Information Processing and Management 52 (2016) 258–272

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# Improving patient record search: A meta-data based approach

Iman Amini [a,*], David Martinez [b], Xiaodong Li [a], Mark Sanderson [a]

[a] RMIT Department of Computer Science and NICTA, Australia
[b] MedWhat.com and the University of Melbourne, Australia

ARTICLE INFO

*Article history:*
Received 16 June 2014
Revised 29 July 2015
Accepted 31 July 2015
Available online 11 November 2015

*Keywords:*
Information storage and retrieval
Information search and retrieval
ICD classification
Pseudo relevance feedback

ABSTRACT

The International Classification of Diseases (ICD) is a type of meta-data found in many Electronic Patient Records. Research to explore the utility of these codes in medical Information Retrieval (IR) applications is new, and many areas of investigation remain, including the question of how reliable the assignment of the codes has been. This paper proposes two uses of the ICD codes in two different contexts of search: Pseudo-Relevance Judgments (PRJ) and Pseudo-Relevance Feedback (PRF). We find that our approach to evaluate the TREC challenge runs using simulated relevance judgments has a positive correlation with the TREC official results, and our proposed technique for performing PRF based on the ICD codes significantly outperforms a traditional PRF approach. The results are found to be consistent over the two years of queries from the TREC medical test collection.



## 1. Introduction

Electronic Patient Records (EPR) also referred to as Electronic Health Records (EHR) are the digitally stored medical notes written by doctors and practitioners during patients' visits. The volume of EPR is ever increasing. Demand for their use beyond simply recording the health of a person is changing. One potential new application is to use a collection of such records as a source for finding patients for a medical trial. For this task it is necessary to search over large numbers of EPRs to find patients matching certain criteria, such as suffering a given disease, or belonging to a demographic group. However, because of the structure and vocabulary of the records, search over such content presents new research challenges. In order to start exploring this problem, TREC[1] (Text REtrieval Conference) organized medical IR tracks in 2011 and 2012, where the goal was to identify patient records that fulfill the characteristics of given queries (e.g. "Patients with hearing loss"). The queries were built by targeting a list of research areas that the U.S. Institute of medicine has considered priorities for comparative effectiveness research. Participation in these tracks was strong with 54 research groups submitting runs over the two years it lasted.

One of the differentiating characteristics of search over EPRs is that the target documents contain associated meta-data, such as codes belonging to the International Classification of Diseases (ICD), which are manually assigned to each report by health administration workers.

The ICD is a system for classification of health care, providing a system of diagnostic codes with a large diversity of symptoms, signs and medical findings. The codes are used to help with health informatics processes such as billing for health insurance reimbursement (Puckett, 2011).

* Corresponding author. Tel.: +61421498523.
*E-mail addresses:* iman.amini@rmit.edu.au, iman.amini@gmail.com (I. Amini), david.martinez@nicta.com.au (D. Martinez), xiaodong.li@rmit.edu.au (X. Li), mark.sanderson@rmit.edu.au (M. Sanderson).

[1] http://trec.nist.gov.

http://dx.doi.org/10.1016/j.ipm.2015.07.005
0306-4573/

Other usages of ICD codes are to help with statistics related to the general health of a country, monitor the prevalence of diseases and to be used for the compilation of the national mortality and morbidity statistics. ICD codes have been shown to have problems of completeness and bias (Roque et al., 2011), and this could harm IR effectiveness. The codes are also challenging to work with, as they have a hierarchical structure with different levels of specificity. For instance *hearing loss* can be linked to many ICD codes, including but not limited to 389.03 (middle ear), 389.0 (conductive hearing loss), and 380.01 (external hearing loss).

Exploiting the presence of the ICD codes in records has not been extensively explored for IR, and our focus here is to enhance ranking methods for medical IR by relying on those codes. The main reason for examining their use, is that queries in the domain tend to refer to diseases, and the ICD codes carry information summarizing the patient's diseases and health conditions.

Since the workers who assign the codes to EPRs are required to follow strict guidelines, the use of ICD codes could help to alleviate some of the imprecision present in a bag-of-words representation of such records. This is particularly important in patient records, as often the free text part of the record will contain speculation, negations (e.g. "the patient does not have X"), references to past conditions, family history of the patient, etc. ICD codes, on the other hand, refer to the current conditions of the patient.

Apart from being used to enhance retrieval effectiveness, they have also been studied as a source of evidence for building test collections for medical IR (Koopman et al., 2011a). Here researchers have speculated that ICD codes can accurately summarize the content of queries and documents, and can be used as a proxy for relevance judgments (*qrels*) in IR test collections. However, a limitation of this previous work is that assessments have never been compared to real clinical queries.

In this paper, we exploit ICD codes for medical IR in two ways. We perform the first systematic analysis of the use of ICD codes for pseudo-relevance judgment (PRJ), by comparing the ranking of runs submitted to TREC based on qrels and the ranking based on qrels built from ICD codes. We use runs submitted to the first Medical TREC track (Voorhees & Tong, 2011) in 2011 (*TREC-M1*), and the second track (Voorhees & Hersh, 2012) in 2012 (*TREC-M2*). This analysis intends to address the following research question: *How reliable are the ICD codes for automatic judgment of medical records?*

For our second contribution, we explore whether the direct approach of simply mapping ICDs into their textual representation is the most effective way of using this resource. To answer this question we introduce a novel IR method that relies on ICD codes for a form of pseudo-relevance feedback (PRF), and we compare search based on this system with a common approach for exploiting ICD codes.

The rest of the paper is organized as follows. We present the prior work related to this research with a short survey on techniques used in medical IR, and also previous approaches for PRJ. Next, we describe the test collection and methodology to build our PRJ framework, followed by our proposed approach to model the ICD based PRF. The paper concludes with the analysis of the results, discussion of the limitations, and the future work.

## 2. Background

The ability to conduct research on the retrieval of clinical records has been limited in previous years, due to the lack of a publicly available dataset of appropriate size. The bulk of the work in this area has been focused on Natural Language Processing challenges, such as extracting specific information from a small number of clinical records (Özlem Uzuner, 2012), while the IR research on biomedical text has focused on searching the literature. However in 2011 the TREC medical retrieval track was introduced, and this generated much interest in the IR challenges of search over EPRs (Voorhees & Hersh, 2012; Voorhees & Tong, 2011). We describe here the main medical IR approaches that are related to our work, as well as previous work on PRJ, which is the other research focus of this paper.

### 2.1. Information retrieval for patient records

The TREC Medical Challenges of 2011 and 2012 give the best picture of the state of the art at this point, since several research groups participated on a shared task over the same patient repository. The best runs in 2011 (King et al., 2011) and 2012 (Zhu & Carterette, 2012) focused on different aspects of the search. King et al. (2011) relied heavily on text processing and information extraction. They tuned their system using their own manually created relevance judgment of approximately 190 reports per query. In 2012 Zhu and Carterette relied on evidence aggregation, external query expansion and Markov Random Fields. They employed 3 levels for merging the results of IR systems by evaluating visits, based on the best evidence from the reports, aggregation of reports to a visit, and finally the combination of both approaches. The 2012 winning system benefited from the availability of training data from 2011, and they performed optimization of parameters over the early query set. Both groups gained improvement by using external knowledge sources for query expansion, however, many other configurations contributed to the performance of their final systems.

Our group participated in both editions of the challenge. In TREC-M1 we mainly focused on external knowledge sources, such as the UMLS, and DBpedia for query expansion (Amini et al., 2011). In 2012 we took a different approach by locally expanding the queries with the collection (using pseudo-relevance feedback based on ICD codes), and by detecting and modifying the negated text in the reports (Amini et al., 2012). Our 2012 submission is the basis of this article, and here we extend the system by exploring the use of ICD as pseudo-relevance judgments, present diverse ways of mapping queries into ICD codes, and evaluate its performance systematically over the 2011 and 2012 medical TREC collections.