Health Card Retrieval for Consumer Health Search

An Empirical Investigation of Methods

Jimmy[1,2], Guido Zuccon[1], Bevan Koopman[3], Gianluca Demartini[1]
[1]University of Queensland, Brisbane, Australia
[2]University of Surabaya (UBAYA), Surabaya, Indonesia
[3]Australian E-Health Research Center, CSIRO, Brisbane, Australia
jimmy@uqconnect.edu.au,g.zuccon@uq.edu.au,bevan.koopman@csiro.au,g.demartini@uq.edu.au

ABSTRACT

This paper investigates methods to rank health cards, a domain-specific type of entity cards, for consumer health search (CHS) queries. A key challenge in this context is which card(s) should be presented to the user. In particular, little evidence exists to determine the effectiveness of retrieval and ranking methods for health cards in CHS. CHS is a challenging domain, where users lack domain expertise and thus are often unable to formulate effective queries, and to interpret the retrieved results. In addition, unlike in other contexts, CHS presents the opportunity to exploit a number of domain specific characteristics and features.

In this paper, we focus on difficult queries with self-diagnosis intents. Our study makes the following contributions: (1) it assembles and releases the first test collection of health cards for research purposes, and (2) it empirically evaluates a large range of entity retrieval methods adapted to health cards retrieval, including features specific to health cards for learning to rank. This is the first study that thoroughly investigates methods to rank health cards.

KEYWORDS

health cards, consumer health search, entity retrieval, learning to rank

ACM Reference Format:
Jimmy[1,2], Guido Zuccon[1], Bevan Koopman[3], Gianluca Demartini[1]. 2019. Health Card Retrieval for Consumer Health Search: An Empirical Investigation of Methods. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3357384.3358128

1 INTRODUCTION

An entity card is an information object within a search engine results page (SERP) that encapsulates a variety of information on a particular entity [3, 7, 12, 15, 17]. Entity cards have been used to support user search activities in the context of a user's formulated query [7, 12], as well as in proactive systems where relevant entity cards are shown before any query is submitted [15]. Presenting relevant entity cards is known to increase user engagement with the search results and reduce the number of queries issued, thus improving the overall user experience [7].

In the context of consumer health search (CHS), a health card is a specific type of entity card which presents information about a specific health concept in a coherent and easy to read form [2, 10]. Health cards are beneficial to users in the context of self-diagnosis and health decision making [10]: they reduce the effort and workload required to complete the search task, and help less knowledgeable users to take well informed health decisions.

However, while previous work has considered whether the display of a health card allowed users to make better health decisions, and if so under which conditions. Our own previous research [11] has shown that the display of multiple health cards (up to 4) better supported health search users in specific circumstances, e.g., to make differential diagnosis, when the relevance of a single health card was uncertain [11].

Research on retrieving and ranking entity cards (or even deciding if displaying a card at all) is limited – even more so in the context of health cards and health search. No previous work has directly investigated the retrieval and ranking of health cards. While related to the general problem of retrieving entity cards, health cards and health search present its own challenges. Effective search is hindered by vocabulary mismatch and lack of domain expertise by users. These issues affect both health search in terms of query formulation and result interpretation and health card retrieval and ranking in terms of matching and deciding the utility of the health card (and thus whether to display the card) [5, 6, 18]. This may explain why, for example, commercial web search engines currently limit the display of health cards to queries that explicitly contain the card's title (i.e., the entity name) [2], although health cards have been shown to be valuable also in the presence of less "navigational" and more explorative health queries [10]. For example, health cards are commonly triggered when the query is a condition such as "meningitis"; while they are not displayed when the user queries using observed symptoms (e.g., "headache fever neck stiff light sensitivity"). This is despite the search results do suggest a relationship between the queries with the observed symptoms and the health condition, e.g., 70% of the search results in the first Google SERP for the symptoms query above relate to "meningitis".

In this paper we study the problem of ranking health cards in answer to explorative, self-diagnostic consumer health search queries. These are difficult queries for search engines to answer, as they are often underspecified and ambiguous [6]. They are also difficult for users to formulate and appraise results for. Thus, for these queries, the display of relevant high quality health cards could most benefit users' health information acquisition and decision [10].

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358128

Table 1: Example health scenarios and correct diagnoses.

Topic 5, Diagnosis: Deep vein thrombosis
Scenario (Topic): Your 65-year-old aunt has had leg pain and swelling over the last 5 days. She has had a high blood pressure, mild congestive heart failure, and recent hospitalization for pneumonia. She had been recovering from the pneumonia at home but when beginning to move around and walk, her right leg became painful, tender, and swollen. Her veins in the right leg are enlarged and her right leg is slightly redder than her left. The back of her knee also fells tender.
Topic 10, Diagnosis: Meningitis
Scenario (Topic): Your 18-year-old nephew is experiencing a very bad headache and fever over the last 3 days. He also complaints of light sensitivity and neck stiffness.

Table 2: Query variations for topic id 10.

migraine	headache fever light sensitive stiff neck
migraines	light sensitivity neck stiffness headache
migraine symptoms fever	light sensitivity neck stiffness is it the flu
bad headache with fever cause	light sensitivity, headache, stiff neck, fever
headache. fever, neck stiffness	headache fever light sensitivity neck stiffness
headache and fever self treatment	headache fever light sensitivity neck stiffness feber
how often do migraines cause fevers	headache fever light sensitivity neck stiffness fever
headache and fever, light sensitivity	bad headache and fever light sensitivity and neck stiffness

On the other hand, these are queries for which identifying the correct health card(s) to display is particularly challenging. Showing multiple cards may appear to be relevant to the user's ambiguous query as multiple possible diagnoses may be relevant to a medical case, until further analysis of evidence does not lead to the iterative exclusion of the least likely, i.e., the process of differential diagnosis. In this context, we make the following contributions:

(1) We assemble and release the first collection of health cards for research purposes. The considered health cards are related to a large set of queries for medical cases used for evaluation, associated ground truth diagnoses, and other possible diagnoses a user may hypothesise on the basis of (real) search results for the queries.

(2) We empirically evaluate four general entity retrieval methods adapted to the problem of ranking health cards, providing the first quantitative evaluation of entity retrieval techniques in this context. In doing so, we consider specialisations of such techniques to the specific settings of health card retrieval.

2 METHOD

2.1 Creation of Topics and Query Variations

We used 45 standardised patient vignettes from a survey of symptoms checkers [8] as basis of our health search topics. These vignettes were compiled from various medical sources such as education material for health professionals and a medical resource website. Each vignette contained age, gender, symptoms, correct diagnosis and correct category of triage urgency for a given condition. Of the 45, we discarded 4 vignettes since there were no health cards in our collection that match their correct diagnosis.

For each vignette, we created a health search task scenario by removing clinical observations which would be not possible for a user to know (e.g., imaging findings, chest auscultations, etc.). We also replaced medical terms with their appropriate layman terms (e.g., "rhinorrhea" was replaced with "runny nose"), to make the scenarios more realistic. Table 1 shows 2 of the 41 scenarios used in this study. Each scenario constituted a topic in our experiment. Each topic has only one relevant health card based on the correct diagnosis for the topic's scenario.

Amazon Mechanical Turk (AMT) workers were recruited to complete HITs based on the 41 topics. In each HIT, we asked workers to use a custom web search engine interface we developed, which mimicked a typical web search engine and allowed users to enter queries and retrieve web pages. Search engine results and snippets were acquired using the Bing Web Search API; only organic results were shown (no advertisement and entity cards). AMT workers were asked to identify the most likely health condition for the given topic and indicate what should the person in the scenario do next (e.g., self-treat, seek attention from a medical professional, etc.). At minimum, workers were asked to submit a query before they could submit their answers. We enforced this by asking workers to mark a search result that they considered as most useful when completing a task. We allowed workers to submit as many queries as required to complete the task.

For each topic, we recruited U.S. based AMT workers; they were paid \$0.2 for completing a HIT (1 HIT = 1 topic), with a bonus of \$0.5 for submitting a correct diagnosis[1]. For quality control, initially, we offered 12 HITs for each of the first 10 topics. Then, we evaluated the worker submissions. Workers who submitted poor results were blocked and their submissions discarded. Finally, we sent invitations to the remaining workers to complete the 31 topics left (10 HITs per topic). In total, we accepted 372 submissions for 41 topics (average of 9.1 submissions per topic). Some topics have less than 10 submissions since poor results were discarded and some HITs were left unfinished. 45.58% of the accepted submissions have the correct diagnosis.

From the 372 submissions, we collected 626 query variations of which 586 were unique, but we removed two queries which failed to retrieve any health cards ("rantidine" and "excema"), thus in total we considered 584 query variations. The average length for unique query variations is 6.78 ± 5.31 words. Table 2 shows query variations for topic 10 (see Table 1).

41 out of 584 query variations had "navigational" intent targeting a specific health condition that is not the correct condition for the topic. While these queries are part of the user attempts to identify the correct diagnosis (e.g., "meningitis"), the correct health cards for these queries (query-based relevance), e.g., "migraine", may not be the one for the correct diagnosis (topic-based relevance), e.g., "meningitis". Thus, we considered two types of relevance judgements: topic-based and query-based. For topic-based, the relevant health card is the one that matches the correct diagnosis for the topic. For query-based, the relevant health card is the one that matches the health concept for the query – query-based relevance judgements were considered only for the 41 navigational queries. The query-based judgement results are available online[2].

2.2 Creation of the Health Cards Collection

To create the health cards collection, we crawled pages within the "Diseases and Conditions" sections from Mayo Clinic[3], as of April 16, 2019. In total, we harvested information for 1,142 health conditions, resulting in as many health cards. Our collection is comparable in size to that of web search engines like Bing[4]. For each condition, we extracted its name (title), aliases, overview, symptoms, and

[1] Paid after the diagnosis was verified against the ground truth.

[2] http://ielab.io/publications/jimmy-2019-healthcardretrieval

[3] https://www.mayoclinic.org/diseases-conditions

[4] To verify this, 377.942 health condition phrases extracted from the UMLS were submitted to Bing: 38.958 phrases retrieved a total of 1.330 health cards.

treatments. This ensured that each card contained fielded information similar to those health cards that are commonly shown by commercial search engines, e.g., Google health cards.

2.3 Considered Ranking Methods

The health cards catalogue was indexed with respect to the fields title, aliases, overview, symptoms, and treatments. Stop words were removed and Porter stemmer applied.

Then, we considered three common entity ranking models [9, 12, 16]: BM25F, LM with Dirichlet smoothing and Fielded Sequential Dependence Model (FSDM). In addition, we investigated the effectiveness of Learning to rank (LTR). For this, we used LambdaMart, a listwise ranking algorithm as implemented by QuickRank [4]. As features for LTR, we adopted all term-based features for entity ranking listed by Balog [12]. Then, we added health-entity-based features which are akin to the term-based features. However, instead of terms, we used health concepts to represent queries and health cards. For example, for query length for the health entity feature, instead of counting the number of words in a query, we counted the number of health concepts. To obtain health entities, we used QuickUMLS [14] to map terms in queries and health cards to UMLS concept identifiers. Because of the primarily self-diagnosis intent of the considered topics and search queries, we retained only health entities for which the concept identifiers belonged to the following semantic groups: disorders, chemicals & drugs, devices, procedures, and anatomy. Finally, we considered features that specifically exploit the characteristics of health cards:

- Sum of the similarity between health entities in query & field f
- Sum of the similarity between health entities in query & card

These features attempt to measures the similarity between health entities in queries and in cards. We used word2vec clinical concept embeddings [1] to measure the similarity between health entities.

For all methods, we performed parameter tuning using the same 5-fold cross validation split at topic level. For LTR, for each fold we used 80% of the training data for learning and 20% for validation and tuning, with the held out fold used for testing. For BM25F, LM and FSDM, we tuned the field weights between 0 (ignored), 1, and 2 (twice more important). For BM25F, we tuned the field length normalisation b for long fields (overview, symptoms, and treatments), considering the values 0.25, 0.5, and 0.75. We did not tune for title and aliases fields as b has minimum impact on short fields. We did not tune k as the term frequency statistic across health cards are unlikely to significantly vary. For LM and FSDM, we tuned μ for long fields with values 500, 1000, and 1500.

2.4 Measures and Statistical Analysis

We used success at rank 1 (S@1), i.e., how many times the correct health card was retrieved at rank 1, as primary evaluation metric because it is common practice for current commercial search engines to display only one health card for a query. We also considered S@4 to investigate effectiveness when multiple health cards (4 cards) were shown: this was because of recent work showing the display of multiple cards better supported health search users in exploratory tasks like those considered here [11]. In addition, we considered reciprocal rank (RR) to track at which rank position the correct health card was retrieved: this provides a measure of the gap between the current and the expected performance (i.e., RR=1 means the correct card was successfully retrieved at rank 1). Finally, we also computed session-based success measures; i.e., whether the correct health card for the search task was displayed at the targeted ranks at any point throughout a user's search session.

Table 3: Experiment results for user queries and scenarios. Win, Tie and Loss show the number of queries that performed better, equal, or worse than BM25F.

	User Queries (n=584)			**Win/Tie/Loss**		
	S@1	S@4	RR	S@1	S@4	RR
$BM25F^a$	$.2252^{bCd}$	$.4426^{BC}$	$.3411^{BC}$	-	-	-
LM^b	$.1798^{aCD}$	$.3493^{ACD}$	$.2774^{ACD}$	22/512/50	36/457/91	164/132/288
$FSDM^c$	**.2928**^{AB}	**.4991**^{ABD}	**.4003**^{ABd}	67/488/29	55/507/22	269/159/156
LTR^d	$.2748^{aB}$	$.4332^{BC}$	$.3610^{Bc}$	69/476/39	51/475/58	189/112/283

	Scenario Query (n=41)			**Win/Tie/Loss**		
	S@1	S@4	RR	S@1	S@4	RR
$BM25F^a$	**.3415**	.4634	**.4362**	-	-	-
LM^b	.2927	.4390	.3909	2/35/4	1/38/2	12/13/16
$FSDM^c$	.2195	**.5122**	.3796	1/34/6	5/33/3	13/11/17
LTR^d	.2195	.3658	.3015	3/30/8	2/33/6	6/6/29

Statistical significant differences were measured using pairwise t-test with Bonferroni correction. These are reported in results tables using lower case superscripts for $\alpha < 0.05$ and upper case for $\alpha < 0.01$.

3 EXPERIMENT RESULTS

3.1 Retrieval performance

Table 3 reports the retrieval results obtained by the different methods. Along with the results for user queries, we also report (for completeness) the results obtained when the verbose (layman) scenario was used as the query. While it is not expected users would query with the complete scenario in real situations, we observe that the best methods on the scenarios obtain higher effectiveness than the best methods on the user queries. This is understandable as the user queries generally contain less details than the scenarios, highlighting both the difficulties by users in formulating effective queries and by systems in retrieving the correct cards.

In terms of overall retrieval effectiveness on realistic user queries, methods based on FSDM and LTR are comparable (with FSDM being generally better than LTR); both significantly better than LM, while only FSDM exhibits statistical significant differences with BM25F. This may be because FSDM does explicitly address expressions with sequential words that are commonly seen in the user queries (e.g., "neck stiffness" in Table 2), while BM25F does not.

Table 3 also reports a summary of query-by-query comparison between each method and BM25F (e.g., a *win*: better than BM25F on one query). This shows that while FSDM and LTR achieve a similar number of wins over BM25F for S@1 and S@4 (for user queries), LTR does have more losses.

3.2 Search session based evaluation

Within the search sessions obtained via AMT, users submitted multiple queries when completing a task (Section 2.1). Different search styles were displayed. For example, Table 5 reports a querying session of a user, displaying progressive reformulation of queries as the session goes on. In this example, the user concluded the search task with a clear query and obtained the information which lead them making the correct diagnosis.

Table 4: Session-based results: success is measured over the whole search session for a single user, rather than for each query separately.

	User Sessions (n=372)		Win/Tie/Loss	
	S@1	S@4	S@1	S@4
$BM25F^a$	$.2997^{bd}$	$.5269^{BC}$	-	-
LM^b	$.2406^{aCD}$	$.4462^{ACd}$	18/313/41	22/297/53
$FSDM^c$	$.3548^{B}$	$\mathbf{.5860}^{ABD}$	49/294/29	34/326/12
LTR^d	$\mathbf{.3683}^{aB}$	$.5134^{bC}$	57/286/29	34/298/40

Table 5: A query session and simulated card retrieval effectiveness for Scenario 38 (correct diagnosis: "canker sore").

	q_1 mouth ulceration	q2 mouth herpes	q3 mouth herpes ulcer	q4 canker sore
S@1	0.0000	0.0000	0.0000	1.0000
S@4	0.0000	0.0000	1.000	1.0000
RR	0.1667	0.1429	0.5000	1.0000

Table 6: Experiment results for user queries using LTR with only basic features (42 Features) vs. LTR with also health specific features (90 features). * and ** indicates statistical significance difference with regard to LTR-basic at $\alpha < 0.05$ and $\alpha < 0.01$ levels.

	User Queries (n=584)			Win/Tie/Loss		
	S@1	S@4	RR	S@1	S@4	RR
$LTR - basic$	.2166	.3947	.3187	-	-	-
LTR	$\mathbf{.2748}^{**}$	$\mathbf{.4332}^{*}$	$\mathbf{.3610}^{**}$	66/488/32	65/478/43	174/116/296

This motivated us to investigate session-based evaluation – i.e., consider the effectiveness of methods over a whole session for each single users. We report session-based evaluation results in Table 4 for S@1,4. These results show that over a session, LTR is better than other methods at retrieving the correct card at rank 1 (although significant differences are found only with respect to LM); if multiple (4) cards were displayed, then FSDM is best.

3.3 Features importance for Learning to Rank

RankEval [13] was used to study the effect different features had on LTR performance; the tool ranks features based on their importance and usage statistics. Figure 1 reports the top ten most important features for each query set based on importance gain. This analysis suggests that for LTR, the features we introduced based on specific health cards characteristics (entity based representation and entity similarities) are amongst the features that contribute most to the LTR effectiveness.

We further investigate the gains obtained in LTR from including health cards specific features as opposed to considering only the generic features for entity retrieval [12](LTR-basic). Table 6 shows that the health card features we introduced in this work provide significant gains in performance, with the LTR using basic features even performing worse than BM25F.

4 CONCLUSION

In this paper, we empirically investigated methods to retrieve health cards in the context of consumer health search with self-diagnosis intents. Queries were often ambiguous and generally difficult, as they often did not contain an explicit mention of the target health card title. As part of this investigation, we assembled the first test collection of health cards containing information for 1,142 health conditions. Furthermore, we collected 373 search session for a total of 626 queries (586 unique) for 41 self-diagnosis search tasks. This data is released to the research community[5].

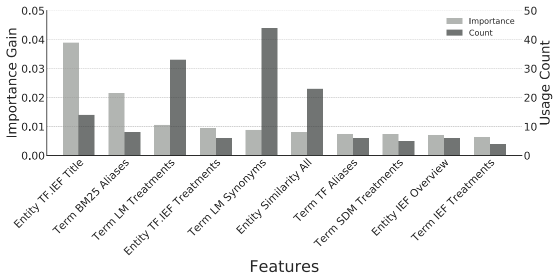

Figure 1: Feature importance for LTR on user queries.

The retrieval methods considered were BM25F, LM, FSDM and LTR. Our results suggest that FSDM and LTR are comparable and performed best across a large set of realistic user queries for the task at hand. For LTR, we introduced features specific to health cards (health entities and health entities similarities), which were found to have a statistically strong impact on the effectiveness of this method.

Acknowledgments: Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP) (20151022014644). Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award; both these grants funded this study.

REFERENCES

[1] Beam AL, Kompa B, Fried I, Palmer NP, Shi X, Cai T, and Kohane IS. 2018. Clinical concept embeddings learned from massive sources of medical data. *arXiv preprint arXiv:1804.01486* (2018).
[2] Gabrilovich E. 2016. Cura Te Ipsum: answering symptom queries with question intent. In *Second WebQA workshop, SIGIR 2016 (invited talk)*.
[3] Radlinski F, Craswell N, Billerbeck B, Shokouhi M, Ahari S, Agrawal N, Hoad T, Zhou S, and Awan MA. 2015. Entity detection and extraction for entity cards. US Patent 9,158,846.
[4] Capannini G, Lucchese C, Nardini FM, Orlando S, Perego R, and Tonellotto N. 2016. Quality versus efficiency in document scoring with learning-to-rank models. *IP&M* 52, 6 (2016), 1161–1177.
[5] Zuccon G and Koopman B. 2018. SIGIR 2018 Tutorial on Health Search (HS2018): A Full-day from Consumers to Clinicians. In *SIGIR'18*. ACM, 1391–1394.
[6] Zuccon G, Koopman B, and Palotti J. 2015. Diagnose this if you can. In *ECIR'15*.
[7] Bota H, Zhou K, and Jose JM. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *CHIIR'16*.
[8] Semigran HL, Linder JA, Gidengil C, and Mehrotra A. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351 (2015).
[9] Chen J, Xiong C, and Callan J. 2016. An empirical study of learning to rank for entity search. In *SIGIR'16*. ACM, 737–740.
[10] Jimmy, Zuccon G, Koopman B, and Demartini G. 2019. Health Cards for Consumer Health Search. In *SIGIR'19*.
[11] Jimmy, Zuccon G, Demartini G, and Koopman B. 2019. Health Cards to Assist Decision Making in Consumer Health Search. In *AMIA'19*.
[12] Balog K. 2018. *Entity-oriented search.* Springer.
[13] Claudio L, Cristina IM, Franco MN, Raffaele P, and Salvatore T. 2017. RankEval: An Evaluation and Analysis Framework for Learning-to-Rank Solutions. In *SIGIR'2017*.
[14] Soldaini L and Goharian N. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR Workshop*.
[15] Shokouhi M and Guo Q. 2015. From Queries to Cards: Re-ranking Proactive Card Recommendations Based on Reactive Search History. In *SIGIR'15*.
[16] Zhiltsov N, Kotov A, and Nikolaev F. 2015. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *SIGIR'15*.
[17] Thomas P, Moffat A, Bailey P, Scholer F, and Craswell N. 2018. Better Effectiveness Metrics for SERPs, Cards, and Rankings. In *ADCS '18*.
[18] Zeng Q, Kogan S, Ash N, Greenes RA, and Boxwala AA. 2002. Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine* 41, 4 (2002), 289–298.

[5] http://ielab.io/publications/jimmy-2019-healthcardretrieval