Streamlining a Transplantation Survival Prediction Program with a RDF Triplestore

Dennis Medved*, Johan Nilsson•, and Pierre Nugues*

(*) Department of Computer Science, Lund University, Lund, Sweden
(•) Department of Clinical Sciences, Cardiothoracic Surgery, Lund University and Skåne University Hospital, Lund, Sweden
{dennis.medved,pierre.nugues}@cs.lth.se,johan.nilsson@med.lu.se

Abstract. In this paper, we describe the conversion of a heart transplantation dataset in the RDF format and its application in a survival prediction program. The International Society for Heart & Lung Transplantation (ISHLT) maintains a registry of heart transplantations that it gathers from operations performed worldwide. The US United Network for Organ Sharing (UNOS) and the Scandinavian Scandiatransplant are contributors of this registry although they use different data models. We designed a unified graph representation covering the three data models and we converted the ISHLT tables into RDF triples. We used the resulting triplestore as input to a survival prediction program for transplanted patients [3, 2]. Recipient and donor properties are essential to predict the mortality or survival of the patients. In contrast with the manual techniques we used to extract data from the tabulated files, the RDF triplestore enables us to experiment quickly and automatically with combinations of features that influence most the survival that we will demonstrate at the conference.

Keywords: RDF triplestore; transplantation data set; feature sets

1 Introduction

The International Society for Heart & Lung Transplantation (ISHLT) maintains a registry of heart transplantations it collects from national or regional organizations across the world. ISHLT aggregates the data submitted by the contributing organizations. The US United Network for Organ Sharing (UNOS) and the Scandinavian Scandiatransplant are two such contributing organizations. In total, ISHLT contains about 100,000 recorded heart transplantations. The data is normally redistributed to researchers after a request in a tabular format.

Although ISHLT could be a superset of all the data it receives, it merely includes the variables that are frequently recorded by the regional registries. The three data sources we considered, ISHLT, UNOS, and Scandiatransplant, have then a different structure, different variables, and use different variable names. ISHLT, for instance, does not feature the variable `crossmatch_done`, a HLA compatibility test, that is documented by UNOS.

Patient and donor factors are essential to predict the mortality of heart transplantations [4]. Domingos (2012) provides an eloquent advocacy of the importance of such factors, or features, in the success of machine-learning projects [1]. As we wanted to mine automatically the feature sets from the data sets and integrate data from all our sources, including Scandiatransplant, we designed a unified, extendible, RDF representation. Our goal was to make the data extraction easier, possibly based on the data from different organizations, and automate the machine learning tasks, notably the feature selection.

2 Creation of a RDF Triplestore

The ISHLT, UNOS, and Scandiatransplant data sets are normally distributed to the researchers as SAS or CSV files. We started from the CSV files and we created the RDF triplestore using Google Refine[1] and RDF Refine[2]. The facet functions of Google Refine, notably, were instrumental to clean the data and build unified value names.

The CSV files represent the transplants as rows, where each column is a variable for the transplant. In the RDF conversion, we mapped each row to a head node and we created leaf nodes for the selected variables.

The data sets use different names to denote the variables. For example, the most recent creatinine value for the recipient patient is *Most rec. Creat.* in Scandiatransplant, *creat* in ISHLT, and *creat_trr* in UNOS (Figure 1). We created unified names for about 140 of the variables, such as `aaot:creatinine` for the creatinine value, where the `aaot` prefix stands for *Algorithms and Applications for Organ Transplantation*. UNOS has more variables than ISHLT and Scandiatransplant. We took the UNOS names as is when they had no counterpart in the two other registries.

We finally added metadata about the variables containing the original variable name as well the new one, the description of the variable, the source form of the data, as well as comments and variable start and end date. Figure 2 shows the metadata on `aaot:creatinine`.

3 Generating Data Sets from the Triplestore

We created a SPARQL endpoint using the openRDF Sesame framework[3]. Compared with the tedious copy-and-paste techniques we used to create data sets to test our survival predication programs, SPARQL offers an easier way to extract relevant data samples. The extraction of the survival duration for transplants matching the conditions:

– The recipient is a male older than 17 with blood group A;

[1] `http://code.google.com/p/google-refine/`
[2] `http://refine.deri.ie/`
[3] `http://www.openrdf.org/`

Fig. 1. An unification of the variable representing the most recent creatinine level of the recipient.

- The donor is female with blood group A;
- From Scandiatransplant and ISHLT registries

is concisely expressed using the SPARQL query:

```
SELECT ?transplant ?survival_time
FROM <file://Scandia.ttl>
FROM <file://ISHLT.ttl>
WHERE {
    ?transplant aaot:gender "M" .
    ?transplant aaot:age ?age .
    ?transplant aaot:ABO "A" .
    ?transplant aaot:gender_donor "F" .
    ?transplant aaot:ABO_donor "A" .
    ?transplant aaot:survival_time ?survival_time .
    FILTER (?age > 17)
}
```

4 Testing Survival Hypotheses

The *Virtual recipient-donor match program* is an artificial neural network (ANN) that models survival curves for pairs of recipients and donors [3, 2]. This model can be used to predict the outcome for each potential recipient in a waiting list. The program then produces a ranking list, from low-risk to the high-risk recipients, that can be presented to the transplant physicians.

The system demonstration includes a Java program that queries the triple-store and extract sets of variables to evaluate in the survival hypotheses. This

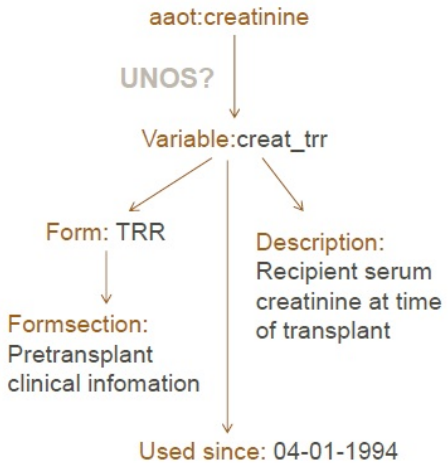

Fig. 2. `aaot:creatinine` metadata for the UNOS part of the database.

enables us to generate matrices systematically and easily and have a comprehensive experimental setup. We apply a set of transformations to replace possible missing values though a random assignment from the set of existing values and the conversion of nominal data into vectors of numbers. The virtual match program then predicts the survival curves for the patients in a waiting list.

References

1. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM 55(10), 78–87 (Oct 2012), http://doi.acm.org/10.1145/2347736.2347755
2. Nilsson, J., Ohlsson, M., Höglund, P., Ekmehag, B., Koul, B., Andersson, B.: Artificial neural networks - relative importance of different recipient-donor characteristic combinations on survival after heart transplantation. The Journal of Heart and Lung Transplantation 30, S68 (2011)
3. Nilsson, J., Ohlsson, M., Höglund, P., Ekmehag, B., Koul, B., Andersson, B.: Virtual recipient donor match – a new way to increase the number of heart transplantations. Submitted (2013)
4. Weiss, E.S., Allen, J.G., Arnaoutakis, G.J., George, T.J., Russell, S.D., Shah, A.S., Conte, J.V.: Creation of a quantitative recipient risk index for mortality prediction after cardiac transplantation (impact). The Annals of Thoracic Surgery 92(3), 914 – 922 (2011), http://www.sciencedirect.com/science/article/pii/S0003497511009350