MINING OF HEALTH INFORMATION FROM ONTOLOGIES

Maja Hadzic
Research Lab for Digital Health Ecosystems, Curtin Universityof Technology
m.hadzic@curtin.edu.au

Fedja Hadzic, Tharam Dillon
Digital Eciossytems and Business Intelligence Institute (DEBII), Curtin Universityof Technology
fedja.hadzic@postgrad.curtin.edu.au, t.dillon@curtin.edu.au

Keywords: Data mining, Tree mining, Ontology, Ontology mining, Human disease ontology, Human disease study, Health information system

Abstract: Data mining techniques can be used to efficiently analyze semi-structured data. Semi-structured data are predominantly used within the health domain as they enable meaningful representations of the health information. Tree mining algorithms can efficiently extract frequent substructures from semi-structured knowledge representations. In this paper, we demonstrate application of the tree mining algorithms on the health information. We illustrate this on an example of Human Disease Ontology (HDO) which represents information about diseases in 4 'dimensions': (1) disease types, (2) phenotype (observable characteristics of an organism) or symptoms (3) causes related to the disease, namely genetic causes, environmental causes or micro-organisms, and (4) treatments available for the disease. The extracted data patterns can provide useful information to help in disease prevention, and assist in delivery of effective and efficient health services.

1 INTRODUCTION

New modern techniques are providing huge, rapidly accumulating amounts of information. To extract and analyze the data poses a much bigger challenge for researchers than to generate the data (Holloway *et al.*, 2002). Experienced scientists and doctors are overwhelmed with this situation.

There is a need for an intelligent and efficient system to make use of all the available information. The true value of this information can be significantly increased through smart information processing and analysis. Such systems could play a crucial role in filtering the flood of data to the point where human experts could apply their knowledge sensibly.

Information technologies must be effectively implemented within health domain. In their paper, Horvitz-Lennon *et al.* (2006) state that we need to fully embrace information technology and its potential for improving service efficiency and develop a better information infrastructure for the patient's care.

Data mining is a set of processes that is based on automated searching for actionable knowledge buried within a huge body of data. Data mining techniques extract information and find hidden patterns and behaviors, and support making predictive models for decision making and new discoveries. Within the biomedical and health field, data mining techniques have been predominately used for tasks such as text mining, gene expression analysis, drug design, genomics and proteomics (Zaki *et al.*, 2003). The data analysis necessary for microarrays has necessitated data mining (Piatetsky-Shapiro & Tamayo, 2003). Recently, use of data mining methods has been proposed for the purpose of mapping and identification of complex disease loci (Onkamo & Toivonen, 2006). However, the proposed methods are yet to be implemented. Frequent pattern analysis has been a focused theme of study in data mining. A lot of algorithms and methods have been developed for mining frequent sequential and structural patterns (Han & Kamber, 2001; Agrawal & Srikant, 1994; Tan *et al.*, 2006a). Implementation of data mining techniques within health domain could help in disease prevention and assist in delivery of effective and efficient health services.

Within the data mining field, tree mining has recently attracted lots of interest. Our work in the tree mining field is characterized by a Tree Model Guided (TMG) (Tan *et al.*, 2005; Tan *et al.*, 2006b) candidate generation approach. This non-redundant systematic enumeration model uses the underlying tree structure of the data to generate only valid candidates. Using the general TMG framework a number of algorithms were developed as follows. MB3-Miner (Tan *et al.*, 2005) mines ordered embedded subtrees while IMB3-Miner (Tan *et al.*, 2006b) can mine both, induced or embedded ordered subtrees by using the level of embedding constraint. Razor algorithm (Tan *et al.*, 2006c) was developed for mining of embedded subtrees where the distances of nodes relative to the root of the original tree need to be considered. UNI3 algorithm (Hadzic *et al.*, 2007a) mines induced unordered subtrees. Our algorithms were applied on large and complex tree structures and their scalability was experimentally demonstrated (Tan *et al.*, 2005; Tan *et al.*, 2006b). From the application perspective, in (Hadzic *et al.*, 2006) we have applied our tree mining algorithm for the analysis of Protein Ontology (Sidhu *et al.*, 2004) database for Human Prion proteins which was represented in XML format. In this paper, we explain how tree mining techniques can be applied within the health domain for deriving useful knowledge patterns that can help disease prevention and management.

2 HUMAN DISEASE ONTOLOGY

We designed Human Disease Ontology (GHDO) (Hadzic & Chang, 2005) to have the following four branches or subontologies:

- *disease types*, describing different types of a disease;
- *phenotype*, describing disease symptoms;
- *causes* responsible for that disease that can be genetic, environmental and/or microorganism;
- *treatments*, providing an overview of all treatments possible for a particular disease;

Top-level hierarchy of the HDO is illustrated in Figure 1.

The information presented in Figure 1 state that a disease may have different types, and associated subtypes and sub-subtypes. For each disease, there is a corresponding phenotype (or observable characteristics of an ill individual), namely symptoms of a disease. Disease cause can be genetic (genotype), environmental or a microorganism. Genetic causes can be a mutated gene, a complex of genes or DNA region of interest. DNA region of interest is a region in the DNA sequence that potentially contains a gene responsible for the disease. This region needs to be further examined in order to correctly locate the mutated gene. Environmental causes of a disease can be stress, climate, drugs misuse, family conditions or economic condition. Microorganisms that may cause a disease may be virus or bacteria. Possible treatments for a disease can be drug therapy, chemotherapy, surgery, psychotherapy or physiotherapy.

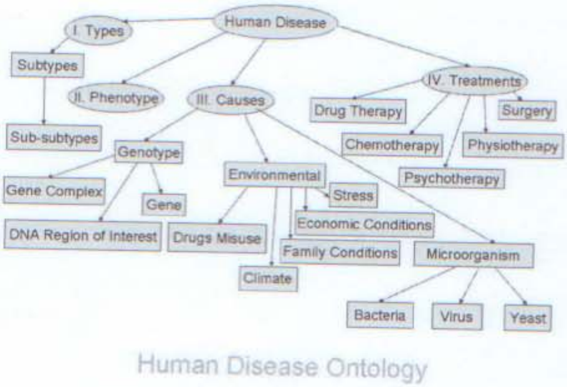


Figure 1: Top-level hierarchy of the HDO

Researchers in the medical ontology-design field have developed different terminologies and ontologies in many different areas of medical domain. In order to obtain some uniformity across different ontologies, definitions from other published and consensual ontologies can be reused (Noy & Musen, 2000). We can use other ontologies such as LinkBase (Montyne, 2001) and UMLS (Bodenreider, 2004). These ontologies contain over million concepts, and we do not require all of them. The four different branches (subontologies) of the GHDO ontology can serve as a reference point against which the concepts from the existing ontologies can be reorganized, aligned and merged. Bimolecular ontologies such as TAMBIS ontology (Stevens *et al.*, 2002) that represents general knowledge in regard to proteins and associated genes may be suitable to cover the ontology part in regard to genetical disease causes. So, we can use terminology from existing ontologies but select and organize the concepts in a way that can be used in our application.

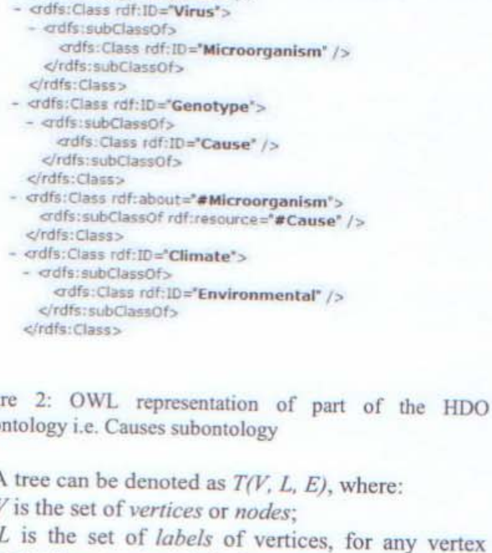
Note that the HDO has a tree (hierarchy) structure which allows it to be analyzed using some available tree mining techniques. The current ontology languages allow the use of graph structures to represent the domain knowledge in an ontology. A large portion of current ontologies have predominantly hierarchical structures. Furthermore, it is often the case that the graph-structured knowledge representation can be modelled using tree structures without losing too much semantics. The root of the complexity of mining graph structures is in the existence of cycles, and in many cases the number of cycles in graph instances is limited. The complexity of processing tree structures tends to be more manageable and is one of the promising directions towards automatic analysis of ontologies.

We assume that the available health information can be represented according to the four HDO 'dimensions' or subontologies. In the rest of our paper, we will use the term 'HDO instance' to refer to a specific record found within a given database/application that can be represented using the HDO structure. In our previous publications such as (Hadzic & Chang, 2005), this 'HDO instance' corresponds to Specific Human Disease Ontology.

3 TREE MINING CONCEPTS

The aim of this section is to provide the definitions of some basic tree concepts necessary for understanding the current work. Please refer to (Tan *et al.*, 2005; Tan *et al.*, 2006b; Hadzic *et al.*, 2007a) for a more extensive overview of the tree mining area including the discussion of implementation issues and algorithm comparisons. A tree is a special type of graph where no cycles are allowed. It consists of a set of *nodes (or* vertices) that are connected by *edges*. Each edge has two nodes associated with it. A *path* is defined as a finite sequence of edges and in a tree there is a single unique path between any two nodes. The *length of a path p* is the number of edges in p. A rooted tree has its top-most node defined as the *root* that has no incoming edges and for every other node there is a path between the root and that node. A node u is said to be a *parent* of node v, if there is a directed edge from u to v. Node v is then said to be a *child* of node u. Nodes with no children are referred to as *leaf* nodes and otherwise they are called *internal nodes*. If for each internal node, all the children are ordered, then the tree is an ordered tree. The problem of frequent subtree mining can be generally stated as: Given a tree database T_{db} and minimum support threshold (σ), find all subtrees that occur at least σ times in T_{db}.

A HDO instance can be captured by an OWL document analogous to the one shown in Figure 2. In our example, the actual information content can be viewed as an ordered labeled tree. This specific instance aims to capture the information about causal factors of a human disease and corresponds with the Causes subontology shown in Figure 1.



```
  <?xml version="1.0" ?>
- <rdf:RDF xmlns="http://www.owl-ontologies.com/cause
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#" )
    xmlns:p1="http://www.owl-ontologies.com/assert.ow
    <owl:Ontology rdf:about="" />
  - <rdfs:Class rdf:ID="Virus">
    - <rdfs:subClassOf>
        <rdfs:Class rdf:ID="Microorganism" />
      </rdfs:subClassOf>
    </rdfs:Class>
  - <rdfs:Class rdf:ID="Genotype">
    - <rdfs:subClassOf>
        <rdfs:Class rdf:ID="Cause" />
      </rdfs:subClassOf>
    </rdfs:Class>
  - <rdfs:Class rdf:about="#Microorganism">
      <rdfs:subClassOf rdf:resource="#Cause" />
    </rdfs:Class>
  - <rdfs:Class rdf:ID="Climate">
    - <rdfs:subClassOf>
        <rdfs:Class rdf:ID="Environmental" />
      </rdfs:subClassOf>
    </rdfs:Class>
```

Figure 2: OWL representation of part of the HDO subontology i.e. Causes subontology

A tree can be denoted as $T(V, L, E)$, where:

(1) V is the set of *vertices* or *nodes*;

(2) L is the set of *labels* of vertices, for any vertex $v \in V$, $L(v)$ denotes the label of v; and

(3) $E = \{(x,y) | x,y \in V\}$ is the set of edges in the tree.

In Figure 3, we represent the OWL document from Figure 2 as a tree with "Cause" as the root node.

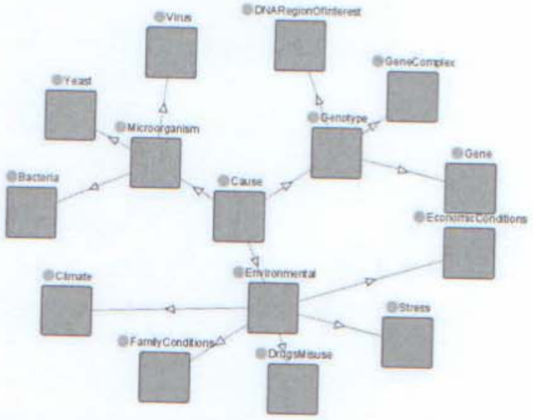


Figure 3: Causes subontology of HDO (viewed in Protégé)

Most of the current tree mining algorithms are mainly focused on extracting induced and embedded subtrees. An induced subtree preserves the parent-child relationships of each node in the original tree. In addition to this, an embedded subtree allows a parent in the subtree to be an ancestor in the original tree and hence ancestor-descendant relationships are preserved over several levels. Formal definitions follow:

A tree $T'(V', L', E')$ is an **induced** subtree of a tree $T (V, L, E)$ iff (1) $V' \subseteq V$, (2) $E' \subseteq E$, (3) $L' \subseteq L$ and $L'(v)=L(v)$, (4) $\forall v' \in V'$, $\forall v \in V$, v' is not the root node, and v' has a parent in T', then parent(v')=parent(v).

A tree $T'(V', L', E')$ is an **embedded** subtree of a tree $T(V, L, E)$ iff (1) $V' \subseteq V$, (2) if $(v_1,v_2) \in E'$ then $parent(v_2) = v_1$ in T', only if v_1 is ancestor of v_2 in T and (3) $L' \subseteq L$ and $L'(v)=L(v)$.

The subtrees can further be distinguished depending on the order of sibling nodes. An **ordered** subtree preserves the left-to-right ordering among the sibling nodes in the original tree. In an **unordered** subtree the left-to-right ordering among the sibling nodes does not need to be preserved. The order of the sibling nodes (and the subtrees rooted at sibling nodes) can be exchanged and the resulting subtree would be considered the same. Examples of different subtree types related to the Figure 3 are given in Figure 4.

The available support definitions are transaction based, occurrence-match, and hybrid support (Tan *et al.*, 2006b; Hadzic *et al.*, 2007b). **Transaction based support** only checks for the existence of items in a transaction while **occurrence-match** support takes the repetition into account and counts the item occurrences in the database as a whole. Using **hybrid support** with a threshold set to $x|y$, a subtree will be considered as frequent iff it occurs at least y times in x number of transactions.

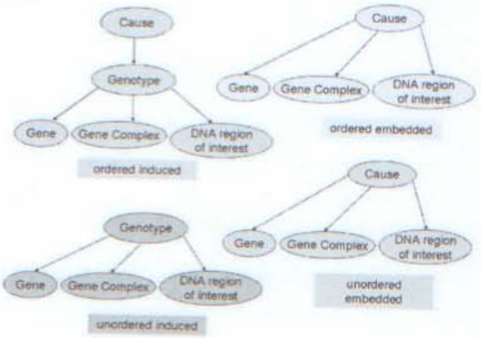


Figure 4: Example of different subtree types

4 MINING OF HEALTH IMFORMATION

In order to apply tree mining to the problem, the following five phases take place: data selection and cleaning, data formatting, tree mining, pattern discovery, and knowledge testing and evaluation.

In the first phase, we focus on the Data Selection and Cleaning. Most of the databases also contain information that is not needed by the application. The irrelevant information, as well as noise and inconsistent data, should be removed from the data set to be mined. Extra care should be taken during this process since some data may appear to be noisy but in fact represents a true exceptional case.

In the second phase, Data Formatting, all the collected data needs to be represented using the same format that is understandable by the tree mining algorithm used.

The third phase is concerned with applying the tree mining algorithms for extracting interesting patterns from now a clean and correctly formatted dataset. One needs to consider what particular type of subtree is the most suitable one to mine for satisfying the application needs. When the data to be mined comes from one organization, the format and ordering of the presented information is expected to be the same, and hence mining of ordered subtrees will be sufficient. On the other hand, if the collected data originate from separate organizations, then mining of unordered trees would be more suitable. The organizations could order their concepts differently, but since an unordered subtree with different order among the sibling nodes is still considered as the same candidate, the common characteristics of a particular illness would still be found. Another choice to be made is whether induced or embedded subtrees should be mined. When mining health related data it is important that particular information stays in the context where it occurred. With respect to tree patterns this implies that the relationship of nodes in the extracted subtrees should be limited to parent-child relationships. Allowing ancestor-descendant relationships would result in information loss about the context where a particular disease characteristic occurred. Some attributes in the dataset may have a similar set of values and by mining induced subtrees there the attribute to which each value belongs will be indicated in the extracted pattern.

As a final consideration the support definition chosen should be dependant on the way that the data is organized. Next we provide three common ways in which the data could be presented, and indicate the suitable support definition that should be used.

Case 1: Each HDO instance is stored as a separate subtree in the OWL document and HDO instances describing different diseases are stored in separate documents. In this case both, occurrence match or transaction based support would be suitable.

Case 2: Each HDO instance is stored as a separate subtree in the OWL document but now one OWL document contains all HDO instances for all investigated diseases. Here the transactional support would be more appropriate.

Case 3: A collection of HDO instances related to one particular disease is always contained in a separate subtree of an OWL document. Hybrid support definition is most suitable in this case.

In cases 1 and 2, the minimum support threshold should be close to the number of HDO instances that the dataset contains about a particular disease, in order to find the commonality among all the records related to that particular disease. However, since noise is often present in the data, the support can lowered but not too much so that irrelevant factors are not picked up as important. In case 3, the number of diseases described would be used as the transactional part of the hybrid support, while the minimum occurrence of a subtree within each transaction should reflect the number of HDO instances.

As a common data mining practice, the data set at hand could be split into two subsets. One is used for deriving the knowledge model ('internal data' from Figure 5) and while the second one is used for testing the derived knowledge model ('external data' from Figure 5). When possible the data collected by another organization can be used as external data. During Pattern Discovery phase, new knowledge about specific disease(s) emerges. For example, these results may help to associate precise combinations of genetic and environmental factors with a specific disease type. The results could increase the understanding of the disease under study and make a breakthrough in the research, control and prevention of this disease. Knowledge Testing and Evaluation is illustrated in Figure 5. The 'external data' is used to verify a formed hypothesis before it can extend the current body of knowledge. The choice of the tree mining parameters often affects the nature and granularity of the obtained results. In cases where the hypothesis is not supported by the 'external data', the parameters will be adjusted and the previous steps alternated.

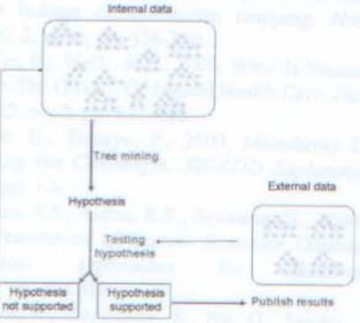


Figure 5: Testing and evaluation of the derived knowledge

5 CONCLUSION

Data mining systems in general could play a crucial role in deriving knowledge and assisting in the prevention, diagnosis, treatments and control of human