Characterizing Health-Related Information Needs of Domain Experts

Eya Znaidi[1], Lynda Tamine[1], Cecile Chouquet[2], and Chiraz Latiri[3]

[1] IRIT, University of Toulouse
[2] Institute of Mathematics, University of Toulouse
[3] Computer Sciences Department, Faculty of Sciences of Tunis

Abstract. In information retrieval literature, understanding the users' intents behind the queries is critically important to gain a better insight of how to select relevant results. While many studies investigated how users in general carry out exploratory health searches in digital environments, a few focused on how are the queries formulated, specifically by domain expert users. This study intends to fill this gap by studying 173 health expert queries issued from 3 medical information retrieval tasks within 2 different evaluation compaigns. A statistical analysis has been carried out to study both variation and correlation of health-query attributes such as length, clarity and specificity of either clinical or non clinical queries. The knowledge gained from the study has an immediate impact on the design of future health information seeking systems.

Keywords: Health Information Retrieval, Information Needs, Statistical Analysis.

1 Introduction

It is well known in information retrieval (IR) area that expressing queries that accurately reflect the information needs is a difficult task either in general domains or specialized ones and even for expert users [14,17]. Thus, the identification of the users' intention hidden behind queries that they submit to a search engine is a challenging issue. More specifically, according to the Pew Internet and American Life Project, health-related queries are increasingly expressed by a wide range of age groups with a variety of backgrounds [23]; consumer health information through online environments support a variety of needs including the promotion of health and wellness, use of health care services, information about disease and conditions, and information about medical tests, procedures and treatment. Unfortunately, it reveals from the literature that despite of the diversity of the available health IR systems and the diversity of the used information sources, users still felt in retrieving relevant information that meet their specific mental needs [2,21]. To answer this issue, several studies focused on the analysis of health searchers' behaviour, including attitudes, strategies, tasks and queries [11,16,18]. These studies involved large numbers of subjects within general web search settings, with uncontrolled experimental conditions, making it difficult to generalize their findings to expert searches involved by physicians.

N. Peek, R. Marín Morales, and M. Peleg (Eds.): AIME 2013, LNAI 7885, pp. 48–57, 2013.

Moreover, most of these studies focused on search behaviour through search strategies and tactics. Unlike previous work, we address more specifically in this paper, domain expert health search through the analysis of query attributes namely length, specificity and clarity using appropriate proposed measures built according to different sources of evidence. For this aim, we undertake an in-depth statistical analysis of queries issued from IR evaluation compaigns namely Text REtrieval Conference (TREC)[1] and Conference and Labs of the Evaluation Forum (CLEF)[2] devoted for different medical tasks within controlled evaluation settings. Our experimental study includes a statistical pair-wise attribute correlation analysis and a multidimensional analysis across tasks.
The remainder of this paper is structured as follows. Section 2 presents related work on health information searching. Section 3 details the query attributes and section 4 describes the tasks and query collections analysed in the study. In section 5 we present and discuss the results analysis. Finally, section 6 summarizes the study findings, highlights design implications and concludes the paper.

2 Related Work

The increasing amount of health information available from various sources such as government agencies, non-profit and for-profit organizations, internet portals etc. presents opportunities and issues to improve health care information delivery for medical professionals [1], patients and general public [10]. One critical issue is the understanding of users' search strategies and tactics for bridging the gap between their intention and the delivered information. To tackle this problem, several studies investigated mainly the analysis of consumer health information behaviour in one side and their query formulations in the other side. Regarding consumer's health information behavior, several aspects have been investigated such as: (1) pattern of health information searching [16]: findings highlight in general that health IR obey to a trial-and-error process, or can be viewed as a serie of transitions between searching and browsing, (2) access results [16]: studies revealed that the majority of users access to top documents in the ranked outcome list of results, (3) goals, motivation and emotions particularly in social environments [13]: the authors emphasize that motivation is the main factor leading to the success or failure of health searches. More close to our work, the second category of research focused on query formulation issues by analysing query attributes such as length and topics. Several studies [11,14,20] highlighted that queries are short containing less than 3 terms with an average of 2 terms. For instance authors in [20] studied health related information searches on MedlinePlus and hospitals and revealed that queries lengths were in the range 1-3. The same general finding has been reported in [11] regarding queries submitted to Healthlink on the basis of 377000 queries issued from search logs. [14] reported quite analoguous results from health web searches studies. Through other observations at the topic level [8,19,22], where topics where

[1] `http://trec.nist.gov/`
[2] `http://www.clef-initiative.eu/`

identified using linguistic features or medical items, it seems that users may do not, in general, make use of terminologies and taxonomies; they use in contrast terms of their own-self leading to mispelling ones or abbreviations. However the above studies were conducted in the context of general web search involving participants with a variety of skills, backgrounds and tasks. Other studies looked at the differences in search strategies between domain experts and novices in the medical domain [3,4] or focused on expert information search behaviors [9,18]. In studies conducted in [3,4], the authors oberved significative differences in the way the users explore the search, beginning from key resources viewed as hubs of their domain for expert domain users rather than starting from general resources for novices. In [18], the author examined the search behavior of medical students at the beginning of their courses, the end of the courses and then six months later. The results suggest that search behaviour changes in accordance with domain expertise gain. In the work presented in this paper, we focus on the analysis of query formulations expressed by medical experts. The main underlying objective is not to explicitly compare expert health searches from novices, but to highlight the pecularities of expert search queries in attempt to customize the search which in turn, can impact medical education and clinical decisions. We address the following research questions: (1) How expert query attributes are correlated within a medical task and across different medical tasks? (2) Are clinical queries significantly different from non clinical queries?

3 Query Attributes

In our study, we consider a health-IR setting where an expert submits a query Q to a target collection of documents C. We propose and formalize in what follows various query attributes and justify their construction.

- ***Query Length.*** We retain two facets of query length: (1) length as the number of significant words, $LgW(Q)$, and (2) length as the number of terms referencing preferred entries of concepts issued from MESH[3] terminology, $LgC(Q)$. Our choice of MESH terminology is justified by its wide use in the medical domain. For this aim, queries are mapped to MESH terminology using our contextual concept extraction technique [6,7].
- ***Query Specificity.*** Specificity is usually considered as a criterion for identifying index words [12]. In our study, we are interested in two facets:
  1. *Posting specificity* $PSpe(Q)$: expresses the uniqueness of query words in the index collection; the basic assumption behind posting specificity is that less documents are involved by query words, more specific are the query topics. It is computed as follows:

$$PSpe(Q) = \frac{1}{LgW(Q)} \sum_{t_i \in words(Q)} -\log(\frac{n_i}{N}) \quad (1)$$

[3] MEdical Subject Headings.

where $LgW(Q)$ is the query length in terms of words, $words(Q)$ is the set of query words, n_i is the number of documents containing the word t_i, N is the total number of documents in the collection C.

2. *Hierarchical specificity* $HSpe(Q)$: it is based on the query words deepness of meaning defined in MESH terminology. The basic underlying assumption is that a child word is more specific than its parent word in the terminology hierarchy. Hierarchical specificity is given by:

$$HSpe(Q) = \frac{1}{LgC(Q)} \sum_{c_i \in Concepts(Q)} \frac{\text{level}(c_i) - 1}{\text{Maxlevel(MESH)} - 1} \tag{2}$$

where $LgC(Q)$ is the query length in terms of concepts, $Concepts(Q)$ is the set of query concepts, $level(c_i)$ is the MESH level of concept c_i, $Maxlevel(MESH)$ is the maximum level of MESH hierarchy.

– ***Query Clarity.*** Broadly speaking, a clear query triggers a strong relevant meaning of the underlying topic whereas an ambiguous query triggers a variety of topics meanings that do not correlate each other. We propose to compute two facets of clarity:

1. *Topic based clarity* $TCla(Q)$: The clarity score of a query is computed as the Kullback-Leiber divergence between the query language model and the collection language model, given by [15]:

$$TCla(Q) = \sum_{t \in V} P(t|Q) log_2 \frac{P(t|Q)}{P_{coll}(t)} \tag{3}$$

where V is the entire vocabulary of the collection, t is a word, $P_{coll(t)}$ is the relative frequency of word t and $P(t|Q)$ is estimated by: $P(t|Q) = \sum_{d \in R} P(w|D)P(D|Q)$ where d is a document, R is the set of all documents containing at least one query word.

2. *Relevance based clarity* $RCla(Q)$: a query is assumed to be as much clear as it shares concepts with relevant documents assessed by experts. This assumption is the basis of IR models. Accordingly, $RCla(Q)$ is computed as:

$$RCla(Q) = \frac{1}{|R(Q)|} \sum_{d \in R(Q)} \frac{|Concepts(Q) \cap |Concepts(d)|}{LgC(Q)} \tag{4}$$

where $R(Q)$ is the set of relevant documents returned for query Q as assessed by experts, $|Concepts(d)|$ (resp. $|Concepts(Q)|$) is the number of document concepts (resp. query concepts).

– ***Query Category.*** We are interested in both clinical and non clinical queries. For this aim, we used the PICO model to classify queries [5]: P corresponds to patient description (sex, morbidity, race, age etc.), I defines an applied intervention, C corresponds to another intervention allowing comparison or

control and O corresponds to experience results. According to this definition, we manually annotated all the test queries as clinical (C) if they contain at least 3 PICO elements, non clinical (NC) otherwise.

4 Data Sources

To perform this study, we used data issued from TREC and CLEF. We exploited queries (number is noticed $Nb.Q$), documents (number is noticed $Nb.D$) and physicians relevance assessments data with respect to various medical IR tasks described below:

- *TREC Medical records task* ($Nb.Q = 35, Nb.D = 95.701$): the retrieval task consists in identifying cohorts for comparative effectiveness research. Queries describe short disease/condition sets developed by physicians; documents represent medical visit reports.
- *TREC Genomics series task*: The TREC Genomics task was one of the largest and longest running challenge evaluations in biomedicine. This task models the setting where a genomics researcher entering a new area expresses a query to a search engine managing a biomedical scientific literature namely from Medline collection. TREC genomics queries evolved across years: gene names in 2003 ($Nb.Q = 50, Nb.D = 525.938$), information needs expressed using acronyms in 2004 ($Nb.Q = 50, Nb.D = 4.591.008$) and question-answering in the biomedical domain in 2006 ($Nb.Q = 28, Nb.D = 162.259$).
- *ImageCLEF case-based task* ($Nb.Q = 10, Nb.D = 55.634$): The goal of the task was to retrieve cases including images that a physician would judge as relevant for differential diagnosis. The queries were created from an existing medical case database including descriptions with patient demographics, limited symptoms, test results and image studies.

5 Results

5.1 Query Characteristics

To highlight the major differences between the collections (medical tasks), we first performed a descriptive analysis. Figure 1 shows the distributions of the six query attribute facets per collection and for all the queries presented by boxplots. Analysis of variance or non-parametric Kruskal-Wallis tests (adapted to small samples) were performed to compare attributes averages and to detect significative differences between the different collections (indicated by $p - value < 0.05$). From figures 1.(a) and 1.(b), it is interesting to notice that similar trends are observed between the two facets of length. Moreover, the query length attribute is significantly different across the five query collections ($p - value < 0.0001$), despite the fact that they all represent experts' information needs. The highest query length was observed for ImageCLEF queries with 24 terms and 5 concepts in average, versus lowest values for TRECGenomics2003 queries (4.6 terms and 1.4 concepts on average).

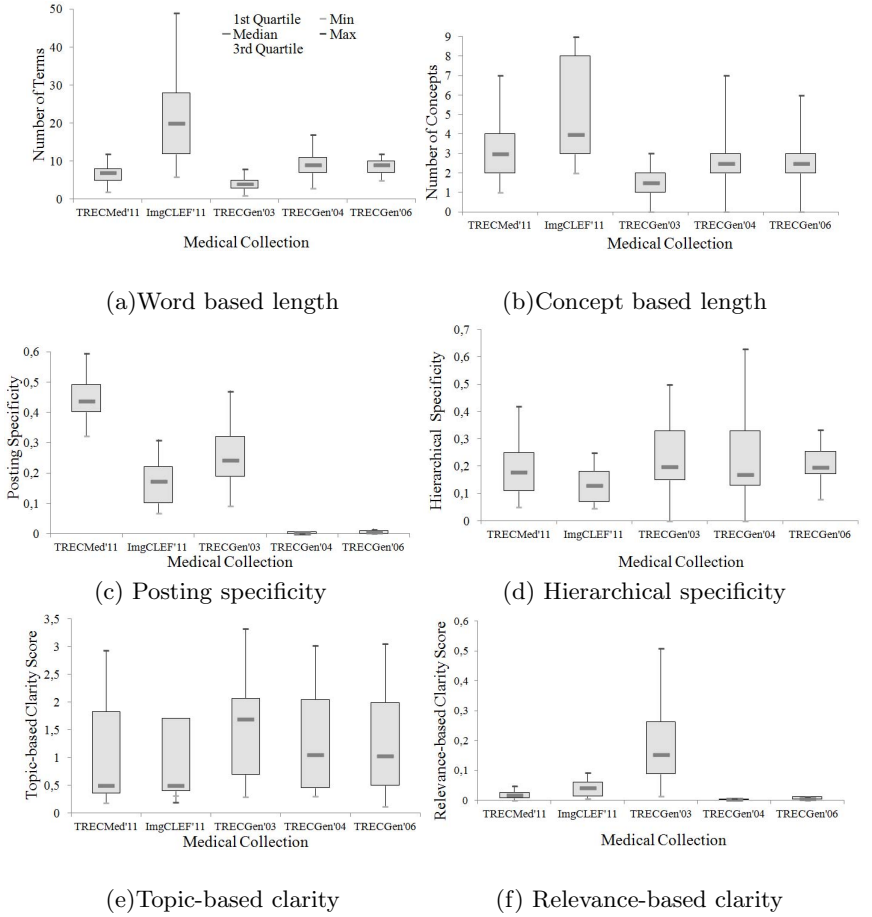

Fig. 1. Query attribute facet distributions per collection

This can be explained by the main differences in the related tasks. Indeed, while in ImageCLEF, physicians express long technical descriptions of patient cases including images, biomedicine experts in TRECGenomics2003 express queries as gene names leading to relative short queries and consequently to a few number of concepts. Figures 1.(c) and 1.(d) represent respective distributions of query posting specificity scores and hierarchical specificity scores based on MESH terminology. As expected, given the definitions of these two facets, resulted scores are different. Significative differences of posting specificities were observed between the collections ($p-value < 0.0001$), whereas hierarchical specificities are not significantly different between the collections. As shown in Figure 1.(c), TRECGenomics 2004 and 2006 collections are characterized by relatively low values of posting specificity, compared with the three other collections. This can be also explained by the nature of the task: in TRECGenomics2004 collection, experts abuse of acronyms and abbreviations that are poorly distributed

in Medline documents. In TRECGenomics2006, queries were expressed as specific entity-based questions about genes and proteins. Regarding hierarchical specificity, we observe that value ranges are wider for TRECGenomics2003 and TRECGenomics2004 (and potentially TRECMedical 2011) indicating that medical experts tend to make use of specialized terms through terminologies. Moreover, higher values of specificities indicate how much experts make use of their domain knowledge to point on their specific information needs. Analysing the clarity attribute from figures 1.(e) and 1.(f) conduct us to observe that the differences between the collections are more emphasized for the relevance based clarity score ($p-value < 0.0001$) than for the topical based one ($p-value > 0.05$), probably due to the larger variability of the latter score, as shown by the range of boxplots in figure 1.(e). However, some trends are similar: highest clarity scores are identified for TRECGenomics2003 queries, in opposition to ImageCLEF ones. This indicates that searching for genes and proteins favors the expression of unambiguous queries whereas medical case patient descriptions trigger various expert intents. We previously identified that queries on genes and proteins are short, so we can assume that even short expert queries can be clear depending on the task involved.

5.2 Correlation Analysis of Query Attributes

To study correlations between query attributes, we computed Spearman correlation coefficient (ρ) between the six quantitative query attribute facets. Only highly significative correlation pairs for each collection were displayed in Table 1. Given the differences highlighted above between the collections, we study the correlations between the query facets for each collection. In the four larger collections, we observe systematically a strong positive correlation between query length in number of words and query length in number of concepts ($p - value < 0.0001$). This was expected for two different reasons: the first one is related to the fact that a biomedical concept entry is generally, by definition, a set of words. The second reason, is as stated above, related to the search strategy of medical experts in searching for health information that favours the use of concepts relying on their domain knowledge. We can also notice that all significative correlations involve query length in number of words, reflecting the importance of this feature to characterize expert information needs. Other significative correlations are highlighted but not systematically on all the collections. In TRECMedical and TRECGenomics06 collections, a significative positive correlation is observed between the query length in number of words and posting specificity. This can be partly explained by the fact that, according to formula (1), longer is the query, higher is its posting specificity in general, the correlation is particularly higher for the two above tasks because of their comparable nature regardless of the form of the need: simple query or factual question related to a specific patient case. However, query length in terms and hierarchical query specificity facets are negatively correlated for TRECGenomics2003 ($p - value < 0.0001$). This is explained by the nature of the underlying task and the used terminology, namely MESH for concept recognition; it is probably not appropriate such as GENE

Table 1. Results of two-way significantly correlated facets

Collection	Facet1	Facet2	ρ	$p-value$
TREC Medical2011 ($N = 35$)	$LgW(Q)$	$LgC(Q)$	0.69	< 0.0001
	$LgW(Q)$	$PSpe(Q)$	0.39	< 0.02
TREC Genomics2003 ($N = 50$)	$LgW(Q)$	$LgC(Q)$	0.55	$< .0001$
	$LgW(Q)$	$HSpe(Q)$	-0.54	< 0.0001
TREC Genomics2004 ($N = 50$)	$LgW(Q)$	$LgC(Q)$	0.47	< 0.001
TREC Genomics2006 ($N = 28$)	$LgW(Q)$	$LgC(Q)$	0.67	< 0.0001
	$LgW(Q)$	$PSpe(Q)$	0.58	< 0.001

Ontology to point on specific terms. Finally, concerning ImageCLEF collection, no significative correlation is identified between query attributes.

5.3 Comparative Analysis of Clinical Vs. Non Clinical Queries

This part of our study aims to identify the attributes that can differentiate clinical queries from non clinical queries. According to TRECGenomics2004 and TRECGenomics2006 tasks, our manual annotation of clinical Vs. non clinical queries confirms that they do not include any clinical queries. Thus, given the low number of identified clinical queries this analysis will focus on all the 5 pooled collections (173 queries including 27 clinical ones). We modeled each attribute facet according to the query category and the collection by a two-way analysis of variance. No significative interaction between the collection and the clinical category was detected, indicating that the differences between clinical and non clinical queries, if they exist, are similar in the five collections, and justifying pooling collections for this analysis. Results displayed in Table 2 present mean (noted by m) and standard deviation (noted by $s.d.$) of each query attribute facet, for clinical and non clinical queries. In addition, a p-value (corresponding to the query category effect from the two-way analysis of variance) is given, allowing to assess differences between clinical and non clinical queries as significative ($p-value < 0.05$) or not. The results of the two-way analysis of variance revealed significative differences between clinical and non clinical queries. For length in words, the average number of words in clinical queries is estimated at 10 against 8 in non clinical queries ($p-value < 0.01$). As expected, there is also a significative difference in length in concepts ($p - value < 0.02$): clinical queries have on average three identified concepts against two for non clinical queries. The average score of relevance clarity is slightly higher for clinical queries ($p-value = 0.05$). For this attribute, the differences between clinical and non clinical queries are more highlighted per collection. Another comment relates to the posting specificity: the results of the analysis of variance with two attribute facets are in favor of no difference between specificity of clinical queries and those of non clinical queries ($p - value > 0.05$). But considering only query category factor in the model (without collection factor), we detect a highly significative effect of the clinical category on the posting specificity ($p - value < 0.0001$), more in accordance with average values of 0.32 for clinical queries and 0.15 for non clinical

Table 2. Clinical Vs. non clinical queries analysis

Facet	Clinical (n=27) m (s.d.)	Non clinical (n=146) m (s.d.)	p-value
LgW	10.0 (10.0)	8.0 (5.4)	< 0.01
LgC	3.2 (2.0)	2.3 (1.5)	< 0.02
PSpe	0.32 (0.13)	0.15 (0.18)	> 0.05
HSpe	0.19 (0.11)	0.24 (0.19)	> 0.05
TCla	1.27 (0.89)	1.31 (0.84)	> 0.05
RClar	0.07 (0.10)	0.065 (0.12)	0.05

queries. This can be explained by the fact that the TRECGenomics2004 and TRECGenomics2006 collections that have no clinical queries are characterized by very low values of specificity, as described above.

6 Findings Summary and Concluding Remarks

The analysis results issued from our study provide a picture of the pecularities of medical experts' queries with respect to different tasks. Our findings demonstrate, that unlike web health searchers [20], physicians' queries are relatively long and that length depends on the task at hand: medical case based search lead to longer queries than entity based search; moreover physicians make use of both their domain knowledge and semantic ressources for formulating queries, being specific particularly in medical case related information search. It has also been highlighted that clinical queries, compared to non clinical ones, are longer in both words and concepts, clearer according to the relevance facet, and more specific according to the posting facet. These findings suggest for the design of novel functionalities for future health IR systems including: query suggestion or query reformulation using appropriate levels of terminology to improve query clarity, search results personalization based on expertise level, query category, and task. Before designing such tools, we plan in future to undertake first, a large-scale user study that highlights the differences between expert search and novice search in the medical domain.

References

1. Andrews, J.E., Pearce, K.A., Ireson, C., Love, M.M.: Information-seeking behaviors of practitioners in a primary care practice-based research network (PBRN). Journal of the Medical Library Association, JMLA 93(2), 206–212 (2005)
2. Arora, N., Hesse, B., Rimer, B.K., Viswanath, K., Clayman, M., Croyle, R.: Frustrated and confused: the american and public rates its cancer-related information-seeking experiences. Journal of General Internal Medicine 23(3), 223–228 (2007)
3. Bhavnani, S.: Important cognitive components of domain-specific knowledge. In: Proceedings of Text Rerieval Conference TREC, TREC 2001, pp. 571–578 (2001)
4. Bhavnani, S.: Domain specific strategies for the effective retrieval of health care and shopping information. In: Proceedings of SIGCHI, pp. 610–611 (2002)

5. Boudin, F., Nie, J., Bartlett, J.C., Grad, R., Pluye, P., Dawes, M.: Combining classifiers for robust pico element detection. BMC Medical Informatics and Decision Making, 1–6 (2010)
6. Dinh, D., Tamine, L.: Biomedical concept extraction based on combining the content-based and word order similarities. In: Proceedings of the 2011 ACM Symposium on Applied Computing, SAC 2011, pp. 1159–1163. ACM, New York (2011)
7. Dinh, D., Tamine, L.: Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 375–386. Springer, Heidelberg (2011)
8. Dogan, R., Muray, G., Névéol, A., Lu, Z.: Understanding pubmed user search behavior through log analysis. Database Journal, 1–19 (2009)
9. Ely, J.W., Osheroff, J.A., Ebell, M.H., Chambliss, M.L., Vinson, D.C., Stevermer, J.J., Pifer, E.A.: Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ 324(7339), 710 (2002)
10. Eysenbach, G.: Consumer health informatics. Biomedical Journal (3), 543–557 (2012)
11. Hong, Y., Cruz, N., Marnas, G., Early, E., Gillis, R.: A query analysis of consumer health information retrieval. In: Proceedings of Annual Symposium for Biomedical and Health Informatics, pp. 791–792 (2002)
12. Jones, S.: A statistical interpretation of term specificity and its application to retrieval. Journal of Documentation 28(1), 11–20 (1972)
13. Oh, S.: The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. Journal of the American Society for Information Science and Technology 63(3), 543–557 (2012)
14. Spink, A., Jansen, B.: Web Search: Public Searching of the Web. Kluwer Academic Publishers (2004)
15. Steve, C.R., Croft, W.: Quantifying query ambiguity. In: Proceedings of the Second International Conference on Human Language Technology Research, HLT 2002, San Francisco, CA, USA, pp. 104–109 (2002)
16. Tomes, E., Latter, C.: How consumers search for health information. Health Informatics Journal 13(3), 223–235 (2007)
17. White, R., Moris, D.: How medical expertise influences seb search behaviour. In: Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 791–792 (2008)
18. Wildemuth, B.: The effects of domain-knowledge on search tactic formulation, vol. 55, pp. 246–258 (2004)
19. Zeng, Q., Crowell, J., Plovnick, R., Kim, E., Ngo, L., Dibble, E.: Research paper: Assisting consumer health information retrieval with query recommendations. Journal of American Medical Informatics Associations 13(1), 80–90 (2006)
20. Zeng, Q., Kogan, S., Ash, N., Greenes, R., Boxwala, A.: Characteristics of consumer technology for health information retrieval. Methods of Information in Medicine 41, 289–298 (2002)
21. Zeng, Q., Kogan, S., Plovnick, R., Croweel, J., Lacroix, E., Greens, R.: Positive attitudes and failed queries: An exploration of the conundrums of health information retrieval. International Journal of Medical Informatics 73(1), 45–55 (2004)
22. Zhang, J., Wolfram, D., Wang, P., Hong, Y., Gillis, R.: Visualization of health-subject analysis based on query term co-occurrences. Journal of American Society in Information Science and Technology 59(12), 1933–1947 (2008)
23. Zickuhr, K.: Generations 2010. Technical report, Pew Internet & American Life Project (2006)