Development and Evaluation of a Case-Based Retrieval Service

Emilie PASCHE [a,b,1], Marcello CHINALI [c], Julien GOBEILL [a,b], Patrick RUCH[a,b]
[a] *BiTeM Group, Information Science Department, University of Applied Sciences of Western Switzerland (HES-SO, HEG), Switzerland*
[b] *SIB Text Mining, Swiss Institute of Bioinformatics, Switzerland*
[c] *Department of Pediatric Cardiology and Cardiac Surgery, "Bambino Gesù" Children Hospital, Italy*

Abstract. Identifying similar patients might greatly facilitate the treatment of a given patient, enabling to observe the response and outcome to a particular treatment. Case-based retrieval services dealing with natural language processing are of major importance to deal with the significant amount of unstructured clinical data. In this paper, we present the development and evaluation of a case-based retrieval (CBR) service tested on a collection of Italian pediatric cardiology cases. Cases are indexed and a search engine is proposed. Search functionalities, such as interactive MeSH normalization and relevance feedback, are proposed. While the qualitative evaluation aims to provide feedback and recommendations, the quantitative evaluation enables to estimate the precision of the system. In more than half of the cases and for up to two thirds of them, the system is able to suggest a similar episode of care at first rank. With an improvement of the feedback relevance strategy, we can expect an improvement of the precision. The CBR can be expanded to multilingual EHR and other fields.

Keywords. Natural language processing, ontology, electronic health record

1. Introduction

Physicians, who are facing complex diseases, show a great interest in finding populations of patients similar to their patients. Thus, they can observe the response of a particular treatment and learn about the outcomes at different points in time in a given clinical pathway. However, a substantial part of the clinical information is stored in unstructured textual contents. Therefore, tools are essential to enable the retrieval of similar cases. While case-based retrieval (CBR) tools based on structured data are numerous, less systems are managing unstructured data. Miotto *et al.* [1] describe a CBR system aiming to identify eligible patients for clinical trials. This system is based on structured data (i.e. diagnosis, medications and laboratory results) and unstructured data (i.e. clinical notes). Hsu *et al.* [2] present a CBR system dedicated to patients with intracranial aneurysm. This system uses various modalities, such as free-text clinical reports and structured data. A model-driven visualization enables to facilitate the

[1] Corresponding author, Emilie Pasche, Haute Ecole de Gestion, Campus Battelle Bâtiment B, Rue de la Tambourine 17, 1227 Carouge, Switzerland; E-mail: emilie.pasche@hesge.ch.

understanding of the output. Mourão *et al.* [3] report on a CBR system based on multimodal data. The system uses images to enrich the queries.

As part of the MD-PAEDIGREE project, a case-based retrieval (CBR) service has been developed. This service aims to find similar episodes of care based on different modalities: unstructured data (i.e. discharge summaries) and structured data (i.e. gender and age). In this paper, we present the development and evaluation of this CBR service, on a test collection of Italian pediatric patients with cardiac diseases.

2. Method

2.1. Data

The CBR is based on a set of 47,433 episodes of care, corresponding to 33,674 distinct patients consulting for cardiac pathologies. The source data originate from two Italian hospitals. Extracted episodes of care contain a discharge summary, called *clinical synthesis* in the following. Textual contents are in Italian. Demographic data (i.e. gender and age) are also retrieved.

2.2. Development of the case-based search engine

The system harvests electronic health records (EHR) from the MD-PAEDIGREE infostructure with a secured API developed by GNúBILA. Normalized descriptors (i.e. MeSH) are automatically assigned to each case using MHIta [4], a service developed by HES-SO to normalize clinical texts written in Italian with MeSH descriptors. Selection of MeSH descriptors is based on a dynamic threshold strategy. Cases are then indexed using Apache Solr. At query time, MeSH descriptors are also interactively being assigned to the query. The Solr retrieval engine outputs similar cases in EHR. A weight of 0.001 is attributed to the age and gender, while the MeSH descriptors and unstructured text receive a weight of 1. The user can then assess retrieved episodes of care as relevant or not relevant. These judgements are used to reformulate the query with additional keywords based on a Rocchio algorithm [5]. Therefore, the user can obtain refined results.

2.3. Quantitative and qualitative evaluation

While two out of five electronic information systems are abandoned or do not respect the requirements [6], the testing and validation of a service is of major importance. Three dimensions are commonly considered: the usefulness of a medical system, its robustness and its facility of use.

The qualitative evaluation is based on the usability testing methodology [7]. The system is tested by an end-user (i.e. a MD specialized in pediatric cardiology), who performs tasks (i.e. to search for similar cases to a given case). During the whole process, the end-user is asked to verbalize his thoughts. An evaluator (i.e. a researcher) is observing and recording his comments. The data are then coded and classified by the evaluator and recommendations to improve the systems are proposed.

The quantitative evaluation is based on a benchmark, thus following the standard practice in the information retrieval domain [8]. A set of 40 queries is created. A query

corresponds to the clinical synthesis of an episode of care randomly selected among the 47,433 episodes of care of the collection. An expert in cardiology manually acquires the relevance judgments. The expert executes each of the queries on the CBR and assesses the top-10 results with one of the following categories: relevant (i.e. similar to the input case) or irrelevant (i.e. judged as not similar to the query). Because this task is precision-oriented (i.e. the CBR does not aim at retrieving all the similar cases, but rather at retrieving some similar cases in order to extract useful information), we focus on precision metrics. Precision is the proportion of retrieved instances that are correct.

3. Results

3.1. GUI

The service can be accessed through the MD-PAEDIGREE portal but is restricted to allowed users due to the use of confidential data. The CBR service is a 5-step process. First, the user describes a patient with natural language, and can optionally add the age and gender of the patient. Second, the query can be interactively refined with additional keywords automatically suggested: MeSH concepts or keywords obtained by relevance feedback (only available after a first iteration). Third, the user can filter the results based on the structured data (e.g. show only boys from 3 to 10 year-old). Fourth, the similar episodes of care are displayed, ranked by relevance. To facilitate the processing by the physician, following information is displayed: demographic information (i.e. gender and age), MeSH terms automatically attributed to the clinical synthesis, clinical synthesis, a relevance score, link to the full patient history. In addition, a radio button is proposed, representing the relevance judgement. The user can then iterates to refine his query and thus obtain more relevant results, or he can expand his query to external resources (e.g. literature).

3.2. Qualitative evaluation

The end-user appreciated the simplicity of use of the CBR service. Nevertheless, a few technical problems have arisen during the evaluation. The two main problems observed were truncated reports and the failure to answer to some queries. Those technical problems were fixed right after the evaluation session.

The automatic MeSH normalisation triggered a strong interest from the evaluator, which is familiar with the terminological resources as it is used by the MEDLINE digital library – the legacy reference for healthcare literature.

The Rocchio relevance feedback feature showed some limitations during the evaluation session. The suggested terms were reported as too general (i.e. common Italian words) or not clinically relevant. However, data analysis showed that for more than 90% of the queries, a few terms were selected.

Regarding the similar episodes of care suggested by the CBR, it was reported that the system was very efficient to retrieve similar cases when the input case was a regular case. However, the system showed difficulties to deal with the detection of the grade (e.g. normal, minor, severe, etc.), with the detection of negation or nuance (e.g. may be, unlikely, etc.), or with long and complex queries.

The evaluators also tested the preliminary version of the Rocchio-based relevance service. The results were very diverse: for a few queries, some additional relevant

documents were retrieved, for others irrelevant documents were added, while for some queries, the additional keywords did not bring any change in the ranking of the cases.

3.3. Quantitative evaluation

Among the 40 queries, two queries were excluded due to technical failure. Eight queries returned no similar case among the top-10. Table 1 shows different measures of precision, for all queries, and for queries with at least a relevant identified answer. In more than half of the cases and for up to two thirds of them, the system is able to suggest a similar episode of care at first rank. Further, Table 2 presents the results obtained with the relevance feedback algorithm. We observe a slight improvement of the P5 and P10 with the Rocchio-based results.

	All queries (38)	**Queries with at least a relevant case (30)**
P0	0.5	0.63
P5	0.44	0.55
P10	0.42	0.54

Table 1. Evaluation of the first round of results returned by the CBR

	All queries (24)	**Queries with at least a relevant case (19)**
P0	0.5	0.63
P5	0.52	0.65
P10	0.45	0.56

Table 2. Evaluation of the Rocchio-based results returned by the CBR

4. Discussion

For eight queries, no similar case was found in the top-10. There are two hypotheses that might be considered to explain such phenomena: the system was not able to find relevant documents for these queries; the collection did not contain any relevant documents for these queries, meaning the case is so rare that there is no similar case. It is useful to highlight in this respect, that being the experiment conducted among patients with rare diseases (pediatric congenital cardiac malformations), unique cardiac phenotypes are often encountered in clinical practice, possibly explaining the lack of similarity match. If this second explanation is valid then such queries are artificially decreasing the precision of the search engine. The real precision of the system is therefore located between these lower (i.e. including the eight queries with no relevant document identified) and upper boundaries (i.e. excluding the eight queries).

The relevance feedback functionality is worth being further explored. Indeed, despite its very basic tuning at the moment of the evaluation, the quantitative evaluation showed a small positive impact on the precision. Several options are envisaged: 1) filtering of the terms suggested by the Rocchio algorithm to clinical terms only; 2) investigating negative feedback; 3) filtering words with a high document frequency using IDF (Inverse Document Frequency).

While this approach has been tested on a collection of documents in Italian, it can be expended to other languages. Indeed, the cases are automatically normalized with a terminology available in multiple languages: the MeSH terminology. Developments are being made to integrate episodes of care in English.

A limitation of our quantitative evaluation study is first the limited number of results assessed for each query and second the single expert who evaluated the results. A higher number of results assessed would enable the possibility to tune the system (i.e. to try to maximize the number of relevant results in the first positions). Indeed, as we do not know if results from position 11 are relevant, any of these results pushed in a top position after tuning would decrease the precision.

5. Conclusion

We have thus developed a case-based retrieval service dealing with several modalities (e.g. structured data, unstructured data, ontologies) and proposing various functionalities to search for similar cases (e.g. search in EHR, search in literature, relevance feedback, etc.). A methodology to develop and monitor the progress of the CBR prototype has been implemented and tested. The feedback obtained from the qualitative evaluation, despite the known rarity of the group of diseases analysed, was sufficient to improve the application regarding usability. From a quantitative point of view, the current results are already regarded as fair to support a case-based retrieval application, although several components, such as the relevance feedback service, needs fine-tuning to convince the end-users.

Acknowledgement.

This experiment has been supported by the MD-PAEDIGREE project, partially funded by the European Union under the Information Communication Technologies Programme (contract number 600932). We would like also to acknowledge David Manset, Sébastien Gaspard and Nicolas Mugnier for their help with the extraction of EHRs and integration of the CBR within the MD-PAEDIGREE portal.

References

[1] R. Miotto, C. Weng, Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials, *J Am Med Inform Assoc* **22** (2015), e141-50.
[2] W. Hsu, R.K. Taira, F. Viñuela, A.A. Bui, A Case-based Retrieval System using Natural Language Processing and Population-based Visualization, *Proc IEEE Int Conf Healthc Inform Imaging Syst Biol* **2011** (2011), 221-228.
[3] A. Mourão, E. Martins, J. Magalhães, Multimodal medical information retrieval with unsupervised rank fusion, *Comput Med Imaging Graph* **39** (2015), 35-45.
[4] P. Ruch, Automatic assignment of biomedical categories: toward a generic approach, *Bioinformatics* **22(6)** (2006), 658-64.
[5] P. Ruch, I Tbahriti, J. Gobeill, A.R. Aronson, Argumentative feedback: A linguistically-motivated term expansion for information retrieval, *Proceedings of the COLING/ACL* (2006):675-82.
[6] J. Horsky, K. McColgan, J.E. Pang, A.J. Melnikas, J.A. Linder, J.L. Schnipper, B. Middleton, Complementary methods of system usability evaluation: surveys and observations during software design and development cycles, *J Biomed Inform* **43(5)** (2010):782-90.
[7] A. Kushniruk, Evaluation in the design of health information systems: application of approaches emerging from usability engineering, *Comput Biol Med* **32(3)** (2002):141-49.
[8] S.E. Robertson, S. Walker, M. Beaulieu, Experimentation as a way of life: Okapi at TREC, *Information Processing & Management* **36(1)** (2000):95-108.