An Adaptive Evidence Weighting Method for Medical Record Search

Dongqing Zhu and Ben Carterette
Department of Computer & Information Sciences
University of Delaware
Newark, DE, USA 19716
[zhu | carteret]@cis.udel.edu

ABSTRACT

In this paper, we present a medical record search system which is useful for identifying cohorts required in clinical studies. In particular, we propose a query-adaptive weighting method that can dynamically aggregate and score evidence in multiple medical reports (from different hospital departments or from different tests within the same department) of a patient. Furthermore, we explore several informative features for learning our retrieval model.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

medical record search; EMR; information retrieval; cohort identification; language models

1. INTRODUCTION

The rich health information contained in electronic medical records (EMR) is useful for improving quality of care. One important application is to search EMR to identify cohorts for clinical studies, which requires retrieval systems specifically designed with medical domain knowledge.

To promote research on medical information retrieval, particularly for EMR retrieval, the Text REtrieval Conference (TREC) organized a Medical Records track in 2011 and 2012 [11, 10]. The task is an ad hoc search task for patient visits based on unstructured text in EMR. One particular problem in EMR search is how to aggregate and score evidence that distributes across multiple documents. This is because a patient can have multiple medical reports generated from several hospital departments or even from different tests within a single department.

In this paper, we propose a novel weighting method that can adaptively weight evidence with respect to different queries. We evaluate our algorithm on TREC test collections. The cross-validation results show that our weighting method is better than a fixed-weighting method across several evaluation metrics. Though the improvement is not statistically significant, we believe that our method has the potential to be further improved when more test collections are available.

Our work makes the following contributions: 1) we propose a novel adaptive weighting method for aggregating and scoring evidence in medical records, 2) we propose and explore several features that are based on semantic similarity between medical concepts for predicting the weights of our adaptive weighting method.

2. RETRIEVAL TASK AND DATA

We use the official test collection of the TREC 2011 & 2012 Medical Records Track [11, 10] for our experiments. The test collection contains 100,866 de-identified medical reports, mainly containing clinical narratives, from the University of Pittsburgh NLP Repository.

The retrieval task[1] is an ad hoc search task for patient visits. A patient *visit* to the hospital usually results in multiple medical reports, meaning there is a 1-to-n relationship between visits and reports.

ID	Topic
107	Patients with ductal carcinoma in situ (DCIS)
118	Adults who received a coronary stent during an admission
109	Women with osteopenia
112	Female patients with breast cancer with mastectomies during admission

Table 1: Example topics of medical records track.

NIST released 81 information needs (or "topics" in TREC terminology) which were designed to require information mainly from the free-text fields (i.e., topics are not answerable solely by the diagnostic codes). Topics are meant to reflect the types of queries that might be used to identify cohorts for comparative effectiveness research [11]. Table 1 lists several TREC topics as examples. The topic specifies the patient's condition, disease, treatment, etc. Relevance judgments for the topics were also developed by TREC assessors based on the pooled results from TREC participants.

3. ADAPTIVE EVIDENCE AGGREGATION

Evidence in a visit can have different forms of distribution. Generally, there are two extreme cases: 1) Strong evidence exists in only one report of the visit; 2) Evidence spreads almost evenly across the majority of reports associated with the visit.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

[1] http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html

Report-based Retrieval
For the first case, we can estimate the relevance of a visit based on its most relevant report. Thus, we use reports as the initial retrieval units (i.e., building an index for reports and applying the retrieval model to each report), and then transform a report ranking into a visit ranking based on the strongest report-level evidence, which is equivalent to using the following report score merging method for ranking visits:

$$\text{score}_\text{R}(V, Q) = \text{MAX}(\text{score}(R_1^V, Q), \text{score}(R_2^V, Q), ...), \quad (1)$$

where R_j^V is a report associated with visit V based on the report-to-visit mapping, $\text{score}(R_j^V, Q)$ is the relevance score of the report with respect to query Q.

Visit-based Retrieval
Eq.1 cannot handle case 2 well. For example, if visit V_1 has strong evidence in multiple reports and visit V_2 has strong evidence in only one report, V_1 and V_2 will have the same relevance score by using Eq.1. To deal with this problem, we aggregate evidence by merging reports from a single visit field by field into a single visit document V, and then performing retrieval from an index of visits.

Like report-based retrieval, this visit-based retrieval has its own disadvantages since it cannot handle case 1 well. For example, if visits V_1 and V_2 both have strong evidence in only one of their reports but V_1 has three times more reports than V_2, the strong evidence in V_1 will be weakened after merging, resulting in V_1 receiving a lower relevance score than V_2.

3.1 A Novel Scoring Function

The comparison of the report-based and visit-based retrievals shows that these two strategies complement each other. Thus, we propose a new query-adaptive scoring function as shown below:

$$\text{score}(V, Q) = \alpha_Q \cdot \text{score}_\text{R}(V, Q) + (1 - \alpha_Q) \cdot \text{score}_\text{V}(V, Q), \quad (2)$$

where $\text{score}_\text{R}(V, Q)$ and $\text{score}_\text{V}(V, Q)$ are the relevance scores of document V from report-based and visit-based retrievals respectively, and α_Q is the query-adaptive coefficient for scoring merging. If we can adjust α_Q appropriately, Eq.2 should be able to deal with all the evidence distribution cases mentioned above.

3.2 Learning Algorithm

In this paper, we propose to adaptively set α_Q with respect to different queries by learning the weight α_Q based on a set of features.

In particular, we can view α_Q as a mixing probability: the probability that the evidence clusters in only one report rather than spreads across multiple reports. Then, assuming the log-odds of that probability can be expressed as a linear combination of feature values, we may write:

$$\log \frac{\alpha_Q}{1 - \alpha_Q} = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \epsilon_Q$$

where β_0 is a model intercept (or bias term), x_i is the value of feature number i, β_i is the weight coefficient of that feature, and ϵ_Q is a slack variable.

This is essentially a logistic regression model[2]. Logistic regression is fit using iteratively reweighted least squares to find the values of the β coefficients that are the best fit to training data. Given feature values and their β coefficients, we can then predict the mixing probability α_Q for new queries.

3.3 Features

We propose 14 features that are possibly related to the evidence distribution in visits, and can be used to predict the weight α_Q in Eq.2. All these features are based on characteristics of the medical concepts contained in the query. We detect these medical concepts using MetaMap [1], a medical NLP tool developed at the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS) Metathesaurus. The concepts are represented by the Concept Unique Identifier (CUI) in UMLS Metathesaurus. Thus, we use Q_C to represent a concept query that is converted from the original text query Q and contains only CUIs. Next, we describe these 14 features in detail:

1. Length of the query
Intuitively, evidence is more likely to resides across reports for long queries. Thus, we use the length of query $|Q|$ as the feature to estimate the evidence distribution. It is defined formally as $|Q| = \sum_{w \in Q} \text{cnt}(w, Q)$, where $c(w, Q)$ is the count of term w in Q.

2. Number of concepts in the query
Similarly, if a query contains more medical concepts, it is more likely to find that the evidence distributes across multiple reports. We define this feature formally as $|Q_C| = \sum_{w \in Q_C} \text{cnt}(w, Q_C)$, where $\text{cnt}(w, Q_C)$ is the count of term w in Q_C. Q_C a better feature than Q because if the query contains a medical concept whose name is very long then Q might not be a good indicator of the evidence distribution.

3. Broad/narrow query concepts
A text query can contain several medical concepts, for each of which the MetaMap program will return 1 to 10 candidates. We hypothesize that a concept with more candidates is less specific, and thus more likely to be a broad concept and appears in multiple reports. Thus, the average number of returned MetaMap candidates for concepts in a query may be a good indicator of evidence distribution. We define this feature as $R_C = \frac{\sum_{w \in Q_C} |\text{Meta}(w)|}{|Q_C|}$, where $|Q_C|$ the original concept query length (i.e., the length before expansion), $|\text{Meta}(w)|$ is the number of concept candidates returned by MetaMap for term w in concept query Q_C.

4. Semantic similarity among query concepts
Intuitively, if Q_C contains concepts that are semantically close, the associated evidence in a visit may also co-occur in a single report. However, if the concepts are semantically distant, the corresponding evidence may tend to distribute across reports. Thus, we use the semantic distance among query concepts to estimate how the evidence distributes.

We use YTEX[3] to measure semantic similarity. Given a pair of UMLS concepts, YTEX can produce knowledge based and distributional based similarity measures. The former uses knowledge sources such as dictionaries, taxonomies,

[2] While logistic regression is often used for 0/1 classification problems, it can also be used when the target variable is a real number between 0 and 1. In this case it is sometimes called a "quasibinomial" model.

[3] http://code.google.com/p/ytex/wiki/SemanticSim_V06

Type	Method	Notation	Name
Knowledge-based	Path-Finding	WUPALMER	Wu & Palmer
		LCH	Leacock & Chodorow
		PATH	Path
		RADA	Rada
	Intrinsic IC based	IC_LIN	Lin
		IC_LCH	Leacock & Chodorow
		IC_PATH	Jiang & Conrath
		IC_RADA	Rada
		JACCARD	Jaccard
		SOKAL	Sokal & Sneath
Distributional-based	Corpus IC based	CIC_LIN	Lin

Table 2: Semantic similarity measures.

and semantic networks, while the latter mainly uses the distribution of concepts within some domain-specific corpus [3].

We use the 11 measures listed in Table 2 as our features. Due to the limited space, we will not describe these features; Garla and Brandt provide a detailed overview [3].

For each query and each specific measure, we take the mean of the semantic similarity scores for all UMLS concept pairs in the query. This averaged semantic similarity score will be the feature score.

4. EXPERIMENTAL SETUP

We use the Indri[4] retrieval system for indexing and retrieving. In particular, we use the Porter stemmer to stem words in both text documents and queries, and use a standard medical stoplist [4] for stopping words in queries only.

Our retrieval model is a linear combination of the Markov random field model (MRF) [8] and a mixture of external collection-based relevance models (MRM) [2] for query expansion. Our collections for expansion are the ClueWeb09 Category B (excluding the Wikipedia pages) corpus, the 2009 Genomics Track corpus, 2012 Medical Subject Headings (MeSH), and the medical records corpus itself. Both report and visit-based retrievals use this system.

Because the focus of this work is to evaluate the adaptive scoring function as shown in 2, we will set the parameters of the MRF and MRM models to some default values. We use the same set of parameter values for both the report and visit-based retrievals. We set the Dirichlet smoothing parameter μ to 2500. For MRF model, we follow Metzler and Croft [8] and set the feature weights $(\lambda_T, \lambda_O, \lambda_U)$ to (0.8, 0.1, 0.1). For MRM model, we take take the top-weighted 10 terms from the top-ranked 50 documents for each expansion collection. More detail about our model is presented in recent work [13].

To evaluate our learning algorithm as described in Section 3, we first obtain the optimal coefficient $\alpha_{Q\text{-opt}}$ for each topic Q by sweeping [0, 1] at a step size of 0.1. Then we conduct leave-one-out cross-validation (LOOCV), in each iteration of which the system predicts α_Q for one new topic based on $\alpha_{Q\text{-opt}}$'s for the other 80 topics. With limited topics available for learning a relatively complex prediction model, using LOOCV can maximize the size of training data we can use in each iteration of the cross-validation, and lead to a better estimate for each feature weight.

We train our systems on MAP. This is because: 1) training on MAP is most commonly used in IR to improve retrieval performance; 2) we find that training on MAP improves the retrieval performance on other evaluation metrics as well while training on other evaluation measures does not improve the overall performance. Thus, MAP will be the primary evaluation measure in this work. In fact, MAP correlates well with other evaluation measures as we will show in the Section 5.

Feature	Significance	Feature	Significance
IC_RADA	0.0112	R_C	0.0654
WUPALMER	0.0299	SOKAL	0.0671
RADA	0.0368	IC_LIN	0.0824
JACCARD	0.0647	IC_PATH	0.0876

Table 3: Features in the pruned set using LOOCV, sorted by their statistical significance scores.

To access the statistical significance of differences in the performance of two systems, we perform one-tailed paired t-test for MAP (since we train systems on MAP). We report scores for MAP, R-precision (Rprec), bpref, and precision at rank 10 (P10).

5. RESULTS AND ANALYSIS

5.1 Feature Selection

To choose a good feature combination, we use a greedy feature elimination approach in which we start with a full set of features and iteratively eliminate exactly one feature at a time that has the greatest negative impact on the retrieval performance until when further removing any feature will degrade the performance.

After the above feature set pruning step, there are 8 features left as shown in Table 3. We further study the importance of each feature by analyzing the prediction model trained in a randomly selected iteration of LOOCV using these 8 features. Based on the statistical significance of each feature as shown in Table 3, we can infer that:

1) All the intrinsic IC based features except IC_LCH are in the pruned feature set, indicating that these types of similarity measures are generally more effective for predicting α_Q than other measures. In fact, the intrinsic IC similarity measure incorporates taxonomical evidence explicitly modeled in ontologies (such as the number of leaves/hyponyms and subsumers), which are not captured by the path-finding based measure. Furthermore, the intrinsic IC similarity measure avoids dependence on the availability of domain corpora, thus is considered more scalable and easily applicable than the distributional-based measure [9].

2) R_C is a good feature though it only uses similarity information about each query concept and its neighbors (rather than other query concepts) in the semantic network.

3) Neither $|Q|$ nor Q_C is in the pruned set, indicating that non-semantic-similarity-related features are generally not useful for estimating the evidence distribution.

4) RADA is a feature that might worth further exploration because both the Path-finding based and the intrinsic IC based RADA features are in the pruned set.

5.2 Adaptive Weighting

Fixed Weighting

We first evaluate the performance of Eq. 2 when α is fixed (i.e., not adaptive). In each iteration of the LOOCV, we obtain the best value setting for α on the 80 training topics by sweeping [0, 1] at a step size of 0.1, and then apply the trained α value to the single testing topic. We show the results in the 'Fixed-weighting' row of Table 4. Note that

[4] http://www.lemurproject.org/indri/

System	**MAP**	**R-prec**	**bpref**	**P10**	Pred. MSE
Visit-based	0.4122	0.422	0.499	0.619	–
Report-based	0.4354^{V}	0.435	0.511	0.607	–
Fixed-weighting	$0.4472^{V,R}$	0.443	0.520	0.631	0.128
Adaptive-weighting	$0.4485^{V,R}$	0.447	0.523	0.642	0.125
Optimal-weighting	$0.4639^{V,R,F,A}$	0.457	0.539	0.656	0.000

Table 4: Performance comparison. A superscript on the MAP score of system X corresponds to the initial of system Y, and indicates statistical significance (p < 0.05) in the MAP difference between X and Y. The last column is the mean square error of the predicted weights. 'Fixed-weighting' corresponds to one of the top-ranked TREC systems as mentioned in Sections 4 and 5.2.

this system is a better version of system udelSUM [13] which is one of the top-ranked 2012 Medical Records track systems.

Optimal Weighting

We also obtain the optimal $\alpha_{Q\text{-opt}}$ for each topic separately by sweeping α from 0 to 1 with a step size of 0.1. Then, we use the $\alpha_{Q\text{-opt}}$'s to compute the best retrieval performance (i.e., an upper-bound) Eq. 2 can possibly achieve, as shown in the 'Optimal-weighting' row of Table 4.

Performance Comparison

Table 4 shows performance comparison of our adaptive merging method with fixed-weighting, optimal-weighting, and two other baselines (report-based retrieval and visit-based retrieval). Our adaptive merging method is better than the fixed weighting method on all the evaluation metrics. The improvement is not statistically significant ($p = 0.191$), possibly because 81 topics may not be enough to train a good prediction model for our adaptive weighting method. In addition, the data are slightly skewed as Figure 1 showing that $\alpha_{Q\text{-opt}} = 1$ or 0.9 on about one third of the topics.

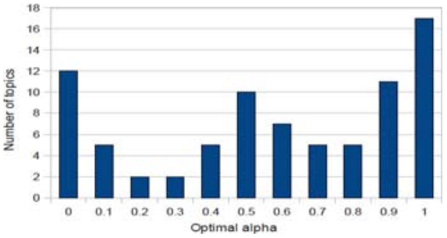

Figure 1: Distribution of topics against $\alpha_{Q\text{-opt}}$.

6. RELATED WORK

Due to the sensitivity of patient data, methods emerging from research on information retrieval for EMR retrieval have not been well explored by academic researchers. Fortunately, the Text REtrieval Conference (TREC) organized the Medical Records track in 2011 & 2012 making a set of real medical records and human judgments of relevance to search queries available to the research community.

Some interesting work have been done using the TREC collection. Limsopatham et al. [5] proposed an effective term representation to handle negated phrases in clinical text. They also incorporated dependence information of the negated terms into the term representation and achieved significant improvement over a baseline system that had no negation handling mechanism.

More recently, Limsopatham et al. [7] proposed an effective representation for EMR retrieval, in which medical records and queries are represented by medical concepts that directly relate to symptom, diagnostic, test, diagnosis, and treatment. We have built on their work, combining a concept representation with text-based retrieval to improve on both and provide a base in which additional medical knowledge can be incorporated easily.

Among more relevant works, Limsopatham et al. [6] explored using the type of medical records for enhancing retrieval performance. They demonstrated that incorporating department level evidence of the medical reports in their extended voting model and federated search model could improve the retrieval effectiveness. Their work opens another interesting direction for exploring evidence distribution and score merging. Zhu and Carterette's system [12] aggregated report-level evidence and visit-level evidence, and achieved significant improvement over a strong baseline.

7. CONCLUSION AND FUTURE WORK

In this paper, we present a medical record search system which is useful for identifying cohorts required in clinical studies. In particular, we propose a query-adaptive weighting method that can dynamically aggregate and score evidence within multiple medical reports. We show by cross-validation that our weighting method is better than a fixed-weighting method across several evaluation metrics. Though the improvement is not statistically significant, we believe that our method has the potential to be further improved by incorporating other useful features or by using advanced prediction models. Furthermore, we explore several informative features for weight prediction. We believe these features might be useful for improving medical IR systems.

8. REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. *Proceedings of AMIA Symposium*, pages 17–21, 2001.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, pages 154–161, 2006.

[3] V. Garla and C. Brandt. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*, 13:261, 2012.

[4] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Health Informatics. Springer, 3rd edition, 2009.

[5] N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis. Exploiting term dependence while handling negation in medical search. In *Proceedings of SIGIR*, pages 1065–1066, 2012.

[6] N. Limsopatham, C. Macdonald, and I. Ounis. Aggregating evidence from hospital departments to improve medical records search. In *Proceedings of ECIR*, 2013.

[7] N. Limsopatham, C. Macdonald, and I. Ounis. A task-specific query and document representation for medical records search. In *Proceedings of ECIR*, 2013.

[8] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, page 472, 2005.

[9] D. Sánchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, 44(5):749 – 759, 2011.

[10] E. M. Voorhees. DRAFT: Overview of the TREC 2012 medical records track. In *TREC*, 2012.

[11] E. M. Voorhees and R. M. Tong. DRAFT: Overview of the TREC 2011 medical records track. In *TREC*, 2011.

[12] D. Zhu and B. Carterette. Combining multi-level evidence for medical record retrieval. In *Proceedings of SHB*, 2012.

[13] D. Zhu and B. Carterette. Exploring evidence aggregation methods and external expansion sources for medical record search. In *Proceedings of TREC*, 2012.