# HSD 2013

## ACM SIGIR Workshop on Health Search & Discovery

### Helping Users & Advancing Medicine

August 1, 2013 – Dublin, Ireland

http://research.microsoft.com/hsd2013


Organizers:

**Ryen White**, Microsoft Research
**Elad Yom-Tov**, Microsoft Research
**Eric Horvitz**, Microsoft Research
**Eugene Agichtein**, Emory University
**Bill Hersh**, Oregon Health & Science University

# Table of Contents

## Oral Presentations

Pg

Learning Attribution Labels for Disorder Mentions in Online Health Forums ..... 03
*Patricia Driscoll, Sharon Lipsky Gorman, and Noémie Elhadad*

HealthTrust: A PhD Dissertation on the Retrieval of Trustworthy Health Social Media ..... 07
*Luis Fernandez-Luque*

Assisted Query Formulation for Multimodal Medical Case-based Retrieval ..... 11
*André Mourão, Flávio Martins, and João Magalhães*

Towards Discovery-Oriented Patient Similarity Search ..... 15
*Haggai Roitman, Sivan Yogev, Yevgenia Tsimerman, and Yardena Peres*

Why Is It Difficult to Detect Outbreaks in Twitter? ..... 19
*Avaré Stewart, Nattiya Kanhabua, Sara Romano, Ernesto Diaz-Aviles, Wolf Siberski, and Wolfgang Nejdl*

## Poster Presentations

A Tool for Monitoring and Analyzing HealthCare Tweets ..... 23
*Ahmed Ali, Walid Magdy, and Stephan Vogel*

Test Collections for Medical Information Retrieval Evaluation ..... 27
*Lorraine Goeuriot, Liadh Kelly, and Gareth J.F. Jones*

Khresmoi Professional: Multilingual Semantic Search for Medical Professionals ..... 31
*Liadh Kelly, Gareth Jones, Allan Hanbury, Lorraine Goeuriot, and Henning Müller*

Query Expansion using open Web-based SKOS Vocabularies ..... 35
*Flávio Martins, Bernhard Haslhofer, and João Magalhães*

Systems for Improving Electronic Health Record Note Comprehension ..... 39
*Balaji Polepalli Ramesh and Hong Yu*

Towards a Gold Standard for the Evaluation of Health Recommender Systems ..... 43
*Martin Wiesner, Monika Pobiruchin, and Daniel Pfeifer*

Towards Intelligent and Socially Oriented Query Recommendation for Electronic Health Records Retrieval ..... 47
*Danny T.Y. Wu, Lei Yang, Qiaozhu Mei, David A. Hanauer, and Kai Zheng*

Clinical Information Retrieval with Split-layer Language Models ..... 51
*Stephen Wu, Dongqing Zhu, William Hersh, and Hongfang Liu*

Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs ..... 55
*Andrew Yates, Nazli Goharian, and Ophir Frieder*

# Learning Attribution Labels for Disorder Mentions in Online Health Forums

Patricia Driscoll
Department of Biomedical Informatics
Columbia University
New York, NY 10032
pdriscoll@gmail.com

Sharon Lipsky Gorman
Department of Biomedical Informatics
Columbia University
New York, NY 10032
srg7002@dbmi.columbia.edu

Noémie Elhadad*
Department of Biomedical Informatics
Columbia University
New York, NY 10032
noemie@dbmi.columbia.edu

## ABSTRACT

Members of online health communities post and discuss valuable information about both their health and their peers' health. When mining the content of health forums, one step that has received little attention thus far is to detect whether a disorder mentioned in a post can be attributed to its author. This information has the potential to enhance tasks such as adverse drug event detection and more generally association rule discovery over health forums. This paper presents a supervised method to learn disorder attribution labels in forum posts.

## Keywords

online health forums, disorder attribution, information extraction, text mining

## 1. INTRODUCTION

As online health communities become increasingly relevant as a source for biomedical knowledge [12, 5, 16, 18], extracting modifiers associated with health term mentions can improve the reliability and performance of text mining activities. Examples of such modifiers include negation (e.g., the disorder "nausea" is negated in the statement "I never had any nausea when I had chemo"), uncertainty (the presence of the disorder "lymphedema" is uncertain in the statement "This could be lymphedema"), and temporality (the disorder "breast cancer" started approximately two years from the post's time in "I was diagnosed with breast cancer two years ago"). In this paper, we focus on the task of attribution of disorder mentions. We define the attribution of a disorder mention to be either `Personal` (i.e., the disorder can be attributed to the author of the post), `Someone Else` (i.e., the author of the post mentioned someone else's disorder), or `General` (i.e., the post mentions a disorder as an abstract entity, not instantiated for any individual).

Earlier work on attribution labeling in general-domain natural language processing has addressed tasks like identifying opinion holders [2, 9] and distinguishing between subjective and objective sentences [14]. In the clinical domain, regular expressions and classification systems have been used to categorize conditions mentioned in patient notes as experienced by a patient or not (e.g., a family member) [7, 17, 10]. Recent work in public health informatics has looked at extracting flu infection-related Twitter data and classifying it into Self- vs. Other-related using Twitter-specific features and part-of-speech templates [11]. Within health forums, because there is much dialogue amongst community members, the prevelance of disorder mentions not attributed to the members themselves can be quite high. We argue disorder attribution is a critical step when learning association rules across community members. We present a novel method of identifying attribution of individual mentions of disorders within a post that relies on lexical and syntactic features and incorporates information gathered from unlabeled data into its training.

## 2. METHODS

The goal of the attribution labeling is as follows. Given a span of text representing a disorder mention, classify its attribution as `Personal`, `Someone Else`, or `General`. We cast the problem as two supervised tasks: (i) classify a disorder mention as `General` vs. `Specific` (where `Specific` means either `Personal` and `Someone Else`); and (ii) classify a disorder mention as `Personal` vs. `Someone Else`. In this section, we describe the dataset and annotation process used for the task, and the features used in the classification tasks.

### 2.1 Data Collection and Annotation

#### *2.1.1 Forum Data*

Following ethical guidelines for processing online patient data [6], the data used in these experiments was collected from `breastcancer.org`, a publicly available breast cancer forum with a large number of participants. Data was collected from the 17 most popular subforums of the site as of 2010, resulting in approximately 26,000 threads and 524,000 posts [8]. A subset of 1,000 posts was selected for manual annotation. The 1,000 posts were chosen from all 17 subforums, and as such represent a wide range of topics,

*Corresponding author. 622 W. 168th Street, New York, NY 10032, USA. (212) 305-0509.


*HSD 2013*, August 1, 2013, Dublin, Ireland

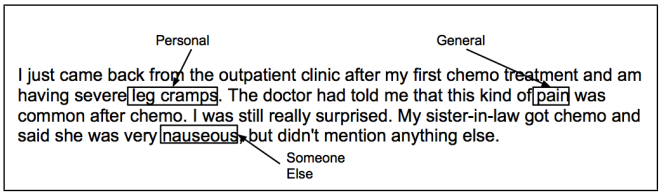



**Figure 1: Disorder mentions annotated with one of three labels: `General`, `Personal`, or `Someone Else`.**

from treatments and decision making to emotional support to discussions about daily life as a cancer patient and survivor.

### 2.1.2 Disorder Span Annotation

A disorder was defined as any disease, symptom, or finding that can be found in the Unified Medical Language System (UMLS) [1] under the Disorder semantic group [3]. Because the language of online health communities contains many lexical variants of UMLS disorders that are not covered by the UMLS, we allowed for other spans of texts to be annotated as disorders whenever necessary, as long as they were consistent with the definition of disorder, as provided by the UMLS. Signs of health and normal test results were included (e.g., "I got a *normal mammogram*", "I am *normal weight*"). We added a constraint that disorders be parts of noun phrases, (e.g., a noun but also standalone adjectives and gerunds). Modifiers were only included in the case of type, test result indicator, parts of official findings (e.g., "*ER/PR +*", "*stage 1 infiltrating ductal carcinoma*") and idiomatic prepositional phrases (e.g., "*shortness of breath*"). Importantly, disorder mentions did not need to refer to specific patients (e.g., "*cancer* clinic", "patients with *angioedema*").

### 2.1.3 Disorder Attributions Annotation

There were three labels for attributions: `General`, `Personal`, and `Someone Else`. The `General` label was used for any disorder mention not tied to a specific individual (e.g., "cancer clinic"). `Personal` refers to a disorder experienced by the author of the post (e.g., "I have breast cancer"). The label `Someone Else` was used for any other disorder mentions tied to specific individuals other than the author of the post (e.g., "My husband had terrible leg pain"), as well as questions posed to other members (e.g., "Is there anyone here who has had leg pain with Tamoxifen?")

The labels `Personal` and `Someone Else` were not used where a disorder mention was negated. For example, all three statements "I don't have lymphedema" and "I don't think I have lymphedema", and "She doesn't have any lymphedema" were labeled `General`. Mentions of disorders with uncertainty were labeled according to their regular categories. For example, "I wonder if I have lymphedema" and "I think she has lymphedema" were annotated as `Personal` and `Someone Else` respectively.

In a small subset of the data, users posted to the forum as a proxy for an identified patient, often a mother or sister who did not have access to the forum. In these cases, to preserve the integrity of the labels, the `Personal` label was used when talking about the identified patient, and the `Someone Else` label applied to anything the poster said about themselves. If the poster referred to having breast cancer herself, the post was not considered a proxy post.

Figure 1 shows an example of annotated disorder mentions in context. The 1,000 posts were annotated by two annotators independently and then adjudicated in the open for the annotators to resolve any differences, both in the disorder spans and the attribution labeling. Overall, 762 of these posts had at least one disorder mention, for a total of 2,845 total annotations: 1,106 `General`, 1,446 `Personal`, and 293 `Someone Else`.

## 2.2 Features

We experimented with three sets of features:

**Bag-of-Word Window** bag of words for any unigram appearing within a 5-word window on either side of the disorder span (the disorder span itself is not part of the feature set).

**Word Clusters** word cluster assignments within the sentence/paragraph/post of disorder mention. The word clusters are obtained from a class-based n-gram language model over the unlabeled forum dataset.

**Lexico-syntactic Features** features derived from a parsed version of the sentence where a disorder mention occurs.

### 2.2.1 Word Clusters

The goal of word clusters is to incorporate a dimension reduction step and a generalization approach on top of the bag-of-word features. The word clusters are obtained through mining a large dataset (in our case, the entire unlabeled dataset, which consists of 523,000 posts and over 10 million words after removing stop words and punctuation marks). We rely on the clustering algorithm proposed in [4]. The method, inspired by language modeling, seeks to group together words that appear in similar local contexts. As such, it is highly relevant for extraction tasks, and has been used for news tagging [13] as well as biomedical named entity recognition [15].

Each cluster starts with a single word; clusters are merged so as to maximize the probability of the language model. The clustering generates a dendogram, where the leaves correspond to the words and intermediate nodes are merged sub-clusters. As such, one can leverage the clusters at different depths, corresponding to different granularities. In this work, we experimented with depths 8, 10, and 12 as in [13]. Only words with frequency over 10 in the overall dataset were considered.

The produced word clusters are useful in identifying and grouping lexical variants of similar concepts, and as such provide higher-level representation over bag of words, all the while incorporating variants from a large dataset (523,000 posts) and mitigating the small size of the annotated dataset (1,000 posts). For instance, one word cluster contained misspellings of different medication terms (e.g., abraxene, abraxane, arimadex, arimedex, arimidex, armidex, carbo, carboplatin, falsodex, faslodex, taxotere and taxoterrible).

For a given disorder mention appearing in a unit (we experimented with units as sentence, paragraph, and post), cluster-based features are computed as the normalized number of words in the unit assigned to particular clusters.

### 2.2.2 Lexico-Syntactic Features

To augment word cluster data for what was expected to be the less context-dependent task of classifying `Personal`

| Cluster Size | Words in Cluster |
|---|---|
| 18 | dad, grandmother, law's, maternal, uncles, aunts, paternal, cousins, father, brother, parents, father's, aunt, passed, cousin, law, dad's, died |
| 5 | ovarian, history, cancers, factors, factor |
| 1 | love |
| 1 | family |
| 98 | y'all, pls, sista, xxxx, teresa, kate, buddy, jax, jodi, carole, alice, ... |
| 12 | wants, recently, figure, concerned, question, worried, mention, wondering, mentioned, line, similar, wonder |
| 18 | hoax, nccn, gls, violate, petition, pi, 1988, mbc, itunes, lop, timtam, sarcoma, pdf, surveillance, $ed, app, launched, seer |
| 27 | brca1, predictive, mutations, engineering, genetics, carriers, mutation, genetic, tripple, inherited, braca, accuracy, 'triple, hx, brca2, genes, predisposition, variations, chemosensitivity, brac, brca, gene, ... |
| 10 | survivor, events, men, alive, mark, survivors, fought, survived, survive, event |
| 8 | ladies, boards, thread, advice, posts, helpful, board, posted |

**Table 1: Ranked most discriminative clusters for distinguishing `Personal` vs. `Someone Else` when restricted to cluster features at the sentence level with cluster tree depth of 10. Direct reference to another person appears to be a key distinguishing feature.**

| Cluster Size | Words in Cluster |
|---|---|
| 8 | findings, data, study, evidence, reported, patients, studies, researchers |
| 33 | lowers, drastically, associated, significantly, reductions, reduce, rate, decrease, reducing, increased, preventing, loading, interferes, decreases, increase, reduction, decreased, increasing, upping, higher, greater, ... |
| 9 | prevent, non, form, addition, certain, avoid, including, known, example |
| 17 | impact, function, protect, improved, ability, lead, affects, lack, blame, immune, kill, causes, hormones, affect, thyroid, bodies, improve |
| 3 | diagnosis, dx, diagnosed |
| 1 | risk |
| 27 | limited, alternatives, carefully, appropriate, individual, names, directly, proper, specifically, setting, various, correctly, adequate, refer, kinds, lacking, familiar, sufficient, vaguely, properly, uses, particular, vividly, procedures, types, specific, base |
| 9 | believe, point, reason, important, sense, making, difference, fact, mean |
| 4 | need, make, want, people |
| 3 | did, got, went |

**Table 2: Ranked most discriminative clusters for distinguishing `General` vs. `Specific` when restricted to cluster features at the sentence level with cluster tree depth of 10.**

vs. `Someone Else`, the annotated dataset was parsed using OpenNLP English Treebank parser. A set of lexico-syntactic features was computed as follows:

- Value of PRP$ in same NP or VP as item
- Subject of clause if PRP (NP inside same S as item)
- Total number of PRP$ of each type in the same S as item
- Total number of PRP of each type in the same S as item
- Yes/no: RB={no, not, n't} in same VP as item
- Yes/no: RB = {no, not, n't} in same S as item

## 3. EXPERIMENTAL RESULTS

Two SVM [1] classifiers were trained and evaluated using 10-fold cross-validation. We constructed test and train splits by randomly assigning items in order to reach roughly balanced folds across labels. The clusters at tree depths 8, 10, and 12 at the sentence, paragraph, and post levels produced 9,464 features.

Tables 3 and 4 shows the performance across each subtask for three systems (along with 95% CI): a baseline model with bag of word features, a model augmented with the word clusters as described above, and a model augmented with parse features. In both subtasks, the full set of features yielded significant improvement in overall accuracy, as well as label-specific precisions and recalls.

Tables 1 and 2 show for informative purposes the top-10 cluster-based features discriminative towards each classification according to a chi-square feature selection step (there was no feature selection step carried out in the SVM) for the cluster-based features when limited to the sentence level with a tree depth of 10. We note that for the `General` vs. `Specific` classification, clusters of words which were typically used in research articles discussed in the forum have a high discriminatory power towards `General`, while words like "did" and "got" are more indicative of a `Specific` disorder mention. For the `Personal` vs. `Someone Else` classification, direct reference to another person was highly discriminative (either through a family reference like "mom" or addressing peers as in "y'all").

As expected, the lexico-syntactic features contributed to the `Personal` vs. `Someone Else` classification, and to a lesser extent the `General` vs. `Specific` classification. Classification without clustering (i.e., baseline+parse) did not yield any improvement over the baseline classification.

[1]We used the SVMlight package.

| | accuracy | precision | recall |
|---|---|---|---|
| baseline | 73.32 ± 3.46 | 66.17 ± 3.56 | 64.65 ± 3.20 |
| +clustering | 77.46 ± 1.15 | 70.84 ± 1.76 | 71.61 ± 1.93 |
| +parse | 77.68 ± 1.58 | 70.81 ± 2.32 | 72.69 ± 2.58 |

**Table 3: `General` vs. `Specific` performance with 95% confidence interval.**

| | accuracy | precision | recall |
|---|---|---|---|
| baseline | 84.01 ± 2.10 | 91.09 ± 1.36 | 89.55 ± 1.91 |
| +clustering | 88.78 ± 1.52 | 93.21 ± 1.46 | 93.36 ± 1.15 |
| +parse | 90.39 ± 1.93 | 93.98 ± 1.54 | 94.53 ± 1.34 |

**Table 4: `Personal` vs. `Someone Else` performance with 95% confidence interval.**

# 4. CONCLUSION

Sophisticated data mining on the vast range of drug, disease, and side effect information contained in online health forum data will require the ability to organize this data by the patients it refers to and separate the general from the specific. As a first step, we classify disorder mentions (including side effects, symptoms, and findings) in two subtasks: `General` vs. `Specific`, and `Personal` vs. `Someone Else`. SVM classifiers augmented with word clusters obtained from unlabeled data and with lexico-syntactic features significantly outperform a bag-of-words SVM baseline. The system's performance indicates attribution can be uncovered from health forum data and is accurate enough to be leveraged for further analysis.

# 5. ACKNOWLEDGMENTS

This work is supported by an NSF award #1027886 (NE) and NCI award R21CA143642-01 (NE). Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the funding organizations. We are grateful to Deb Roy and Jethran Guinness from the MIT Media Lab for providing us with their implementation of the word clustering algorithm.

# 6. REFERENCES

[1] Unified medical language system. `http://www.nlm.nih.gov/research/umls/`.

[2] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, page 2224, 2004.

[3] O. Bodenreider and A. McCray. Exploring semantic groups through visual approaches. *J Biomed Inform*, 36(6):414–432, 2003.

[4] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Comput Linguist*, 18(4):467–479, 1992.

[5] B. W. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. In *Proc AMIA Annu Symp*, page 217, 2011.

[6] G. Eysenbach and J. E. Till. Ethical issues in qualitative research on internet communities. *BMJ*, 323(7321):1103–1105, 2001.

[7] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*, 42(5):839–851, 2009.

[8] M. Jha and N. Elhadad. Cancer stage prediction based on patient online discourse. In *Proc BioNLP Workshop*, pages 64–71, 2010.

[9] S.-M. Kim and E. Hovy. Identifying opinion holders for question answering in opinion texts. In *Proc Workshop on Question Answering in Restricted Domains*, pages 1367–1373, 2005.

[10] Y. Kim, E. Riloff, and S. M. Meystre. Improving classification of medical assertions in clinical notes. In *Proc ACL-HLT*, pages 311–6, 2011.

[11] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proc NAACL-HLT*, pages 789–795, 2013.

[12] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proc BioNLP Workshop*, pages 117–125, 2010.

[13] S. Miller, J. Guinness, and A. Zamanian. Name tagging with word clusters and discriminative training. In *Proc HLT*, pages 337–342, 2004.

[14] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proc EMNLP*, pages 105–112, 2003.

[15] S. K. Saha, S. Sarkar, and P. Mitra. Feature selection techniques for maximum entropy based biomedical named entity recognition. *J Biomed Inform*, 42(5):905–911, 2009.

[16] H. Sampathkumar, B. Luo, and X.-W. Chen. Mining adverse drug side-effects from online medical forums. In *Proc HISB*, pages 150–150, 2012.

[17] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556, 2011.

[18] C. C. Yang, L. Jiang, H. Yang, and X. Tang. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *Proc SIGKDD Workshop on Health Informatics*, 2012.

# HealthTrust: A PhD Dissertation on the Retrieval of Trustworthy Health Social Media

Luis Fernandez-Luque
Norut & University of Tromsø
Tromsø Science Park
Tromsø, Norway
+47 934 21 287
luis.luque@norut.no

## ABSTRACT

There is a global trend towards the use of Internet to find health information. Health consumers nowadays have access to a wide range of online resources spread over traditional websites and social media. Finding trustworthy social media is increasingly challenging, despite the increased effort of health authorities and other trustworthy individuals and organizations to produce high quality social media. Misleading and harmful content, such as promoting anorexia as lifestyle or macabre amputations, is often very popular hindering the visibility of high quality online resources. In this dissertation, I studied the challenges that need to be faced to facilitate the retrieval of trustworthy health social media. This paper summarizes seven sets of studies that have been published in 10 papers over the last six years. In addition, I discuss future research areas for the retrieval of high quality health social media.

## Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; H.3.3 [**Information Search and Retrieval**]: information filtering, Content Analysis and Indexing;

## General Terms

Algorithms, Performance, Design, Economics, Reliability, Experimentation, Human Factors.

## Keywords

eHealth, Telemedicine, Social Media, YouTube, Information Retrieval, Social Networks

## 1. INTRODUCTION

This dissertation is addressing the problem of health social media overload, especially with regards of online videos. The use of social media to disseminate health information is not new since online forums of patients have been used since late 1990s. However, in the recent years there has been an explosion of the amount of health social media. For example, YouTube was created in 2005 and nowadays hundreds of US and European hospitals have published hundreds thousands of health videos [2, 14]. This increase on the rise of online health information is driven by the demand. Most online adults in Europe and USA are using Internet as a source of health information [17, 20], and the trend is also emerging in countries such as China and Brazil [19].

Health consumers can be overwhelmed by the amount of available information, and also by the difficulty of identifying misleading content. The concept of quality content is very complex in the health domain since it includes many aspects (e.g., technical aspects, esthetics, relevance, credibility) [11]. One of the major concerns is the appearance of social media that provides harmful information, such as advice for loosing weight by means of water fasting [18]. That misleading content can be of very good technical quality. It is also common to find videos with "doctors" justifying that vaccination causes autism. What is more, that misleading content is often more popular in social media platforms than content from trustworthy health authorities [15].

General web information retrieval approaches are not specifically designed for the complexity of the health domain. The quality of results retrieved by major search engines has been widely contested, including search of content in social media platforms such as YouTube [1, 15]. Normally, generic web information retrieval systems are trying to accommodate the information needs and curiosity of the more general public. For example, videos with conspiracy theories about vaccination (e.g. vaccination for population control) are highly ranked in search results about vaccination. There are also web information retrieval approaches designed for the health domain, such as the WRAPIN participated by the Health on the Net Foundation [12]. However, those rely mainly on the manual evaluation of the content providers and therefore they face major scalability issues [3].

As a health consumer who actively had to search for online health information for my loved ones, I was thrilled to understand why "good" health social media was very hard to find. As a computer science researcher, I was also thrilled to explore new approaches to find high quality health social media. In order to conduct my PhD, I had to address the following research gaps.

The novelty of health social media: this PhD project started in the year 2007 when most social media platforms were still emerging. For example, YouTube was only two years old. At that time very little was known about the use of YouTube in the health domain. When I published my first paper on YouTube (2008) only 13 papers were published about the subject, nowadays more than 85 papers have been indexed in PubMed (medical research database). Similar lack of knowledge was common in all the aspects of health social media.

The lack of technical studies focusing on finding health social media: there is a wide range of possible technical solutions to facilitate the retrieval of social media (e.g., NLP, Link Analysis, etc.). However, very few studies have been addressing how to apply those techniques to facilitate the retrieval of health social media. Nevertheless, advanced computing techniques have been used to model and monitor health aspects from social media (e.g. detecting flu epidemics from social media streams [6]).

Trust-based metrics for retrieving health social media: in the health domain trustworthiness has been normally referring to the credibility of the messenger in terms of whether he or she is qualified. That approach in the online context has been extrapolated to "quality labels and certificates". Here the research

gap is related to the lack of studies addressing the influence of social network metrics as indicator of the trustworthiness of health social media. Noteworthy, there are some studies showing that online patients' communities are able to filter misleading health information posted within their communities [5].

# 2. RESEARCH QUESTIONS

The overall research question is "How computing techniques can be adapted to find trustworthy health social media?" This broad research question can be subdivided in the following questions. The first three research questions were aiming at increasing the understanding of the problems related to the retrieval of health social media, while the last research question is aiming at evaluating the feasibility of a possible approach:

**RQ1.) Which are the current challenges of finding health social media and videos in particular?**

This research question is aiming at characterizing aspects related to the current use of health social media for the retrieval of trustworthy content. Among other aspects, it includes the study of quality metrics and human factors.

**RQ2.) Which are the technical solutions for modeling and characterizing health social media?**

This research question is addressing the need of an overview of the technical solutions available for modeling health social media, which is a crucial part for the retrieval process. These technical solutions are also constrained by socio-ethical aspects such as privacy.

**RQ3.) How can Social Network Analysis be used to extract information about the characteristics of health social media?**

This research question was defined to explore how social network analysis could be applied to online health communities. The goal is to use social network analysis to extract metrics and characteristics about the social influences, such as trust, among the communities' members.

**RQ4.) Can trust-based metrics improve the retrieval of social videos about diabetes?**

This research question explored how a trust-based metric derived from the social network analysis can be used to enhance the retrieval of health social media. As case study, I tested the metric to retrieve diabetes videos from YouTube.

# 3. RESEARCH DESIGN

The research carried out in this dissertation had a major multidisciplinary component. Since the area of health social media is immature and continuously evolving, it was nearly impossible to gather all the background knowledge without setting up multiple studies. These studies required in many cases the collaboration of experts from different disciplines (e.g. anthropologists, public health experts, psychologist, medical doctors, data mining experts, patients' advocates, etc.). Not surprisingly, there is a wide range of research methodologies applied in the studies presented in this dissertation, which are discussed in the dissertation papers [4, 7–11, 13, 16, 21, 23].

# 4. Studies and Key Results

## 4.1 RQ1.): Characterization of the use of health social videos.

The study of the characteristics of the metadata of health social media is crucial for any research aiming at improving the retrieval process. The first study (**RQ1.Study 1**) looked at the disclosure of private health information in YouTube videos about multiple sclerosis [7], that information could potentially be used for modeling users and videos. That study was complemented with another experiment with regards of the use of medical terminology in YouTube´s surgery videos [16].

Another major challenge addressed in this research question was the lack of consensus with regards of the definition of quality within health social media and videos in particular. In the study **RQ1.Study 2**, we performed a systematic review of published papers about online health videos in order to identify the most commonly mentioned quality features [11].

The understanding of human factors involved in the creation and dissemination of social media is essential. For that reason, we performed the **RQ1.Study 3** consisting the analysis of videos from patients who are sharing about their disease in YouTube [13]. The authors of those videos were asked to report their motivations. I analyzed those videos with the collaboration of several psychologist researchers. Finally, the **RQ1.Study 4** was aiming at comparing the features of misleading videos promoting anorexia against informative videos [21].

### *4.1.1 Key results*

**RQ1.S1**: the 20% of comments on Multiple Sclerosis videos contained personal health information

**RQ1.S1**: 5% of the tags of surgery videos on YouTube were exact match to standardized medical thesauri.

**RQ1**.**S2**: we identified 17 quality features of online health videos. The most common criteria were related to the credibility of the content.

**RQ1.S3**: patients were mainly motivated to start publishing videos online due to lack of patient-oriented content. The main motivation to keep posting was the social support related to belonging to a community.

**RQ1**.**S4**: informative videos about anorexia had more views and were more prevalent than misleading content. However, misleading videos had better ratings and higher viewer engagement (e.g., more favorites, more ratings per view).

## 4.2 RQ2.): Extracting information from Health Social Media

A very wide range of technologies can be used to extract information from health social networks and eventually be used for modeling content and users. The **RQ2.Study 1** was designed as multidisciplinary review study aimed at identifying different technical solutions for extracting information from health social network [10], including also socio-ethical aspects.

### *4.2.1 Key Results*

**RQ2.S1**: there are many technical approaches that can be used to model users and content within health social networks. These include artificial vision, NLP, link analysis, etc. Many of those techniques will face major socio-ethical challenges due to privacy concerns (e.g. modeling health risk behaviors). The use of social network analysis seems to be the most promising way to extract information about trustworthiness within health social networks without major privacy concerns.

### *4.2.2 Health Social Network Analysis*

Based on the results of the RQ2, I decided to further explore how social network analysis could be used to extract characteristics

from online social networks, especially those related to trust. The **RQ3.Study 1** aimed at studying the network dynamics and influences within antagonist social networks related to anorexia, including sub-communities promoting information and misinformation [23]. This study was complemented with the **RQ3.Study 2** aimed at studying the characteristics of diabetes communities. In this study the focus was on the structure of the networks and the features of the most trusted members [4].

#### *4.2.3 Key Results*

**RQ3.S1**: The social network analysis of communities promoting anorexia and their informative counterpart showed that those networks are clearly separated but intermingle. The best parameters to predict the pertinence to misleading or informative sub-communities were the social ties (e.g. contacts). Textual content (e.g. tags) had limited predictive value since both communities used in many cases same tags.

**RQ3.S2**: In the study of the structure of several online communities about diabetes, we found a similar network structure. The combination of social network analysis with the available health and demographic information of the members showed that the most experienced diabetes (i.e. time from diagnosis) users were the top influencers.

### 4.3 RQ4.): Trust-based retrieval of diabetes videos

I decided to explore if using a metric related to the trustworthiness of the content provider within a health community could be used to improve the retrieval of diabetes videos. In the **RQ3.Study 1** [8, 9]**,** we extracted the authoritative scores (using HITS) of YouTube's members publishing videos about diabetes. The authorities' scores were integrated into a metric called HealthTrust. With the help of diabetes patients and professionals, I compared the search results for diabetes related videos ranked by YouTube's relevance against another ranking based on HealthTrust.

#### *4.3.1 Key Results*

**RQ4.S1**: the use of the metric HealthTrust to rank search results about diabetes videos performed significantly better than using the relevance ranking of YouTube. HealthTrust performed better by avoiding the most misleading videos (e.g. gruesome amputations) that were very popular on YouTube but could have been shocking for most patients.

## 5. DISCUSSION

The use of social networks metrics to enhance the retrieval of trustworthiness in social media appears to be effective. This is not very surprising since research shows that online communities of patients can be very effective at filtering misleading information [5]. However, this approach will face major challenges related to the identification of sub-communities promoting misleading information. In the feasibility evaluation of HealthTrust, we clearly identified that there is room for improvement of the currently available information retrieval tools, since they are not correctly satisfying the information needs of users on need of trustworthy health information.

It also appears that in many cases popularity metrics can promote misleading health information. In the study of pro-anorexia videos it was found that popularity is higher in harmful videos. In the study of HealthTrust, it was also found that gruesome videos were highly ranked by YouTube, and again very popular in terms of views. The success of HealthTrust appears to have been caused by anchoring the social network analysis to the health community.

Another finding of this dissertation is a mismatch between the quality of informative health social media and the trustworthiness of its content. Misleading or harmful content in many cases had better non-health-related quality features (e.g., engagement with viewership, descriptions, metadata, and esthetics). There is a need for better health communication strategies adapted to the nature of social media, as it is already addressed by some public health authorities [22].

### 5.1 Limitations

There are many small limitations in each of the studies presented in this dissertation. I present here the major limitations (see published papers for detailed description of the limitations). The characterization of the use of health social media is still a work-in-progress and will require a major research effort in the following years. Our knowledge of the driving forces is still very limited and this dissertation only covers a small part of the phenomena. There are many studies warning about the quality issues of finding health social media, but very few studies focus on proving computing solutions to that challenge.

The studies presented in this dissertation are addressing a wide range of health areas: anorexia, multiple sclerosis, diabetes and surgery. Since each health problem is completely different we need to be very careful not to draw general conclusions. Another limitation of this dissertation is the lack of studies addressing the improvement of health outcomes. Future research needs to be conducted to prove if better access to health social media can improve health.

The main contribution, the HealthTrust metric for improving retrieval of diabetes videos, was tested using a controlled setting were the patients and professionals were asked to rate the research results of several queries selected beforehand. That evaluation is not the optimal one, which would have required the development of a web portal, so subjects can search in real life settings.

### 5.2 Future Work

The next step in this research effort will be the integration of the metric HealthTrust in a real web portal in order to better evaluate the performance of the algorithm. That research is already on going but preliminary results are not yet available.

The study of how misleading health communities have high popularity and visibility in social media is a very promising area for future research. That research will not only help to filter out that misleading content, but also provide guidelines to health authorities that are using social media for health promotion. The societal implications of that research are enormous since misleading information enhanced by social media is already affecting the vaccination rates in many countries, which is ultimately leading to increased mortality in children.

The metric HealthTrust was only evaluated in the case of diabetes where there are not clear sub-communities promoting misleading health information. The next step will be to design trust-based metrics that take into consideration the complete health communities ecosystem. In addition, it will be possible to explore the personalization of trust-based metrics based on multi-level health interests of the users (e.g. cardiac nurse who is the mother of a child with diabetes).

## 6. ACKNOWLEDGMENTS

This project has been supported by the Tromsø Telemedicine Laboratory (co-funded by the Research Council of Norway, project 174934) and the project HealthTrust (Tromsø

Forskningsstiftelse).This project would have not been possible without the help of my supervisor Randi Karlsen, co-supervisors, co-authors and others.

## 7. REFERENCES


[1] Backinger, C.L. et al. 2011. YouTube as a source of quitting smoking information. *Tobacco control*. 20, 2 (Mar. 2011), 119–22.

[2] Van de Belt, T.H. et al. 2012. Use of social media by Western European hospitals: longitudinal study. *Journal of medical Internet research*. 14, 3 (Jan. 2012), e61.

[3] Bernstam, E. V et al. 2008. Commonly cited website quality criteria are not effective at identifying inaccurate online information about breast cancer. *Cancer*. 112, 6 (Mar. 2008), 1206–13.

[4] Chomutare, T. et al. 2013. Inferring Community Structure in Healthcare Forums. An Empirical Study. *Methods of information in medicine*. 52, 2 (Feb. 2013).

[5] Esquivel, A. et al. 2006. Accuracy and self correction of information received from an internet breast cancer list: content analysis. *BMJ (Clinical research ed.)*. 332, 7547 (Apr. 2006), 939–42.

[6] Eysenbach, G. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research*. JOURNAL MEDICAL INTERNET RESEARCH.

[7] Fernandez-Luque, L. et al. 2009. An analysis of personal medical information disclosed in YouTube videos created by patients with multiple sclerosis. *Studies in health technology and informatics*. 150, (Jan. 2009), 292–6.

[8] Fernandez-Luque, L. et al. 2012. HealthTrust: A Social Network Approach for Retrieving Online Health Videos. *Journal of Medical Internet Research*. 14, 1 (Jan. 2012).

[9] Fernandez-Luque, L. et al. 2011. HealthTrust: trust-based retrieval of YouTube's diabetes channels. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* (New York, New York, USA, Oct. 2011), 1917.

[10] Fernandez-Luque, L. et al. 2011. Review of extracting information from the Social Web for health personalization. *Journal of medical Internet research*. 13, 1 (Jan. 2011), e15.

[11] Gabarron, E. et al. 2013. Identifying Measures Used for Assessing Quality of YouTube Videos with Patient Health Information: A Review of Current Literature. *Interactive Journal of Medical Research*. 2, 1 (Feb. 2013), e6.

[12] Gaudinat, A. et al. 2004. WRAPIN: new generation health search engine using UMLS knowledge sources for MeSH term extraction from health documentation. *Medinfo*.

[13] Gómez-Zúñiga, B. et al. 2012. ePatients on YouTube: Analysis of Four Experiences From the Patients' Perspective. *Medicine 2.0*. 1, 1 (Apr. 2012), e1.

[14] Hospital Social Network List: 2010. *http://ebennett.org/hsnl/*. Accessed: 2011-11-09.

[15] Keelan, J. et al. 2007. YouTube as a source of information on immunization: a content analysis. *JAMA : the journal of the American Medical Association*. 298, 21 (Dec. 2007), 2482–4.

[16] Konstantinidis, S. et al. 2013. The Role of Taxonomies in Social Media and the Semantic Web for Health Education. A Study of SNOMED CT Terms in YouTube Health Video Tags. *Methods of information in medicine*. 52, 2 (Feb. 2013), 168–179.

[17] Kummervold, P.E. et al. 2008. eHealth trends in Europe 2005-2007: a population-based survey. *Journal of medical Internet research*. 10, 4 (Jan. 2008), e42.

[18] Lau, A. et al. 2012. Social media in health–what are the safety concerns for health consumers? *Health Information Management Journal*. 42, 2 (2012), 30–35.

[19] McDaid, D. and Park, A. 2011. *Online Health: Untangling the Web*.

[20] Susannah Fox, S.J. 2009. The Social Life of Health Information. Pew Internet. American Life Project.

[21] Syed-Abdul, S. et al. 2013. Misleading Health-Related Information Promoted Through Video-Based Social Media: Anorexia on YouTube. *Journal of medical Internet research*. 15, 2 (Jan. 2013), e30.

[22] The Health Communicators Social Media Toolkit: *http://www.cdc.gov/socialmedia/Tools/guidelines/pdf/SocialMediaToolkit_BM.pdf*. Accessed: 2012-04-06.

[23] Yom-Tov, E. et al. 2012. Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes. *Journal of medical Internet research*. 14, 6 (Jan. 2012), e151.

# Assisted query formulation for multimodal medical case-based retrieval

André Mourão
a.mourao@campus.fct.unl.pt

Flávio Martins
flaviomartins@acm.org

João Magalhães
jm.magalhaes@fct.unl.pt

Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Caparica, Portugal

## ABSTRACT

Medical information retrieval systems support health care experts in diagnostic and treatment decisions through the management of large amounts of clinical data. However, the ever growing data produced in medical environments and the proficiency of non-professional users pose several challenges to a retrieval system.

In this paper, we propose a medical retrieval system, supporting semantic multimodal queries for medical case-based search. The system explores many of the commodities available in commercial search engines and provides the user with key tools to support medical information discovery: multimodal queries and semantic suggestions of medical terms. It is build upon state-of-the-art information retrieval and data fusion techniques. We also propose a new data-fusion technique, which we called Inverted Squared Rank (ISR), to better deal with the combination of ranked lists from various heterogeneous systems: similar to Reciprocal Rank Fusion approaches like RR [10] and RRF [3]. The proposed rank fusion method outperforms RR and RRF in most measures and is particularly better on the top 10 results.

The system is at `http://medical.novasearch.org/`

## Keywords

assisted query, search interfaces, multimodal fusion, multimodal medical retrieval

## 1. INTRODUCTION

Search engines are often a key tool for both healthcare professionals and laypeople when investigating medical cases. Case-based retrieval systems can support healthcare users in many ways - suggesting new publications, exploring similar symptoms/conditions, confirming a diagnosis, etc.

We propose a search engine focused on usability and usefulness, not only for health professionals but also accessible to laypeople. At the heart of a case-base retrieval system is the support for rich queries with heterogeneous data. Textual queries can include a long descriptions of the patient condition and images often provide additional information that is difficult to convey in textual queries. For instance, medical images capture the actual exams (e.g. x-ray, MRI) providing exact visual information about the patient (e.g. position of a mass on an MRI). Thus, we designed the system in a flexible way to support multiple data-fusion techniques (e.g. CombMNZ, RR, RRF, CombSUM).

The proposed system also provides an intuitive and simplified way of accessing large medical knowledge bases. It identifies medical terms in real-time and suggests related terms based on medical ontologies. This provides a glimpse of related conditions/diagnostics which can assist users in the formulation of a more targeted query.

In general, the system combines the simplicity of web search engines (text queries, semantic autocomplete, and the general look and feel) with automatic query expansion and image query using simple drag and drop.

In this article we present the framework, focusing on the fusion techniques and the user interface for case-based search. The underlying search framework was evaluated on the case-based retrieval task of the ImageCLEF 2013 medical dataset[1].

## 2. RELATED WORK

Several systems designed for medical retrieval (textual or visual) are available online. The MedGIFT group [5] designed two search engine interfaces to demonstrate their work in medical retrieval: a text based case retrieval search engine[2] and a visual medical image search[3]. The visual medical image interface allows the upload of query images and searching for similar images to the ones in the articles found. Although these two systems work well in their domains, either text or images, they are independent. For instance, it is not possible to search for images using a text query or combine image and text in the query.

An example of a content based image retrieval (CBIR) system is img(Anaktisi)[4] [9]. It was created to demonstrate the CEDD and FCTH image features [2] for image retrieval in multiple datasets. It includes the IRMA medical dataset

[1] `http://www.imageclef.org/2013/medical`
[2] `http://fast.hevs.ch:8080/MedSearch/faces/Search.jsp`
[3] `http://fast.hevs.ch:8080/MedSearch/faces/ImageSearch.jsp`
[4] `http://orpheus.ee.duth.gr/anaktisi/`


*HSD 2013*, August 1, 2013, Dublin, Ireland.


**Figure 1: Search interface**

and allows searching using corpus images as queries.

In our approach, images and text are completely integrated and all combinations are possible. This means, it allows searching for cases using only images, search for images using only text or use both text and images in the same query for either result type.

Outside the medical domain, we find the MMretrieval search engine[5] [8], a multimodal and multilingual search engine for retrieval of Wikipedia images. The system searches for each modality separately (including text in different languages) and applies a variety of late fusion algorithms to combine the results.

Our search engine relies on the fusion of multiple modalities (text and images). Thus, a fusion technique must be applied. Early fusion relies on combining the features from multiple modalities before searching in a common index. Late fusion relies on searching each modality index separately and combining the results in a final step. There are various data fusion methods such as Condorcet [6], CombSUM and its variants [1] (score-based), and Reciprocal Rank approaches like RR [10] and RRF [3] (rank-based) that are among the most applied but there is no clear off-the-shelf solution for all search tasks and modalities. Rank based approaches are gaining momentum, and we implemented a variant of combMNZ and RR that shows potential for multimodal combination on search engines.

We based our image retrieval on features that obtained good results on previous editions of ImageCLEF [7]. We included CEDD and FCTH (texture and color), Local Binary Pattern histograms (contours) and color histograms (color) for retrieval. Textual retrieval is based on Apache Lucene[6], with BM25L as the retrieval function.

Query expansion is useful to increase IR systems performance, making queries match more relevant documents that might not contain the exact query terms entered. Automatic query expansion (AQE) adds terms to the query without user intervention. This is already being performed on major commercial search engines. Most of the times, the expanded terms are synonyms or highly related terms and the user does not receive any feedback on the expanded terms. Interactive query expansion (IQE) gives the user the power to decide what terms are expanded however it is often an interface offered after the initial query at the cost of a more complicated interaction. Our approach is a mixture of IQE and AQE. The user can visualize what terms will be added to the query and opt-out expansion if incorrect or not desirable.

## 3. MEDICAL QUERY FORMULATION

Nova MedSearch is a multimodal (text and image) medical search engine that can retrieve either similar images or related medical cases. These tasks are from the medical ImageCLEF 2013 evaluation campaign. The results are displayed in an ranked list with basic information (e.g. title, keywords, images (if available)) and a link to the corresponding article details. The interface in Figure 1 takes into account both the relevancy of the images and text similarity.

### 3.1 Multi-part queries

Our interface aims at simplifying the inclusion of images and text data in the medical query (a screenshot of the different components of the interface is in Figure 1). For instance, we add support for drag-and-drop functionality for custom medical image queries. The free text query box allows entering a textual description of the patient, and the system automatically expands a recognized term into its related terms. In the example, we see the terms that are related to the search term "spiral computed tomography". The search results contain a visual presentation of the submitted query and the retrieved examples. In addition to general article information (title with link to full article, authors and abstract), we also display the images of the article that are most related to the query images.

### 3.2 Assisted query expansion

The main novelty of the search interface is the assisted query semantic-expansion. Since medical terminology is part

[5] http://www.mmretrieval.net/
[6] http://lucene.apache.org/

of natural language, terms are not unique, and multiple definitions of the same symptom/medication/disease are available. For example, our system returns "acetylsalicylic acid" and "2-(acetyloxy)benzoic acid" as terms related to "aspirin".

We implemented a guided query expansion system that interactively provides auto-complete suggestions and expansion feedback sourced from a SKOS version of the MeSH indexing terms. Medical SKOS provides domain specific expert knowledge regarding the relationships between terms. We decided to use a SKOS version of MeSH to provide two functionalities:

- word based auto-completion with terms
- automatic term expansion with semantically related terms

The process works as follows:

1. when the user starts typing a word, a dropdown box appears with the terms that match the query;
2. if the user selects a term from the list, the browser retrieves the synonyms from our framework and adds them to the query implicitly;
3. the user can then see the expanded terms by putting the mouse over the words. The user can opt-out the suggestions by clicking the ⊗ mark.

Since our system uses a SKOS representation for the terms expansion process, we can also support the SNOMED Medical ontology.

# 4. SEARCH-RESULTS FUSION

In this section we shall describe the search-results fusion methods that combine the rank from the multimodal information sources.

## 4.1 Text retrieval

The text is indexed using Lucene and the BM25L retrieval function is used. The indexed fields depend on the task. For image retrieval, we achieved good results indexing and searching only on the title, abstract and image captions. For case retrieval, we searched on the full document (title, abstract, chapters and captions).

## 4.2 Image retrieval

For image retrieval, we extracted a set of features that are known to be effective in medical images retrieval (CEDD, FCTH, Local Binary Pattern histograms and color histograms - see the related work section). The features of all images in the corpus are stored in a fast $L_2$ index. The image retrieval results are sorted by their similarity, with the score being the $L_2$ distances between the query image and the result images. For case-base retrieval, an additional step must be performed: the image id (IRI), must be converted into a document id (DOI) (Figure 2 (a)) and the duplicate results must be merged to have an unique document list (Figure 2 (b)). More details are present in section 4.3.

## 4.3 Fusion

Result fusion aims at combining ranked lists from multiple sources into a single combined ranked list. Consider these

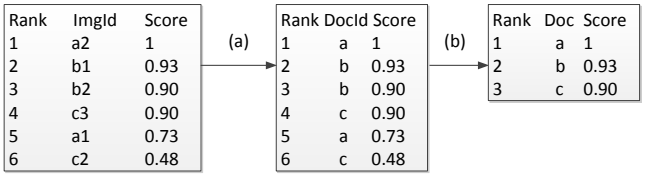



**Figure 2: Case based retrieval step. (a) get document id (DOI) from image id (IRI); (b) combine multiple document results into one (unique) document list. The example uses CombMAX in fusion because it is easier to visualize.**

two use cases: combine the results from queries with multiple images and combine the results from text and images queries. Query images can have different modalities (e.g. x-rays, PET scans, CT scans) and represent the same concept in multiple ways (e.g. hepatolenticular degeneration images can be represented by a photo of an eye or by a light microscopy of the affected cells). Thus, we took a late fusion approach, combining the results from multiple image queries into a single image result list. We found that late fusion of the results was also useful for heterogeneous queries (e.g text only, single image, text and 3 images), as the combination of the image and text search can be ignored if the query does not contain images.

There are two main approaches for late fusion: score based and rank based. Score based approaches (CombSUM, CombMAX and CombMNZ) combine the normalized scores given by the individual searches (e.g. image search, textual search) as a basis to the create the new ranked list. The studied variant that achieves the best performance [1] is CombMNZ, but ranked based fusion is gaining momentum, and can outperform score based fusion under most conditions [3, 4]. For each document $i$, the score after fusion can be computed as:

$$\text{combSUM}(i) = \sum_{k=1}^{N(i)} S_k(i), \tag{1}$$

$$\text{combMAX}(i) = \max(S), \forall S \subset D_i, \tag{2}$$

$$\text{combMNZ}(i) = N(i) \times \text{combSUM}(i), \tag{3}$$

where $S_k(i)$ is the score of the $i$ document on the $k$ result list.

$N(i)$ refers to the number of times a document appears on a results list. A result list $k$ does not contain all documents. Documents with a zero score or a very high rank can be safely ignored. Thus, $N(i)$ varies between 0 (the document $i$ does not appear on any list) and the total number of results list (the document $i$ appears on all lists). For example, in our experiments, there are two results lists: one for image search and other for textual search, limited to 1000 results each.

Rank based fusion methods consider the position of each document in each one of the individual ranks. Reciprocal Rank and Reciprocal Rank Fusion are the two methods we evaluated:

$$\text{RR}(i) = \sum_{k=1}^{N(i)} \frac{1}{R_k(i)}, \tag{4}$$

$$\mathrm{RRF}(i) = \sum_{k=1}^{N(i)} \frac{1}{l + R_k(i)}, \text{with } l = 60. \quad (5)$$

where $R_k(i)$ is the rank of document $i$ on the $k$ rank.

After analyzing both score and rank based approaches, we combined elements from both to improve precision. Inverted Squared Rank (ISR) combines the inverse rank approaches of RR and RRF (using the squared rank to improve precision at top results) with the frequency component of combMNZ (results that appear on multiple lists are boosted):

$$\mathrm{ISR}(i) = N(i) \times \sum_{k=1}^{N(i)} \frac{1}{R_k(i)^2}. \quad (6)$$

# 5. EXPERIMENTS

To assess the proposed methods, we tested the search-results fusion on the Medical case-based search task of the ImageCLEF 2013 evaluation campaign.

## 5.1 Dataset

Our dataset is composed of the data released for medical ImageCLEF 2013. It is a subset of over 70,000 PubMed articles with over 300,000 images. Each article is identified with an unique identifier (DOI) and is divided into title, abstract, chapters and image captions. All images on the dataset have an unique identifier (IRI) and can be associated with the corresponding article and caption.

## 5.2 Results

We compared the performance of the fusion algorithms using the best textual and visual runs. Our methodology was as following: for all (36) multimodal queries present in the ImageCLEF medical 2013, we searched text and images separately and combined our image and text runs using multiple fusion algorithms. Performance was evaluated using trec_eval and the relevance judgments provided.

**Table 1: Fusion comparison for the medical ImageCLEF case based queries**

| Comb | MAP | GM-MAP | bpref | P@10 |
|---|---|---|---|---|
| **ISR** | **0.1608** | 0.0779 | **0.14** | **0.1800** |
| RRF | 0.1597 | **0.0787** | 0.13 | 0.1571 |
| RR | 0.1582 | 0.0779 | **0.14** | 0.1771 |
| combSUM | 0.0804 | 0.0039 | 0.09 | 0.1429 |
| combMNZ | 0.0794 | 0.0035 | 0.08 | 0.1371 |
| combMAX | 0.0292 | 0.0013 | 0.03 | 0.0457 |

With our data, rank-based approaches outperformed score based approaches by a factor of 2. One of the reasons is the differences between the scoring of the text and images. Even though both visual and text scores have the same normalization, the interval [0...1], the distribution of the results in the score space is different. Rank based approaches can handle multi-modality better, because the scores are not used.

Regarding the differences between RR, RRF and ISR: ISR performed better in our experiments in most of the measures, with a significant performance boost on P@10. This metric is particularly important for search engines, because most users won't browse beyond the first page of results ( 10 first). The polynomial component promotes top ranking results to the top of the list, offering a better user experience.

# 6. CONCLUSIONS

Our system combines a powerful framework based on state-of-the-art image and text processing algorithms with a simple yet powerful multimodal search interface to provide a valuable tool to retrieve medical data. In addition to the interface, we introduced ISR, a variant of RR and RRF aimed at increasing relevance of the results at the top of the list. We believe that it will help users to get relevant information, reducing frustration.

The system is still a work in progress. We are planing on testing the system with health care professionals to test usability and improve it.

# 7. REFERENCES


[1] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, 31(3):431–448, May 1995.

[2] S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis, and N. Papamarkos. Accurate Image Retrieval Based on Compact Composite Descriptors and Relevance Feedback Information. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 24(2):207 – 244, 2010.

[3] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR '09*, pages 758–759, New York, NY, USA, 2009. ACM.

[4] D. Frank Hsu and I. Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retr.*, 8(3):449–480, May 2005.

[5] MedGIFT Group. medSearch - Medical search engine by HES-SO Valais, 2009.

[6] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 538–548, New York, NY, USA, 2002. ACM.

[7] H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF 2012 working notes*, 2012.

[8] K. Zagoris, A. Arampatzis, and S. A. Chatzichristofis. www.mmretrieval.net: a multimodal search engine. In *Proceedings of the Third International Conference on SImilarity Search and APplications*, SISAP '10, pages 117–118, New York, NY, USA, 2010. ACM.

[9] K. Zagoris, S. A. Chatzichristos, N. Papamarkos, and Y. S. Boutalis. img(Anaktisi): A Web Content Based Image Retrieval System. In *2009 Second International Workshop on Similarity Search and Applications*, pages 154–155. IEEE, Aug. 2009.

[10] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, and L. Zhao. Expansion-based technologies in finding relevant and new information: Thu trec2002 novelty track experiments. In *the Proceedings of the Eleventh Text Retrieval Conference (TREC*, pages 586–590, 2002.

# Towards Discovery-Oriented Patient Similarity Search

Haggai Roitman, Sivan Yogev, Yevgenia Tsimerman, Yardena Peres
IBM Research - Haifa
Haifa 31905, Israel
{roitman,sivany,yevgenia,peres}@il.ibm.com

## ABSTRACT

In this paper we address the Patient Similarity task. We provide an overview of existing solutions for this task and shortly discuss their pros and cons. We then propose a solution for patient similarity search, implemented using a novel query language for semi-structured entity-relationship data. Using our intuitive solution, the patient similarity task can be reduced into a patient discovery problem that can be efficiently implemented.

## Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical information systems; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Design

## Keywords

Social-Medical Discovery, Patient Similarity, IBM MedICS

## 1. INTRODUCTION

Social media technologies for the healthcare domain have recently attracted a lot of attention, resulting in the emergence of new socially flavored medical services. Utilizing the data openness and sharing through social channels [3], such services now offer new types of discovery options; to name a few are online PHR services such as Microsoft HealthVault[1] that allows users to manage their medical records; online social-medical communities such as Patients-Like-Me [2] and Cure-Together[2] that allow patients to discover other patients who share similar medical characteristics, such as similar disorders or symptoms; and online discovery services such as TrialX[3] for clinical trials and Medify[4] for medical treatments.

One of the most basic social-medical discovery tasks is the task of *Patient Similarity* (also known as the *Patients-Like-This* task), roughly defined as follows: *given a patient in mind as an input, return a list of patients ranked by their similarity to the patient* [6]. Patient similarity can be utilized for implementing various social-medical services, e.g., recommending treatments [4], discovering relevant clinical trials [4, 1], etc. The similarity between two patients can be measured by various metrics and usually depends on the specific sub-medical domain in mind [6]; e.g., searching for similar diabetic patients would require focusing on a different set of similarity features than searching for patients with similar hematologic disorders.

Existing solutions for patient similarity can be classified into two main types. The first type of solutions, commonly used today, use machine learning (ML) techniques [6, 2, 11]. The main advantage of such solutions is that they can be used for weighing patient similarity features and metrics (e.g., using regression models [6], SVM [2], ensemble learning [11], etc).

In spite of the elegancy of ML solutions, they may still impose some expensive computational limitations, mostly due to scalability issues. This in turn usually results in solutions that can only be run in an offline manner and in batches. In addition, ML solutions may require continuous model maintenance every time the patient data repository is updated. Therefore, while these solutions might be well suited for analysis purposes, they can not be efficiently used for online discovery (search).

The second type of solutions are based on more "traditional" information retrieval (IR) techniques. Compared to the ML solutions, IR solutions are easy to implement and explain, can be efficiently executed online, and do not require expensive maintenance. Even though, to date, most IR based social-medical discovery solutions provide very simple search interfaces over social-medical data [7]. Such solutions are usually very limited, supporting only keyword search with some basic categorical search over few social or medical facets such as patient demographics, treatments, symptoms, etc (e.g., Patients-Like-Me [4]).

[1] `http://www.microsoft.com/en-us/healthvault`
[2] `http://curetogether.com`
[3] `http://trialx.com`
[4] `http://medify.com`


*HSD 2013* August 1, 2013, Dublin, Ireland.

The more general problem of entity similarity search has received a lot of attention lately by the IR community, and several new solutions have been proposed (e.g., [5, 10]). Such solutions usually represent entities (e.g., patients, medications, allergies) and their relationships (e.g., consumed by, has allergy, etc) using an Entity-Relationship Graph (ERG). Similar entities are then discovered using various random-walk methods on the ERG graph [5, 10]. Yet, similarly to the ML solutions, these solutions still require either an expensive or offline computation every time the ERG graph may change, and therefore, they do not scale well and cannot be easily used for social-medical discovery.

In some earlier works [7, 9, 13, 12], we have introduced the IBM Medical Information and Care System (IBM MedICS) social-medical discovery service that supports unified search over social-medical data with rich data exploration capabilities. IBM MedICS fuses data from various social and medical sources, e.g., PHR and EMR data and social data such as the list of patient's physicians, family members, friends, similar patients, etc. Both social and medical data are represented by entities and their relationships [7] and searched over using a novel entity search and data exploration technology [13], while preserving the patients' privacy [8]. In this paper we discuss the details of IBM MedICS patient similarity search solution. Patient similarity search is implemented by utilizing a novel query language that can be used for searching any semi-structured social-medical entity-relationship data [12, 13]. The advantage of such a solution is that it enjoys the intuitiveness and efficiency of IR solutions, while having enough expressive power to fully utilize social-medical data, "imitating" the capabilities of ML solutions. The rest of this paper describes our patient similarity search solution and its usage within IBM MedICS.

## 2. SOLUTION OVERVIEW

Our social-medical discovery model is based on our previous work in [7, 9]. Built on foundations of conceptual modeling [12], social data and medical data are fused together using a uniform representation in the form of a rich entity-relationship graph (ERG) [7]. As a result, social discovery can be augmented with medical discovery and vice-versa. Formally, we assume that each social or medical entity e has a type (e.g., `Patient`, `Medication`, etc), a set of one to many attributes, each attribute further has a name and a value (e.g., `Patient.age:20`, `Medication.name:Warfarin`, etc). A relationship r includes a name and captures an association between two or more entities and may further have some attributes of its own, termed the relationship context.

As an example, Figure 1 depicts an instance ERG graph that can be searched using our discovery system; this instance includes three entities (of type `Patient`, `Medication`, and `Disease`) that are combined together using the ternary relationship `Treatment`. Further note that, an attribute on the treatment relationship is used to capture the consumed medication dosage amount.

Our social-medical discovery solution is implemented using a unique inverted-index structure based on faceted search foundations [13]. Within this solution, entities are serialized as multi-field documents, while relationships are captured by a new type of facet serialization termed *category sets* [12].

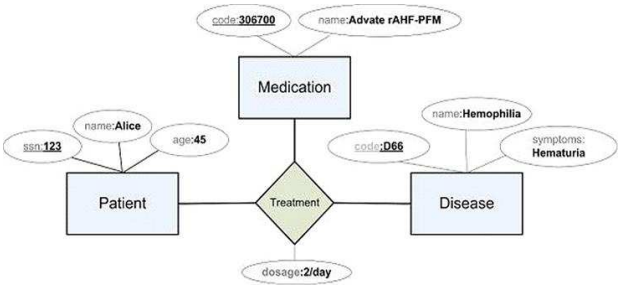



**Figure 1: An example ERG instance with a single medical treatment relationship.**

Entities or relationships are further represented by a category (facet) and various relationships among entities and their attributes (context) are captured using category sets which are efficiently serialized within the inverted index [13].

Our social-medical patient similarity solution is based on an extension to the query language we previously proposed in [13] to further allow efficient and intuitive discovery of similar entities over semi-structured entity-relationship (ER) data. Such a query language enables to discover similar patients based on very complex entity-relationship patterns. For example, the following query can be used to retrieve all patients in the system (including the one depicted in Figure 1) that were treated with the Advate medication for some blood related diseases:

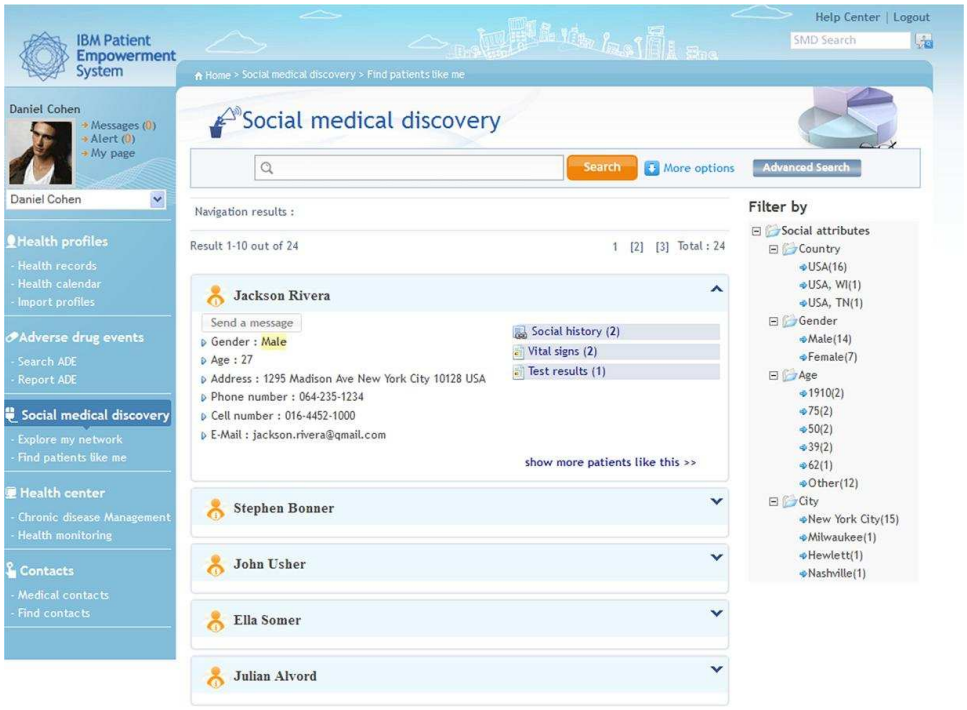
```
Patient.  AND (Treatment WITH Disease.name:Hemo*
AND WITH SIM(Medication.name:Advate))
```

The notation `SIM`(entity) in our extended query language denotes the similarity operator. Such an operator can be used to retrieve also medications that are (approximately) similar to the Advate medication, e.g., Advate rAHF-PFM, Helixate FS, Kogenate FS, etc; based for example, on word similarity, synonymy, coding system mapping, ontological proximity, etc. The `WITH` notation is another special operator that can be used to define various relationship participation patterns; used in this example to specify the treatment relationship that should be (approximately) satisfied.

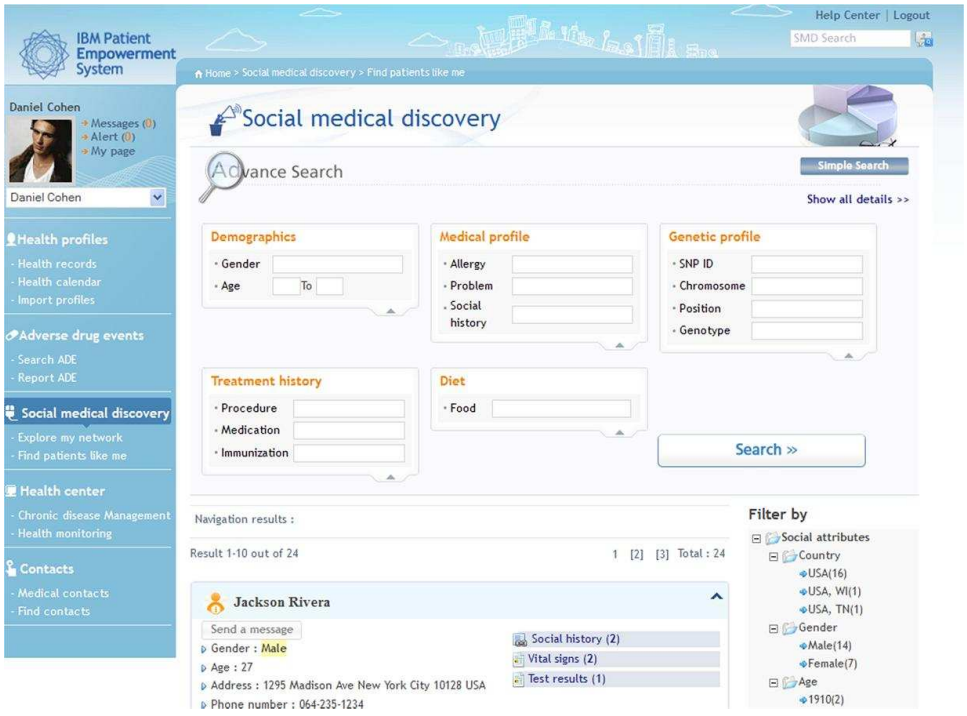
Next we describe our patient similarity search solution, based on our query language. We now define any patient's attribute type (e.g., `gender`), relationship type (e.g., `Treatment`), or relationship members and attributes (e.g., `Medication`, `dosage`) as potential patient similarity features (or features for short). Given a patient p, our similarity search algorithm aims at maximizing the following top-k similarity function:

$$sim(p, p') = \sum_{a \in A(p)} w_a \cdot sim(p, p'|a) + \sum_{r \in R(p)} w_r \cdot sim(p, p'|r) \quad (1)$$

where $A(p)$ and $R(p)$ denote the set of patient $p$'s attributes and relationships and $w_a$ and $w_r$ denote weights.




(a) Social-medical discovery main UI




(b) Advanced patient similarity search UI

**Figure 2: IBM MedICS social-medical discovery UI with various patient similarity search options.**

Therefore, given a patient p, we wish to discover up to k patients $p'$ that are most similar to patient $p$ given all similarity features of interest; either based on direct patient attributes, relationship similarity or indirect similarity that can be inferred from patients' relationships with other similar entities; similarly to the SimRank concept [5]. The weights $w$ can be further used for controlling the relative importance of various similarity features during discovery time; e.g., for diabetic patients, similarity based on features such as sugar level and diabetes type may be more important than similarity based on features such as gender or age. Such weights for various features can be manually provided by the searcher, capturing the searcher's preferences.

Our patient similarity search algorithm discovers similar patients in two main steps. At first, we extract the patient's entity, its relationships and related entity members from the ERG maintained by the social-medical system. Next, different patient features are combined together by their type (e.g., medications, allergies, foods, etc). Each feature type is then represented by a single query predicate that is constructed from its feature instances. As an example, assuming a given patient that consumed two medications, say Warfarin and Advate, the following query predicate will be generated for representing the treating Medication feature, where `w[Treatment]` further denotes the treatment feature importance:

```
Patient.  AND (Treatment WITH SIM(Medication.name:Warfarin)
OR SIM(Medication.name:Advate))^w[Treatment]
```

Finally, the similarity ranking function is realized using a single similarity search query that combines the various feature queries predicates together using their disjunctive form which can be efficiently evaluated by our system [13].

## 3. IMPLEMENTATION

The proposed patient similarity solution was implemented and integrated with the social-medical discovery service of IBM MedICS [7, 9]. IBM MedICS is a novel clinical decision support system (CDSS) that empowers the patients and helps to increase patient safety by assisting patients and their medical providers with daily medical decision-making.

IBM MedICS provides its users with three different options for discovering similar patients. The first two options are depicted in Figure 2(a) which illustrates the main social-medical discovery UI of IBM MedICS. The first option allows the current searcher to discover patients that are directly similar to her (using the *Find patients like me* link in Figure 2(a)). For this option, the current patient searcher is provided as an input to the similarity search algorithm. The second option allows the searcher to focus the similarity search on any patient result that is displayed in the list of search results (using the *Show more patients like this* link in Figure 2(a)). Using this option, the selected patient is provided as an input to the similarity search algorithm. The last and most sophisticated way to discover similar patients is using the advanced search UI depicted in Figure 2(b). Using this UI, "expert" users such as physicians and researchers can explicitly define the list of patient features (and their weights) including specific values that should be searched over in order to discover relevant patients. Built on exploratory search foundations, the UI further provides advanced social-medical discovery options, including, among others, a unique combination of faceted search and ER graph navigation [7, 13].

## 4. SUMMARY

In this paper we focused on the patient similarity task. Although several solutions already exist in the literature for this task, none currently seems to be satisfactory enough for discovery purposes. To address the challenges, we described a novel patient similarity solution, implemented within the IBM MedICS system. We believe that this paper opens new avenues for this task, challenging for seeking solutions that provide a more intuitive and efficient discovery, yet expressive enough for capturing diverse patient similarity features for various sub-medical domains. As future work, we plan to add sophisticated evidence capabilities to our similarity search solution, based on a new combination of data visualization, data summarization and provenance techniques.

## 5. REFERENCES


[1] M. Allison. Can web 2.0 rebot clinical trials? *Journal of Nature Biotechnology*, 27(10), 2009.

[2] L. Chan, T. Chan, L. Cheng, and W. Mak. Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 467 –470, dec. 2010.

[3] G. Eysenbach. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3), 2008.

[4] B. Jed, D. Bern, C. Lustig, and G. Hayes. Patientslikeme: Empowerment and representation in a patient-centered social network. In *Proceedings of Workshop on Research in Healthcare, CSCW*, 2010.

[5] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.

[6] S. Klenk, J. Dippon, P. Fritz, and H. Gunther. Determining patient similarity in medical social networks. In *Proceedings of First International Workshop on Web Science and Information Exchange in the Medical Web (MedEx)*, 2010.

[7] H. Roitman, Y. Messika, Y. Tsimerman, and S. Yogev. A unified approach for social-medical discovery. In *Proceedings of the 23rd International Conference of the European Federation for Medical Informatics (MIE), Oslo, Norway*, 2011.

[8] H. Roitman, Y. Tsimerman, S. Yogev, and Y. Peres. On the support of flexible patient privacy policies in social-medical discovery. In *Proceedings of the 25th International Conference of the European Federation for Medical Informatics (MIE), Pisa, Italy*, 2012.

[9] H. Roitman, S. Yogev, Y. Tsimerman, D. W. Kim, and Y. Mesika. Exploratory search over social-medical data. In *Proceedings of CIKM*, pages 2513–2516, New York, NY, USA, 2011. ACM.

[10] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[11] F. Wang, J. Sun, and S. Ebadollahi. Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. In *SDM'11*, pages 59–70, 2011.

[12] S. Yogev and H. Roitman. Bridging the gaps towards advanced data discovery over semi-structured data. In *Proceedings of the ER*, pages 156–165, Berlin, Heidelberg, 2012. Springer-Verlag.

[13] S. Yogev, H. Roitman, D. Carmel, and N. Zwerdling. Towards expressive exploratory search over entity-relationship data. In *Proceedings of WWW*, pages 83–92, New York, NY, USA, 2012. ACM.

# Why Is It Difficult to Detect Outbreaks in Twitter?

Avaré Stewart[1], Nattiya Kanhabua[1], Sara Romano[2],
Ernesto Diaz-Aviles[1], Wolf Siberski[1], and Wolfgang Nejdl[1]

[1]L3S Research Center / Leibniz Universität Hannover, Germany
{stewart, kanhabua, diaz, siberski, nejdl}@L3S.de

[2]Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione / University Federico II Naples, Italy
sara.romano@unina.it

## ABSTRACT

In this paper, we present an event-based Epidemic Intelligence (EI) system framework leveraging social media data, e.g., Twitter messages (or tweets) for providing public health officials the necessary tools to survey and sift through relevant information, namely, disease outbreak events. There exists three main research challenges in gathering epidemic intelligence from social media streams: 1) *dynamic classification* to enable message filtering, 2) *signal generation* producing reliable warnings based on observed term frequency changes in the filtered messages, and 3) providing *search and recommendation* functionalities to domain experts, for better assessment of the potential outbreak threats associated with the generated signals. We outline possible approaches to solve these important challenges as well as discuss areas where further research is required. The aim of this paper is to provide guidance for similar endeavors, and to give prospective event-based Epidemic Intelligence system builders a more realistic view on the benefits and issues of social media stream analysis.

## 1. INTRODUCTION

Social media, e.g., Facebook or Twitter messages, are valuable sources for providing real-time information, such as, status updates, opinions or news. Numerous real-time Web applications increasingly use Twitter for tasks, such as, detecting natural disaster [19], political persuasion [16, 20], or present trends [15]. In the medical domain, it has been shown that Twitter is capable of transmitting information faster than traditional media channels [8, 14], thus giving human experts a head start in dealing with health-related information. To exploit this timeliness potential, we present an event-based Epidemic Intelligence (EI) system, which has emerged as a type of intelligence gathering aimed to detect events of interest to the public health from unstructured text on the Web. In our proposed EI system, we detect public health events by mining and analyzing tweets; as well as provide support for public health officials to retrieve and explore the *signals* of infectious disease outbreaks. Signals represent a very dynamic type of information object, which are generated for each temporal anomaly found in time series data that occur when an infectious disease or its impact is above an expected level, for a particular time and place. Signals are monitored by public health authorities and help them assess the need for action, in response to potential threat. Note that, there are existing EI systems, such as, the BioCaster Global Health Monitor[1] or HealthMap [2]. However, they differ from our proposed system in the level of analysis, information sources, language coverage and visualization.

Although numerous approaches successfully detect relevant seasonal influenza outbreak events from Twitter [1, 5, 15], it seems that the challenges in building an EI system are easily underestimated, especially when it comes to detecting *emerging* (*unseen* or *non-seasonal*) health events from social media streams. Inspired by the outcome of collaborations conducted as part of the European research project Medical Ecosystem: Personalized Event-based Surveillance [3], with medical domain experts and epidemiologists, the aim of this paper is to describe an EI system that is better targeted towards the needs of real-world users to access public health information. This includes describing three main challenges associated with social media stream analysis systems, outlining possible approaches to handling such challenges and pointing out issues where more research is needed in order to achieve high-quality results coupled with wide-spread acceptance within the public health domain. Specifically, we have identified the following core challenges in event detection on Twitter data streams:

- **Adaptive Message Filtering.** Although the detection for well-known, recurring events (e.g., influenza) is mature, the detection of *novel* and *aperiodic* public health events requires *adaptive approaches* which take into account feature change over time, i.e., to enable the identification of new relevant terms.
- **Signal Generation from Noisy Data.** Time series data created from Twitter is usually noisy, incomplete and sparse. Given the imperfect data, it is important to consider measures for *assessing the reliability of signals*, i.e., the extent to which we can actually trust signals that have been generated for early warning.
- **Threat Assessment Support.** End users need assistance to cope with the cognitive challenges of *search and exploration* of outbreak signals. The effectiveness of straight-forward approaches to retrieval and collaborative filtering can be unsatisfied, given the dynamics of streaming data and the limited context of detected signals as well as their corresponding tweets.

[1]http://biocaster.nii.ac.jp/
[2]http://www.healthmap.org/en/
[3]http://www.meco-project.eu/


*HSD 2013,* August 1, 2013, Dublin, Ireland.

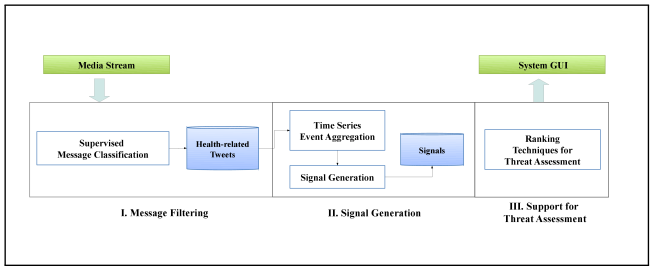



**Figure 1: Event-based Epidemic Intelligence system.**

## 2. SYSTEM OVERVIEW

The overview of our EI system is illustrated in Figure 1. We gather tweets relevant for outbreak surveillance using of multi-lingual list of terms consisting of not only infectious disease names and pathogens, but also, their synonyms and symptoms (provided by the experts).

The `Supervised Message Classification` module is responsible for filtering tweets irrelevant to the medical domain. Although existing works [4, 15, 18] have already addressed this problem in static settings. In this work, we propose an adaptive feature change detection method as we will discuss in more detail in Section 3.

In the `Signal Generation` module, signals (i.e., outbreak warnings) are generated using time series data, as was done in previous work [13], consisting of aggregated counts over the common entity tuples of the relevant messages. In general, input data used for anomaly detection is noisy (contain spurious events), incomplete (under-reporting of an event or using of acronyms and abbreviates not recognized) and sparse (low aggregation counts). In Section 4 we consider the impact these aspects for signal generation.

Even though successive stages of filtering are carried out within each module, domain experts may still be faced with potentially many signals and their associated tweets. Therefore, these signals are presented to the user via the `Support for Threat Assessment` module. In particular, we seek to employ time-aware ranking and recommendation techniques to tackle the problem of information overload (cf. Section 5).

## 3. CHALLENGE I: MESSAGE FILTERING

The online message classification continues to be a complex and challenging task for long term surveillance and intelligence gathering, in general. One reason for this is that given the evolution of real-world events, the variable to observe cannot always be known a priori. In EI, one such example is in the detection of food-borne illness, in which the contaminated food item is not known in advance.

**Feature change detection.** We need a way to: 1) dynamically detect when *new* and *relevant* terms in a stream appear; and then 2) subsequently incorporate the tweets containing these terms into the classification model.

**Dynamic labeling.** As new terminology evolves, the criteria for defining relevant tweets is also changing. However, expert labeling of classifier training instances is expensive and in practice difficult to obtain, especially for the rate and volume needed to build and maintain a good classifier.

**Ambiguous and noisy data.** When a tweet is matched with the defined keywords, the tweet itself may not refer to a public health event due to polysemy. For example, the term "fever" being used to express excitement, e.g., *Justin Bieber Fever* or *Royal Wedding Fever*. In addition, noise can be caused by spurious events in which an entity is correctly detected, but its role is not, namely: 1) *"A two hour train journey, Love In the Time of Cholera."* or 2) *"I liked a @YouTube video http://youtu.be/... a Metallica, Megadeth, & Anthrax - Helpless"*. Both mention infectious diseases Cholera and Anthrax, but their context is literature and music, respectively.

### 3.1 Proposed Approach

In order to capture aperiodic events with a high impact on accuracy over time, we propose the use of a method that takes into account that the natural language of the tweets in the stream changes constantly in response to the *temporal dynamics of real-world events*. Our dynamic classification consists of two main steps: 1) incorporating the use of an orthogonal vector, which is learned by a Support Vector Machine (SVM), as a description of the feature change; and 2) computing a novelty score that lets the system identify those tweets that contribute to the feature change, so that their true labels can be obtained. In this paper, we only focus on *text-based analysis* of Twitter messages. We plan to investigate the use of Web resources shared in social media, e.g., posted images and videos, as future work.

### 3.2 Implications

Identifying non-relevant tweets is difficult for multiple reasons. For example, automatically filtering a sarcastic and metaphoric tweets correctly is hard for limited context and remains to be tackled with this domain. Feature change detection should be able to identify new (entity and non-entity) keywords in the Twitter stream that are related to outbreaks of infectious diseases. These keywords could be used for automatically updating the list of search term that is used to collect the tweets. Consideration will also be given to when these newly discovered terms should be subject to decay. To this end, we plan to increase the scale of the analysis to include: the classification of more symptoms as well as polysemy terms for diseases. To get a better understanding of the impact of detected feature change on the classification accuracy, a larger set of expert labeled tweets for experimentation would be useful to further improve the significance of the results. Nonetheless, doing so, would still not address the need to experts to re-label each time feature change was detected and in practice, the overhead of such a task is too expensive and not timely enough.

## 4. CHALLENGE II: SIGNAL GENERATION

Health-related tweets obtained from the previous stage will be leveraged in order to generate an early warning signal (so-called *signal generation*). Signals represent each temporal anomaly found in time series data occurring when the impact of an infectious disease is above an expected level, and it is a difficult task because of the following challenges:

**Incompleteness and sparsity of data.** This implies that instances of an event are missing or under-reported. This may occur due to: 1) the presence of processing errors - an acronyms or abbreviations not recognized as medical conditions; 2) the fact that people who are actually suffering do not tweet; 3) the tweets which contain these mentions have not been collected by the system, i.e., based on the imbalance between the type of tweets collected (e.g., personal versus news tweets); and 4) the minimum required entity types are not present. Sparse time series data refers specifically to low aggregation counts, which impact the anomaly detection algorithm.

**Temporal and spatial dynamics of diseases.** The characteristics of infectious diseases are highly dynamic in time and space, and their behavior varies greatly among different regions and the time periods of the year. Some infectious diseases can be rare or aperiodic, while others occur more periodically. In addition, various diseases have different transmission rates and levels of prevalence within a region. For example, cholera infections vary greatly in frequency, severity, and duration. On the one hand, in some regions historically, only sporadic outbreaks occur in areas, such as, parts of South America and Africa. On the other hand, even in areas where cholera infections are endemic (the South Asian countries of Bangladesh and India) the epidemic levels change dramatically from one year to the next [9].

Given imperfect time series, we need to know the extent to which we can actually trust signals that have been generated for early warning. To this end, we aim at answering the question: *Are there ideal algorithms and/or parameter settings for signal generation using Twitter?*

## 4.1 Proposed Approach

Studying the usefulness of Twitter data in the medical domain requires real-world outbreak statistics. We previously built outbreak ground truths (historical baselines) by relying upon ProMED-mail[4], a global reporting system providing information about outbreaks of infectious diseases. We collected 3,056 ProMED-mail reports and identified 14 different outbreaks occurring during year 2011 as ground truths [12]. An important aspect of our work is that we consider the duration of each outbreak by analyzing temporal expressions mentioned in a ProMED-mail document, unlike aforementioned work [3] that assumes the publication date of a document as the estimated relevant time of an outbreak. The reason is that the events in ProMED-mail undergo moderation, so there is often a delay between the time of the actual outbreak and the publication date of the related report.

A basic approach to detect anomaly in health-related time series data is to exploit different state-of-the-art biosurveillance algorithms [2, 10]. These algorithms are already widely used in the existing Biosurveillance systems, so they can be used for assessing the reliability of signals from the perspective of the domain experts. The metrics used to assess generated signals are sensitivity, predictive positive value and F-measure. Sensitivity refers to the proportion of true signals correctly detected by a surveillance algorithm.

In our recent work [11], we sought a new feature by moving beyond using only keywords/medical conditions. We proposed to analyze the diversity metrics of tweets over time, so-called *temporal diversity*. The diversity statistics can capture a broad spectrum of topics, communities and knowledge that are evolving over time. In particular, analyzing temporal diversity can shed light on two aspects. First, an increase of content diversity over time indicates that a community is broadening its area of interest. Second, negative peaks in diversity can additionally reveal a temporary focus on specific events. To address an efficiency issue, we employed an algorithm based on sampling [6]. We performed a correlation analysis of the temporal diversity of 14 real-world events with their estimated event magnitudes during the known outbreak periods. Our analysis showed that correlation results are varied greatly among outbreaks reflecting the characteristics (severity and duration) of outbreaks.

[4]`http://www.promedmail.org`

## 4.2 Implications

As we aimed at detecting outbreak events for general diseases that are not only seasonal, but also sporadic diseases that occur in low tweet-density regions, some difficulties in constructing the outbreak ground truth still remain, which resulted in a dataset that was limited in terms of the number of outbreaks and their diversity. Particularly, the smaller the number of outbreaks we analyzed, the harder it was to generalize our solution. The process of creating the ground truth for disease outbreaks requires information extraction techniques, namely, different NLP tools for extracting relevant information. Unfortunately, the accuracy of such tools are not nearly 100%, which has a severe impact to the coverage (number) and quality of outbreak ground truth found. For example, place names are ambiguous and can be wrongly determined as the country of an outbreak as illustrated in this sentence *The **Uganda** Virus Research confirms Ebola virus Sudan species*. In addition, the accuracy of information extraction techniques as well as the noisiness of ProMED-mail data have also limited the coverage and quality of ground truth. For instance, there are many near-duplicate reports of outbreaks and many of irrelevant reports related to disease vaccines instead of outbreaks. Moreover, a report on historical statistics of a disease outbreak is irrelevant information, which should be carefully excluded from the ground truth. Similar to information about updates on an outbreak situation that must be avoided.

# 5. CHALLENGE III: SUPPORTING THREAT ASSESSMENT

For detected events, public health experts participating in its investigation face the overwhelming task of analyzing the large number of tweets associated to the corresponding signals [21]. The real-time nature of Twitter, on the one hand makes it attractive for public health surveillance; yet, on the other, the volume of tweets also makes it harder to: 1) capture the information transmitted, 2) compute sophisticated models on large pieces of the input, and 3) store the input data, which can be significantly larger than the algorithm's available memory [17].

## 5.1 Proposed Approach

To reduce this information overload and support the task of threat assessment, we explored to what extent recommender systems techniques can help to filter information items according to the experts' context and preferences. Our previous work [7] has shown the effectiveness of *Personalized Tweet Ranking for Epidemic Intelligence* for a case study of the 2011 EHEC outbreak in Germany. In particular, we focused on a personalized learning to rank approach that ultimately offers the user the most relevant and attractive tweets to support the task within her/his context. Our approach extended a learning to rank framework by considering a personalized setting that exploits a user's individual *context*. We considered such context as implicit criteria for selecting tweets of potential relevance, and guiding the recommendation process. We used the terms in the expanded context that correspond to medical conditions, locations or complementary context (that corresponds to the set of nouns, which are neither locations nor medical conditions) in order to to build a set of tweets by querying our collection. This step helped us to filter irrelevant tweets. Next, we elicited judgments from experts on a subset of the tweets retrieved in order to build a ranking function model. We then obtained

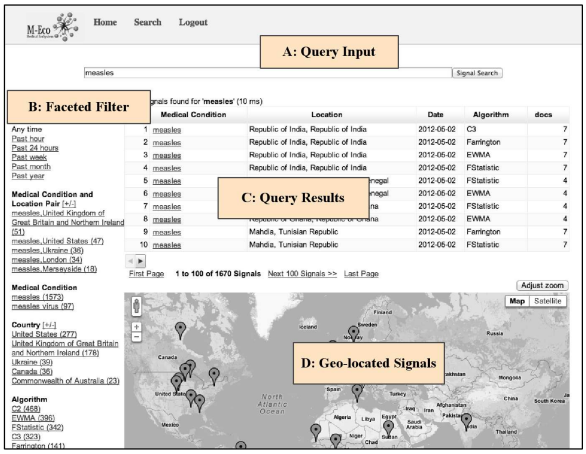



**Figure 2: User Interface showing *A. Query Input*: for entering a search term, e.g., "measles", *B. Faceted Filter*: options for filtering signal search results by metadata, *C. Query results*: result set of signals, and *D. Geo-located Signals*: a map for visualizing signals' geo-location.**

for each labeled tweet, a feature vector that help us to train our personalized ranking model. Finally, we applied a learning to rank algorithm to obtain the ranking function for the given user context.

In addition to the personalized learning to rank approach, we present the first prototype for support search and retrieval of signals. We envision the functionality of *signal-based* retrieval, that is, returning signals as results of a given query instead of only documents. Once the desired signals are obtained, the user is able to access the original tweets associated to each of them. Having signals as basic unit of information allows us to perform a focused indexing of only the tweets relevant to a particular signal. Figure 2 shows the user interface along with a brief description of its main panels. A possible solution is to implement a ranking model that: (1) extends a learning to rank framework by considering a personalized setting that exploits a user's individual *context*; (2) answers user's query by providing a list of relevant tweets ordered from newest to oldest, starting from the time the query was issued. When selecting tweets to include in the list, systems should favor both the *relevance* and *recency* of tweets.

## 5.2 Implications

However, the current load of experts in assessing these signals can be reduced significantly by employing personalized ranking techniques. Given that experts' interactions and explicit feedback are scarce in EI systems, the application of standard recommender system algorithms is not straightforward making it harder to build effective models for ranking or recommendation. By exploiting complementary context information, extracted from the social hash-tagging, and the latent topics discovered within the tweets, an effective ranking mechanism for messages associated with signals can be achieved. As a plan for future work, supporting temporal analytics for public health events will bring EI systems a big step forward, and also can provide useful guidance for other systems based on using social media data.

**Acknowledgments** This work was partially funded by the European Commission Seventh Framework Program (FP7 / 2007-2013) under grant agreement No.247829 for the Medical Ecosystem Project (M-Eco).

# 6. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2011.

[2] M. Basseville and I. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice-Hall information and system sciences series. Prentice Hall.

[3] N. Collier. What's Unusual in Online Disease Outbreak News? *Journal of Biomedical Semantics*, 1(1):2, 2010.

[4] N. Collier and S. Doan. Syndromic classification of twitter messages. In *eHealth*, pages 186–195, 2011.

[5] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[6] F. Deng, S. Siersdorfer, and S. Zerr. Efficient jaccard-based diversity analysis of large document collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012.

[7] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd. In *International AAAI Conference on Weblogs and Social Media*, 2012.

[8] M. Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[9] M. Emch, C. Feldacker, M. S. Islam, and M. Ali. Seasonality of cholera from 1974 to 2005: a review of global patterns. *International Journal of Health Geographics*, 7(1), 2008.

[10] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society*, 159(3):pp. 547–563, 1996.

[11] N. Kanhabua and W. Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd international conference on World Wide Web companion*, 2013.

[12] N. Kanhabua, S. Romano, and A. Stewart. Identifying relevant temporal expressions for real-world events. In *SIGIR Workshop on Time-aware Information Access*, 2012.

[13] N. Kanhabua, S. Romano, A. Stewart, and W. Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 2010.

[15] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, September 2012.

[16] C. Lumezanu, N. Feamster, and H. Klein. #bias: Measuring the tweeting behavior of propagandists. In *Proceedings of the Sixth International Conference on Weblogs and Social Media*, 2012.

[17] S. Muthukrishnan. *Data streams: algorithms and applications*. Now Publishers, 2005.

[18] M. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[20] E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, 2012.

[21] G. Shmueli and H. Burkom. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics*, 52(1):39–51, February 2010.

# A Tool for Monitoring and Analyzing HealthCare Tweets

Ahmed Ali, Walid Magdy, Stephan Vogel
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
{amali, wmagdy, svogel}@qf.org.qa

## ABSTRACT

The amount of data exchanged over social media is witnessing a major growth in the last few years. Various studies in different domains investigated extracting useful information from this exchanged data. Less attention was directed toward studying healthcare in social media compared to other topics such as politics and marketing. In this paper, we present a platform for monitoring healthcare tweets on social media in different regions. The platform offers a solution to governments or healthcare providers to monitor public health and measure public satisfaction with healthcare services from what people post on Twitter. It helps in the early detection of disease outbreak and healthcare public view. Our platform uses an automatic classification method for detecting healthcare related tweets. It presents comprehensive reports that provide the most popular topics people share, sentiment analysis of tweets, multimedia content related to health. The platform is tested and demonstrated for three different locations: London, Boston, and Dublin. In addition, its effectiveness is tested for Arabic and English tweets with no specific location.

## 1. INTRODUCTION

Social media is currently playing a fundamental role in the life of many internet users. Public posts on a social website such as Twitter include personal status, opinion sharing, discussions, marketing, campaigning … etc. Among the material users share on Twitter are tweets related to health and healthcare. Some users share information and updates on their health or the health of loved ones. This is a common behavior by many people who seek support during difficult times. These social posts can be of tremendous value if detected and monitored by healthcare providers: they provide indicators about the general public health, they can help in early detecting of an epidemic, or can alert about ongoing concerns with healthcare [1-3]. In addition, people often share their experience with healthcare facilities like hospitals, clinics, and health centers. People can recommend specific doctors or clinics, or they can complain about particular aspects of health facility, such as queuing in emergency department, food service in hospital, competence and attitude of caregivers … etc. In general, healthcare related tweets can be utilized as an indicator about the quality of services provided by these health facilities, and can help to improve the provided services to patients. Unfortunately, studying healthcare material posted on social media has received less attention compared to other topics [2, 4].

In this paper, we present a tool that monitors and analyzes health related tweets and presents a comprehensive report for what the public shares about health in specific locations. The presented work in this paper addresses the following questions:

1. How to collect health related posts from social media, such as Twitter?
2. What analysis is required to extract indicative information from the collected tweets?
3. How extracted information could be presented?

Our proposed approach for collecting healthcare tweets uses a semi-supervised approach to identify relevant tweets based on a set of pre-defined keywords. Sentiment analysis is applied to obtain an indication of people's satisfaction and/or feelings. In addition, the collected tweets are parsed and analyzed to extract the most shared videos, images, and links in these tweets. Finally, a comprehensive report about the extracted information is generated and presented in a web interface to allow experts or care givers to monitor the public health in specific locations. This report includes: most circulated tweets, videos, images, and articles; tag-cloud of top used terms; and sentiment graph. We test this platform on three different locations: London, Boston and Dublin. In addition, it is tested on English and Arabic tweets coming from unspecified locations.

## 2. HEALTHCARE AND SOCIAL MEDIA

Social network started to have visible presence in health care in 2008. "*Hello-Health*" [5] is an initiative to think of the Electronic Health Record (EHR) as a social network, which acts as a concierge practice. The main shift is from large hospital networks to patient support groups and news media tools, such as weblogs, instant messaging, video chat and social networking. Patient could use the *Hello-Health* network to send a message to the physician describing the symptoms and asking for advice. A quick e-mail from the physician to follow up, and if needed a guaranteed hospital visit scheduled in 24 hours. Another example, *PatientsLikeMe* [6] is a platform to use online personal information for patients to share their experience using patient-reported outcomes, find other patients like them matched on demographic and clinical characteristics, and learn from the aggregated data reports of others to improve their outcomes. The main goal for *PatientsLikeMe* is to help patients answer the question: "Given my status, what is the best outcome I can hope to achieve, and how do I get there?".

Patients have moved beyond searching and asking towards sharing and interacting. The *PewInternet&American 2011* study showed that 23% of those with chronic health issues, such as cancer, diabetes, or heart disease, have gone online to look for patients with similar conditions; while only 15% of patients with no chronic conditions have sought peer-to-peer information [7].

The Centers for Disease Control (CDC) has been leading the efforts to use twitter as a channel to reach out to the public to deliver more information about infection, disease. During the H1N1 outbreak of 2009, the CDC decided to communicate with patients and caregivers across the US though Twitter. The CDC got more than 1.2 million followers for emergency information and 46,000 following the flu [8, 9].

HealthVault is a Microsoft ongoing initiative to integrate health records into a unified solution, enabling patients to store, and share health information online[10]. Google Health was a similar


*HSD 2013*, August 1, 2013, Dublin, Ireland.

platform launched in 2008 [11]. Both HealthVault and Google Health were developed as platforms to facilitate easy access to and management of the Electronic Public Health Record EPHR and sharing details amongst user specified networks.

In [12], it was shown that the shift of accessibility in healthcare is no longer just about getting an appointment with a physician or scheduling a treatment; it goes to community outreach, since social network sites can help hospitals to communicate with patients they serve. However social media in healthcare is still in its infancy. A study carried out in the US including 5,800 hospitals, showed that only 965 hospitals, which is less than 17%, are using social media to reach out to patients [13]

Several studies investigated the use of Twitter to analyze ongoing health related events. A study for detecting the influenza epidemic was carried out over 300 million tweets [14]. They used support vector machines (SVM) and showed the feasibility of their approach with a correlation of 0.89 to the gold standard from the Infection Disease Surveillance Center IDSC, thus outperforming the Google flu trend. Another recent study used Twitter to track the public concern in the US during the influenza A H1N1 [1]. It showed the capability of Twitter feeds not only to describe, but to track users' interest and concern about the development of the H1N1 epidemic and track disease activity. The study used a data set of 5 million tweets.

The aforementioned studies focused on sampling and analyzing healthcare related data on social network to have representative sampling with enough confidence in the given results. However, to the best of our knowledge, no attention was directed to providing a generic healthcare platform for detecting health related social posts, analyzing them, then presenting them in a comprehensive way for capturing trends in people's perception and reaction towards health related issues.

# 3. TWEETS MONITORING METHODOLOGY

In this section, the method underlying our study is presented. First we describe the data collection process, then the analysis criteria are introduced, and finally we describe how the healthcare tweets are modeled in both Arabic and English tweets.

## 3.1 Microblog Data Streams

Our platform is monitoring tweets coming from 5 streams; three streams coming from specific locations: London, Dublin, and Boston; and two language streams that are not necessary geo-tagged: English stream and Arabic stream. The purpose of the location-tweets streams is to monitor health related tweets in specific regions, while the purpose of language-specific streams is to monitor the trend in health related tweets globally over Arabic and English tweets in general.

Location-tweets are streamed based on geo-location of tweets by specifying longitude, latitude, and radius, which are specified to cover the targeted cities. Language-tweets are streamed by searching Twitter for "lang:en" and "lang:ar" for the English and Arabic tweets respectively. For each stream, health related terms were used to identify potentially health-related tweets. Table 1 shows the details of how each stream is collected and the average number of healthcare tweets identified per day. Table 2 presents an example of the terms used to identify the potentially health related tweets for English and Arabic.

Text of streamed tweets is pre-processed using state-of-the-art normalization techniques for microblogs in English [15] and Arabic [16] .

**Table 1: Tweets Streams**

| Stream | English | Arabic | Boston | London | Dublin |
|---|---|---|---|---|---|
| Location | Any | Any | Boston | London | Dublin |
| Language | English | Arabic | English | English | English |
| Longitude<br>Latitude | Null | Null | 42.36,<br>-71.05 | 51.51,<br>-0.1241 | 53.25,<br>-6.36 |
| Radius | | | 50 km | 50 km | 50 km |
| avg. # tweets collected/day | 50,000 | 20,000 | 3000 | 7000 | 800 |

**Table 2: Example of terms used to identify potential health-related tweets**

| | |
|---|---|
| English | *hospital clinic disease infection virus #health #healthcare #nhs #disease #emergency #medicine #ObamaCare* |
| Arabic | مستشفي عياد مريض مرض الم فيرس فيروس وجع تعبان مصاب |

## 3.2 Modeling Healthcare in Twitter

Terms used in Table 2 lead to the collection of tweets that are potentially related to healthcare. However, we noticed that some of the collected tweets that contain any of these terms are irrelevant to healthcare because of the multiuse of these terms in different domains. For example, in English word 'virus' has a potential to exist in healthcare tweets, however it occurs more often in the context of computer virus. Similarly, in Arabic we found out that the word "**كورونا**" (corona) appears so often in the first 2 weeks of May 2013 and this is due to virus corona in Saudi Arabia, and such a word is unlikely to stay as a healthcare top frequent keyword after the virus outbreak. Therefore, a more reliable classification is required to distinguish between healthcare relevant and irrelevant tweets. For this purpose, we used more restricted queries to select the tweets among the potentially identified healthcare ones that are very likely to be relevant healthcare. Then we used the narrow identified tweets to train a support vector machine (SVM) with linear kernel to classify the remaining identified tweets as relevant or irrelevant to healthcare.

The SVM models are trained and updated automatically in an unsupervised manner. Figure 1 shows our classification algorithm. Using the same tweets stream we build three data sets, a positive and a negative set for the training the models, and the set of potentially relevant tweets, which would be classified by the SVM classifier. Our approach works as follows:

- Tweets $Set_P$ P (positive): the positive tweets are tweets related to health and healthcare and we select them by having a restricted precise Boolean regular expression. For example in English, we search for tweets with the most frequent healthcare hash tags such as #healthcare #ObamaCare #nhs and the tweets itself have at least one of the healthcare keywords such as: infection, virus, hospital, clinic ... etc. Out of theses tweets, we select the most retweeted 1000 tweets and use them as positive training examples.
- Tweets $Set_T$ (potentially relevant tweet): This set contains potentially relevant tweets. They are identified using the general health related terms shown in Table 2. Intuitively, this set is larger than the $Set_P$ since it includes any tweet, which matches one or several of the given keywords. However the precession is low, since this set contains irrelevant tweets such as the computer virus tweets mentioned earlier. A similar pattern happens in Arabic tweets, where religious tweets often have one or more health-related keywords. The Tweet set T suffers from low precision, and this database should be filtered to identify the truly healthcare related tweets.
- Tweet $Set_N$ (negative): This set is used as negative examples in the SVM training process. We select randomly 1000 tweets that

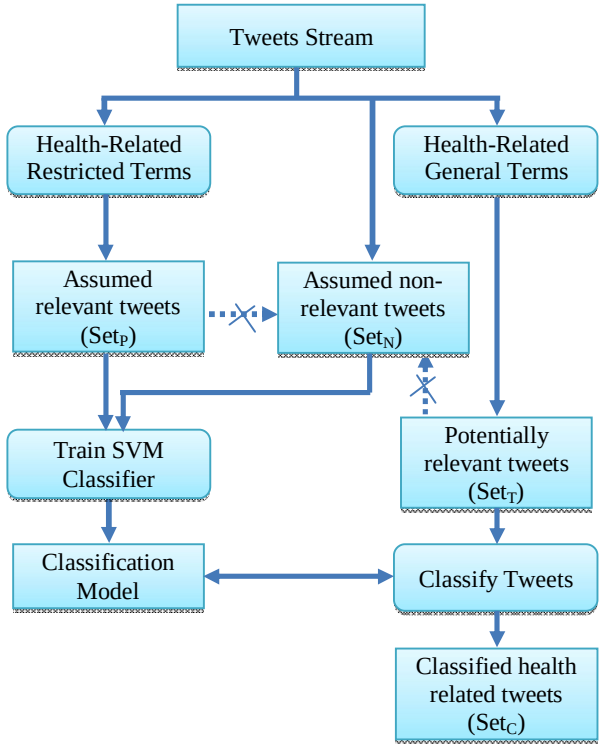



**Figure 1: Healthcare tweets expansion using SVM classifier**

do not match any of the healthcare terms. $Set_P$ and $Set_N$ are used together for training the SVM model, which is then used to classify the tweets of $Set_T$ to either relevant or irrelevant tweets to healthcare.

The process of training the SVM model is applied every 30 minutes using the identified tweets from the past 48 hours. Classified tweets as relevant from $Set_T$ are added to $Set_P$ and a comprehensive report is generated from them and presented in a web interface to users.

## 3.3 Healthcare Tweets Identification Results

The healthcare tweets retrieval is done for each of the tweet feeds separately. Table 3 shows the results studied over 3 feeds; Arabic, London and Boston. For each one of the 3 feeds, 200 tweets randomly chosen over 2 days to measure the robustness of the selection algorithm, the evaluation has been done manually by labeling the tweet either relevant or irrelevant to healthcare.

As shown in Table 3, the first step is to start by $Set_P$ in the first row. It has high precision and low coverage, which is used as positive examples to train the SVM models. The second set ($Set_T$) is larger and potentially relevant as shown in the second row. This set is the one that matches the general health terms, and which is classified later by the SVM model. Finally, the classified set ($Set_C$) in the third row is used for monitoring and analysis. The results show the robustness the classification approach, where large number of healthcare tweets is identified with high precision.

**Table 3: Accuracy for healthcare related tweets**

| Tweets Set | Arabic Stream | | Boston Stream | | London Stream | |
|---|---|---|---|---|---|---|
| | #tweets | Prec | #tweets | Prec. | #tweets | Prec. |
| $Set_P$ | 3000 | 0.90 | 400 | 0.95 | 550 | 0.97 |
| $Set_T$ | 20,000 | 0.65 | 3000 | 0.71 | 7000 | 0.75 |
| $Set_C$ | 12,000 | 0.82 | 1900 | 0.89 | 4500 | 0.87 |

# 4. HEALTH-RELATED TWEETS ANALYSIS AND VISUALIZATION

Extensive analysis is applied to the identified health-related tweets from each stream. A web interface has been developed to select any of the data streams and display the results of the analysis. The presentation of content is inspired from previous work in [17, 18], which presents a comprehensive report about relevant tweets to a given topic or entity. However, we use additional visualization tools that relate more to the analysis we use for our task. The analysis is applied to the identified tweets in the last 48 hours and the presented reports are updated every 30 minutes.

The analysis applied and presented consists of:

- **Extracting Most Popular Tweet:**

Identified health-related tweets of each stream are grouped by aggregating similar tweets into the same group. For a fast and robust matching between tweets, we keep only the main content text of tweets by removing all hashtags, name mentions, URLs, punctuations, symbols, emoticons, and retweet symbols. Tweets that match exactly in their main content are grouped together. Groups are presented in ranked (descending) order by their size with the most common tweet form as the representative of the group along with the number of tweets in the group.

- **Extracting Top Circulated Videos, Images, and Links:**

Since URLs in tweets are typically shortened and URLs may have multiple shortened forms, we expand all URLS to find the original URLs. We used URLs pointing to YouTube videos to obtain a ranked list of the most popular videos and embed them in the report separately. Also links pointing to Twitter images are presented for the most circulated images. Other URLs are extracted and their titles are presented with links to the pages. Links are ranked by the number of appearances in tweets.

- **Sentiment Analysis:**

Applying sentiment analysis on the healthcare related tweets can be used as an indicator about public satisfaction or feelings towards health. We use the SentiStrength [19] tools to estimate the strength of the positive and the negative sentiments in each tweet using the scale from -5 (extremely negative) to +5 (extremely positive). Tweets are then sorted according to the sentiment score to identify the most negative and most positive tweets. Also, we show the sentiment score on the top frequent tweets, which indicates the public view especially with the most popular (re-tweed) posts.

- **Tag Cloud:**

The top frequent words, terms, and hash tags (excluding the stop words) are used to draw a tag cloud, where font size used for the different terms indicates their frequencies. The tag-clouds helps to summarize the most popular terms in the tweets, which in turn indicate the most popular topics people are interested in.

# 5. SAMPLES OF THE EXTRACTED INFORMATION FROM TWEETS

In this section we present examples of the information that was extracted from the identified healthcare tweets to demonstrate the benefit of our system.

The top frequent tweets show the public awareness and concern with healthcare services. For example the third line in the most popular tweets is very important to professionals in healthcare sectors, government, and probably insurance companies to be aware with the concern of such a big difference in price for the

**Table 4: Examples of the extracted top tweets, articles, videos, and images related to healthcare from different streams**

| Most Popular Tweet | Stream |
|---|---|
| nurses fighting to save nhs for patients, 150,000-strong london march largely ignored by uk media | London |
| it is so painful. 135,000 health professionals applied for 1,760 jobs in alicante some months ago | Dublin |
| how can a procedure cost $297,000 at one hospital and $84,000 at another? disparities in #healthcare | Boston |

| Most Circulated Article title | Stream |
|---|---|
| Cap on number of GP visits being considered by Tories | London |
| IMF to have power over Irish healthcare spending | Dublin |
| 5 Jedi Mind Tricks to Help Yourself Get Healthy | Boston |
| فايروس كورونا الجديد وخطر واعراضه (lecture on virus corona) | Arabic |

| Most Circulated Videos (showing titles only) | Stream |
|---|---|
| My Spinal Cord Injury Story in pictures - Motivation -Bradley Hill Fitness | London |
| The Epidemiology of In-Flight Medical Emergencies | Dublin |
| Corona virus and the risk of iron and its symptoms | Arabic |

| Most Circulated Images (showing caption only) | Stream |
|---|---|
| Think again about going barefoot, a viewer got this nasty virus http://t.co/9rreT7Deys | English |
| كيف الإصابة بـ #فيروس_كورونا ؟وماهي أعراضه ؟ وكيف تقي نفسك؟وماهو العلاج ؟ http://t.co/3JBiLkgqhj | Arabic |

same procedure in two different hospitals. The first post is important for the hiring in healthcare sector.

The example of most circulated articles show the public concern about both healthcare (first two examples) and public health (in the last 2 examples). Interesting findings result from comparing the different data streams: while common health related topic like healthy life style and cuts for the healthcare are discussed in Dublin, London, and Boston,, we find that in the Arabic tweets, mainly in Saudi Arabia, the virus corona was the hot topic at the same time, due to the outbreak in. So, it is really useful to get the proposed automatic update about the health public concern.

The examples of videos are mainly focusing on public health issues, such as diseases outbreaks, healthy life, and lecture on how to deal with a certain virus (corona) … etc.

The section displaying the most circulated images is a very dynamic part in the generated report, as people are sharing images very frequently. For example the tweet about a woman contracting a virus after walking barefoot has been posted more than 1,100 times, but disappeared again within 2 days. Being able to address these findings will be quite useful for caregivers.

## 6. CONCLUSION AND FUTURE WORK

This paper presented a platform for monitoring and analyzing public tweets that relate to healthcare. We proposed an automatic approach for collecting relevant tweets to healthcare. We showed that our approach leads to the retrieval of reasonable number of healthcare tweets with high precision. The platform analyzes these collected tweets and extracts some useful information to be presented in a web interface for healthcare experts. Our platform was tested on 5 streams of tweets from specific and general location in two different languages. We believe that this platform is considered a good start for similar tools that helps in monitoring flowing information about health in social media in an attempt to utilize this information for improving the health services and the detection of public health threats.

For future work, we plan to apply different user studies on our platform to get feedback about its practicality for healthcare institutes. Furthermore, automatic alerting systems could be developed within the system to trigger alerts without the need for checking the web interface. A simple alert can be the number of identified health related tweets in a given window of time. Also, it can be a more advanced alert based on analyzing the appearance of new terms that indicate a healthcare disasters or threats.

## 7. REFERENCES


1. Signorini, A., A. Segre, and P. Polgreen, *The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S during the Influenza A H1N1 Panademic.* PLoS ONE, 2011. **6**(5): p. 1-10.
2. Prier, K.W., et al., *Identifying Health-Related Topic on Twitter An Exploration of Tobacco-Related Tweets as a Test Topic*, in *Springer-Verlag Berlin Heidelberg*2011. p. 18-25.
3. Paul, M.J. and M. Dredze. *You Are what you Tweet: Analyzing Twitter for Public Health.* in *The Fifth International AAAI conference on Weblogs and Social Media*. 2011.
4. Adams, S.A., Blog-based applications and health information: Two case studies that illustrate important questions for Consumer Health Informatics (CHI) research. International Journal of Medical Informatics, 2010. 79: p. e89-e96.
5. Hawn, C., Take Two Aspirin And Tweet Me In the Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Helath Care. Health Affairs, 2009. **28**: p. 361-368.
6. Wicks, P., et al., *Sharing Health Data for Better Outcomes on PatientsLikeMe.* Journal of Medical Internet Research, 2010.
7. Fox Susannah, *The social life of Health Information, 2011*, in *Pew Internet & American Life Project*2011: http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx. p. 1-45.
8. Aikin, A. *Case Study: Social Networking at the CDC.* 2010.
9. Eytan, T., et al., *Social Media and the Health System.* The Permanente Journal, 2011. **15**: p. 71-74.
10. Mircrosoft, *Healthvault: Connected Continuous Care How technology will transform chronic disease management*, 2011.
11. Aramaki, E., S. Maskawa, and M. Morita. *Twitter catches the flu: detecting influenza epidemics using Twitter*. in *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011.
12. Williams, J., *A New Road Map for Healthcare Business Success.* Healthcare Financial Management, 2011: p. 63-69.
13. Bennett, E., *A New Home for the Hospital Social Network List.* 2012.
14. ARAMAKI, E., S. MASKAWA, and M. MORITA, *Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter*, in *EMNLP*2011. p. 1568–1576
15. Han, B. and T. Baldwin, *Lexical Normalisation of Short Text Messages: Makn Sens a #twitter*, in *ACL-HLT*2011. p. 368–378.
16. Darwish, K., W. Magdy, and A. Mourad, *Language Processing for Arabic Microblog Retrieval*, in *CIKM*2012.
17. Bennett, E., *Hospital Social Network List.* 2011.
18. Magdy, W., A. Ali, and K. Darwish. *A summarization tool for time-sensitive social media*. in *CIKM*. 2012.
19. Yerva, S.R., et al., *TweetSpector: Entity-based retrieval of Tweets*, in *SIGIR* 2012.
20. Thelwall, M., et al., *Sentiment strength detection in short informal text.* Journal of the American Society for Information Science and Technology, 2010. **61**(12): p. 2544–2558.

# Test Collections for Medical Information Retrieval Evaluation

Lorraine Goeuriot, Liadh Kelly, Gareth J. F. Jones
Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin 9, Ireland
{lgoeuriot, lkelly, gjones}@computing.dcu.ie

## ABSTRACT

The web has rapidly become one of the main resources for medical information for many people: patients, clinicians, medical doctors, etc. Measuring the effectiveness with which information can be retrieved from web resources for these users is crucial: it brings better information to professionals for better diagnosis, treatment, patient care; and helps patients and relatives get informed on their condition. Several existing information retrieval (IR) evaluation campaigns have been developed to assess and improve medical IR methods, for example the TREC Medical Record Track [11] and TREC Genomics Track [10]. These campaigns only target certain type of users, mainly clinicians and some medical professionals: queries are mainly centered on cohorts of records describing a specific patient cases or on biomedical reports. Evaluating search effectiveness over the many heterogeneous online medical information sources now available, which are increasingly used by a diverse range of medical professionals and, very importantly, the general public, is vital to the understanding and development of medical IR. We describe the development of two benchmarks for medical IR evaluation from the Khresmoi project. The first of these has been developed using existing medical query logs for internal research within the Khresmoi project and targets both medical professionals and general public; the second has been created in the framework of a new CLEFeHealth evaluation campaign and is designed to evaluate patient search in context.

## 1. INTRODUCTION

The web is now used as one of the main resources for medical information by multiple user groups seeking to address many different classes of information need. Information Retrieval (IR) aims to provide results in response to user queries which address these information needs. Improving Medical IR constitutes a great challenge, as health is prevalent in everyone's life. Evaluating the effectiveness of IR for medical search tasks is key to developing effective systems and technologies. To date several IR evaluation campaigns have been developed in order to assist assess and improve medical IR methods, for example TREC Medical Record Track [11] or TREC genomics track [10]. However, these campaigns only target certain types of users, mainly clinicians and some medical professionals; and have only examined search of health records and evaluation search queries have mainly been centered on cohorts of records describing a specific case or on biomedical reports.

Some analysis of user query logs in the medical domain show that representative queries would be much shorter, whether they come from experts or non-experts [12, 1]. Thus existing benchmarks have not explored the type of online heterogeneous medical content typically searched by both professional and non professional searchers, they have done this using laboratory style queries which are not representative of the observed querying behaviour of real users.

We describe benchmark creation for medical IR evaluation within the Khresmoi project[1]. Khresmoi aims to develop a multilingual and multimodal search and access system for biomedical information and documents [5]. The project targets three user groups: general public, general practitioners and consultant radiologists. In this paper we focus on medical IR using text search over crawled resources, and hence on the first two user groups. In so doing, we describe two generated benchmarks: the first one has been created from existing query logs for internal research within the Khresmoi project and targets both medical professionals and general public; the second one has been created in the framework of the new CLEFeHealth evaluation campaign as part of the CLEF 2013 benchmark laboratories, and it targets patients only.

This paper is structured as follows: Section 2 presents a brief overview of past and present medical IR evaluation campaigns and the benchmarks used; Section 3 provides an overview of the Khresmoi project; the benchmarks created within the project are described in Section 4 and the future work using these benchmarks is briefly introduced in Section 5.

## 2. MEDICAL IR EVALUATION TO-DATE

In this section, we describe past and current medical IR evaluation campaigns and developed benchmarks. We show that, while these campaigns have been important in facilitating great progress in medical IR, they are very limited in the scope of the medical search tasks addressed and that the behaviour of end users has been overlooked.

### 2.1 Existing Benchmarks

OHSUMED, published in 1994, was the first collection containing medical data used for IR evaluation [6]. The collection contained around 350,000 abstracts from medical journals on the MEDLINE database over a period of five years and two sets of topics: a manually created one and one based on the controlled vo-


*HSD 2013*, August 1, 2013, Dublin, Ireland.

[1] http://khresmoi.eu/

cabulary thesaurus of the Medical Subject Headings[2] (MeSH). The collection was created for the TREC 2000 Filtering Track but also used for other research on health IR [2, 8].

The TREC Genomics Track, which ran between 2003 and 2007, investigated IR systems on biomedical genomics data [10]. This included tasks ranging from ad-hoc retrieval to document categorisation, passage retrieval, and entity-based question-answering. The test collection contained publications from medical journals and clinical reports related to genes and genomics.

Thus while these tasks were important in exploring search for scientific medical purposes they did not address the needs of less scientifically trained searchers.

The ImageCLEFmed Track on medical image retrieval, which ran between 2003 and 2013, provided several tasks supporting evaluation of medical image search [7, 9]. This included tasks on language-independent methods for the automatic annotation of images with concepts; multimodal IR based on the combination of visual and textual features; and multilingual image retrieval methods. The medical task in ImageCLEF concentrated on access to biomedical images in the literature and on the web. Several challenges of automatic image analysis were tackled in this benchmark by a sometimes large variety of participating research groups. While very important in areas where medical images form a vital part of the search data, these activities have again not addressed the more general medical search needs of many users.

The TREC Medical Records Track ran in 2011 and 2012 [11]. This task was based on a collection of de-identified medical records, queries that resembled eligibility criteria of clinical studies, and associated relevance judgements. Records were grouped into visits, corresponding to a patient admission in the hospital; visits ranged in length from a few hours to in excess of a year. The goal of the track was to find patient cohorts that are relevant to the criteria for recruitment as populations in comparative effectiveness studies. Again while an important search task, this activity did not address search over many of the useful and important resources now available.

Recently, NTCIR (NII Test Collection for IR Systems) launched a new campaign, called MedNLP, which aims to extract specific information from Japanese medical reports, written by physicians about imaginary patients[3]. This includes two identification tasks (i.e., personal health information (e.g., name or gender) and complaints or diagnoses) and a "free task", where participants are invited to submit practical or creative solutions to other tasks. This is currently an exploratory activity, and while related to information access, this does not directly address search.

In summary, these previous campaigns have provided resources for evaluating various health IR techniques, aiming to support clinicians and other healthcare workers. Examples include identifying patient cohorts, searching medical images, and coding diagnoses. However, to date evaluation campaigns have not considered more general information needs such as patients and general practitioners information needs.

## 2.2 Towards Representing User Needs in Benchmarks

As shown in the previous section, existing medical IR evaluation benchmarks are highly oriented towards clinicians. Firstly, the datasets are very specialised: either health records, genomics articles, medline abstracts, etc. To our knowledge, existing benchmarks do not provide general health information that would meet the information needs of patients or GPs. Secondly, the queries themselves describe patient cases or are extracted from medical thesauri such as MeSH. As has been observed [12], medical queries tend to be much shorter than those used in existing benchmarks. The lack of resources representing patients and GPs information needs is motivated by several factors. First, it is much more difficult to target their diverse information needs than those of a community of practice such as clinicians due to differences in, for example, their health knowledge and computer skills. Second, they represent a much wider and more heterogeneous subject population than the populations focused on in other campaigns: patients and their relatives may have different interests, different abilities to interpret health information, and different health profiles. For example, diabetes patients may have more health knowledge on this chronic disease than patients with short-term diseases, and diabetic children will most likely wish to retrieve different types of information than their parents. However, finding documents that solve these information needs of laypeople is critical because of the effect incorrect information may have: cybercondria, self-medication, etc.

# 3. KHRESMOI PROJECT

As background to the development of our evaluation tasks, in this section we provide a brief overview of the Khresmoi project of which these form a part. Khresmoi aims to develop a multilingual and multimodal search and access system for biomedical information and documents [5].

The Khresmoi system is composed of multiple interacting component technologies that aim to help a user retrieve valuable medical information adapted to their requirements - preferred language, medical knowledge, etc. System components include machine translation, information retrieval, summarisation, semantic enrichment, spell checking, etc.

## 3.1 Use Cases

The Khresmoi project targets users who speak different languages, have different medical knowledge levels and different levels of knowledge of the language of the documents. Three use cases have been defined and studied in detail: two groups with general medical interests (general public and general practitioners); and one group of clinicians with specialized expertise (radiologists). Each of these groups have been studied within the project and their information needs and search behaviours have been classified through surveys and concrete scenarios [1].

## 3.2 Khresmoi System Evaluation

A major part of the Khresmoi project is the evaluation of Khresmoi technologies as used by our target user groups in order to assess the success and efficiency of Khresmoi project outcomes. Two types of evaluations are being carried out: user-centred evaluation, involving subjects performing predefined tasks on Khresmoi prototypes; and empirical evaluations, for automated assessment of system performance, both in terms of the effectiveness of individual components and the components in combination, and specifically how they interact in combination. Datasets are created to conduct all of these evaluations in a comprehensive and consistent manner.

## 3.3 CLEFeHealth

In order to extend our investigation of medical IR beyond the scope of the Khresmoi project itself, members of the Khresmoi project team are also participating in the organisation of a health

[2] http://www.ncbi.nlm.nih.gov/mesh

[3] http://mednlp.jp/medistj-en

related evaluation workshop: CLEFeHealth[4] as part of the CLEF 2013 benchmarking laboratories. The goal of CLEFeHealth is to evaluate systems that support laypeople in searching for and understanding their health information. CLEF eHealth is comprised of three specific tasks related to information access.

The specific use case considered is as follows: Before leaving hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in hospital.

The first task considered in the workshop aims at extracting names of disorders from the discharge summaries, while the second task requires normalisation and expansion of abbreviations and acronyms present in the discharge summaries. The use case then postulates that, given the discharge summaries and the diagnosed disorders, patients often have questions regarding their health condition. The goal of the third task is to provide valuable and relevant documents to patients, so as to satisfy their health-related information need. One of the features of this scenario is that we are able to identify the patient context in which the search is made from the contents of the discharge report. The role of Khresmoi within CLEF eHealth is as part of the team running the third task.

# 4. MEDICAL IR BENCHMARK CREATION

In this section we describe the benchmarks for medical IR evaluation, developed within the Khresmoi project and the CLEF eHealth workshop. A more detailed description of the benchmark developed for CLEF eHealth is described in [4]. These benchmarks are composed of a document collection, a set of queries and a list of relevant documents for each query. The document collection is shared across both benchmarks and described next. This is followed by details on the query set and relevant document set generation process used in the Khresmoi project test collection.

## 4.1 Khresmoi Document Collection

The Khresmoi document collection consists of a large web crawl of health resources, containing about 1.5 million documents. This collection consists of web pages covering a broad range of health topics, targeted at both the general public and healthcare professionals. These domains consist predominantly of health and medicine websites that have been certified by the Health on the Net (HON) Foundation[5] as adhering to the HONcode principles[6] ( 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[7], Diagnosia[8] and Trip Answers[9].

## 4.2 Khresmoi Query Set

In order to perform some of the evaluations mentioned in Section 3.2, queries have been gathered for two use cases: general public and physicians. To obtain a set of queries representative of what our potential end-users would enter in a search system, we collected queries from existing query logs. For the general public, queries have been gathered from Health on the Net (HON) search engine. This query log contains queries issued in various languages, only the English ones were considered here. The physicians queries come from the Trip database[10] query logs. A set of 50 short general public (1-2 words in length), 50 long general public queries (>2 words in length), and 50 general practitioner (average 3 words in length) queries have been created for each use case. They have been manually selected by medical professionals to be representative of Khresmoi end-users. Moreover, they have been manually corrected (if they contained spelling errors) and translated into Czech, French and German. Classical IR dataset provide a description with each of the query in order to support the relevance assessment process. However, a description of the query can only be given by the author of the query when she is performing the search. As this information cannot be retrieved from query logs, it had to be generated by medical experts from selected queries, by estimating or inferring the likely search context based on their experience, and on the Khresmoi user requirements [1]. A category and a description are added manually to each query, as shown in the following example:

```
<query>
<title lang="en"> involuntary trembling or quivering
</title>
<title lang="fr">Tremblement et palpitation involontaires
</title>
<title lang="ge">unwillkürliches zittern oder zucken
</title>
<title lang="cz">neúmyslný třes a chvění</title>
<category>Symptoms</category>
<desc>results should provide possible health conditions
for which this symptom is known and also treatment
options</desc>
</query>
```

## 4.3 CLEFeHealth Query Set

The queries used in this task aim to model those used by patients to find out more about their disorders, once they have examined a discharge summary. The discharge summaries used for the task originate from the de-identified clinical free-text notes of the MIMIC II database, Version 2.5. Disorders have been identified within discharge summaries and linked to the matching UMLS (Unified Medical Language System) concept.

A query is generated for a given disorder and a discharge summary by nursing medical experts. Medical experts were used in this query generation process to overcome issues with patient privacy and recruitment. We believe that, being in daily contact with patients receiving treatments and discharge summaries, nurses are familiar with patients information needs and patient profiles.

65 disorders were randomly selected from a set of 1,006 disorders identified in CLEF eHealth Task 1. For each disorder, a discharge summary containing the disorder itself has been randomly selected. Using the pairs of disorder and associated discharge summary, the medical experts developed a set of patient queries (and criteria for judging the relevance of documents to the queries, for use in the relevance assessment task described in the next section). Queries are generated in the standard TREC format, consisting of a topic title (text of the query), description (longer description of what the query means), and a narrative (expected content of the relevant documents). A field describing the patient profile has also been added. The following example outlines a query:

```
  <query>
<title> thrombocytopenia treatment corticosteroids length
</title>
<desc> How long should be the corticosteroids treatment
to cure thrombocytopenia?  </desc>
<narr> Documents should contain information about
treatments of thrombocytopenia, and especially
```

[4] http://nicta.com.au/business/health/events/clefehealth_2013
[5] http://www.healthonnet.org
[6] http://www.hon.ch/HONcode/Patients-Conduct.html
[7] http://www.drugbank.ca
[8] http://www.diagnosia.com
[9] http://www.tripanswers.org
[10] http://www.tripdatabase.com/

```
corticosteroids.  It should describe the treatment,
its duration and how the disease is cured using it.
<scenario> The patient has a short-term disease, or
has been hospitalised after an accident (little to
no knowledge of the disorder, short-term treatment)
</scenario>
<profile> Professional female </profile>
</narr>
</query>
```

With this approach, five training and fifty test queries have been generated for use in the task. 65 disorders have been selected (i.e. more than the targeted number of queries) because some disorders/-queries may not be answerable using web pages from the document collection. During the query generation process, the experts manually removed disorders from the list of 65 that do not allow for realistic query generation. CLEF eHealth task participants were allowed to use the discharge summaries along with the query as contextual information.

### 4.4 Relevance Assessments

Relevance assessments for the Khresmoi query set were formed based on pooled sets generated using a combination of existing retrieval approaches. Documents in the pooled result sets have been rated as relevant or irrelevant to the queries by medical experts using details of document relevance given in the description field of each query topic. The relevance of each document was assessed by one expert.

Relevance assessments were conducted for the CLEF eHealth query set after task participants submitted their runs. Each participant was required to submit a baseline run that does not incorporate any advanced techniques (e.g., sophisticated annotation, query expansion, etc. techniques), and could submit up to three additional runs generated using the discharge summaries associated with the queries, and up to three runs using techniques of their choice which do not use the discharge summaries. To add diversity, while keeping the relevance assessment load as light as possible, pooled sets for relevance assessment were generated by merging the top 10 documents from participants baseline run, the best run using discharge summaries and the best run without using them, with duplicates removed. Relevance assessment was conducted on a 4-point scale (3: highly relevant, 2: somewhat relevant, 1: on topic but unreliable, 0: not relevant). Two qrel files were created: one which maintains this graded 4-point scale and one which maps this 4-point scale to a binary scale ($\{3, 2\} \rightarrow 1$: relevant, $\{1, 0\} \rightarrow 0$: not relevant).

## 5. FUTURE WORK

In this paper we described the creation of two new medical IR evaluation benchmarks. These benchmarks are rich resources representative of patients and general practitioners information needs. This benchmark generation also allowed us to investigate the creation of realistic query sets and useful contextual descriptions. This has been done either for existing queries, where the context has to be inferred, and made-up queries, where the context was set by real discharge summaries. While there are no other benchmarks covering such a context, their release represents great potential for improvement of medical IR.

In that sense, CLEF eHealth dataset has been released and 9 teams submitted runs to this campaign. Results were promising and their analysis is described in [3]. Participants results, outputs of the CLEF workshop and other fora will be used to improve the design of the task and the datasets for the 2014 lab.

Within Khresmoi, evaluation of the IR system will be conducted using the Khresmoi test collection described in this paper. Moreover, a set of global empirical evaluations will be performed using this same dataset, in order to evaluate the components interactions and the influence of their performances on each other.

## 6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n$^o$ 257528 (KHRESMOI). The relevance judgements were funded by the ESF project ELIAS.

## 7. REFERENCES

[1] C. Boyer, M. Gschwandtner, A. Hanbury, M. Kritz, N. Pletneva, M. Samwald, and A. Vargas. Use case definition including concrete data requirements (D8.2). public deliverable, Khresmoi EU project, 2012.

[2] V. Claveau. Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain. In *Proceedings of COLING*, 2012.

[3] L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, H. MÃijller, S. SalanterÃd', H. Suominen, and G. Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*, 2013.

[4] L. Goeuriot, L. Kelly, G. J. F. Jones, G. Zuccon, H. Suominen, A. Hanbury, H. Müller, and J. Leveling. Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In *Proceedings of Evaluating Information Access Workshop (EVIA 2013), NTCIR-10 Conference*, 2013.

[5] A. Hanbury, C. Boyer, M. Gschwandtner, and H. Müller. Khresmoi: towards a multi-lingual search and access system for biomedical information. In *Med-e-Tel*, Luxembourg, 2011.

[6] W. R. Hersh, C. Buckley, T. J. Leone, and D. H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR '94*, pages 192–201, 1994.

[7] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, 2011.

[8] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of CIKM 2012*, 2012.

[9] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer, 2010.

[10] P. M. Roberts, A. M. Cohen, and W. R. Hersh. Tasks, topics and relevance judging for the trec genomics track: five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, 12:81–97, 2009.

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

[12] R. W. White, S. T. Dumais, and J. Teevan. How medical expertise influences web search interaction. In *SIGIR*, pages 791–792, 2008.

# Khresmoi Professional: Multilingual Semantic Search for Medical Professionals

Niraj Aswani[i], Thomas Beckers[g], Erich Birngruber[k], Célia Boyer[j], Andreas Burner[k], Jakub Bystroň[h], Khalid Choukri[d], Sarah Cruchet[j], Hamish Cunningham[i], Jan Dědek[h], Ljiljana Dolamic[j], René Donner[k], Ondřej Dušek[h], Sebastian Dungs[g], Ivan Eggel[b], Antonio Foncubierta[b], Norbert Fuhr[g], Adam Funk[i], Alba Garcia Seco de Herrera[b], Arnaud Gaudinat[b], Georgi Georgiev[e], Julien Gobeill[b], Lorraine Goeuriot[f], Paz Gomez[c], Mark A. Greenwood[i], Manfred Gschwandtner[l], Allan Hanbury[a], Jan Hajič[h], Jaroslava Hlaváčová[h], Markus Holzer[k], Gareth Jones[f], Blanca Jordan[c], Matthias Jordan[g], Klemens Kaderk[k], Franz Kainberger[k], Liadh Kelly[f], Sascha Kriewel[g], Marlene Kritz[l], Georg Langs[k], Nolan Lawson[j], Johannes Leveling[f], David Mareček[h], Dimitrios Markonis[b], Iván Martinez[c], Vassil Momtchev[e], Alexandre Masselot[j], Hélène Mazo[d], Henning Müller[b], Michal Novák[h], Johann Petrak[i], João Palotti[a], Pavel Pecina[h], Konstantin Pentchev[e], Deyan Peychev[e], Natalia Pletneva[j], Martin Popel[h], Diana Pottecher[c], Angus Roberts[i], Rudolf Rosa[h], Patrick Ruch[b], Alexander Sachs[l], Matthias Samwald[a], Priscille Schneller[d], Veronika Stefanov[a], Aleš Tamchyna[h], Miguel Angel Tinte[c], Zdeňka Urešová[h], Alejandro Vargas[j], Dina Vishnyakova[b]

[a] Vienna University of Technology, Austria
[b] University of Applied Sciences Western Switzerland
[c] ATOS, Spain
[d] Evaluations and Language Resources Distribution Agency, France
[e] Ontotext, Bulgaria
[f] Dublin City University, Ireland
[g] University of Duisburg-Essen, Germany
[h] Charles University in Prague, Czech Republic
[i] The University of Sheffield, United Kingdom
[j] Health on the Net Foundation, Switzerland
[k] Medical University of Vienna, Austria
[l] Association of Physicians in Vienna, Austria

## ABSTRACT

There is increasing interest in and need for innovative solutions to medical search. In this paper we present the EU-funded Khresmoi medical search and access system, currently in year 3 of 4 of development across 12 partners. The Khresmoi system uses a component-based architecture housed in the cloud to allow for the development of several innovative applications to support target users' medical information needs. The Khresmoi search systems based on this architecture have been designed to support the multilingual and multimodal information needs of three target groups: the general public, general practitioners and consultant radiologists. In this paper we focus on the presentation of the systems to support the latter two groups using semantic, multilingual text and image-based (including 2D and 3D radiology images) search.

## Categories and Subject Descriptors

J.3 [LIFE AND MEDICAL SCIENCES]: Medical information systems

## Keywords

Multilingual, multimodal, medical search system


*HSD 2013*, August 1, 2013, Dublin, Ireland.


*Figure 1 - The Khresmoi Concept*

## 1. INTRODUCTION

The Khresmoi project[1] is developing a multilingual multimodal search and access system for medical and health information and documents. It addresses the challenges of searching through huge amounts of medical data, including general medical information available from various online sources via the internet, as well as 2D and 3D radiology images in hospital archives. The system

[1] http://khresmoi.eu/

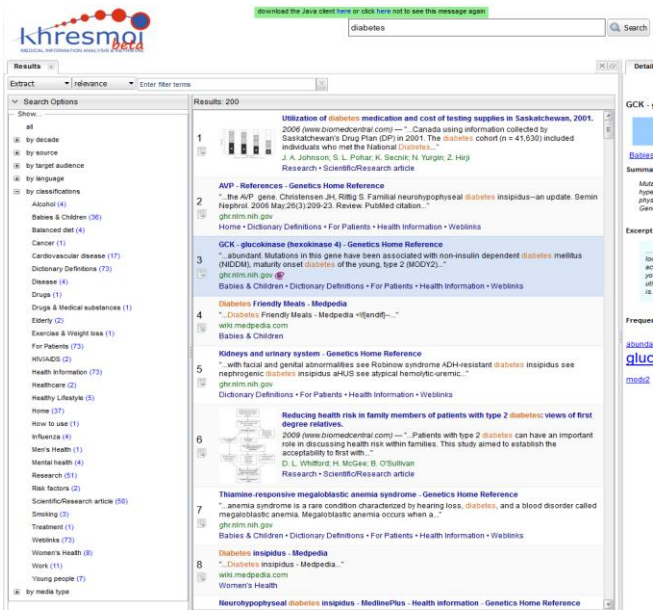



*Figure 2: The web frontend*

allows text querying in several languages with query translation for cross language searching, in combination with image queries. Extensive medical knowledge bases support semantic search. Results can be translated using a machine translation tool specifically trained on medical text.

The system is aimed at three main end user groups: members of the general public, general practitioners and consultant radiologists (a group for whom image search is crucial). An outline of the Khresmoi concept is shown in Figure 1. In this paper we focus on innovative search functionality for general practitioners and consultant radiologists search.

## 2. KHRESMOI TECHNOLOGY

The Khresmoi system has been developed using a Service Component Architecture (SCA) supported by a cloud infrastructure [7]. Components in this system include a search component, a knowledge-base component, a machine translation component, a query disambiguation component, a spell-checking component, etc. An overview of how these components are combined in the architecture is shown in Figure 3. The sources of the documents indexed are shown on the left of the figure (in purple). Text is crawled from websites aimed at the general public and physicians, while images and text are extracted from open access medical journals. 3D images (MR/CT) and associated radiology report text are exported from a Picture Archiving and Communication System (PACS) and indexed to be accessed by radiologists working within the hospital in which this data is stored.

The components that process and index the data are shown in blue. Key components in this architecture are built upon open source software, developed by project partners, which has been significantly advanced by work in Khresmoi to meet the retrieval requirements of medical search systems.

**GATE**[2]: The General Architecture for Text Engineering (GATE) is used to annotate documents at word, section and document levels. Through work in Khresmoi, its capabilities for annotating medical documents have been expanded. The use of cycles of manual correction of the annotations to allow the automatic annotation software to improve the annotation results by learning to correct its errors has also been extensively tested.

**Mimir**[3] uses GATE annotations to perform semantic search. The latest version of Mimir (Mimir 4) includes the ability to rank returned documents.

**ezDL**[4] is a framework for interactive search applications. It has been extended with the capability to display image search results, as well as extensive tools to facilitate collaborative search, such as the ability to share documents and queries between users.

**ParaDISE** is a new visual search engine developed in Khresmoi as a successor to the GNU Image Finding Tool (GIFT). It is more scalable than GIFT due to the use of Hadoop/MapReduce, and contains state-of-the-art image features and visual similarity calculation.

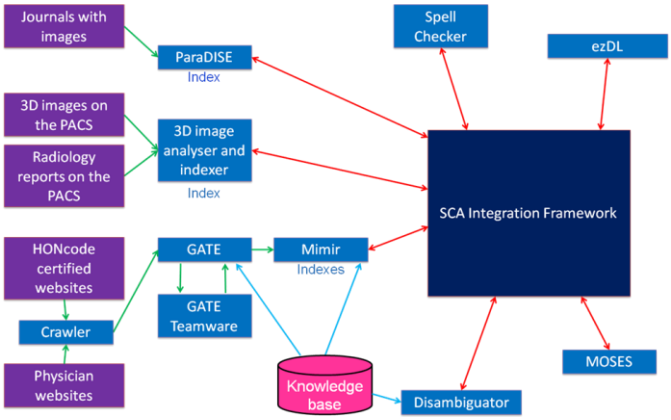



*Figure 3: Overview of the Khresmoi architecture*

Furthermore, the **MOSES** statistical machine translation software[5] has been trained on extensive collections of medical documents to obtain domain-adapted statistical machine translation in the medical domain. The **OWLIM** semantic repository[6] has received performance and functionality upgrades, and has also had its medical knowledge base expanded through the addition of new medical vocabularies and new links between the medical vocabularies.

Finally, technology for analysing 3D CT and MR images is being developed. This allows structures in the images to be automatically identified and mapped to a standard vocabulary. It

[2] https://gate.ac.uk/

[3] https://gate.ac.uk/mimir/

[4] http://ezdl.de/

[5] http://www.statmt.org/moses/

[6] http://www.ontotext.com/owlim

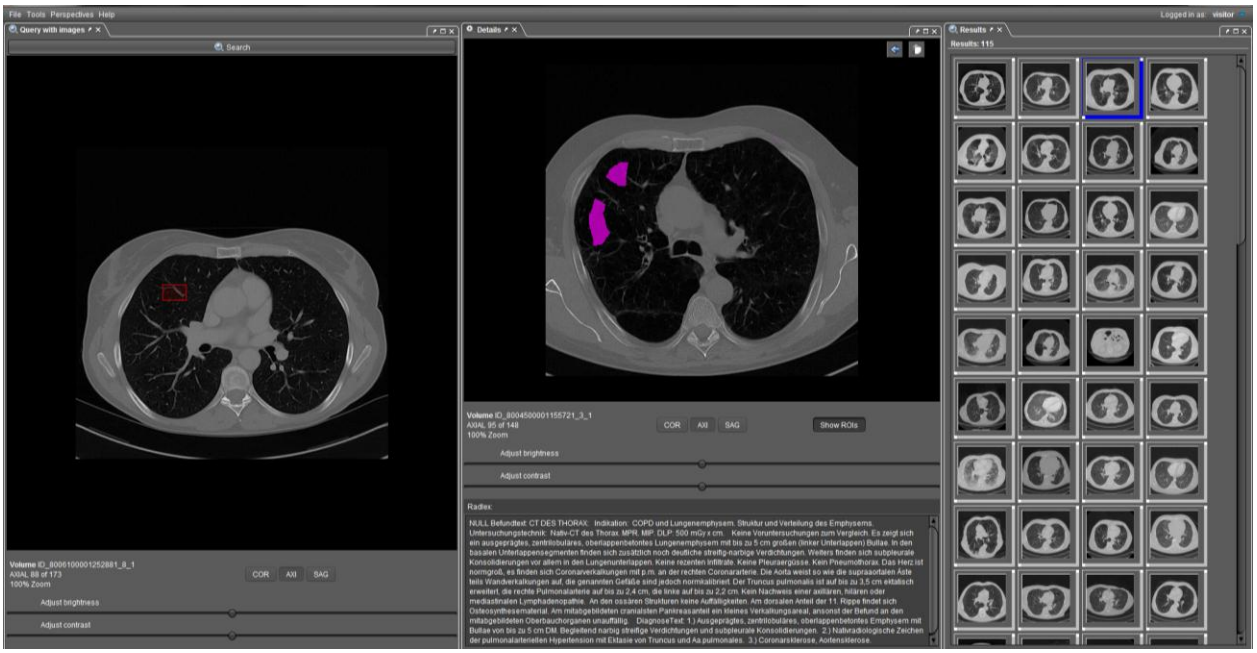


*Figure 4 - Khresmoi interface for radiologists*

also allows retrieval of images to be done based on the visual similarity between images.

The modular integration of multiple software technologies in the system architecture allows for easy development of the required medical search applications, such as the general practitioner and radiologist applications described next.

## 3. MEDICAL SEARCH FOR GENERAL PRACTITIONERS

The Khresmoi search prototype for general practitioners combines technologies developed within the Khresmoi project into an integrated platform. It currently provides two user interfaces. One is a browser based web application[7] written in GWT while the other is a Java Swing desktop application[8]. Both share a common backend service infrastructure also written in Java. A third user interface for Android devices is currently under development.

Crawled websites with trustworthy medical information targeted at practitioners are semantically annotated using GATE technology, and then indexed by Mimir, described in previous section. In addition, images from medical publications are also indexed (using ParaDISE, described in previous section).

The web frontend (see Figure 2) features basic functionality including running text-based searches, filtering and sorting of result sets. It also includes a facet explorer, which provides a means to quickly filter the result set using the available metadata attributes. The system offers spelling corrections and disambiguation suggestions while a user is typing a query. The system marks documents that the user has viewed with an eye icon (visible in the highlighted result in Figure 2) to allow the user to more easily keep track of the progress of the search process. Result sets may include images which can be used to trigger searches for visually similar images.

Users can store retrieved documents in a tray for later inspection. The personal library is a permanent storage for documents of various formats and is available to all registered users. Queries are recorded and can easily be reissued by utilizing a separate view in the interface.

The interface consists of several components each containing an aspect of the system's functionality, which are arranged in predefined layouts suitable for the most common user tasks. All components can be (un-)hidden from the perspective, re-sized and moved in the interface.

The Swing interface includes all features of the web prototype. In addition, the desktop client has collaborative features, which registered users can use to share documents with other users or user groups. For scientific work import and export of document Bibtex records is supported.

Both interfaces are fully internationalised for all Khresmoi project languages, including English, German, French, Spanish and Czech, as well as for Chinese and Vietnamese. The system also offers translation suggestions for search terms – for example, a German-speaking physician wishing to search in the English literature can type in the search terms in German and selected the proposed translations into English. The user may also machine translate the summaries of documents back into their selected language.

[7] http://khresmoi.is.inf.uni-due.de:8182/
[8] http://khresmoi.is.inf.uni-due.de/khresmoi.jnlp

The prototype for radiologists was created based on the Swing version of the interface. It shares the same technological basis and is described in the next section.

## 4. MEDICAL SEARCH FOR RADIOLOGISTS

Similar to the Khresmoi search for general practitioners, the Khresmoi search system for radiologists combines technologies developed within the Khresmoi project into an integrated platform. This system is for use by radiologists in medical institutions, allowing for the search and comparison of 2D and 3D radiology images. Given the sensitive nature of the medical imagining data of patients, the system is not publicly available. However, a demo can be viewed[9].

Figure 4 presents the interface instantiated for use by radiologists (also noticeable by the colour scheme adapted to the radiology requirements). Note that this is the same interface framework shown in Figure 2, but with different tools visible. Here the query is the selected area of the image slice shown in the left panel. The images in the panel on the right are returned based on their visual similarity to the region marked in the query. In the central panel, the selected image is shown with the region corresponding best to the query region highlighted. The associated radiology report is shown below the central image. For this application, only the images stored in the archives of the hospital in which the system is used are indexed. However, the possibility to do a visual search of 2D images from the medical literature is also provided.

A use case for this system is that a radiologist faced with an unusual or unknown structure in an image can query the hospital archives for images containing a similar structure, and use the (anonymised) radiology reports associated with these images to guide the reading of the image.

## 5. CONCLUDING REMARKS

Medical search systems are required by different classes of individuals – from members of the general public with differing levels of medical knowledge and a range of search and language skills, to numerous classes and types of medical professionals. The Khresmoi project satisfies this need through the development of different instantiations of the multilingual multimodal medical search systems for different classes of users. In this paper we present the Khresmoi system, with a particular focus on *Khresmoi Professional* which currently provides multilingual semantic, text and image based medical search applications for two classes of users: general practitioners and consultant radiologists.

These systems were developed in a holistic way, taking into consideration users' needs and requirements as determined by extensive questionnaires and analyses conducted within the Khresmoi project [3, 5]. Rounds of user-centered evaluation at both the interface component and interface system level have been, and continue to be, used for iterative system refinement [1, 2, 4]. These evaluations use general practitioners and radiologists to perform realistic tasks, to enable development of systems which function for the target users in the most useful ways.

In addition, the backend system components are empirically evaluated using document and image collections and generated search test collections from the Khresmoi project, which represent real users' information needs and querying behaviours [8]. As part of this analysis a novel global empirical evaluation is being conducted to measure the impact of components of the system on each other, and importantly how the performance of these components in isolation and combination impact on the information displayed to users and on the end user experience [6].

[9] http://youtu.be/UnPs7NSet1g

## 6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 257528 (KHRESMOI).

## 7. REFERENCES

[1] Baroz F, Boyer C, Gschwandtner M, Goeuriot L, Hajic J, Hanbury A, Kritz M, Palotti J, Pletneva N, Ruiz de Castañeda R, Sachs A, Samwald M, Schneller P, Stefanov V, Uresova Z, Report on user tests with initial search system, Khresmoi Deliverable D10.1, khresmoi.eu/assets/Deliverables/WP10/KhresmoiD101.pdf (last accessed: May 2013)

[2] Beckers T, Dungs S, Fuhr N, Goeuriot N, Ignalski J, Jordan M, Kelly M, Kriewel S, Report on results of the WP3 first evaluation phase, Khresmoi Deliverable D3.2, khresmoi.eu/assets/Deliverables/WP3/KhresmoiD32.pdf (last accessed: May 2013)

[3] M. Kritz, M. Gschwandtner, V. Stefanov, A. Hanbury, M. Samwald, Utilization and Perceived Problems of Online Medical Resources and Search Tools among Different Groups of European Physicians, Journal of Medical Internet Research, 15(6):e122, 2013.

[4] Markonis D, Holzer M, Baroz F, Luis Ruiz De Castaneda R, Langs G, Boyer C, Mueller H, Report on the results of the initial user test of the radiology search system, Khresmoi Deliverable D10.2, khresmoi.eu/assets/Deliverables/WP10/KhresmoiD102.pdf (last accessed: May 2013)

[5] Müller H. Report on image use behavior and requirements. Khresmoi Deliverable D9.1, khresmoi.eu/assets/Deliverables/WP9/KhresmoiD91.pdf (last accessed: May 2013)

[6] Meta-analysis of the first phase of empirical and user-centered evaluations. Khresmoi Deliverable D7.2, August 2013 (to appear)

[7] Prototype and evaluation of the "Full Cloud Infrastructure" Khresmoi Deliverable D6.4.3, August 2013 (to appear)

[8] L. Goeuriot, L. Kelly, G. J. F. Jones, G. Zuccon, H. Suominen, A. Hanbury, H. Müller, J. Leveling, Creation of a New Evaluation Benchmark for Information Retrieval Targeting Patient Information Needs, Proc. 5th International Workshop on Evaluating Information Access (EVIA), 2013.

# Query Expansion using open Web-based SKOS Vocabularies

Flávio Martins
Universidade Nova de Lisboa
Fac. Ciências e Tecnologia
Lisbon, Portugal
flaviomartins@acm.org

Bernhard Haslhofer
University of Vienna
Computer Science
Vienna, Austria
haslhoferb@acm.org

João Magalhães
Universidade Nova de Lisboa
Fac. Ciências e Tecnologia
Lisbon, Portugal
jm.magalhaes@fct.unl.pt

## ABSTRACT

Achieving high precision in search engines is particularly difficult when dealing with specific technical domains, such as the medical domain, where vocabulary mismatch problems are more prone to occur. Query expansion using lexical or semantic relations is a well-known technique that can attenuate the vocabulary mismatch problem as demonstrated by earlier research. However, existing approaches do not exploit potentially connected open Web-based vocabularies that are increasingly being expressed using the Simple Knowledge Organization System (SKOS).

In this paper, we propose a query expansion technique that exploits term labels and semantic relationships in such vocabularies to improve search results. We evaluated this technique using a SKOS representation of the Medical Subject Headings (MeSH) and the TREC-9 Filtering datasets. Our results show that our SKOS-based query expansion technique improves the P@10, nDCG@10 and MAP metrics across various retrieval models.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [retrieval models, query formulation, search process]; H.3.1 [**Content Analysis and Indexing**]: [thesauruses]

## General Terms

Algorithms, Experimentation, Performance

## Keywords

Search, Query Expansion, Thesaurus

## 1. INTRODUCTION

An information retrieval system can automatically refine user queries by exploiting the semantic relationships between terms and either reformulate queries on the fly or expand terms when documents are being indexed. Modern IR systems already allow the inclusion of thesauri such as WordNet or support query expansion via automatic thesauri generation. However, thesauri used in existing approaches are not Web-based and do not distinguish between the notion of identifiable *concepts* and associated *terms*.


*HSD* 2013, August 1, 2013, Dublin, Ireland.

Query expansion using a knowledge organization system allows a solid representation of expert knowledge containing real-world concepts that are referred to using different terms and with multiple different relations. Hundreds [10] of knowledge organization systems needed for query expansion are now expressed using the Simple Knowledge Organization System (SKOS) [11] and become available as Linked Data on the Web. As a result, definitions of real-world concepts are represented as connected Web resources, which are identified by dereferencable URIs and have term definitions attached. One of such cases is the Medical Subject Headings (MeSH)[1], which is used in large publications such as PubMed, and is now accessible in SKOS containing over 24,626 concepts, and 150,617 concept-labels. This leads us to our research question: *how can SKOS vocabularies be exploited for query expansion in information retrieval systems.*

We propose an approach that uses open, Web-based vocabularies for term expansion either at query or indexing time. In addition to well-known label-based expansion techniques it also supports URI-based query expansion, which is useful when documents such as metadata descriptions contain URI references to concepts defined in SKOS vocabularies. Our approach is agnostic to the vocabulary generation approach and supports manually curated or automatically generated thesauri as long as they are expressed in SKOS.

The central contributions of this paper are: (1) a general SKOS-based query expansion method and a description of how it influences different retrieval models; (2) an open-source implementation of a SKOS-based query expansion model[2]; and (3) an evaluation of our proposed method using the TREC-9 Filtering dataset and the Medical Subject Headings (MeSH) SKOS vocabulary. Our results show that the P@10, nDCG@10 and MAP metrics improve, with our expansion method, across various retrieval models.

## 2. RELATED WORK

Typical query expansion techniques [3] take advantage of statistical correlation, synonyms, the knowledge of the morphology of words, and the use of dictionaries to improve the quality of search results. Voorhees [14] found that automatic expansion using synonym sets from WordNet can degrade performance, while hand picking concepts improves short or ambiguous queries. Hersh et al. [4] described OHSUMED, a medical test collection with judgments, based on journals from the National Library of Medicine (NLM). Later in [5],

[1] `http://www.nlm.nih.gov/mesh/`
[2] `https://github.com/behas/lucene-skos`

Hersh et al. expanded with child terms from MeSH because journals in OHSUMED are indexed using the narrowest indexing terms. They also found that simply adding the new expansions to the query is a simple yet effective approach.

Broader techniques have been proposed since, to cover more formal descriptions of domain-specific thesauri. This field has been quite active as is described by Bhogal et al. [2] who reviewed several ontology-based query expansion techniques. Trying to balance between simple thesauri and ontology representations, the Simple Knowledge Organization System (SKOS) [11] is a standard model for sharing and linking controlled vocabularies (thesauri, taxonomies, classification systems, etc.) on the Web. It defines a set of classes and properties to represent vocabularies as Web graphs. The fundamental building block of a SKOS vocabulary is the *Concept*. Each *Concept* has an unique URI identifier and associated preferred indexing terms, alternative terms, and semantic relations. Thus, SKOS provides a mechanism to fully encode domain-specific thesauri.

When expanding a query term, the relationships among terms also play a key role in the process. Bai et al. [1] confirmed the importance of term-relationships expansion, and also proposed to expand queries based on the relationship between a set of terms and a single term. Thus, SKOS also provides multiple types of relationships to better assist in the query-expansion process: the *skos:altLabel* property is used to declare additional synonyms, abbreviations and acronyms for the concept. The use of broader-narrower relationships between concepts creates a hierarchically organized graph of conceptual resources. In addition, the property *skos:related* can be used to assert a non-hierarchical, associative relationships between two concepts. Figure 1, illustrates an expanded query with multiple types of expansions, e.g. *skos:broader*, *skos:prefLabel* and *skos:altLabel*.

More recent approaches, have mastered several techniques for integrating thesaurus-like resources in query expansion. Lin and Demner-Fushman [6] and Zhou et al. [15] investigated the creation of custom similarities to score concept matches with more emphasis than word matches. Zhou et al. used the MeSH thesaurus with one hierarchical level of relationships and synonyms. In our case, we provide support for multiple levels (MeSH SKOS has 12 hierarchical levels) and multiple relation categories. Theobald et al. [12] developed a system that performs expansion only when thematic similarity is above a threshold, but found that it implies a high execution cost and hard-tuning of parameters.

# 3. SKOS-BASED QUERY EXPANSION

An information retrieval system supporting SKOS-based query expansion loads SKOS vocabularies and expands incoming queries based on the term definitions and semantic relationship in these vocabularies. Figure 1 shows an example query "thrombocytopenia in gestation" expanded by MeSH definitions. The system found that "gestation" and "pregnancy", as well as "thrombocytopenia" and "thrombopenia" are defined as being synonyms (*skos:prefLabel, skos:altLabel*). It also found that "blood platelet disorders" is a broader term (skos:broader) for "thrombocytopenia". These terms are then added to the query, causing queries that would not match previously, such as "thrombopenia in pregnancy", to match. To consider the type of expansion for scoring and ranking, it keeps an attribute indicating the type of SKOS property that caused the expansion.

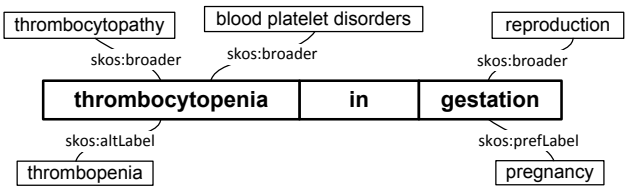



**Figure 1: SKOS Query Expansion example**

Our proposed SKOS-based query expansion approach consist of three major building blocks: *term expansion*, *scoring* and *weighting* based on term expansion types.

## 3.1 Term expansion

Expanding terms requires that SKOS vocabularies are loaded by an information retrieval system. This can be performed either by dereferencing the URIs of concept definitions and following links to related concepts (crawling) or downloading and expanding packaged vocabularies (dump load) from a given URI. In both cases the system requires a single URI to bootstrap SKOS-based query expansion.

Term expansion at query time is part of the query analysis process and performed after the query is tokenized and before any stemming or lemmatization is applied. For each token in the user query the SKOS query expander looks up possible expansions in its internal SKOS representation, which could be represented as an inverted index having SKOS labels in the dictionary and concept URIs in the posting list. Matching expansion terms are then added to the internal query representation in which each expansion term carries an attribute indicating the expansion type, as shown in the previous example. This allows domain experts to fine-tune the influence of expanded terms on scoring of query results according to the needs in their information retrieval scenario.

## 3.2 Scoring

Term expansion can lead to a large number of terms added to the original query. While this should generally increase retrieval effectiveness, it might also cause *query drift*, which means that an expanded query does not reflect the user's initial information need anymore. Instead of applying expensive pruning methods, our proposed scoring model leverages the explicit declaration of expansion types in query representations. This allows the configuration of to what extent certain SKOS definitions contribute to the final score.

Our approach uses regular text retrieval functions, which are minimally modified to enable scoring based on the expansion type. The following formula shows our scoring method applied to the $tf$–$idf$ retrieval model. The score of a query in relation to a given document can be conveyed as follows:

$$Score(q,d) = coord_{q,d} \cdot \left( \sum_{t \in q} (tf_{t,d} \cdot idf_t) + \sum_{c \in expansions(q)} (tf_{c,d} \cdot idf_c \cdot boost_{c_{type}}) \right)$$

$expansions(q)$ denotes the set of expansions found for query $q$, $coord_{q,d}$ is a score factor calculated based on the number of terms matching the document and $boost_{c_{type}}$ is a boost factor that varies with the expansion type of $c$.

With the $coord_{q,d}$ factor, we ensure that a document with more matching query terms will score higher. The score for documents that match more alternative terms, broader, narrower or related concepts resulting from the expansion of the original query will also be higher.

Since not all query terms are equally important, especially when they were not explicitly entered by the user, as it is the case with query expansion, original query terms and terms resulting from expansions have to be weighted differently. This can be controlled by the boost factor $boost_{c_{type}}$. It balances the contributions of the original and expanded query terms to the final score and allows flexible weighting schemes based on expansion types. One could, for instance, weight *skos:related* expansion types less than synonym definitions (*skos:altLabel*). An optimal setup is a combination of boost values, one for each SKOS expansion type.

Similarly, we also implemented this scoring method for the BM25, BM25L, BM25+ and LTC retrieval models. Due to lack of space we omit the description of these functions.

# 4. EVALUATION

We evaluated our approach on the TREC-9 Filtering track dataset, which is a modified version of the OHSUMED dataset. Queries were expanded using the MeSH vocabulary expressed in SKOS. In our tests we restricted the expansion to preferred terms and alternative terms. The overall performance was measured using the average performance over the complete set of queries. We focused on three metrics: precision, normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (MAP). In our evaluation only documents judged as *definitely relevant* were considered for precision and MAP calculations.

## 4.1 Datasets

The **TREC-9 Filtering dataset** is a subset of the MEDLINE database with about 350,000 references from 270 journals covering four years from 1987 to 1991. Each reference contains common bibliographic metadata, including record fields for title, authors and abstract. In addition, each document contains indexing terms from the Medical Subject Headings. The collection includes three-level relevance judgments, which were assigned to documents by human assessors. The TREC-9 Filtering dataset contains 63 queries from the original OHSUMED dataset.

The **Medical Subject Headings (MeSH)** is a controlled vocabulary managed by the U.S. National Library of Medicine (NLM) and is used to index millions of articles in the MEDLINE database. Physicians and medical librarians can use the terms in this vocabulary in their search activities to find the most relevant documents. MeSH is structured in various levels establishing thesaurus-like hierarchical relationships. Van Assem et al [13] made available a SKOS version of the MeSH thesauri.

## 4.2 Results

The evaluation is split into two parts: first, we study the performance of our proposed method for $tf$–$idf$ and BM25; second we extend the evaluation of SKOS-based query expansion to several state-of-the-art retrieval models.

### 4.2.1 Standard retrieval models

The results of our initial evaluation with $tf$–$idf$ and BM25 are presented in Table 1. The performance of the system without expansion is labeled "No Exp." and the best result with SKOS query expansion is labeled "SKOS". The latter is obtained by repeating the benchmarks several times to find the best performing value for the boost parameter, for each retrieval model and each retrieval metric, using a random 30% sample of all queries. This parameter value is then used to benchmark the full set the queries. Using the full set of queries we present charts depicting this procedure in Figure 2(a) and Figure 2(b) where we can see the boost values where performance is best.

For $tf$–$idf$ P@10 improves 6.2% while nDCG@10 is improved by 5%. The performance with BM25 improves 5.2% for P@10 and 5.4% for nDCG@10. The improvements for MAP can be seen in Table 2, for all retrieval models evaluated. The improvements with $tf$–$idf$ and BM25 are of 8.3% and 5.3% respectively.

### 4.2.2 Improved retrieval models

In contrast to Lu et al. [7], we formalized MeSH concepts as SKOS concepts and evaluated other retrieval models besides $tf$–$idf$. First, we chose $tf$–$idf$ with logarithmic term frequency normalization, also known as LTC weighting scheme. Secondly, we also considered BM25L [9] and BM25+ [8] by Lv et al. that aim to eliminate the issues with BM25 when ranking documents with heterogeneous length.

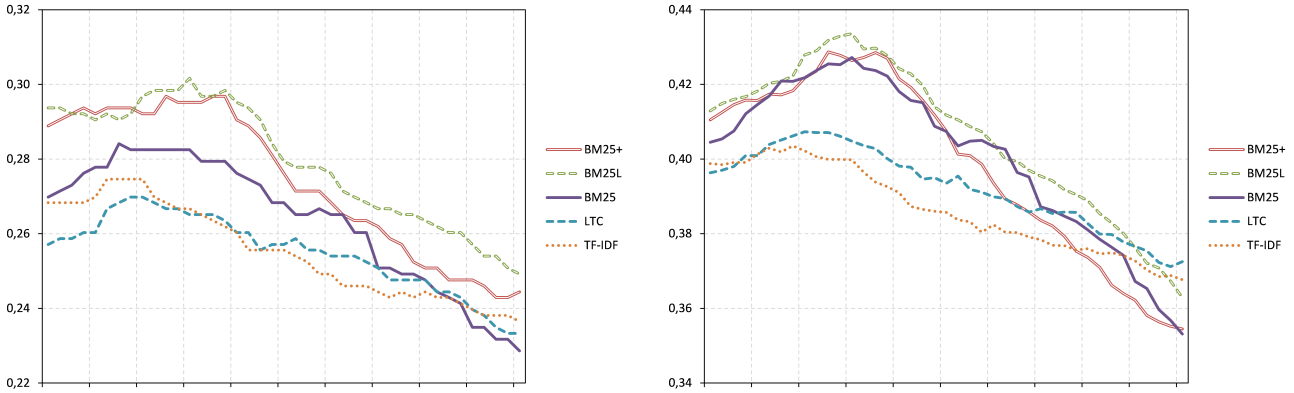
As can be seen in Table 1, P@10 is best with BM25L and BM25+, which reached 0.298 and 0.297 respectively. In Figure 2(a), we can see that all functions have its optimal boost value in the range between 0.3 and 0.8. There is also a steep decline in precision for all BM25-based functions when the boost value passes 0.8. The decline for the $tf$–$idf$-based functions is not as strong. We believe BM25L and BM25+ properties concerning heterogeneous document lengths also provides a better weighting of the long queries generated by our query expansion method.

BM25L marginally edges out BM25+, as can be seen on Figure 2(b), reaching an nDCG@10 result of 0.433 while BM25+ gets 0.426, with a boost of 0.6 and 0.5 respectively. The performance of the BM25-based functions starts to decline for boost values greater than 0.7, and a bit earlier for the $tf$–$idf$-based functions. The BM25 models are more robust as they are able to maintain good retrieval performance across a larger range of boost values. In the end, the best function overall is the BM25L with the best boost around the values of 0.55 and 0.6.

# 5. CONCLUSIONS

Finding semantically related words for query terms and adding them to internal query representations allows information retrieval systems to overcome mismatches between the users' vocabulary and the indexed documents. The proposed SKOS-based query expansion approach offers a straightforward way of making use of Web-based vocabularies to improve retrieval effectiveness across retrieval models. Evaluation with the TREC-9 Filtering dataset, demonstrated that SKOS-based query expansion is best implemented with the BM25L retrieval model, due to its support for the long queries generated by the expansion and for its robustness to expansion weights variations.

In this paper we reported the results we obtained from evaluating our approach and implementation against the TREC-9 Filtering dataset and the MeSH vocabulary. At the moment, we are expanding to further domains.




(a) P@10 for various boost values.

(b) nDCG@10 for various boost values.

**Figure 2: Term weights analysis with SKOS-MeSH expansions.**

| | TF-IDF | | LTC | | BM25 | | BM25L | | BM25+ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@10 | nDCG@10 | P@10 | nDCG@10 | P@10 | nDCG@10 | P@10 | nDCG@10 | P@10 | nDCG@10 |
| No Exp. | 0.259 | 0.382 | 0.260 | 0.386 | 0.270 | 0.405 | 0.294 | 0.413 | 0.289 | 0.411 |
| SKOS | 0.275 | 0.401 | 0.270 | 0.407 | 0.284 | 0.427 | 0.298 | 0.433 | 0.297 | 0.426 |

**Table 1: P@10 and nDCG@10 results.**

| | TF-IDF | LTC | BM25 | BM25L | BM25+ |
|---|---|---|---|---|---|
| No Exp. | 0.172 | 0.180 | 0.202 | 0.215 | 0.218 |
| SKOS | 0.183 | 0.194 | 0.212 | 0.224 | 0.227 |

**Table 2: MAP results.**

## 6. REFERENCES


[1] J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. In *SIGIR'05*, pages 688–695. ACM, 2005.

[2] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886, 2007.

[3] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.

[4] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94*, pages 192–201, Aug. 1994.

[5] W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In *Proceedings of the AMIA Symposium*, page 344, Jan. 2000.

[6] J. Lin and D. Demner-Fushman. The role of knowledge in conceptual retrieval. In *SIGIR '06*, pages 99–106, Aug. 2006.

[7] Z. Lu, W. Kim, and W. J. Wilbur. Evaluation of query expansion using mesh in pubmed. *Information retrieval*, 12(1):69–80, 2009.

[8] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *CIKM '11*, pages 7–16, Oct. 2011.

[9] Y. Lv and C. Zhai. When documents are very long, BM25 fails! In *SIGIR '11*, pages 1103–1104, July 2011.

[10] N. A. A. Manaf, S. Bechhofer, and R. Stevens. The current state of skos vocabularies on the web. In *ESWC*, pages 270–284, 2012.

[11] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. Recommendation, W3C, 2009.

[12] M. Theobald, R. Schenkel, and G. Weikum. Efficient and self-tuning incremental query expansion for top-k query processing. In *SIGIR '05*, pages 242–249, Aug. 2005.

[13] M. Van Assem, V. Malaisé, A. Miles, and G. Schreiber. A Method to Convert Thesauri to SKOS. *The Semantic Web: Research and Applications*, pages 95–109, 2006.

[14] E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94*, pages 61–69, Aug. 1994.

[15] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR '07*, pages 655–662, July 2007.

# Systems for Improving Electronic Health Record Note Comprehension

Balaji Polepalli Ramesh
University of Wisconsin Milwaukee
3200 N Cramer St, Milwaukee, WI
+1-414-229-4677
brpjr@uwm.edu

Hong Yu
University of Massachusetts Medical School, VA Central Western Massachusetts
368 Plantation St, Worcester, MA
+1-508-856-3474
hong.yu@umassmed.edu

## ABSTRACT
Allowing patients access to their physicians' notes has the potential to enhance their understanding of disease and improve medication adherence and healthcare outcomes. However, a recent study involving over ten thousand patients showed that allowing patients to read their electronic health record (EHR) notes caused confusion, especially for the vulnerable (e.g., lower literacy, lower income) groups. This finding is not surprising as EHR notes contain medical jargon that may be difficult for patients to comprehend. To improve patients' EHR note comprehension, we are developing a biomedical natural language processing system called NoteAid (http://clinicalnotesaid.org), which translates medical jargon into consumer-oriented lay language. The current NoteAid implementations link EHR medical terms to their definitions and other related educational material. Our evaluation has shown that all NoteAid implementations improve self-rated EHR note comprehension by 23% to 40% of lay people.

## Categories and Subject Descriptors
**H.1.2 [User/Machine Systems]**: Human information processing

## Keywords
Electronic Health Records, Consumer Health, Information Retrieval, Natural Language Processing

## 1. INTRODUCTION
Allowing patients direct access to their electronic health record (EHR) notes has been shown to enhance medical understanding and improve medication adherence and healthcare outcomes [1]. However, a recent study involving over ten thousand patients showed that allowing patients to read their EHR notes caused confusion, especially for vulnerable (e.g., lower literacy, lower income) groups [1].

The level of a patient's EHR note comprehension depends on his/her level of health literacy. The Institute of Medicine defined health literacy as "the degree to which individuals have the capacity to obtain, process, and understand basic information and services needed to make appropriate decisions regarding their health"[2]. However, the average American reads at or below an 8th grade level, and over 90 million Americans have limited health literacy [3]. Jones et al. (1992) [4] showed that 50% patients do not understand at least one term in their medical problem list. Lober et al. (2006) [5] found that medical terminology presented a barrier to almost one third of population. EHR notes contain complex medical conditions; abbreviations and other medical jargon that make it hard for patients to comprehend. Example 1, below, shows an excerpt from a de-identified EHR progress note. A patient might find it hard to understand the abbreviations "hx", "ED", "SOB" and the medical terms "psoriasis" and "bronchitis".

***Example 1:*** `A patient with` **`hx`** `of active tobacco abuse,` **`bronchitis`**`, and` **`psoriasis`** `presented to` **`ED`** `earlier today with c/o` **`SOB`**`, mild wheezing, chest congestion and chills.`

We are therefore developing a biomedical natural language processing (NLP) system called NoteAid that translates medical jargon into lay language. Studies have shown that patient education is effective in improving health literacy, decreasing disease severity, improving self-management behaviors, and reducing hospitalizations [6-8]. We hypothesize that NoteAid will improve patients' comprehension of their clinical notes and ultimately their healthcare outcome. In our previous evaluation study [9] we reported that NoteAid system improved self-rated EHR note comprehension. In that evaluation, a subject was provided with a clinical note without and with the NoteAid system (in that order) and was asked to report comprehension score of the note. The evaluation design introduced an ordering bias. In this study, we eliminate the ordering bias by randomly assigning EHR note, either with or without the NoteAid system and examine patient EHR comprehension with the systems.

## 2. RELATED WORK
There is a rich literature related to health literacy and comprehension. A substantial amount of work has been done to compile consumer health vocabulary (CHV) [10], [11]. Elhadad (2006) [12] provided definitions of unfamiliar terms and found that such an approach significantly improved reader's comprehension of online news stories. Zeng-Treitler et al. (2007) [13] designed and implemented a text translator that identifies difficult terms and replaces them with easier-to-read synonyms. Approaches have been also been developed to predict term familiarity with linguistic/stylistic features [14], term frequency [12], machine learning [15] as well as machine translation [16]. Tools have also been developed to simplify EHR note content using both syntactic and semantic approaches (e.g.,[17], [18]). Smith et al. (2011) [19] improved coherence by manually rewriting a clinical note and found increased comprehension by lay people.

InfoButtons [20] and the Patient Clinical Information System (PatCIS) [21] provided patients with online information resources and educational material [22]. However, the education material was manually compiled by the researchers after reading the EHR notes. In contrast, the NoteAid system [9] automatically extracts


*HSD 2013*, August 1, 2013, Dublin, Ireland

complex medical jargon from EHR notes and links them to patient education material.

# 3. MATERIALS AND METHODS

## 3.1 The NoteAid System

As shown in Figure 1, the NoteAid system has three modules Concept Identifier (CI), Definition Locator (DL) and Definition Filter (DF). CI processes the input text and maps terms to the corresponding UMLS concepts. DL fetches definitions from Medline Plus, Unified Medical Language System (UMLS) or Wikipedia (Wiki) if the term definition is found. We improved the quality of definitions fetched by Wiki by adding a DF, which fetches a definition if article is health-related. Wiki assigns each article a set of categories, which are organized into a direct acyclic graph. We recognize an article as health-related if any of the assigned category or the corresponding hierarchical categories belong to the following two terms: clinical and health.

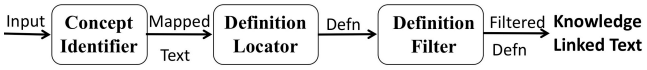



**Figure 1 – Schematic representation of the NoteAid system**

## 3.2 Evaluation Procedure and Metrics

In this study, we evaluated four NoteAid implementations: MedlinePlus (linking EHR concepts to definitions in Medline Plus), UMLS (linking EHR concepts to their synonyms and definitions in the Unified Medical Language System), Wikipedia (Wiki, linking EHR concepts to health related articles in Wikipedia) and the hybrid system that integrates the three aforementioned implementations.

### *3.2.1 Subjects*

With the IRB approval, we recruited subjects from the Amazon Mechanic Turk (AMT). We used AMT because the subjects have various background and qualifications, and therefore are representative in terms of health literacy. Many research studies use AMT for data collection and survey and have proven to be a reliable resource [23].

### *3.2.2 Evaluation Data*

We randomly selected 20 de-identified progress notes (PGN) from the Pittsburgh NLP repository [24]. We measure the readability of notes with Flesch-Kincaid grade level [25]. Lower Flesch-Kincaid grade level implies higher readability.

### *3.2.3 Evaluation Process*

We evaluate a total of five systems: the baseline system in which a clinical note is presented without a NoteAid implementation and four NoteAid implementations where a note is presented with a NoteAid implementation. We recruited 25 subjects, 5 subjects for each of the 5 systems. Each subject was asked to evaluate 20 PGN notes. For each note (either the note alone or with a NoteAid implementation), we asked the subject to read and score his/her level of comprehension on a scale of 1-5. Each subject was given a link to a welcome page describing the study, followed by demographic information page, qualifying question page, pages containing EHR notes to evaluate and finally a thank-you page along with the validation code.

For quality control, we gave each subject a question related to his/her evaluation data. The evaluation was hosted and the evaluation results were stored on a local server. At the end of the evaluation subjects received a code to confirm their participation in the study and receive payment for the task. Each subject spent 30~40 mins to complete the entire evaluation and s/he was paid $5.

Four subjects failed to complete the evaluations. Our results were based on the analyses of the evaluation of the remaining 21 subjects who completed their tasks.

### *3.2.4 Evaluation Criteria*

We report the average self-rated comprehension scores and used the Mann-Whitney-Wilcoxon test to compare the comprehension score with and without a NoteAid implementation and between different NoteAid implementations.

In order to evaluate whether the self-rated comprehension scores represent readability, we report the correlation between the Flesch-Kincaid grade level [25] and self-rated comprehension scores. We also evaluate whether education is correlated with self-rated comprehension scores and report the Pearson coefficient value between the education level (four scales in decreasing order: Master, Bachelor, Associates, and High School).

### *3.2.5 Demographic Information of Subjects*

Twenty-one subjects (9 female and 12 male) completed the evaluation. The number of White American (White), Asian, and Black American (Black) were 15, 4, and 2 respectively. The subjects in the study had a wide range of educational (Edu) backgrounds. Six (28.57%) subjects had a Masters (MA, MS) degree and 6 (28.57%) had Bachelors (BA, BS) degree each, 2 (9.52%) of them had an Associate (Asso) degree and the remaining 7 (33.34%) subjects had a high school diploma (High Sch). Table 1 details the demographic information of subjects as well as their education status.

**Table 1 –Demographic information of the study subjects**

| | Systems | Notes Alone | Medline Plus | UMLS | Wiki | Hybrid |
|---|---|---|---|---|---|---|
| Edu | High Sch | 1 | - | 2 | 3 | 1 |
| | Asso | - | 1 | - | - | 1 |
| | BA, BS | 1 | 2 | - | 1 | 2 |
| | MA, MS | 1 | 1 | 2 | 1 | 1 |
| Se | Female | - | 3 | 3 | 1 | 2 |
| | Male | 3 | 1 | 1 | 4 | 3 |
| Race | White | 2 | 2 | 3 | 4 | 4 |
| | Asian | 1 | 1 | - | 1 | 1 |
| | Black | - | 1 | 1 | - | - |

# 4. RESULTS

The 20 PGNs comprise of 473 sentences, 4862 words and have an average Flesch Kincaid Grade Level of 9.8. Table 2 below shows the average comprehension scores of PGNs without any NoteAid implementation and with each of the four NoteAid implementations. The average comprehension score of subjects and Flesch-Kincaid grade level had a spearman ranked correlation coefficient of rho =-0.77 ($p<0.01$).

As shown in Table 2, all NoteAid implementations improved self-rated PGN note comprehension and the improvements were statistically significant ($p<0.01$, the Mann-Whitney-Wilcoxon test). The difference in comprehension scores between different NoteAid implementation was not statistically significant except for the difference between the MedlinePlus and the UMLS implementations ($p<0.01$, the Mann-Whitney-Wilcoxon test). Table 2 also shows the number of concepts identified by each of the NoteAid implementations.

**Table 2 - The average self-rated comprehension values (average ± std dev) and number of concepts identified by NoteAid implementations. (*p<0.01)**

| System | Notes Alone | Medline Plus | UMLS | Wiki | Hybrid |
|---|---|---|---|---|---|
| Score | 2.95 ±0.67 | 4.12* ±0.33 | 3.63* ±0.57 | 3.85* ±0.47 | 3.92* ±0.40 |
| # conc | NA | 52 | 352 | 436 | 476 |

Figure 2 shows the average self-rated comprehension scores of all NoteAid implementations for every PGN note, and Figure 3 shows a scatter plot of the average self-rated comprehension scores with notes alone and notes with the MedlinePlus implementation. The results as shown in both figures demonstrate a strong and consistent improvement of self-rated comprehension scores with NoteAid implementations for every PGN note.

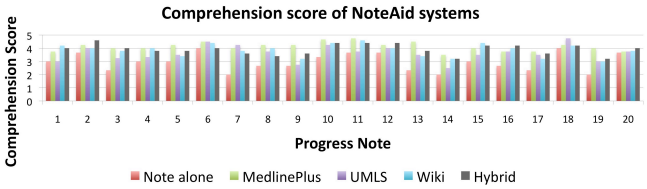



**Figure 2 - The average self-rated comprehension score for each note with different NoteAid system implementations**

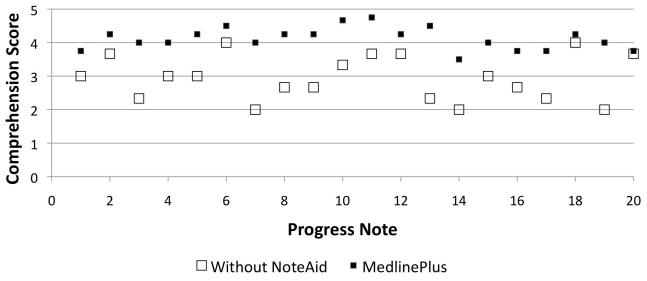



**Figure 3 – Scatter plot of average self-rated comprehension scores with note alone and with the MedlinePlus NoteAid implementation**

The Pearson coefficient values between the subject education level and comprehension scores are: Note Alone: 0.98, MedlinePlus: 0.31, UMLS: 0.71, Wiki: -0.47 and Hybrid: 0.04.

## 5. DISCUSSION AND FUTURE WORK

Our results show that, when the clinical notes are presented alone, the self-rated comprehension scores are highly correlated (0.98 Pearson coefficient) with the education levels of the subjects. The results support the validity of self-rated comprehension scores. In contrast, the correlation results are mixed with different NoteAid implementations. While the UMLS has a correlation value of 0.71, the MedlinePlus and Hybrid implementations decrease to 0.31 and 0.04. The Wiki implementation has a negative correlation: -0.47. Several factors may contribute to the results. First, the definition quality of the UMLS, Wiki, MedlinePlus and Hybrid resources are not yet evaluated and it is unclear whether the definitions correctly represent the semantic meanings of the notes. Secondly, although providing definitions may help comprehension, providing too much or unnecessary information (such as Wiki) may hurt those who have a better education level. In future work, we need to conduct a comprehensive "think aloud" evaluation study to understand the behavior of users. We will also need to evaluate the quality of definitions of different NoteAid implementations and patient comprehension by replacing complex medical jargon with its equivalent lexical lay term variants [26,27] in EHNs.

The significant improvement of MedlinePlus over the UMLS implementation may be attributed to the lower readability of content in UMLS. For example, the definition of "malnutrition" is complex in UMLS and has a Flesch-Kincaid grade level of 19. Whereas the definition of "malnutriton" in Medline Plus has a Flesch-Kincaid grade level of 11. Similarly, Wiki has a grade level of 13 for "malnutrition." Although the improvement in comprehension of Wiki over MedlinePlus implementation was not statistically significant, Wiki content may not be accurate as discussed earlier.

Our results show that all four NoteAid implementations improved EHR note self-rated comprehension significantly over Notes alone. The results are largely consistent with our previous evaluation [9] in which NoteAid implementations were evaluated in a before-and-after fashion but there are differences between the two evaluation results. In our previous study, we found that the Wikipedia implementation had the largest improvement and that the MedlinePlus implementation decreased the self-rated comprehension scores. Such discrepancy can be explained by the limitations of our study.

First, we report subjects' self-rated note comprehension but did not evaluate to what extent they accurately comprehended the note content. The before-and-after evaluation design [9] may be a better model as we force a subject to read the EHR note prior to her/his exposure to the improved note (note+NoteAid). Our before-and-after evaluation results are also consistent with the number of concepts recognized by each NoteAid implementation. The MedlinePlus implementation has the least number of concept recognition, and therefore its comprehension improvement may be small. A randomized design, as we have done in this study, may provide an evaluation subject little incentive for comprehending the note content. In the future work, we will test subjects' comprehension based on content analyses of every clinical note. Furthermore, we will evaluate subjects' health literacy [28].

Secondly, the number of subjects is small in this study. As a result, we can't evaluate the impact of moderators. For example, the data size is not well rounded to conclude that the subjects' education levels impact self-rated comprehension scores. Other limitations of the study include that lay people performed our evaluation but not patients who comprehend their own EHR notes.

## 6. CONCLUSION

Our evaluation results show that NoteAid improves EHR note self-rated comprehension of lay people in a randomized evaluation study and the MedlinePlus implementation demonstrated the highest improvement.

## 7. ACKNOWLEDGEMENTS

We thank Tamsin Maxwell for providing helpful comments. Research reported in this publication was supported in part by 1R01GM095476 to Hong Yu, by a start-up fund from University of Massachusetts Medical School to Hong Yu, and by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR000161. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

# 8. REFERENCES


[1] T. Delbanco, J. Walker, S. K. Bell, J. D. Darer, J. G. Elmore, N. Farag, H. J. Feldman, R. Mejilla, L. Ngo, J. D. Ralston, S. E. Ross, N. Trivedi, E. Vodicka, and S. G. Leveille, "Inviting Patients to Read Their Doctors' Notes: A Quasi-experimental Study and a Look Ahead," *Ann. Intern. Med.*, vol. 157, no. 7, pp. 461–470, Oct. 2012.

[2] Lynn Nielsen-Bohlman, Allison M. Panzer, David A. Kindig, Editors, Committee on Health Literacy, *Health Literacy: A Prescription to End Confusion*. Washington, D.C.: The National Academies Press, 2004.

[3] "The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy," 06-Sep-2006. [Online]. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006483. [Accessed: 22-May-2013].

[4] R. B. Jones, S. M. McGhee, and D. McGhee, "Patient on-line access to medical records in general practice," *Health Bull. (Edinb.)*, vol. 50, no. 2, pp. 143–150, Mar. 1992.

[5] W. Lober, B. Zierler, A. Herbaugh, S. Shinstrom, A. Stolyar, E. Kim, and Y. Kim, "Barriers to the use of a Personal Health Record by an Elderly Population," *AMIA. Annu. Symp. Proc.*, vol. 2006, pp. 514–518, 2006.

[6] R. L. Rothman, D. A. DeWalt, R. Malone, B. Bryant, A. Shintani, B. Crigler, M. Weinberger, and M. Pignone, "Influence of patient literacy on the effectiveness of a primary care-based diabetes disease management program," *Jama J. Am. Med. Assoc.*, vol. 292, no. 14, pp. 1711–1716, Oct. 2004.

[7] D. A. DeWalt, R. M. Malone, M. E. Bryant, M. C. Kosnar, K. E. Corr, R. L. Rothman, C. A. Sueta, and M. P. Pignone, "A heart failure self-management program for patients of all literacy levels: a randomized, controlled trial [ISRCTN11535170]," *BMC Health Serv. Res.*, vol. 6, p. 30, 2006.

[8] D. Schillinger, M. Handley, F. Wang, and H. Hammer, "Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: a three-arm practical clinical trial," *Diabetes Care*, vol. 32, no. 4, pp. 559–566, Apr. 2009.

[9] B. Polepalli Ramesh, T. Houston, C. Brandt, H. Fang, and H. Yu, "Improving Patients' Electronic Health Record Comprehension with NoteAid," presented at the MedInfo 2013, Copenhagen.

[10] B. Smith and C. Fellbaum, "Medical WordNet: a new methodology for the construction and validation of information resources for consumer health," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 371.

[11] *Consumer Health Vocabulary http://www.consumerhealthvocab.org/*. .

[12] N. Elhadad, "Comprehending technical texts: predicting and defining unfamiliar terms," *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, pp. 239–243, 2006.

[13] Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, and D. Rosendale, "Making texts in electronic health records comprehensible to consumers: a prototype translator," *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, pp. 846–850, 2007.

[14] G. Rosemblat, R. Logan, T. Tse, and L. Graham, "Text Features and Readability: Expert Evaluation of Consumer Health Text," *MEDNET*.

[15] Q. Zeng, E. Kim, J. Crowell, and T. Tse, "A text corpora-based estimation of the familiarity of health terminology," *Biol. Med. Data Anal.*, pp. 184–192, 2005.

[16] Q. Zeng-Treitler, H. Kim, G. Rosemblat, and A. Keselman, "Can multilingual machine translation help make medical record content more comprehensible to patients?," *Stud. Health Technol. Inform.*, vol. 160, no. Pt 1, pp. 73–77, 2010.

[17] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A Semantic and Syntactic Text Simplification Tool for Health Content," *AMIA. Annu. Symp. Proc.*, vol. 2010, pp. 366–370, 2010.

[18] G. Leroy, J. E. Endicott, O. Mouradi, D. Kauchak, and M. L. Just, "Improving perceived and actual text difficulty for health information consumers using semi-automated methods," *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, vol. 2012, pp. 522–531, 2012.

[19] C. A. Smith, S. Hetzel, P. Dalrymple, and A. Keselman, "Beyond Readability: Investigating Coherence of Clinical Text for Consumers," *J. Med. Internet Res.*, vol. 13, no. 4, Dec. 2011.

[20] J. J. Cimino, G. Elhanan, and Q. Zeng, "Supporting infobuttons with terminological knowledge," *Proc Amia Annu Fall Symp*, pp. 528–32, 1997.

[21] J. J. Cimino, V. L. Patel, and A. W. Kushniruk, "The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records," *Int. J. Med. Inf.*, vol. 68, no. 1–3, pp. 113–127, Dec. 2002.

[22] J. J. Cimino, V. L. Patel, and A. W. Kushniruk, "What do patients do with access to their medical records?," *Stud. Health Technol. Inform.*, vol. 84, no. Pt 2, pp. 1440–1444, 2001.

[23] J. Proulx, S. Kandula, B. Hill, and Q. Zeng-Treitler, "Creating Consumer Friendly Health Content: Implementing and Testing a Readability Diagnosis and Enhancement Tool," presented at the HICSS, 2012.

[24] W. Chapman, *University of Pittsburgh NLP Repository (http://www.dbmi.pitt.edu/nlpfront)*. .

[25] L. Si and J. Callan, "A statistical model for scientific readability," in *Proceedings of the tenth international conference on Information and knowledge management*, 2001, pp. 574–576.

[26] L. Deléger and P. Zweigenbaum, "Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora," in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, Singapore, 2009, pp. 2–10.

[27] N. Elhadad and K. Sutaria, "Mining a lexicon of technical terms and lay equivalents," in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, Stroudsburg, PA, USA, 2007, pp. 49–56.

[28] R. M. Parker, D. W. Baker, M. V. Williams, and J. R. Nurss, "The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills," *J. Gen. Intern. Med.*, vol. 10, no. 10, pp. 537–541, Oct. 1995.

# Towards a Gold Standard for the Evaluation of Health Recommender Systems

Martin Wiesner, Monika Pobiruchin, Daniel Pfeifer
Dept. of Medical Informatics, Heilbronn University
Max-Planck-Str. 39
74081 Heilbronn, Germany
{martin.wiesner, monika.pobiruchin, daniel.pfeifer}@hs-heilbronn.de

## ABSTRACT

During the last decades computer scientists devoloped systems that collect huge amounts of data in large clinical databases (e.g., laboratory results, disease codes, reports...) representing an individual's state of health. For this reason, the amount of digital information available for decision making has increased drastically.

On the other hand, there is a need for improved personalized delivery of medical content. Yet, a manifold medical vocabulary used in discharge letters for example, poses a major obstacle for laymen. Recommender Systems can be adapted to cope with the special requirements specific for the health domain. Such systems are referred to as Health Recommender Systems (HRS). Thus, it is possible to compute and deliver potentially relevant information items.

However, the evaluation of an HRS remains an open task. In this paper we present an approach to evaluate such a system in a controlled study. The construction of a gold standard for case related recommendations (i.e., test collection) is described. Additionally, a statistical test for our setting is presented and potential risks are discussed.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Selection process; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks; J.3 [**Life And Medical Sciences**]: Medical Information Systems, Health

## General Terms

Algorithms, Theory

## Keywords

Health Recommender Systems, Information Needs, Relevance Computation, Graph Theory, Health Care


*HSD 2013*, August 1, 2013, Dublin, Ireland.

## 1. INTRODUCTION

Increasing health information needs and changes in information seeking behavior can be observed around the globe [1]. According to recent studies more than 80% of the US-American internet users search for health information online [2]. However, information overload and irrelevant information are major obstacles for drawing conclusions on the personal health status and taking adequate actions. Faced with a large amount of medical information on different channels (e.g., news sites, web forums, etc.) users often get lost or feel uncertain when investigating on their own. Improved personalized delivery of medical content could support users in finding relevant information [3], [4].

Though, a manifold and heterogeneous medical vocabulary poses a major obstacle for laymen. Therefore, better suited software systems are needed which interactively support users in finding answers instead of tons of search engine hits. Recommender Systems can be adapted to cope with the special requirements of the health domain. Such systems are referred to as Health Recommender Systems (HRS).

At the same time, huge amounts of medical facts are contained in clinical records and hold a lot of medical evidence. Since the end of the 20th century computer scientists devoloped systems focussed on collecting data in large clinical databases (e.g., laboratory results, disease codes, reports, etc.) representing an individual's state of health. For this reason, the amount of digital information available for decision making has increased drastically [5].

However, the evaluation of an HRS remains an open task. In this paper we describe an approach to evaluate such a system via a gold standard in a controlled study as part of our current research activities.

## 2. RELATED RESEARCH

Different approaches exist to compute personalized medical recommendations. Many of them use methodologies and techniques originating from the fields of Information Retrieval and research on Recommender Systems. These are discussed briefly in the next section.

### 2.1 Information Retrieval (IR)

Traditional IR methodologies [6] rely on term matching strategies. Typically, a similarity coefficent ($sc$) between a set of query terms $Q = \{q_1, .., q_i\}$ and potentially matching documents $D_j$ - represented by characteristic terms - is computed: $sc(Q, D_j)$. In standard approaches like Salton's well known vector space model (VSM, [7]) for an $sc > 0$ some query terms $q_i \in Q$ must be present in $D_j$. When comput-

ing *sc*, the frequency and the specifity of matching terms $q_i$ are taken into account. Yet, classic IR approaches do not provide means to perform semantic context resolution. Term-based search engines and user formulated queries seem to be inappropriate to compute relevant information related to the medical domain [8]. This is because often medical terms from $Q$ cannot be found in a related document $D_j$ even though $Q$ and $D_j$ are semantically close. To overcome such drawbacks, techniques like query expansion, reformulation and suggestion have been introduced in previous studies [9].

## 2.2 Recommender Systems (RS)

Recommender systems suggest items of interest to users of information systems or e-business systems and have evolved in the recent decades. A typical and well known example is Amazon's suggest service for products.

A basic form of content recommendation is provided by consumer-centric web portals for medical information, for example symptoms and diseases. Hereby, there is a risk of information overload for laymen. Additionally, it is difficult to offer relevant information "when users have not specified what they exactly want" [10]. However, if users of web portals have an account and a medical profile (i.e., a health record) linked to it, a RS could provide matching health information artifacts of higher individual relevance.

A lot of research has been invested in recommender systems [11], [12]. Related solutions make intensive use of results from other fields, such as IR. Fernandez-Luque, Karlsen and Vognild found "challenges and opportunities" in terms of using recommender technology as a means to educate the uneducated health individual [13]. They suggested the use of so-called Computer-Tailoring Health Education Systems.

RS which primarily focus on the medical domain are the subject-matter of our research activities [14]. Our implementation of an HRS makes use of a graph data structure related to health concepts derived from *Wikipedia*. This Health Graph $G$ is directed, typed and weighted and represents a semantic network associating semantically close medical terms via edges.

## 2.3 Evaluation of IR Systems

Various methods to evaluate the effectiveness of IR systems exist [15], [16]. Often, IR studies use a test collection which serves as a baseline to map queries to items of interest. Yet, most publicly available test collections rely on news paper articles and are therefore not especially tailored for the use in the assessment of an HRS[1]. Hence, we describe the construction of a medical RS test collection (gold standard for case related recommendations) in Section 4.

# 3. CONCEPTS & SCOPE OF AN HRS

For a successful integration into any health related information system, it is important to consider the system context of an HRS. As depicted in Figure 1, a health professional or patient has access to medical data persisted in a database of a personal health record (PHR) system. A profile-based HRS component is used as an extension of an existing system. Thus, it is possible to compute and deliver potentially relevant information items.

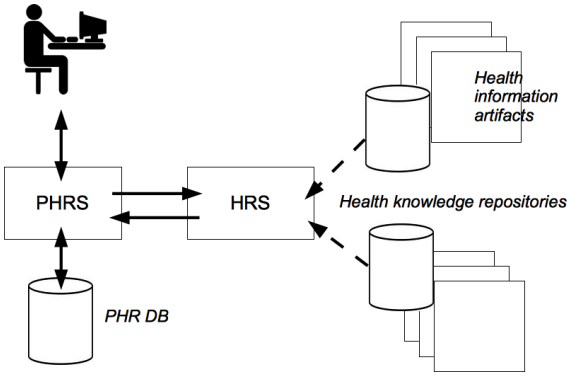



**Figure 1: System context of an HRS-enabled PHR: A set of relevant items is computed from an evidence-based content store (health knowledge repositories).**

## 3.1 Health Information Artefacts

Health information artifacts are found in large health knowledge repositories and can be obtained by various resources. Some resources mainly focus on health professionals (e.g., Pubmed ), yet others are more consumer-centric (e.g., MedlinePlus[2] , InformedHealthOnline[3] or WebMD[4]) and the vocabulary in such resources differs from expert-centric ones. Consumer-centric medical content is often freely available and may consist of:

- expert-proven advisory on how to cope with a disease
- disease definitions in general which support in understanding medical terminology
- care plans which might prevent patients from acting against rules suggested by evidence based medicine
- hints on healthier living

All these facets can be found in consumer-centric content. Therefore, such medical content can be recommended to end-users in a personalized manner.

## 3.2 Challenges & Requirements

Data entries of medical records are frequently stored as unstructured plain text. This creates the following difficulties for IR term matching approaches:

- A retrieval (i.e., recommendation) process must be able to cope with
  (a) imprecise terms (e.g., Hepatitis: chronic viral Hepatitis),
  (b) colloquial terms (e.g., Period: Menstruation) and
  (c) misspellings (e.g., Diabedis: Diabetes mellitus)
- The system must also recognize expert vocabulary and classification system codes primarily used by physicians and other health professionals, such as
  – STEMI for Myocardial Infarction
  – *I21* or *I22* (ICD-10) for 'Myocardial Infarction'

Such obstacles can result in less specific recommendations when integrating classic IR approaches into electronic/personal health record systems.

Therefore, our approach uses semantic query expansion techniques to enrich concept terms found in health record entries to reformulate any query. Further details on the recommendation process and $G$ are found in [17].

[1] i.e., classic test collections are not derived from evidence-based medical records which often contain medical abbreviations, see Section 3.2

[2] see: http://www.nlm.nih.gov/medlineplus
[3] see: http://www.informedhealthonline.org
[4] see: http://emedicine.medscape.com

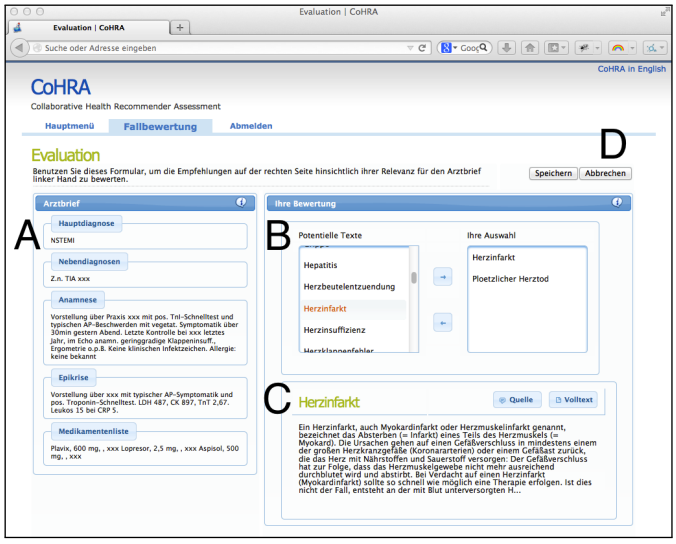



**Figure 2: The main view of our HRS assessment system. Health professionals select matching items based on their expertise.**

# 4. EVALUATION APPROACH

At first, health professionals will have to assess the relevance of potential matches for a certain clinical case $c_i$. During this first phase, a gold standard of human expert recommendations is obtained. The second phase of the evaluation involves at least two implementations of an HRS component:

(1) A traditional HRS implementation based on well known techniques of IR, in particular on a standard implementation of the VSM using the *Apache Lucene* library. Thus, document sets $D_{VSM}$ are obtained.

(2) An advanced HRS implementation which uses query expansion techniques via $G$ and features such as negation recognition. Thus, document sets $D_{HRS}$ are obtained.

Both implementations will compete against each other in matching the recommendations made by the human experts. Thereby, the retrieval precision $\rho$ of both systems will be measured and compared against each other. A statistical test will then reveal which system performed better in computing the ideal set of documents from the content repository $R_c$. First experiments indicate that there might be an improvement provided by our advanced HRS approach. Figure 3 illustrates all phases of the study and related tasks.

## 4.1 Data Sources

On the one hand, about 27,000 fully anonymized, real-world discharge letters (provided by the *Heidelberg University Hospital*) are available in our letter repository $R_l$. The letters revolve around the field of cardiology and typically contain text sections including anamnesis, diagnoses, laboratory results, outcomes of procedures and recommended medication. These medical facts represent the foundation for our experiment, as they provide semi-structured text data which has to be processed by any HRS component.

On the other hand, the medical content (i.e., health information artifacts) to be rated by the group of physicians is provided by the German Institute for Quality and Efficiency in Health Care *IQWiG*[5]. It is an independent publisher of evidence-based consumer health and patient information. This collection of documents comprises a total of about 800 health information artifacts written especially for laymen. It is guaranteed to be of high quality and the result of evidence-based medicine. A subset of 75 documents which are relevant for the field of cardiology is presented to the participants of the study.

[5]see: http://www.iqwig.de

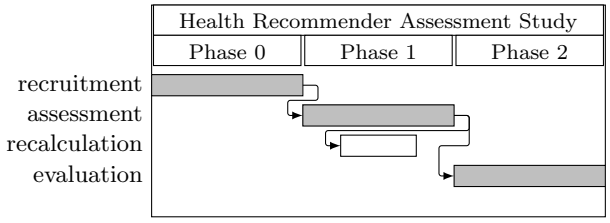



**Figure 3: Timeline of the study phases and related tasks. Grey: mandatory task - White: optional task**

## 4.2 Web-based Assessment System

The primary aim of the study is to capture the recommendations made by the group of health professionals for a set of randomized cases from $R_l$. For this purpose we implemented a browser-based assessment system, as depicted in Figure 2. A cardiologic case $c_i \in C$ is presented (A) on the left. Every physician selects items from a candidate list (B), displayed on the right of the current case. For every candidate item a preview and a fulltext version is available in the box below the list of items (C). Finally, if the health professional has made his selection the current case $c_i$ gets closed (D), selected expert documents are persisted and the next case is presented. Thereby, a set of expert recommendations $D_G$ for all $c_i$ is obtained.

## 4.3 Setting & Statistical Test

During phase two of the evaluation, both implementations will compute matching documents pairwise, i.e., for every $c_i \in R_l$ we will compute the sets $D_{HRS,c_i}$ and $D_{VSM,c_i}$. It is obvious to evaluate the retrieval precision $\rho$ of both approaches compaired against the set of documents $D_{G,c_i}$ which originate from the gold standard recommendations previously obtained. As we will evaluate a large number of cases (i.e., $n_c \geq 100$) a normal distribution can be assumed for this setting. To test if our HRS implementation approach outperforms the basic VSM implementation we postulate:

$$\rho_H > \rho_V \tag{1}$$

Hence, a dependent *t-Test* for paired samples can be formulated for a one-sided case:

$$\begin{aligned} H_0 &: \rho_H \leq \rho_V \\ H_a &: \rho_H > \rho_V \end{aligned} \tag{2}$$

with a significance level of $\alpha = 0.05$ and a target power of $1 - \beta = 0.8$. Equation (2) can be reformulated to:

$$\begin{aligned} H_0 &: \mu_H - \mu_V \leq \omega_0 \\ H_a &: \mu_H - \mu_V > \omega_0 \end{aligned} \tag{3}$$

given that $\mu_H, \mu_V$ are the mean values of $\rho_H$, $\rho_V$. $\omega_0$ represents the expected effect value (i.e., the change in terms of retrieval precision). For the time being we assume an effect size of:

$$\omega_0 = \Delta\rho = \rho_H - \rho_V = 0.1 \tag{4}$$

i.e., an improvement in retrieval precision of 10% is expected.

## 4.4 Sample Size Estimation & Recruitment

A pre-study sample size calculation for $\omega_0 = 0.1$ indicates that a total of $n_c = 620$ cases[6] is needed to ensure that a target power of $1 - \beta = 0.8$ is achieved.

Yet, there is a certain risk for the data collection being biased (e.g., loss of interest of participants, different levels of expertise, physicians being pressed for time, etc.) during phase one of the evaluation. This is especially true if every evaluation case $c_i$ would be assessed by just a single physician. In order to prevent any sort of such biases we plan to assess every $c_i$ at least with an interjudge agreement factor $f_j$ of 2. Thus, the required number of physicians is obtained:

$$n_p = \frac{n_c * f_j}{c_p} \tag{5}$$

As a consequence, we have to recruit at least $n_p = 25$ physicians to achieve the target power of 0.8, for a modest $\omega_0 = 0.1$ and the number of cases $c_i$ assessed per participant $c_p = 50$.

# 5. DISCUSSION

As outlined in Section 4, our HRS assessment study is clearly described. Unfortunately, there is some uncertainty which results from a yet unkown effect size, i.e., we cannot exactly determine what improvement in retrieval precision can be expected. As this effect size influences the absolute number of cases to be evaluated, this problem makes any prior estimation difficult.

The latter poses a risk to the success of the study, as there might be some dropouts during the recruitment phase already. Consequently, a large effort has to be put into the recruitment phase of physicians which tackle the amount of work to be done. However, there is a chance for an early recalculation of the sample size during phase one if the actual effect size of $\omega_0$ should be higher than initially estimated, e.g., $\omega_0 = 0.15$. Thereby, less physicians would be needed to take part in the evaluation. Moreover, $c_p$ could be lowered or the interjudge agreement factor $f_j$ could be increased.

Additionally, a bias might occur as the data in $R_l$ orignate from the field of cardiology. This might have an effect on how physicians decide which items are relevant. A high number of well selected participants and an accompanying study manual should compensate for this.

# 6. CONCLUSION

An approach of integrating RS into electronic health record systems was presented briefly. System requirements and realization concepts for such a component were discussed.

The need to evaluate the quality of the proposed recommendation approach, mandates a controlled experiment in which a sophisticated system is compared to a traditional implementation via a gold standard. In a first phase we plan to leverage the expertise of an expert group of physicians in order to develop the gold standard. For this purpose we implemented a web-based assessment system. Thus, a suitable set of individually tailored recommendations represented by a set of laymen-friendly texts can be derived per case. During the second phase two different HRS approaches will get evaluated against the gold standard. For a large number of cases, the retrieval precision is computed and a dependent *t-Test* for paired samples will be applied. A statistical method for sample size calculation was used to estimate how many cardiologic cases should be evaluated to achieve a target power. Nevertheless, it is difficult to estimate the effect size, here the expected retrieval precision ($\Delta\rho$).

[6]for $\omega_0 = 0.15 \rightarrow n_c = 277$ and $\omega_0 = 0.2 \rightarrow n_c = 156$ accordingly

# 7. REFERENCES


[1] Vahideh Zarea Gavgani. Health information need and seeking behavior of patients in developing countries' context; an iranian experience. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 575–579, New York, NY, USA, 2010. ACM.

[2] Pew Internet & American Life project. Health topics - 80% of internet users look for health information online. http://pewinternet.org/Reports/2011/HealthTopics.aspx, 2011.

[3] Melanie Swan. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *International journal of environmental research and public health*, 6(2):492–525, 2009.

[4] Haggai Roitman et al. Increasing patient safety using explanation-driven personalized content recommendation. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 430–434, New York, NY, USA, 2010. ACM.

[5] H Müller, A Hanbury, N Al Shorbaji, et al. Health information search to deal with the exploding amount of health information produced. *Methods of information in medicine*, 51(6):516, 2012.

[6] David Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series)(2nd Edition)*. Springer, 2nd edition, 2004.

[7] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[8] Qing T. Zeng et al. Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Assoc.*, 13(1):80 – 90, 2006.

[9] Kamran Munir et al. Ontology assisted query reformulation using the semantic and assertion capabilities of OWL-DL ontologies. In *Proceedings of the 2008 intern. symp. on DB engineering applications*, IDEAS '08, pages 81–90, New York, NY, USA, 2008. ACM.

[10] Deepak Agarwal et al. Content recommendation on web portals. *Communications of the ACM*, 56(6):92–101, 2013.

[11] Paul Resnick and Hal Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, 1997.

[12] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[13] L Fernandez-Luque, R Karlsen, and L K Vognild. Challenges and opportunities of using recommender systems for personalized health education. *Stud Health Technol Inform*, 150:903–907, 2009.

[14] Martin Wiesner and Daniel Pfeifer. Adapting recommender systems to the requirements of personal health record systems. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 410–414, New York, NY, USA, 2010. ACM.

[15] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[16] S.E. Robertson and M.M. Hancock-Beaulieu. On the evaluation of IR systems. *Information Processing & Management*, 28(4):457 – 466, 1992. Special Issue: Evaluation Issues in Information Retrieval.

[17] Martin Wiesner, Stefan Rotter, and Daniel Pfeifer. Leveraging semantic networks for personalized content in health recommender systems. In *CBMS, 2011 24th Intern. Symp. on Computer-Based Medical Systems*, june 2011.

# Towards Intelligent and Socially Oriented Query Recommendation for Electronic Health Records Retrieval

Danny T.Y. Wu
School of Information
University of Michigan
tzuyu@umich.edu

Lei Yang
School of Information
University of Michigan
yangle@umich.edu

Qiaozhu Mei
School of Information
University of Michigan
qmei@umich.edu

David A. Hanauer
Department of Pediatrics
University of Michigan
hanauer@umich.edu

Kai Zheng
Department of Health Management and Policy, University of Michigan
Kzheng@umich.edu

## ABSTRACT
In this paper, we describe the functionality of a proof-of-concept electronic health records (EHR) search engine, EHR-SE, which is developed under a National Library of Medicine contract to test the effectiveness of (1) an intelligent medical search query recommendation service that expands or suggests alternative search terms to improve search results and (2) a social search feature enabling the share of EHR search knowledge among medical professionals. We present a summary of the results of evaluating the prototype system through a comprehensive, two-week user experiments. The participants provided positive feedback on the overall performance of the system as well as the relevance of recommended query terms. In addition, the participants were highly satisfied with the social search feature and the usability of the system's user interface. We conclude that the provision of query recommendation and social search features has great potential to enhance the efficiency and effectiveness of searching for medical data stored in EHRs.

## Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Medical Information Systems

## General Terms
Algorithms, Experimentation.

## Keywords
Electronic Health Records, Query Recommendation, Social Search, Collaborative Search.

## 1. INTRODUCTION
Rich, detailed data are being captured through the clinicians' daily use of electronic health records (EHR). The data can be leveraged for various secondary uses such as population health management, epidemic surveillance, and translational research, not only to improve the quality of care, but also to create a "rapid learning" healthcare system [1]. These potential benefits, however, cannot be achieved without being able to search and retrieve the data effectively and efficiently.


*HSD 2013*, August 1, 2013, Dublin, Ireland.

In general, patient care data stored in EHRs are either codified or in a free-text, narrative format. Codified data are usually preferred due to the easiness of use for subsequent computational analysis, while unstructured data remain pervasive due to the freedom that allows clinicians to express more on the patient's medical situation. As a result, building a full-text medical search engine with basic Information Retrieval (IR) techniques is the critical first step to unleash the power of unstructured narratives.

Since commercial EHR packages do not usually include full-text search functionality, we have developed one called the Electronic Medical Record Search Engine (EMERSE) and have successfully integrated it with the institutional EHR systems in 2005 [2]. EMERSE has been shown to augment sensitivity, specificity, and efficiency of chart abstraction [3]. Unfortunately, similar to its commercial counterparts, an analysis of the query log of EMERSE showed that the performance of EMERSE still suffers from the low quality of queries entered by end users and the tremendous amount of redundant effort on exploring similar queries across users [4]. Even though clinicians have had years of training, it is still not easy to form adequate queries that lead to desirable search results due to the sophisticated information needs [5]. On the other hand, the clinical narratives have their own unique linguistic characteristics as discussed in our previous studies [6]. These factors taken together impose significant challenges in medical information retrieval. Two techniques that have been deployed in commercial search engines could be helpful to the issue in the medical context: (1) query recommendations, which provide alternative query terms based upon intelligent algorithms; (2) social search, which promotes the sharing of queries among users to utilize the 'wisdom of crowd.'

To justify the idea, we have learned from the success of EMERSE and built a proof-of-concept prototype of the next generation of EHR search engine (referred to as "EHR-SE" hereafter). EHR-SE is featured with relevance assessment at concept level, a component to recommend alternative query concepts, and a social search component that allows users to share search knowledge with each other. In the following sections, we describe the query suggestion component followed by the user interface design, including a community feature that enables social search. The paper ends with real user studies to evaluate the performance and the usability of the prototype system.

## 2. EHR SEARCH WITH AUTOMATIC QUERY RECOMMENDATION

The EHR search engine with automatic query recommendation includes four major functions (see Figure 1): (1) parsing and indexing clinical documents, (2) query parsing, (3) query recommendation, and (4) document retrieval.

In general, the new EHR-SE advances EMERSE and other existing full-text search engines by assessing document relevance and recommending alternative query terms at the concept level instead of at the word level. That is, a document is retrieved not because one of its terms (or its variations) matches the query, but because one of the medical concepts implied by its terms matches the concepts in the query. Relevant document are ranked using classical retrieval methods extended to the concept level. As a result, the new prototype is able to return documents containing "difficulty of hearing" when receiving the query "hearing loss."

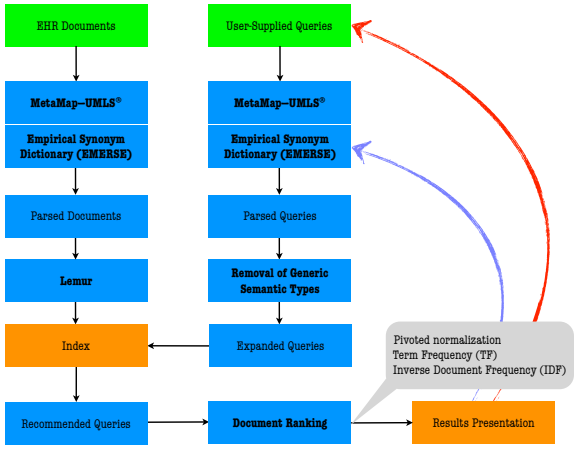



**Figure 1. Workflow of the EHR Search Engine with Automatic Query Recommendation**

### 2.1 Document Pre-Processing and Indexing

As illustrated in the left column in Figure 1, we first construct a corpus of electronic health records consisting of 95,702 EHR documents (17,198 patient visits) from the TREC 2011 Medical Records Track.[1] We parse these documents to extract medical terms and concepts in two ways. Medical terms are identified and mapped to the 'Metathesaurus concepts'[2] in the Unified Medical Language System (UMLS)[3] by MetaMap[4], which provides a numerical score from 0 to 1000 representing the confidence of match. Besides these concepts defined in the professional ontology, we extract medical terms frequently searched by the users and map them to an empirical synonym set (ESS), which is constructed through the analysis of the 4-year search log of EMERSE. Simple maximal string matching is used to identify these user-oriented terms and concepts. The documents are then indexed by either the medical terms in Metathesaurus or those in EES using the Lemur toolkit, resulting in two different indices denoted as *M* and *E* respectively.

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

[2] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

[3] http://www.nlm.nih.gov/research/umls/

[4] http://metamap.nlm.nih.gov/

**Table 1. Mathematical Notations**

| Notation | Meaning |
|---|---|
| *d* | A clinical document |
| *q* | A user-submitted search query |
| *t* | A medical term identified from the query |
| *M(t)* | Synonyms of term *t* under the same Metathesaurus concept(s) |
| *E(t)* | Synonyms of term *t* in the empirical synonym set (EES) |
| *c(t,d)* | Number of occurrence of the term *t* in *d*. |
| *df(t)* | Number of clinical documents that contains term *t*. |
| *N* | Number of clinical documents in the index |
| *avdl* | Average length of clinical documents |

### 2.2 Query Parsing and Recommendation

Table 1 lists the mathematical notations we use in this section. When a user submits a query *q*, the same process for document pre-processing is applied to identify the medical terms in the query and match them to concepts, as shown in the middle column of Figure 1. Metathesaurus concepts with the confidence score 800 or above are kept [5], and generic terms (such as "patient") that are not part of the medical concepts based on UMLS are dropped. For each medical term *t* extracted from the query, the query recommendation algorithm generates all synonyms of the term that are either under the same Metathesaurus concepts, *M(t),* or belonging to the same synonym subset in ESS, *E(t)*. The union of *M(t)* and *E(t)* are presented to expand the original query.

### 2.3 Document Ranking

Once individual documents and queries are both represented by a set of medical terms, the relevance score of a document given a query is calculated based on Pivot Normalization [7], a classical vector space retrieval model:

$$Score(d,q) = \sum_{t\in q\cap d} \frac{1+\ln\left[1+\ln\left(c(t,d)\right)\right]}{(1-s)+s\frac{|d|}{avdl}} . c(t,d) . ln\frac{N+1}{df(t)} \quad (1.1)$$

The query expansion using synonyms imposes new requirements to the retrieval model. We modify the formula such that: 1) the set of synonyms under the same concept are treated as a single term; 2) the metrics are calculated and combined based on two indices, the Metathesaurus index *M* and the ESS index *E*; 3) the metrics are aware of the situation that a medical term may be contained in multiple Metathesaurus concepts of ESS subsets as it could have multiple meanings. As a result, the modified term frequency (TF), inverted document frequency (IDF), and normalized document length can be calculated as:

$$c(t,d) = c_M(t,d) + c_E(t,d) + \sum_{t'\in M(t)} c_M(t',d) + \sum_{t'\in E(t)} c_E(t',d) \quad (2.2)$$

$$df_M(t) = |\{d|t\in d \wedge t\in M\}\cup\{d|t'\in d \wedge t'\in M \wedge t'\in M(t)\}| \quad (2.3)$$

$$df_E(t) = |\{d|t\in d \wedge t\in E\}\cup\{d|t'\in d \wedge t'\in E \wedge t'\in E(t)\}| \quad (2.4)$$

$$idf(t) = \frac{\sum_{t'\in M(t)} idf_M(t') + \sum_{t'\in E(t)} idf_E(t')}{|M(t)| + |E(t)|} \quad (2.5)$$

$$\frac{|d|}{avdl} = \frac{1}{2}\left(\frac{|d|_M}{avdl_M} + \frac{|d|_E}{avdl_E}\right) \quad (2.6)$$

Equation 2.5 provides a treatment for terms that have multiple meanings and are mapped to multiple concepts, with $idf_M(t)$ and $idf_E(t)$ calculated as the classical IDF based on $df_M(t)$ and $df_E(t)$.

Note that in terms of document relevance, this approach based on document indices with medical terms is actually equivalent to indexing the documents directly based on the Metathesaurus and ESS concepts. By doing so, the recommended query terms are automatically adopted in retrieval. We choose to index the documents at a finer granularity (using medical terms instead of concepts) so that a user has the flexibility to adopt, partially adopt, or deny the recommended query.

As a proof-of-concept, we adopted rather straightforward query recommendation and relevance ranking algorithms. Under the same architecture, one could apply more sophisticated query suggestion methods to generate alternatives of $M(t)$ and $E(t)$, and the state-of-the-art retrieval methods to rank the documents. It is our on-going work to explore these advanced methods.

# 3. USER INTERFACE DESIGN

## 3.1 Search Functionality

As a proof-of-concept, we developed a prototype system EHR-SE that retrieves clinical documents through a set of Java-Based web service APIs and presents the results on a web interface. The main workspace of EHR-SE is illustrated in Figure 2.

On the top of the workspace, a general search bar is placed for submitting queries with the ability to constrain specific document types, followed by the automatic query suggestion function and the result list. The query suggestions are organized by identified concepts with different colors. The retrieved documents are organized as a list below the query suggestions. Each record presents a title and a snippet, with the same color-coding scheme applied to indicate how the query terms are matched in the documents. When the title of a document is clicked on, a popup is brought up to show the content. A user may turn on and off the query suggestion to evaluate the performance of the system.

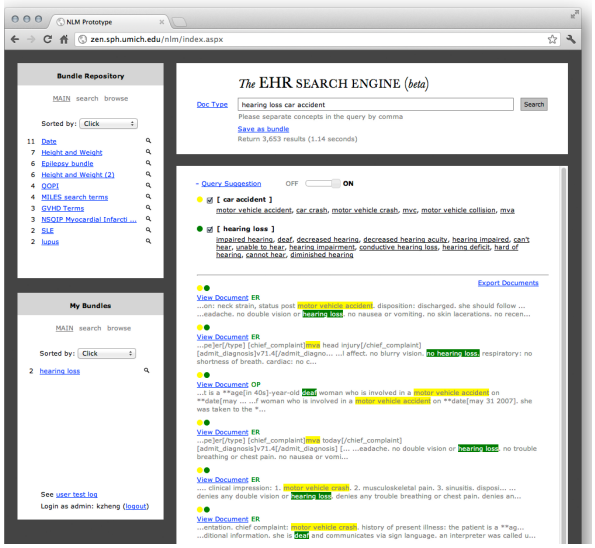



**Figure 2. Main Workspace of EHR-SE**

## 3.2 Social Search Features

One basic problem in information retrieval is that users are always searching for something for which they have limited knowledge. It is challenging for individual users to construct queries well describing their information needs. Social search deals with this problem by encouraging users to contribute their knowledge to query formation, results refinement, and knowledge discovery so that everyone benefits from everyone else and, collectively, the performance of IR systems is improved [8].

We have embodied the idea of social search in our existing search engine, EMERSE. Users can create, modify, and share "search term bundles" that contain collections of keywords and regular expressions to describe the information needs [8]. The commonly used "*Cancer Staging Terms*" bundle, for example, has 202 distinct search terms such as "gleason," "staging workup," and "Tmic." There is, however, no explicit mechanism to show the usefulness and the popularity of bundles. We believe reputation systems could be used to this end [9].

Therefore, we redesigned the bundle functionality of EMERSE and implemented several community features in the prototype. First, users could click the link below the search bar to save a query as a search bundle. Once a bundle has been created, it can be modified by adding or removing terms, privately reused by the owner, or publicly shared with other colleagues. Two bundle repository panels were placed on the left hand side of the workspace to help users browse and search available bundles created by other users by different attributes such as frequency and rating scores. The top repository panel is reserved for all shared bundles while the bottom one is for bundles made by the user herself. Moreover, a five-point rating scale is designed to indicate the reputation of the bundles (see Figure 3). Popular bundles are promoted by being shown on the top of the list in the repository panels by default.

# 4. EVALUATION

User experiments were conducted to evaluate the performance of the prototype. We approached 197 active users of EMERSE and, 43 of them responded to the invitation. Finally, 33 participants were recruited, who came from 10 different departments and 21 different divisions in the University of Michigan Health System.

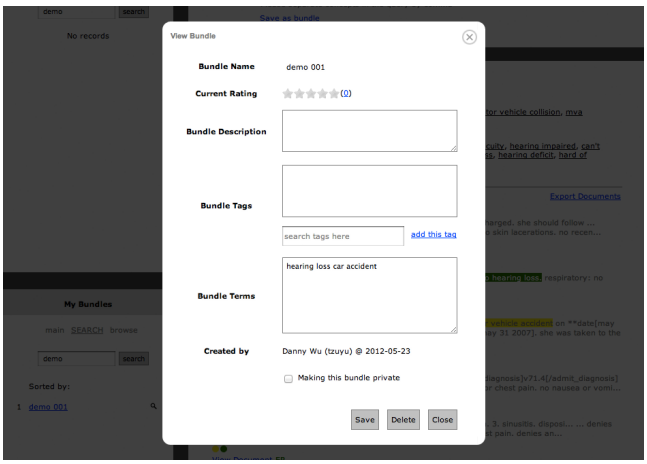



**Figure 3. Community Feature for Preserving Shared Search Knowledge.**

Five standardized scenarios (Table 2) were assigned to each participant to stimulate information needs in reality. In each scenario, the participants could try as many queries as necessary. After obtaining satisfactory search results, each participant

answered the following three questions with a score from 1 to 5 as well as narratives to describe their perceptions of the system performance:

- **Overall Performance:** *"How would you rate the search engine's performance with the automated query recommendation feature turned on, compared to the performance when the feature is turned off;"*
- **Relevance of Search Terms Recommended:** *"How would you rate the relevance of the search terms recommended by the search engine to the keywords you entered initially;"*
- **Usefulness of Query Recommendation:** *"How many of the recommended search terms would you have been able to come up with without the computer's assistance."*

**Table 2. Test Scenarios**

| Sc. | Description |
|---|---|
| 1 | "You are doing a research project in which you want to identify people who have had a concussive episode after being in a car accident." |
| 2 | "You are interested in identifying patients who have the non-invasive form of breast cancer known as DCIS." |
| 3 | "Please try to identify patients who are smokers who have also been diagnosed with PTSD." |
| 4 | "You are interested in how many patients are taking herbal supplements for the purposes of weight loss." |
| 5 | "Someone has asked you to determine if we have many patients diagnosed with mono who had an enlarged spleen." |

The evaluation results are presented in Table 3. The participants gave highly positive feedback on the system performance when automatic query recommendation was turned on compared to the performance when the service was turned off, according to the average score of the first two measures (4.25 and 4.52, respectively). However, these two measures were not judged equally among the scenarios. The AVOVA statistical test showed that the overall performance of scenario 5 was significantly higher than that of scenario 2, while the relevance of suggested terms in scenario 2 and 4 were significantly lower than that in scenario 5. We believe that the sophisticated information needs in the medical setting has resulted in the high variance of system performance. Together with narrative feedback, it appears that the performance could be improved by providing more granular control over the query recommendation such as adding or weighting a term, and handling negation.

**Table 3. Mean Scores of System Performance Evaluation (1: Lowest; 5: Highest)**

| Sc. | Overall Performance | Relevance of Search Term Recommended | Usefulness of Query Recommend (the lower the better) |
|---|---|---|---|
| 1 | 4.24 | 4.73 | 3.15 |
| 2 | 3.94* | 4.21* | 3.00 |
| 3 | 4.42 | 4.58 | 3.09 |
| 4 | 4.09 | 4.18* | 3.09 |
| 5 | 4.55* | 4.88* | 2.73 |
| Avg. | 4.25 | 4.52 | 3.01 |

* these pairs have significant difference between their means, except Scenario 2 & 4 in the 2nd (relevance) measure.

Note that the participants reported that they were not able to come up with many of the system-suggested search terms by themselves (avg. 3.01), which justifies the need to provide automatic query recommendation in EHR search. The statistical test showed no significant difference in this measure among the scenarios.

In addition to the three questions right after each scenario, five questions developed based on the technology acceptance model were given after they finished all the scenarios to solicit feedback on the general satisfaction toward the usefulness and the usability of the prototype system. Two out of five questions were designed to assess each participant's perception on the community features. Table 4 summarizes the results.

**Table 4. Other Feedback of the system (1: Lowest; 5: Highest)**

| | |
|---|---|
| **Usefulness of the Query Recommendation Feature** | 4.67 |
| **Ease of Use of the Prototype System** | 4.58 |
| **Overall Satisfaction** | 4.79 |
| **Usefulness of the Community Feature*** | 4.30 |

* All of them indicated they were also willing to make their own knowledge available to others either via public search bundle sharing (90.6%) or private sharing (9.4%).

# 5. CONCLUSION

We present a comprehensive user study to evaluate a proof-of-concept search engine for electronic health records, which is equipped with an automatic query recommendation module through the identification of medical concepts. Although the algorithms are relatively simple, the feedbacks from the users are promising. Findings of the study indicate a great potential of a next generation of EHR search engines, which combine the practices of concept-level relevance assessment, automatic query recommendation, and search knowledge sharing through social search. It is our future direction to explore and evaluate concrete information retrieval algorithms under such a framework.

# 6. REFERENCES


[1] Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;57:57cm29.

[2] Hanauer DA. EMERSE: The Electronic Medical Record Search Engine. AMIA Annu Symp Proc. 2006;941.

[3] Hanauer DA, Englesbe MJ, Cowan JA Jr, et al. Informatics and the American College of Surgeons National Surgical Quality Improvement Program: automated processes could replace manual record review. J Am Coll Surg 2009;208:37e41.

[4] Hanauer DA, Zheng K, Mei Q, Choi SW. Full-Text search in electronic health records: Challenges and opportunities. In: Kutais BG, ed. Internet Policies and Issues. Volume 7. Hauppauge, NY: Nova Science Publishers; 2011:125–40.

[5] Yang L, Mei Q, Zheng K, Hanauer DA. Query log analysis of an electronic health record search engine. AMIA Annu Symp Proc. 2011:915–24.

[6] Zheng K, Mei Q, Yang L, Manion FJ, Balis UJ, Hanauer DA. Voice-dictated versus typed- in clinician notes: Linguistic properties and the potential implications on natural language processing. AMIA Annu Symp Proc. 2011;1630–8.

[7] Singhal A. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2001;24(4):35–43.

[8] Zheng K, Mei Q, Hanauer DA. Collaborative search in electronic health records. J Am Med Inform Assoc. 2011;18(3):282–91.

[9] Resnick P, Kuwabara K, Zeckhauser R. Reputation systems. Commun ACM. 2000:45–8.

# Clinical Information Retrieval with Split-layer Language Models

Stephen Wu
Mayo Clinic
200 First St SW
Rochester, MN
wu.stephen@mayo.edu

Dongqing Zhu
University of Delaware
101 Smith Hall
Newark, DE 19716
zhu@cis.udel.edu

William Hersh
Oregon Health & Science University
3181 Sam Jackson Park Road
Portland, OR
hersh@ohsu.edu

Hongfang Liu
Mayo Clinic
200 First St SW
Rochester, MN
liu.hongfang@mayo.edu

## ABSTRACT

With the increasing prevalence of electronic medical records (EMRs), search technologies for these systems hold significant promise for improving patient and population care. We present a *split-layer language model* that embeds linguistic layers from existing NLP systems in retrieving medical documents. On the cohort identification task of the TREC Medical Records Track, our approach shows improvement over baselines, with the best performance achieved by mixing in all tested layers of NLP artifacts.

## Categories and Subject Descriptors

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

## Keywords

layered language model, medical records, information retrieval, natural language processing

## 1. INTRODUCTION

With the increasing prevalence of electronic medical records (EMRs), search technologies for these systems hold significant promise for improving patient and population care. The use of EMR text (e.g., progress notes) is particularly important for the clinical domain because some information (e.g., symptoms) is recorded nowhere else. However, raw text alone will not capture some important aspects of meaning. For example, saying that a patient has cancer, history of cancer, may have cancer, or does not have cancer will each have a different effect on relevance to a query. Keyword searches will fail to discover this type of granular information.

Addressing the need for fine-grained analysis, significant work has gone into NLP and information extraction (IE) techniques in the clinical domain, including systems like cTAKES [4], MetaMap [1], and MedTagger [2]. Each of these systems allows medical concepts to be represented. Attributes of those concepts, such as history, uncertainty, and negation, are also typically represented and discovered. These layers of NLP analyses have yet to be fully integrated into clinical retrieval techniques. Thus, the present work is a preliminary step towards this full integration, making NLP-produced concepts and attributes searchable by using language models that account for those artifacts.

In order to evaluate the contribution of this language model, we adopt the retrieval framework of cohort identification from EMR text, as defined in the Text Retrieval Conference (TREC) Medical Records Track [8, 7]. Results show that mixing in different layers of language processing is beneficial in almost every setting, even with naïve weighting schemes.

The rest of this paper proceeds as follows. Section 2 reviews some related work, Section 3 introduces a language model that incorporates layers of NLP output, Section 4 describes the TREC-med evaluation and our system, and Section 5 presents our results.

## 2. RELATED WORK

Unstructured data in IR has been augmented by structured objects (such as NLP artifacts in this work) in several ways. First, searching structured objects has typically been the realm of relational database searches, and some work has been done in embedding text search in these frameworks. Querying both text and its linked NLP artifacts, however, is a challenging and unsolved task. Second, perhaps the primary NLP artifact to be considered in layered language IR is that of named entities (concepts) and their attributes, and significant recent research has focused on entity search and semantic search. Our proposed work will align with these techniques in many cases, but we will take a language modeling approach that weighs NLP artifacts together with these entities (and other structures) in a statistical framework. While overall there were few early successes with using NLP techniques in IR, some approaches were shown to benefit retrieval [9]. Unlike many of those early systems, the split-layer language model is done in a language mod-


*HSD 2013*, August 1, 2013, Dublin, Ireland.

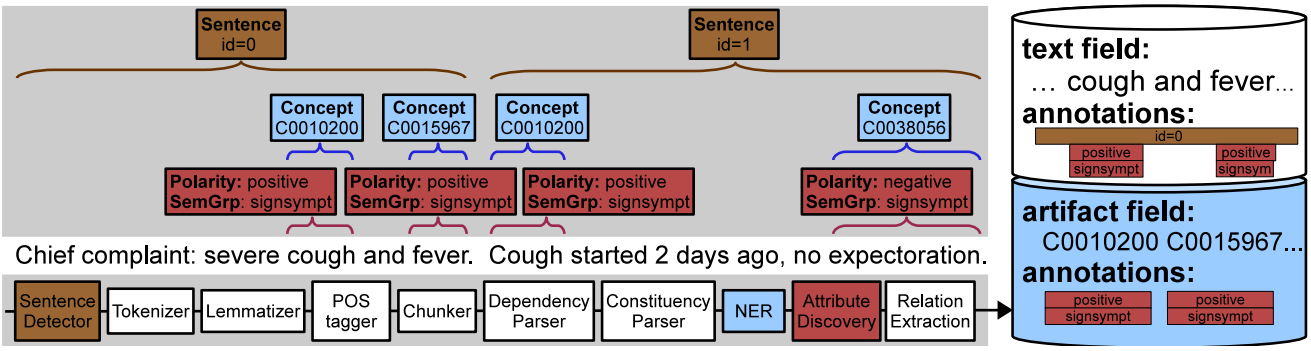



**Figure 1: A medical NLP pipeline with example created types (left) and the resulting layered index (right).**

eling framework, and is perhaps more technically similar to mixing different document representations [3, 10].

Cohort identification is a well-known problem in medical informatics. However, few approaches have used traditional text-based IR approaches. Notable exceptions are the EMERSE (Electronic Medical Record Search Engine) system and the recent work done on the TREC Medical Records Track. EMERSE is a non-commercial EMR search engine that supports free-text queries and has demonstrated the effectiveness of IR techniques in making chart reviews more efficient [5], but it focused on a sound user interface rather than underlying retrieval techniques. The TREC Medical Records track provided resources for significant innovation in medical IR, with a shareable collection of clinical text, information needs, and judgments [8, 7]. We will use the TREC-med setting as our evaluation framework.

# 3. MODEL

We introduce the *split-layer language model* (§3.3), a preliminary instantiation of a general class of *layered language models* that extend the query likelihood language model (§3.1) with NLP-produced artifacts (§3.2). The intuition is that human language can be described in somewhat overlapping linguistic layers: phonology, morphology, syntax, semantics, and discourse. Search technologies often perform quite well with just surface forms (raw text), but especially in the clinical domain, other layers are important. Access to these other layers is provided through standard NLP techniques such as parsing and named entity recognition (NER).

## 3.1 Query likelihood language model

In IR language models, it is common to rank according to $\text{score}(d, q) = \mathsf{P}(d) \cdot \mathsf{P}(q \mid d)$, where the latter conditional probability encapsulates the intuition that an ad hoc user trying to find document $d$ will try to write an effective query $q$. We will represetn the document or query as a list of terms, which we write $T_d$ and $T_q$, respectively. The standard query likelihood language model estimates the conditional probability as a Dirichlet-smoothed maximum likelihood estimate.

$$\hat{\mathsf{P}}_{\text{QL}}(T_q \mid T_d) \stackrel{\text{def}}{=} (1 - \alpha_D) \prod_{t_q \in T_q} \tilde{\mathsf{P}}_d(t_q \mid T_d) + \alpha_D \prod_{t_q \in T_q} \tilde{\mathsf{P}}_D(t_q \mid T_D) \quad (1)$$

where the $\alpha_D$ is related to the Dirichlet smoothing parameter $\mu$ according to $\alpha_D = \frac{\mu}{\mu + |D|}$. Note that $\tilde{\mathsf{P}}_d(t_q \mid T_d)$ actually denotes a probability conditioned on a model $\Theta_{T_d}$, but we will drop the model variable $\Theta$ throughout.

## 3.2 Indexing multiple linguistic layers

In order to incorporate layers of linguistic information, we first assume that these layers can be inferred from text — whether the document or the query. This inference is done by means of standard NLP techniques and stored in index layers. The left side of Figure 1 shows two sentences processed with an NLP pipeline and some output data types; e.g., sentence detection, named entity recognition, and attribute discovery on the text might result in Sentence and Concept types, with Polarity Semantic Group attributes. These output NLP artifacts are explicitly tied to the text with begin and end character offsets. To make the diverse NLP structures retrievable, we build retrieval indexes (Figure 1, right side) using two strategies: annotations and fields.

*Fields* have separate inverted indexes. Figure 1 (right side) shows an index for the original text, labeled "text field," alongside an index for an NLP artifact, labeled "artifact field." NLP artifacts that warrant their own artifact fields are termed *content artifacts*. In the example, the concept unique identifiers (CUIs, from the Unified Medical Language System (UMLS) Metathesaurus) of medically-relevant concepts may be stored as 'text' in a concept index. This is essentially a layer for the shallow semantic representation of NER-produced concepts. The concept-based artifact field at the bottom–right of Figure 1 could, at minimum, be used to support CUI searches.

*Annotations* provide additional layered information whose meaning is inextricably tied to the items in an index. These annotations retain begin and end offsets with respect to the index they are a part of. In the example, Sentence annotations divide the text into different sentences, and are thus annotations on the text field. However, attribute discovery finds Polarity (negation) values on newly-recognized Concepts (named entities), so Polarity values are first stored in the Concept index as annotations (bottom right, Figure 1). Because the Concepts themselves correspond to spans on the original text, it is possible to populate the text space with the corresponding annotations as well (top right, Figure 1). An alternative strategy for future work would be to consider storing a pointer from items in the artifact field to their corresponding spans in the text field. Annotation values may also be indexed (in a manner similar to fields) in

order to support narrower searches. For example, a search for CUIs could be constrained to only non-negated concepts, or could exclude medications. A search specifying sentence annotations gives the expected behavior of exttent retrieval.

### 3.3 Split-layer language model

We define a query according to its index layers as $q = \langle T_q, C_q, A_q \rangle$, where $T$ is the list of text tokens, $A$ is the annotations, and $C$ is the content artifacts associated with the query text. The queries thus deterministically contain all NLP-extracted structures. Similarly, we can write documents as $d = \langle T_d, C_d, A_d \rangle$.

For simplicity, our equations below will make the assumption that there is only one type of annotation and one type of content artifact. We will write annotations ($A$) together with their respective fields ($T$ or $C$). Notationally, we use uppercase variables to represent sets of lowercase variables, so that a corpus $D$ is a set of documents $d$. It is not the case that every type of structure discovered by clinical NLP methodologies will be empirically useful in document retrieval; however, the probabilistic framework must account for the multiple layers of artifacts.

Here, we introduce layered language $T$, $A$, and $C$ components. We make independence assumptions between the text field and any content fields, and treat annotations (on either text or content artifacts) jointly.

$$\hat{\mathsf{P}}_{SL}(q \mid d) \stackrel{\text{def}}{=} f\Big(\hat{\mathsf{P}}_{TBR}(T_q \mid T_d), \quad \hat{\mathsf{P}}_{TBRa}(TA_q \mid TA_d), \quad \hat{\mathsf{P}}_{CBR}(C_q \mid C_d), \quad \hat{\mathsf{P}}_{CBRa}(CA_q \mid CA_d)\Big) \tag{2}$$

where $\hat{\mathsf{P}}(\cdot)$ represents an estimated distribution that will typically include a Dirichlet term. For example, the first term in Eqn. 2 would then be implemented as the query likelihood model $\hat{\mathsf{P}}_{\text{QL}}(T_q \mid T_d)$ of Eqn. 1. The function $f(\cdot)$ represents the way to combine the probabilities from different layers. In this paper, we define $f(\cdot)$ as a simple linear combination function whose effect is similar to ranking by combining different document representations [3, 10].

We have four basic models of the layered query likelihood, each corresponding to the four terms of Eqn. 2:

**TBR.** Text-based retrieval, i.e., the query likelihood model.

**TBRa.** Text-based retrieval with annotations.

**CBR.** Concept (CUI)-based retrieval.

**CBRa.** Concept (CUI)-based retrieval with annotations.

Note that each of the layers essentially implement backoff and smoothing. Because there may be insufficient statistics for concepts with annotations $\hat{\mathsf{P}}(CA_q \mid CA_d)$, the estimates for $\hat{\mathsf{P}}(C_q \mid C_d)$ serve as an backoff model alongside Dirichlet smoothing. The same can be said for the 'text only' and 'text with annotations' layers.

## 4. EVALUATION

### 4.1 Cohort identification task

We evaluated the contribution of different split-level language models on the task of cohort identification, mirroring the Text Retrieval Conference (TREC) Medical Records Track [8, 7]. The retrieval collection was the University of Pittsburgh's BLU repository. Each patient at the University of Pittsburgh would have one or more medical *records* (documents) associated with one or more of his/her *visits* to the hospital. The unit of retrieval was defined as a patient visit, since they were broken by the de-identification procedure that made the records shareable. In total, there were 95,702 records that corresponded to 17,198 visits. The largest visit was 418 records, but the mean visit was 5.56 records.

### 4.2 System and evaluation setup

To isolate the contributions of the split-layer language model, we ignored any metadata (primarily ICD-9 codes, whose availability and reliability are inconsistent) and based retrieval on the text portions only. We processed this text using the MedTagger [2] information extraction system, producing "layers" including: content artifacts (UMLS CUIs), and contextual attributes on those artifacts (semantic group, negation, uncertainty, and experiencer).

These artifacts were indexed in Indri [6] using the offset annotation capabilities. Rather than considering all the layers generated by MedTagger, we focused our evaluation on those that were most likely to be semantically relevant, namely, the CUIs and the contextual attributes listed above. Thus, TBR was a basic query likelihood model; TBRa utilized the concept-based annotations (with the offset spans mapped appropriately); CBR was populated with CUIs; CBRa contained both CUIs and associated annotations.

Our tests did not train any parameters, but tested on all 81 queries from the official evaluations of TREC-med 2011–2012. We use mean average precision (MAP) as our main evaluation metric, since the official metrics for TREC-med were different in 2011 and 2012 (bpref and infAP) but MAP scores corresponded to the official metrics in both cases. We tested the contribution of the Dirichlet smoothing parameter $\mu$ to each of TBR, TBRa, CBR, and CBRa, comparing how each performed. We then examined the possible configurations in which these models could be linearly combined.

## 5. RESULTS AND DISCUSSION

### 5.1 Dirichlet smoothing

Figure 2 shows the effect of the Dirichlet smoothing parameter on the four separate layers. Interestingly, it appears that the performance is inversely proportional to the smoothing parameter. This is surprising given the short length of medical documents, since shorter document lengths typically correspond to more need for collection-level smoothing. These effects may be because medical documents have relevant information concentrated in relatively few notes (such as those within the same specialty or note type); performance is diluted by smoothing with the whole collection. Thus, we use the lowest tested Dirichlet parameters ($\mu = 2500$) in all further analyses.

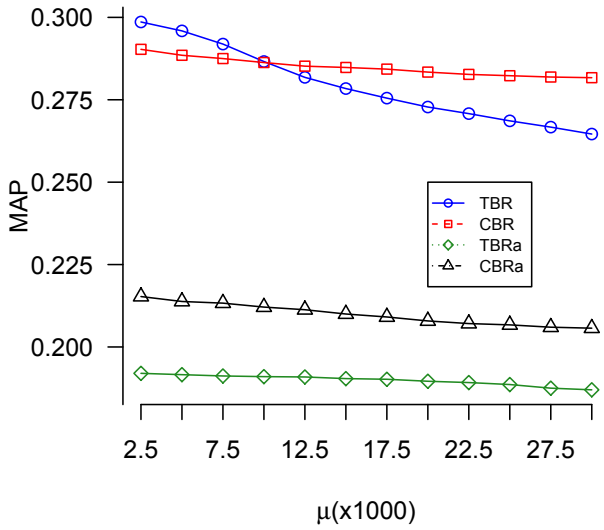



**Figure 2: Four models corresponding to separate layers of NLP processing, showing the effect of Dirichlet smoothing parameter.**

**Table 1: Model combinations of text-based retrieval (TBR), content artifact-based retrieval (CBR), and their combination (TBR+CBR). Subscripts $t$ and $c$ indicate statistical significance ($p < 0.05$ for one-tailed paired t-test) for TBR and CBR respectively.**

| Baseline | alone | +TBRa | +CBRa | +TBRa,CBRa |
|---|---|---|---|---|
| **TBR (t)** | 0.2986 | $0.2996^{t}$ | $0.3022^{t}$ | $0.2996^{t}$ |
| **CBR (c)** | 0.2903 | 0.2905 | 0.2902 | $0.3027^{t}$ |
| **TBR+CBR (s)** | $0.3302^{t,c}$ | $0.3302^{t,c}$ | $0.3302^{t,c}$ | $0.3365^{t,c}$ |

## 5.2 Comparison between layers

Figure 2 also has four lines, corresponding to the TBR, TBRa, CBR, and CBRa models. It is clear that without any mixing of the models, the layer of annotations decreases performance on both the text-based and concept-based retrieval models. When evaluating on the standard metrics such as MAP, this is likely due to the lack of sufficient data to support such specific maximum likelihood models. However, we also hypothesize that the binary relevance judging of TREC-med pools and corresponding metrics do not easily demonstrate the capabilities of annotations — annotations imply weighted relevance for text or content artifacts.

## 5.3 Split-layer combinations

The capabilities of the split-layer language model are evaluated by linearly combining the four basic models in different configurations. In Table 1, we show the effects of a naïve linear combination with equal weights (i.e., arithmetic mean of included components). Moving along each row, we see the positive effect of mixing in the different annotations layers (statistical significance as compared to the baselines is notated by the superscripts). While we noted from Figure 2 that annotations-included layers alone (TBRa or CBRa) underperformed their respective content fields (TBR and CBR), it is clear that *mixing* content layers with annotations layers tends to improve performance. Furthermore, we can see the benefits of mixing text with a content field by moving down the columns; in every case, TBR+CBR outperforms either TBR or CBR alone. Both of these observations are captured in the bottom-right result, in which all four models (TBR + CBR + TBRa + CBRa) yield the best overall performance.

# 6. CONCLUSIONS

We have introduced the split-layer language model, a model that embeds multiple linguistic layers into IR analysis by incorporating the results of core NLP tasks. We have shown that, on the task of cohort identification from medical records, the split-layer language model improves performance over a query likelihood baseline. A model that mixes all text, content artifacts, and annotations performed the best overall.

The split-layer language model is termed such because of the independence assumptions between the text and content artifacts layers. Other layered language models are possible — multiple layers could be integrated into the component models, such that text and content artifacts are estimated together in probability distributions. This is an active area of future work. Additionally, future work will evaluate the effect of smoothing parameters, and will establish a means by which multiple collocated artifacts can be queried at once.

# 7. REFERENCES

[1] A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[2] H. Liu, S. T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P. J. Haug, S. M. Huff, and C. G. Chute. Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2012, page 568, 2012.

[3] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *SIGIR'03*, pages 143–150. ACM, 2003.

[4] G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler, and C. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507, 2010.

[5] L. Seyfried, D. Hanauer, and D. Nease. Enhanced identification of eligibility for depression research using an electronic medical record search engine. *International Journal of Medical Informatics*, 78(12):e13–e18, Dec. 2009.

[6] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[7] E. Voorhees and W. Hersh. Overview of the TREC 2012 medical records track. In *TREC*, 2012.

[8] E. Voorhees and R. Tong. Overview of the TREC 2011 medical records track. In *TREC*, 2011.

[9] W. A. Woods, L. A. Bookman, A. Houston, R. J. Kuhns, P. Martin, and S. Green. Linguistic knowledge can improve information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 262–267. Association for Computational Linguistics, 2000.

[10] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141, 2008.

# Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs

Andrew Yates
Information Retrieval Lab
Department of Computer Science
Georgetown University
andrew@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

Ophir Frieder
Information Retrieval Lab
Department of Computer Science
Georgetown University
ophir@ir.cs.georgetown.edu

## ABSTRACT
Interest in medical data mining is growing rapidly as more health-related data becomes available online. We propose methods for extracting Adverse Drug Reactions (ADRs) from forum posts and linking extracted ADRs to the drugs that users claim are responsible for them. We evaluate our methodology using a corpus of annotated forum posts. We find that our ADR extraction method outperforms a strong baseline in terms of precision at the expense of a similar decrease in recall. When used in conjunction with a strong baseline, our method is able to increase recall by 7% without harming precision.

## Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

## General Terms
Algorithms, Experimentation

## Keywords
Medical data mining, extracting adverse drug reaction, health-related social media

## 1. INTRODUCTION
According to a recent survey, 35% of US adults have attempted to use the Internet to diagnose a medical condition in the past year[1]. As more health-related data appear online, interest in medical data mining grows rapidly. Recent efforts include those that mine symptoms and conditions from query logs [13], mine adverse drug reactions from drug reviews [10, 15], health-related social networks [9], and forum posts [2], and mine the existence of particular [1, 4, 5] or any [11] medical outbreaks.

We focus on extracting Adverse Drug Reactions (ADRs) from forum posts and linking the extracted ADRs to the drug that the user claims, implicitly or explicitly, is responsible for them. Forum posts present unique challenges in that they are relatively long (130 terms on average, in our dataset) and the relationships between drugs and ADRs are not available as structured data. In contrast, drug comments posted to some other forms of social media, such as drug review sites, are explicitly related to a specific drug (i.e., the review or comment is posted on a drug's page) and are relatively succinct (46 terms on average, in the dataset used in [15]).

Specifically, we propose 1) a novel method for extracting ADRs from social media using linguistic dependency relations and a conditional random field (CRF) [8], and 2) a novel method for linking ADRs to the drugs that posting users identify as their cause.

Our contributions are

- A novel method for extracting ADRs from social media
- A novel method for linking ADR mentions to drugs
- A publicly available annotated dataset[2] indicating the ADRs present in forum posts and the drugs that users identified as being responsible for the ADRs.

## 2. RELATED WORK
Several previous efforts focused on extracting adverse drug reactions (ADR) from social media.

Leaman et al. [9] matched terms in a bag-of-words sliding window against known ADRs after correcting for spelling mistakes. Similarly, Benton et al. [2] found ADRs occurring in sets of terms that were more likely to occur together within a bag-of-words sliding window than they were to occur separately. We compare our approach to a bag-of-words sliding window approach.

Li [10] used statistical methods to find terms that were only present in one of two mutually exclusive classes of drug. This method requires that two mutually exclusive drug classes be compared, which requires domain knowledge and is not possible when two such classes do not exist.

Yates & Goharian [15] proposed ADRTrace, a system that found ADRs that exactly matched terms in a list or matched a pattern mined from drug reviews. We compare our approach to ADRTrace.

## 3. METHODOLOGY
We describe our method for extracting adverse drug reactions (ADRs) in Section 3.1. In Section 3.2 we describe our method for associating extracted ADRs with the drug that the user claims is responsible for the ADR.


*HSD 2013*, August 1, 2013, Dublin, Ireland.

[1] http://www.pewinternet.org/Reports/2013/Health-online.aspx

[2] http://ir.cs.georgetown.edu/data/adr_forum_annotations

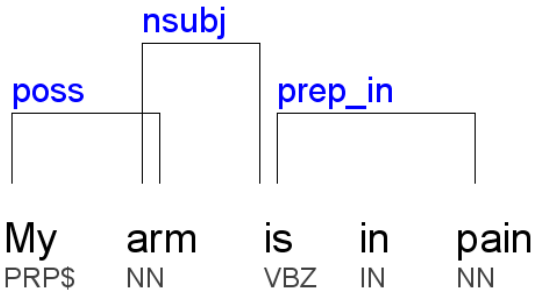



**Figure 1. Example dependency relations**

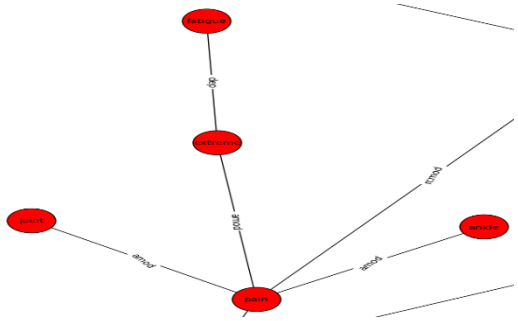



**Figure 2. Dependency relation graph**

## 3.1 Extracting ADRs

Rather than attempting to match all terms in a sliding window against known ADRs as some previous efforts have done, we focus on using dependency relations as a principled way to choose terms to match against known ADRs. We use the Stanford Parser [7] to identify dependency relations, which consist of a relation type (e.g., nominal subject), a head term (i.e., the term which determines the type of phrase), and a modifier term (i.e., a term which modifies the head term). For example, Figure 1 shows the collapsed dependency relations from the sentence "My arm is in pain." A sliding window of at least 4 terms would discover the "arm pain" ADR in this sentence, but it is also possible to discover that ADR by noting that "arm," "is," and "pain" are connected by relations and checking for ADRs composed of a subset of those words.

We find ADRs using dependency relations in two ways. First, we find ADRs by combining every pair of terms that appears in a dependency relation and determine whether the term pair matches a known ADR. If it does the ADR is extracted.

Second, we learn which dependency relation paths can be followed to generate candidate ADRs. To do so, we construct a graph from each forum post. Each vertex corresponds to a term; edges correspond to dependency relations between terms. Figure 2 shows a subset of one such graph. The post contains the "joint pain," "ankle pain," and "fatigue" ADRs, among others. The edges are labeled with the dependency relation types connecting the term vertices (e.g., "amod" and "dep"). It is our method's job to determine when these edges can be followed; we use the relation types (e.g., "amod") later as one of the features used to determine that. This example is kept small so that it may be easily visualized; in actual use ADRs may consist of several terms that are each several hops away from each other.

Each post is then split into individual sentences using the Punkt sentence tokenizer [6]. Each sentence is treated as bag-of-words to find potential ADRs that may exist in the sentence (e.g., "carpal tunnel syndrome" would be found in "syndrome x carpal y tunnel z"). For each of these potential ADRs, we find the shortest path in the graph between each sequential pair of terms (e.g., we find the shortest path between "carpal" and tunnel" and the shortest path between "tunnel and syndrome").

Each sequence of shortest paths in a potential ADR is then turned into binary features for use with a Conditional Random Field (CRF) [8], which are often applied to classification tasks involving natural language, such as named entity recognition. Note that there may not be a single hop path between each term in a potential ADR; thus, term vertices may appear in the path between ADR terms that are not included in the extracted ADR.

The features used by the CRF are:

- Dependency relations present in the path
- Terms present in the path (no distinction is made between head terms and modifier terms)
- Term appearance anywhere in the MedSyn thesaurus [15]
- Head term appearance anywhere in the MedSyn thesaurus
- Modifier term appearance anywhere in the MedSyn thesaurus

These features capture the terms and dependency relations present in the path, and whether each term could be part of an ADR (i.e., whether it exists in MedSyn). This allows the CRF to determine if the path should be followed. When training the CRF, each set of features corresponding to an ADR that is known to exist in a post is given the "FOLLOW" label. Each set of features that corresponds to an ADR that does not exist but could (i.e., a bag-of-words sliding window method with a large window would extract the ADR but the ADR is actually not expressed in the post) is given a "DON'T_FOLLOW" label.

To find ADRs using this method, the CRF is given a set of features for each ADR whose terms exist in the post (i.e., each ADR that would be found using a sliding window approach when the window size is equal to the post's length). The ADR is extracted if the CRF predicts the "FOLLOW" label.

## 3.2 Linking Drugs to ADRs

While it is impossible to determine whether a relationship expressed between a drug and an ADR is true, namely true causation, it is still helpful to detect when such a relationship is expressed. For example, a relationship between the drug "Tamoxifen" and the ADR "hot flashes" is expressed in the sentence "I've been having hot flashes since I started taking Tamoxifen." We detect such relationships by using a CRF to label drug mentions as "DRUG-<drug name>" and ADRs linked to the drug as "<drug name>-ADR." The CRF is run on terms that are each given the following binary features:

- Term itself
- Part-of-speech tag for the three terms before the current term and after the current term (e.g., "VRB@-3"). The tags were found using the Stanford Log-linear Part-of-Speech Tagger [12]. Three terms were used because this performed better than using two terms; increasing the window size to four terms did not improve performance.
- The part-of-speech tag for the current term

- The appearance of the current term as a term anywhere in the thesaurus
- The current term matches the name of a known drug
- The types of the dependency relations that the term appears in as either a head or modifier term anywhere in the post (e.g., "dobj, nsubj")

The CRF is used to label ADR's caused by a drug as "<drug name>-ADR." Any ADR with a term bearing the "<drug name>-ADR" label is linked to the drug <drug name> by our system.

# 4. EVALUATION

We describe our dataset in Section 4.1. We use this dataset to evaluate our ADR extraction performance in Section 4.2 and to evaluate our drug linking performance in Section 4.3.

## 4.1 Dataset

Evaluating our algorithms requires a corpus from which to extract ADRs and the drugs associated with them, a domain-specific thesaurus listing terms and phrases that are equivalent (i.e., refer to the same ADR), and annotations indicating the ADRs expressed in the corpus and the drugs associated with these ADRs.

We used a corpus consisting of 400,000 posts crawled from the Breastcancer.org[3] and FORCE[4] forums. The posts were primarily chosen from sub-forums related to the discussion of ADRs caused by breast cancer drugs. Information on obtaining the corpus is available in [14]. MedSyn [15], a list of synonyms in the medical ADR domain that includes both expert (e.g., "arthralgia") and non-expert (e.g., "joint pain") terms, was used as our thesaurus. MedSyn is derived from a subset of the Unified Medical Language System Metathesaurus (UMLS) [3]; a description of how MedSyn was constructed is available in [15].

The corpus posts were annotated. Non-medically trained annotators were asked to read posts, annotate the ADRs present in the posts, and, if mentioned, indicate the drug that the user associated with each ADR. Annotators were instructed to only annotate first-hand accounts of an ADR and allowed to skip ADRs that were related to a medical procedure (e.g., surgery or chemotherapy). Each post was annotated by three separate annotators; posts and annotations that did not meet this criterion were discarded. In total, the annotators annotated approximately 600 posts with a total of 2,100 annotations. MedSyn was used to treat different terms or phrases that expressed the same ADR as equivalent. Fleiss' Kappa was calculated to be 0.37, indicating fair inter-rater reliability. To use these annotations as ground truth, we discarded any ADRs that were found by only one annotator. When evaluating our ADR extraction performance, we only included the 1,700 annotations that were related to one of the following breast cancer drugs: Arimidex, Aromasin, Femara, Nolvadex, and Taxotere. The annotations and URLs of the forum posts are available on our website[5].

We used the annotations to evaluate our methods rather than using the ADRs listed on drug labels. While we expect that some of the ADRs listed on drug labels are also expressed in the forum posts, other listed ADRs may be infrequent and should not be expected to be found. Annotations are used to perform a more direct evaluation by comparing the ADRs and drug relationships we extract to the annotated ADRs and relationships. Furthermore, the annotations may be used to train and evaluate supervised methods when coupled with the forum posts they correspond to.

[3] http://community.breastcancer.org/

[4] http://www.facingourrisk.org/messageboard/index.php

[5] http://ir.cs.georgetown.edu/data/adr_forum_annotations

## 4.2 ADR Extraction

We used our corpus and annotations to evaluate ADR extraction performance. Five-fold cross-validation was used.

The results, the baselines, and the *DepADR* system described in this paper are shown in Table 1. The percentages in parentheses indicate a system's performance relative to *ADRTrace*'s. An asterisk (*****) indicates a statistically significant change in performance at $p < 0.05$. *ADRTrace+DepADR* indicates that every ADR extracted by either the *ADRTrace* or *DepADR* system is returned. *Window* is a bag-of-words sliding window system with sliding windows of size 25; we chose a large window to establish an upper-bound on recall.

*Window* achieves the highest recall, as would be expected. Using *DepADR* in conjunction with *ADRTrace* yields a 7% increase in recall without harming precision, supporting our hypothesis that dependency relations can be used to extract additional ADRs without returning many more false positives. When used by itself, *DepADR* achieves a 56% improvement in precision at the expense of a 48% reduction in recall. This is unsurprising given *DepADR*'s focus on carefully choosing which terms may compose ADRs. These results suggest that *DepADR* can be paired with an existing system to improve performance when recall is important or used by itself in scenarios where a higher precision is desired.

**Table 1. ADR extraction results**

| | *Precision* | *Recall* |
|---|---|---|
| ADRTrace | 0.39 | 0.61 |
| ADRTrace+DepADR | 0.39 | 0.65 (+7%)***** |
| DepADR | 0.61 (+56%)***** | 0.32 (-48%)***** |
| Window | 0.32 (-18%)***** | 0.74 (+21%)***** |

## 4.3 Drug Linking

We used our corpus and annotations to evaluate our method for linking drugs to ADRs. Five-fold cross-validation was used. We compare our performance to a baseline that returns all (drug, ADR) pairs that exist in a post.

The results are shown in Table 2. The percentages in parentheses indicate a system's performance relative to *Baseline*'s. An asterisk (*****) indicates a statistically significant change in performance at $p < 0.05$. The baseline outperforms our system in terms of recall, which is to be expected given that the baseline returns every possible link. Our system performs 17% better in terms of precision. Coupled with our relatively low recall, these results suggest that our CRF approach is promising but could be improved.

**Table 2. ADR-Drug linking results**

| | *Precision* | *Recall* |
|---|---|---|
| DepADR Linking | 0.63 (+17%)***** | 0.36 (-64%)***** |
| Baseline | 0.54 | 1.0 |

## 5. CONCLUSIONS

We proposed novel methods for extracting ADRs from social media and linking each ADR to the drug that the user identified as being responsible for the ADR. We use supervised methods for both tasks that are trained on our annotated dataset. Our ADR extraction method makes extensive use of dependency relations to precisely choose potential terms to match against ADRs. Our results show that our ADR extraction method statistically significantly outperforms two previously proposed methods, while our drug linking method outperforms a simple baseline.

This work is clearly preliminary; future work will refine the approach described and combine various approaches to improve precision without significantly hampering recall.

## 6. REFERENCES


[1] Aramaki, E. et al. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *Conference on Empirical Methods in Natural Language Processing (EMNLP'11)* (Jul. 2011), 1568–1576.

[2] Benton, A. et al. 2011. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *Journal of biomedical informatics*. 44, 6 (Dec. 2011), 989–96.

[3] Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*. 32, Database issue (Jan. 2004), D267–70.

[4] Corley, C. et al. 2009. Monitoring Influenza Trends through Mining Social Media. *International Conference on Bioinformatics Computational Biology (ICBCB'09)* (2009).

[5] Jamison-Powell, S. et al. 2012. "I can't get no sleep": discussing #insomnia on twitter. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (New York, New York, USA, May. 2012), 1501.

[6] Kiss, T. and Strunk, J. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*. 32, 4 (Dec. 2006), 485–525.

[7] Klein, D. and Manning, C.D. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03* (Morristown, NJ, USA, Jul. 2003), 423–430.

[8] Lafferty, J.D. et al. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Eighteenth International Conference on Machine Learning (ICML'01)* (Jun. 2001), 282–289.

[9] Leaman, R. et al. 2010. Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)* (2010), 117–125.

[10] Li, Y.A. 2011. *Medical Data Mining : Improving Information Accessibility using Online Patient Drug Reviews*. MIT.

[11] Parker, J. et al. 2013. A Framework for Detecting Public Health Trends with Twitter. *IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM'13)* (2013).

[12] Toutanova, K. et al. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (Morristown, NJ, USA, May. 2003), 173–180.

[13] White, R.W. and Horvitz, E. 2012. Studies of the onset and persistence of medical concerns in search logs. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12* (New York, New York, USA, 2012), 265.

[14] Yates, A. et al. 2013. Graded relevance ranking for synonym discovery. *22nd international conference on World Wide Web companion (WWW '13 Companion)* (May. 2013), 139–140.

[15] Yates, A. and Goharian, N. 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)* (2013).