Using Electronic Medical Records and Physician Data to Improve Information Retrieval for Evidence-Based Care

Mengqi Jin
Dept. of Biostatistics
Brown University
Providence, RI

Hongli Li
UpToDate
Waltham, MA

Christopher H. Schmid
Dept. of Biostatistics
Brown University
Providence, RI

Byron C. Wallace
iSchool
University of Texas at Austin
Austin, TX

***Abstract*—Healthcare practitioners are increasingly using search functionality embedded in Electronic Medical Record (EMR) software to search for relevant evidence summaries at point of care. We introduce a *learning to rank* approach that exploits information carried in EMR data and UpToDate user accounts to (significantly) improve ranking results, compared to a comparable model that does not exploit such features.**

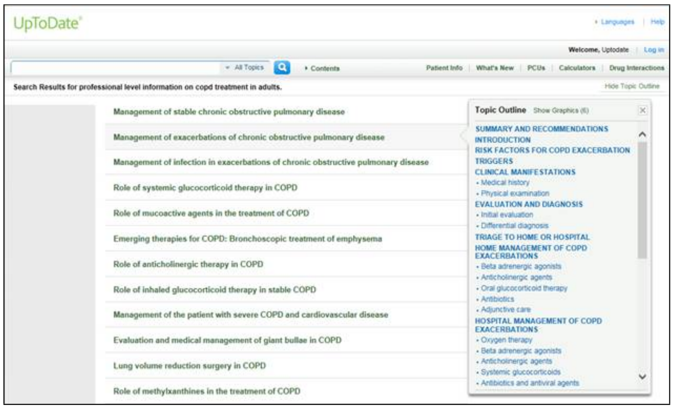
1. Introduction

Healthcare practitioners must be able to rapidly find reliable evidence relevant to their clinical questions if they are to make informed, evidence-based medical decisions. Typically this means consulting online resources via keyword search. We aim to improve this information retrieval process by augmenting free-text queries with contextual information gleaned from Electronic Medical Record (EMR) and UpToDate account data.

UpToDate[1] is an evidence-based clinical decision support resource widely used by healthcare practitioners. UpToDate provides concise reviews of over 15,000 *topics* that summarize published relevant evidence written and reviewed by over 5,700 clinicians. These are accessible via a web interface: users issue keyword searches to find topics relevant to their information needs. Topics contain multiple *sections* (see Figure 1), and the sought after information will likely be found in only a few of these. We aim to build a model that re-ranks sections to facilitate rapid information retrieval.

We exploit the intuition that different types of users will likely be interested in different sorts of information (which will be found in different sections): doctors with different specialities will likely be interested in different facets of topics. Thus users may be looking for different information within the same topic, even if they issued identical queries. Retrieval strategies that rely exclusively on term frequency matching will fail to appropriately tailor results to meet individual healthcare needs.

To address this shortcoming, we present a learning to rank approach that capitalizes on attributes in EMR data and UpToDate user accounts to personalize section ranking results. Using a large corpus of recorded UpToDate searches conducted from EMR software, *we demonstrate that this approach significantly improves ranking performance over a baseline learning to rank model that does not exploit EMR features*.



Search Results for professional level information on copd treatment in adults.

Management of stable chronic obstructive pulmonary disease
Management of exacerbations of chronic obstructive pulmonary disease
Management of infection in exacerbations of chronic obstructive pulmonary disease
Role of systemic glucocorticoid therapy in COPD
Role of mucoactive agents in the treatment of COPD
Emerging therapies for COPD: Bronchoscopic treatment of emphysema
Role of anticholinergic therapy in COPD
Role of inhaled glucocorticoid therapy in stable COPD
Management of the patient with severe COPD and cardiovascular disease
Evaluation and medical management of giant bullae in COPD
Lung volume reduction surgery in COPD
Role of methylxanthines in the treatment of COPD

SUMMARY AND RECOMMENDATIONS
INTRODUCTION
RISK FACTORS FOR COPD EXACERBATION
TRIGGERS
CLINICAL MANIFESTATIONS
- Medical history
- Physical examination

EVALUATION AND DIAGNOSIS
- Initial evaluation
- Differential diagnosis

TRIAGE TO HOME OR HOSPITAL
HOME MANAGEMENT OF COPD EXACERBATIONS
- Beta adrenergic agonists
- Anticholinergic agents
- Oral glucocorticoid therapy
- Antibiotics
- Adjunctive care

HOSPITAL MANAGEMENT OF COPD EXACERBATIONS
- Oxygen therapy
- Beta adrenergic agonists
- Anticholinergic agents
- Systemic glucocorticoids
- Antibiotics and antiviral agents

Figure 1: The UpToDate search interface. We aim to re-rank the topic sections (right) to meet individual needs.

2. Corpus

UpToDate has collaborated with several large EMR vendors to integrate search directly into EMR platforms. This allows healthcare providers to directly perform searches through their EMR software during clinical visits (i.e., at point of care), as is increasingly common [1], [2]. Searches conducted in this way by users registered with UpToDate carry contextual information, including patient and healthcare provider data. Available EMR variables include: patient age and gender, encounter type (e.g., emergency or inpatient), subtopic (diagnosis, dosing, etc.), search performer (provider or patient) and task context (this captures the task being undertaken by the physician at the point of the query, e.g., 'medication list review'). This information follows the Health Level Seven International (HL7) standard for EMR data: hence we refer to it as "HL7 data". We also have access to data stored in provider UpToDate accounts, including professional role (*nurse practitioner*, *physician*, etc.), specialty, and institution type and name. To simplify

[1] http://www.uptodate.com/home

presentation, we group these account attributes with the HL7 data.

UpToDate's search engine is built on top of the open source software framework Apache Lucene.[2] The current system works well for rapid topic identification: in historical search data (90,611 searches in all), users clicked through to only a single topic from the ranked list in 85% of searches conducted through EMR systems. Our focus is therefore meeting more individualized and specific information needs by ranking sections *within* topics for users by exploiting HL7 data.

2.1. Deriving section relevance labels

Our dataset comprises 11 months (July 2013 to May 2014) of browsing logs from UpToDate, which we describe further below. For learning and evaluation we require relevance labels for sections, but we do not have explicit relevance assessments. Instead we use the implicit information recorded in the search logs: the 'dwell time' data. We also capitalize on other types of user browsing behavior, such as scrolling and text highlighting events.

We aim to use patient and provider attributes to improve search results. Thus we will be concerned with triplets of *person-query-documents* (PQDs) in the UpToDate logs. We transform continuous dwell times logged within PQDs into ordinal relevance labels. Specifically, we introduce the ordered relation $5 \succ 4 \succ 3 \succ 2 \succ 1$ (where 5 implies highest relevance). We then derive labels for the sections in the ith PQD, $\mathbf{y_i} = \{y_{i,1}, \ldots, y_{i,n_i}\}$, where n_i is the number of sections in the document. We mapped dwell times onto these labels in accordance with the empirical quintiles into which they fell, e.g., sections with dwell times in the top quintile were designated as 5 (highest relevance). We assigned the lowest relevance label (1) to sections with dwell times of 0.

We allowed more explicit forms of feedback (user *copying* and *highlighting* actions) to override relevance labels assigned based on dwell time alone. In particular, we labeled all sections from which users copied text as highly relevant (5). And we incremented by one the dwell time-based relevance scores calculated for sections in which the user highlighted text (unless said sections had already received the highest relevance level).

To summarize: we derived section relevance scores for documents in the context of a participant/query (PQ) pair as a function of (1) time spent on the section (i.e., the dwell time) and, when available, (2) user actions (highlighting and copying of text) recorded for sections.

3. Methods

We aim to personalize section rankings for queries based on patient and provider attributes contained in the HL7 data. We treat this as a *learning to rank* task and we base our approach on SVM-rank [3], [4], which we next describe. We then provide a detailed account of the novel features we introduce that capitalize on contextualizing information to improve section ranking.

3.1. Learning to rank using Rank-SVM

Rank-SVM is a learning to rank algorithm in which a ranking task is transformed into an equivalent binary classification problem [3], [4]. Consider a set of sections ranked for a specific PQD triple i: $\{(\mathbf{x}_{i,1}, y_{i,1}), (\mathbf{x}_{i,2}, y_{i,2}), \ldots, (\mathbf{x}_{i,n_i}, y_{i,n_i})\}$, where we are denoting feature vectors of sections by $\mathbf{x}$'s and relevance labels by y's. For rank-SVM, we transform this data into training instances $(\mathbf{x}_{i,(a,b)}, y_{i,(a,b)})$ for each pair of examples $\{(\mathbf{x}_{i,a}, y_{i,a}), (\mathbf{x}_{i,b}, y_{i,b})\}$, where $\mathbf{x}_{i,(a,b)} = \mathbf{x}_{i,a} - \mathbf{x}_{i,b}$, and $y_{i,(a,b)} = +1$ if $y_{i,a} \succ y_{i,b}$, $y_{i,(a,b)} = -1$ if $y_{i,a} \prec y_{i,b}$. Rank-SVM minimizes the following hinge loss $\mathcal{L}$ for a weight vector $\mathbf{w}$ [5]:

$$\sum_{i=1}^{m} \sum_{a=1}^{n_i-1} \sum_{b=a+1}^{n_i} max\{0, 1 - y_{i,(a,b)} \mathbf{w} \cdot \mathbf{x}_{i,(a,b)}\}$$

where m is the number of PQDs. To select $\mathbf{w}$, this loss is traded off against model simplicitly via a squared ℓ_2 regularizer scaled by a parameter C.

3.2. HL7 features

A key task in learning to rank is designing a suitable representation for items to be ranked. Here we introduce features that encode EMR and UpToDate account information.

All HL7 fields are categorical. We introduce features that express conditional probability estimates to capture section viewing preferences of users with shared attributes. For example, such features might capture that a given section is specifically relevant to dermatologists (but perhaps not to physicians of other specialties). Intuitively, these carry signal if individuals with shared attributes express similar section preferences for topics. More specifically, denoting by $d_{i,j}$ a section in topic D_i, where $j \in \{1, \ldots, n_i\}$, and HL7 attribute X by HL7_X, we calculate conditional feature values as follows:

$$\frac{count(dwell(d_{i,j}) \geq \tau, HL7\_X = x)}{count(dwell(d_{i,j}) \geq \tau) + 1} \tag{1}$$

where τ is a dwell time threshold (in this work we set this to 3 seconds). This feature expresses an estimated probability that a user will have HL7 attribute $X = x$ given that their dwell time for section j was at least τ seconds.[3] For example, X may be 'specialty' and x may be 'dermatologist': here the feature would capture the frequency with which the section is relevant at some level to dermatologists (normalized by the total section popularity across all users). We refer to these as HL7P features.

[2] http://lucene.apache.org/

[3] We have added 1 to the denominator to avoid division by 0.

Content features	
Level1SecLuceneScore	Lucene score for the query.
QueryLevel1SecTitle5Gram	Average 1-5 gram overlap between the query and the section title.
TopicTitleLev1SecTitle5Gram	Average 1-5 gram overlap between the topic title and the section title.
section_position	Section position in the topic.
hierSecRefProb	$\frac{count(\text{links to section})}{count(\text{all links})}$ where links are observed hyperlinks among sections.
hierSecClickProb	$\frac{count(\text{sub-section}\in\text{section})+1}{count(\text{section'}\in\text{topic})}$ where sub-sections are nested within sections, and all sections have at least 1 sub-section by construction (hence the.
sec_prob	estimate of overall section relevance, i.e., the number of PQDs in which this section was dwelled on for at least τ seconds divided by the total count of PQDs in which the parent topic was opened.
HL7P features	probability estimate of section relevance conditioned on ...
Gender_x	... patient gender.
AgeGroup_x	... (discrete) patient age group.
EncounterType_x	... patient encounter type (e.g., emergency or inpatient).
Performer_x	... the party performing the search.
Subtopic_x	... the specified *subtopic* of the search (e.g., diagnosis).
TaskContext_x	... the task *context* type (current healthcare process step).
InformationRecipient_x	... conditioned on recipient role (patient or provider).
ProfRole_x	... the professional role of the provider (e.g., nurse).
Specialty_x	... the specialty of the provider (e.g., cardiology).
InstitutionType_x	... the institution type from which the search was issued (e.g., teaching hospital).

TABLE 1: Details on the feature sets we use. Content features exploit only data from the query and section and do not capitalize on the information carried in the EMR or UpToDate account data.

We also experimented with language-based "HL7L" features that capture textual similarities between EMR attributes and section texts. Specifically we used Lucene scores, shared n-gram counts and cosine similarities. Due to space constraints we do not explain them in further detail here (these only marginally improved performance).

Table 1 summarizes the Content and HL7P features that we introduced.

4. Experimental Setup and Results

Our aim is to determine whether augmenting instances with contextualizing features derived from EMRs and UpToDate user accounts improves ranking performance. We thus consider two baseline strategies: **baseline I** uses only the section Lucene score (w.r.t. the query) and **baseline II** is a learning to rank approach (SVM-rank) that does not exploit the HL7 features we have proposed. *Our primary experimental question is whether we are able to outperform baseline II by exploiting HL7 features.*

We used the subset of 6,192 Patient-Query-Document pairs (PQDs) in the UpToDate logs that contained topics viewed at least 20 times. We used 75% of these (4,643 PQDs) for training and held the other 25% (1,549 PQDs) out for testing. We can view each PQD as a separate ranking task (over document sections) for which performance statistics may be calculated. We evaluated our approach using two standard metrics for ranking: Kendall's τ [6] and normalized discounted cumulative gain (NDCG) [7]. We tuned the C parameter via cross-fold validation on the training dataset to maximize Kendall's τ.[4]

4.1. Results

Results are shown in Table 2. HL7 features improve performance w.r.t. both metrics. To assess statistical significance, we performed a one-sided Wilcoxon Signed-Rank test for both the Kendalls τ and NDCG metrics. For our paired data, we use the ranking metrics calculated for each test PQD using methods both with and without the HL7 features. We found a statistically significant difference in ranking performances between models that used and did not use HL7 features. That is, the median difference between the Content+HL7P+HL7L model and model that used Content alone was significantly greater than zero ($p << 0.0001$ for Kendalls τ and $p << 0.0001$ for NDCG).

Model	Avg. Kendall's τ	Avg. NDCG
Baseline I (*Lucene*)	0.2013	0.6707
Baseline II (*SVM-Rank; no HL7*)	0.6345	0.8407
Content+HL7P	0.6698	0.8604
Content+HL7L	0.6352	0.8419
Content+HL7L+HL7P	**0.6729**	**0.8647**
HL7L+HL7P	0.4906	0.7796
HL7L	0.0283	0.5793
HL7P	0.4813	0.7758

TABLE 2: Results (averages over the 1,549 heldout PQDs).

Figure 2 describes the relative weights of the ten most predictive features in the best performing model.[5] Features derived from EMRs and UpToDate accounts dominate this plot. An example observation here is that emergency room physicians (2nd most predictive feature) seem to have unique information needs, which squares with intuition.

5. Related Work

Existing personalized ranking strategies focus on exploiting users' search and browsing histories to personalize search result rankings [8], [9], [10]. Relatively little work has been done using personalized health information for ranking. Yadav and Poellabauer [11] did recently propose a ranking model that attempts to take both a user's search query and the user's health profile into account. Their method simply concatenates parameters from Personal

[4] We searched over $C \in \{2^{-6}, \ldots, 2^{12}\}$.

[5] We calculated relative weights by normalizing the raw $\mathbf{w}$ terms by $max(\mathbf{w})$.

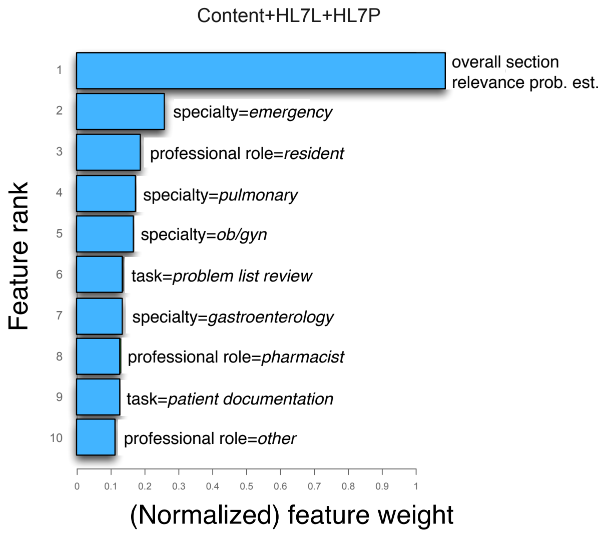


Figure 2: The most predictive features and their relative weights in the best performing model (content+HL7P+HL7L).

Health Records (e.g., previous medications) to the original query and uses this expanded string as input to the Lucene scoring function. In contrast, our approach that does not require building profiles for each user, but instead exploits general contextual parameters.

6. Conclusions

We have introduced a novel learning to rank approach that uses EMR and provider data to significantly improve section rankings within topics for point of care information retrieval, as compared to an equivalent learning to rank model that does not exploit this information. In future work, we hope to refine our feature sets and the way that these are treated within the learning to rank model.

References

[1] K. Natarajan, D. Stein, S. Jain, and N. Elhadad, "An analysis of clinical queries in an electronic health record search utility," *International journal of medical informatics*, vol. 79, no. 7, pp. 515–522, 2010.

[2] L. Yang, Q. Mei, K. Zheng, and D. A. Hanauer, "Query log analysis of an electronic health record search engine," in *AMIA Annual Symposium Proceedings*, vol. 2011. American Medical Informatics Association, 2011, pp. 915–922.

[3] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Advances in neural information processing systems*, pp. 115–132, 1999.

[4] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[5] H. Li, "Learning to rank for information retrieval and natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 4, no. 1, pp. 1–113, 2011.

[6] M. Kendall and J. D. Gibbons, *Correlation methods*. Oxford University Press, 1990.

[7] H. Li, "A short introduction to learning to rank," *IEICE Transactions*, vol. 94-D, no. 10, pp. 1854–1862, 2011.

[8] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui, "Modeling the impact of short-and long-term behavior on search personalization," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 185–194.

[9] J. Teevan, D. J. Liebling, and G. Ravichandran Geetha, "Understanding and predicting personal navigation," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 85–94.

[10] B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 718–723.

[11] N. Yadav and C. Poellabauer, "An architecture for personalized health information retrieval," in *SHB*, C. C. Yang, H. Chen, H. D. Wactlar, C. Combi, and X. Tang, Eds. ACM, 2012, pp. 41–48.