search failures in consumer health information systems. In: AMIA Annual Symposium Proceedings, pp. 430–434 (2003). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479930/
14. NLM: 2012AA consumer health vocabulary source information (2012). http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html
15. Rawson, K.A., et al.: The METER: a brief, self-administered measure of health literacy. J. Gen. Intern. Med. **25**(1), 67–71 (2010). https://doi.org/10.1007/s11606-009-1158-7
16. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending average precision to graded relevance judgments. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 603–610. ACM, New York, July 2010. https://doi.org/10.1145/1835449.1835550
17. Saracevic, T.: Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance. J. Am. Soc. Inf. Sci. Technol. **58**(13), 2126–2144 (2007)
18. Lopes, C.T., Paiva, D., Ribeiro, C.: Effects of language and terminology of query suggestions on medical accuracy considering different user characteristics. J. Assoc. Inf. Sci. Technol. **68**(9), 2063–2075 (2017). https://doi.org/10.1002/asi.23874
19. Toms, E.G., Latter, C.: How consumers search for health information. Health Inform. J. **13**(3), 223–235 (2007). https://doi.org/10.1177/1460458207079901
20. Zarro, M., Lin, X.: Using social tags and controlled vocabularies as filters for searching and browsing: a health science experiment. In: Fifth Workshop on Human-Computer Interaction and Information Retrieval, October 2011
21. Zeng, Q.T., Crowell, J., Plovnick, R.M., Kim, E., Ngo, L., Dibble, E.: Assisting consumer health information retrieval with query recommendations. J. Am. Med. Inform. Assoc. (JAMIA) **13**(1), 80–90 (2006). https://doi.org/10.1197/jamia.m1820
22. Zhang, Y.: Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 210–219 (2010)
23. Zielstorff, R.: Controlled vocabularies for consumer health. J. Biomed. Inform. **36**(4–5), 326–333 (2003). https://doi.org/10.1016/j.jbi.2003.09.015