A Medical Information Retrieval System

Mahsa Kiani, Virendrakumar C. Bhavsar, and Harold Boley
{Mahsa.Kiani, Bhavsar, Harold.Boley}@unb.ca
Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada

Introduction

The first known health record was developed in the fifth century B.C. to express the course of disease, and its probable causes. The first Electronic Health Record Systems (EHRs) began to appear in the 1960s. Previous health records contain valuable information, which could be reused. A retrieval system compares the situation of the current patient (i.e., the query) with the stored records in the database; and it retrieves the most similar records to the query

Health records have been represented using vectors of attributes before. Feature vectors ignore the relations between concepts; however, the relations contain context information as well. Although graphs could represent the binary relationships between different concepts, their comparison could have high computational complexity.

We represent each health record using a hierarchical graph, which is called a generalized tree. The existing unlabeled/ vertex- labeled generalized tree structures have been extended to consider edge labels as well as edge weights. Then, we propose a new similarity approach which integrates the similarity of corresponding attributes with the structure similarity.

Representation of Data

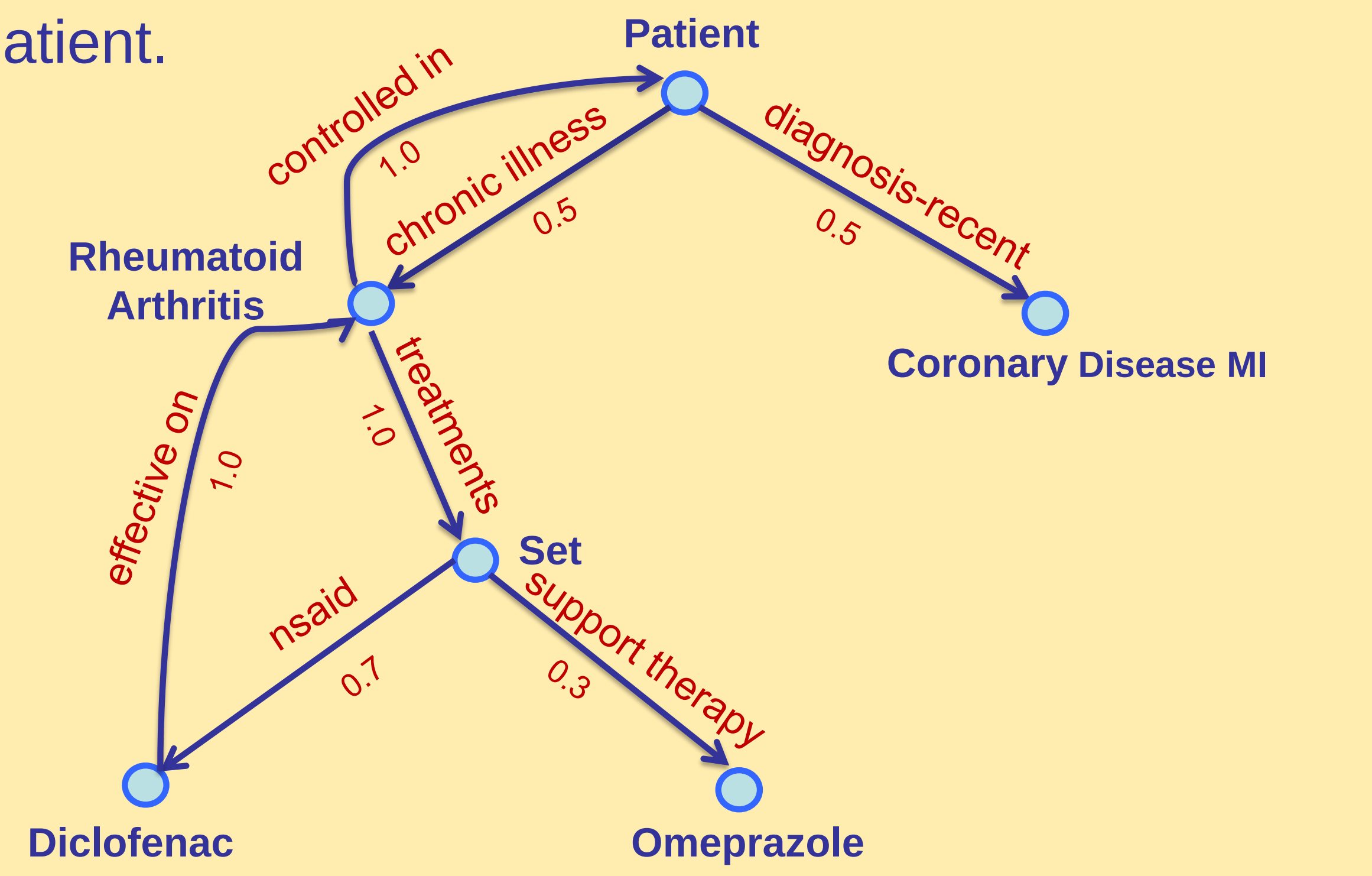
Given query as well as stored structured data are represented as an Attributed Generalized Tree.

Edge weights express users' assessment regarding the relative importance of the attributes represented by edge labels.

➢Labels are unique and appear in lexicographic left-to-right order.

➢Weights are in the real interval [0, 1] and for each generalized tree its edge weights are normalized.

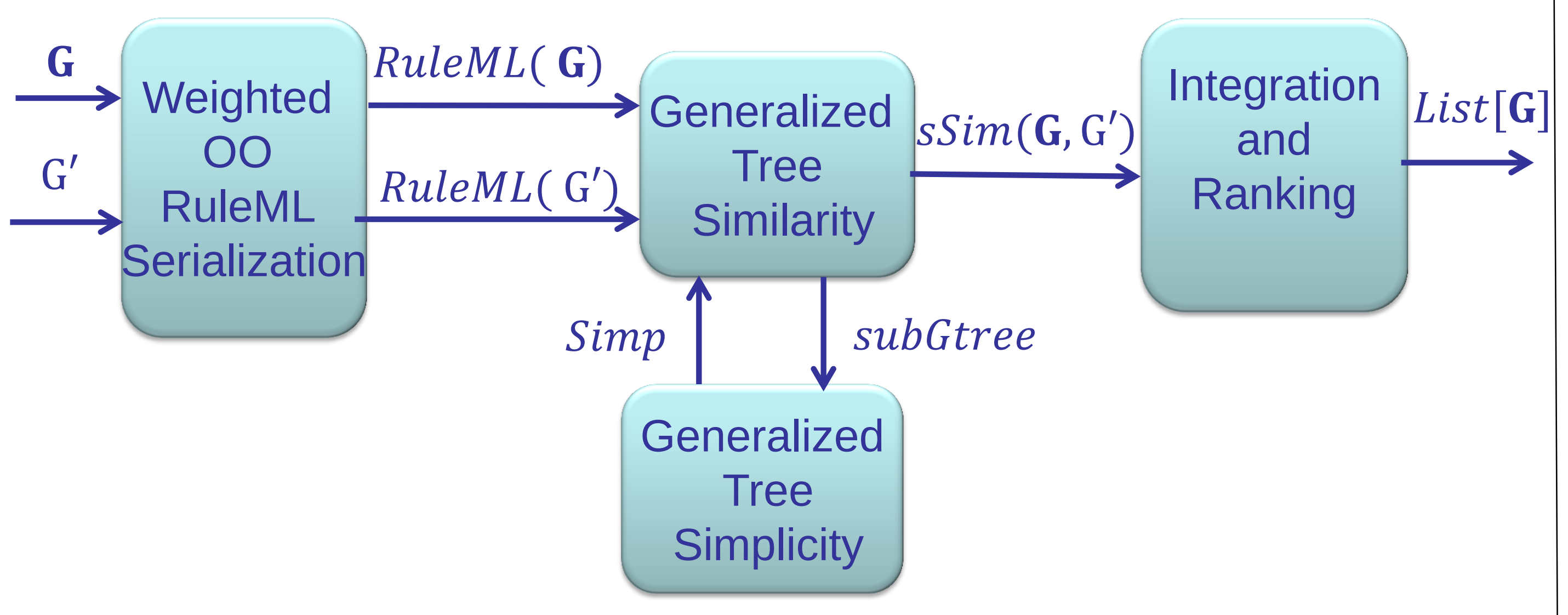
The patient has been recently diagnosed with Coronary Disease MI. She has a chronic Rheumatoid Arthritis, which is treated by Diclofenac and Omeprazole. Diclofenac has been reported to be effective on Rheumatoid Arthritis, and the condition is controlled in the patient.



Edge weights are assigned based on expert knowledge.

Architecture

- $\mathbf{G}$ and G': A set of stored generalized trees extracted from Electronic Health Records, and a given query being matched
- $sSim(\mathbf{G},\mathrm{G}')$: Structure similarity values
- $subGtree$: Missing sub-structure
- $Simp$: Simplicity value
- $subGtree$: Missing sub-structure
- $List[\mathbf{G}]$: Ranked Generalized Trees



System Architecture

Weighted OORuleML Serialization

Given query and each stored generalized tree are uniformly represented and interchanged using a weighted extension of Object Oriented RuleML [1]. This approach preserves all structural information of each generalized tree including its hierarchical structure, vertex labels, edge labels, and edge weights.

Similarity Module

In order to compute structure similarity of two given generalized tree, the vertices of two structures are visited simultaneously, starting from their roots. Two structures are traversed by matching the corresponding edges considering their edge weights. The cycle-detection strategy of the depth-first search algorithm is used to handle cycles in traversing two generalized trees.

Simplicity Module

The contribution of missing sub-structures in the overall structure similarity.is computed using a recursive simplicity algorithm. Increasing the number of vertices and edges in missing sub-structure decreases the simplicity; Decreasing the simplicity value results in decreasing the similarity of the pair of generalized tree structures.

Integration and Ranking Module

The integration and ranking module ranks the generalized trees in $\mathbf{G}$ based on the structure similarity.

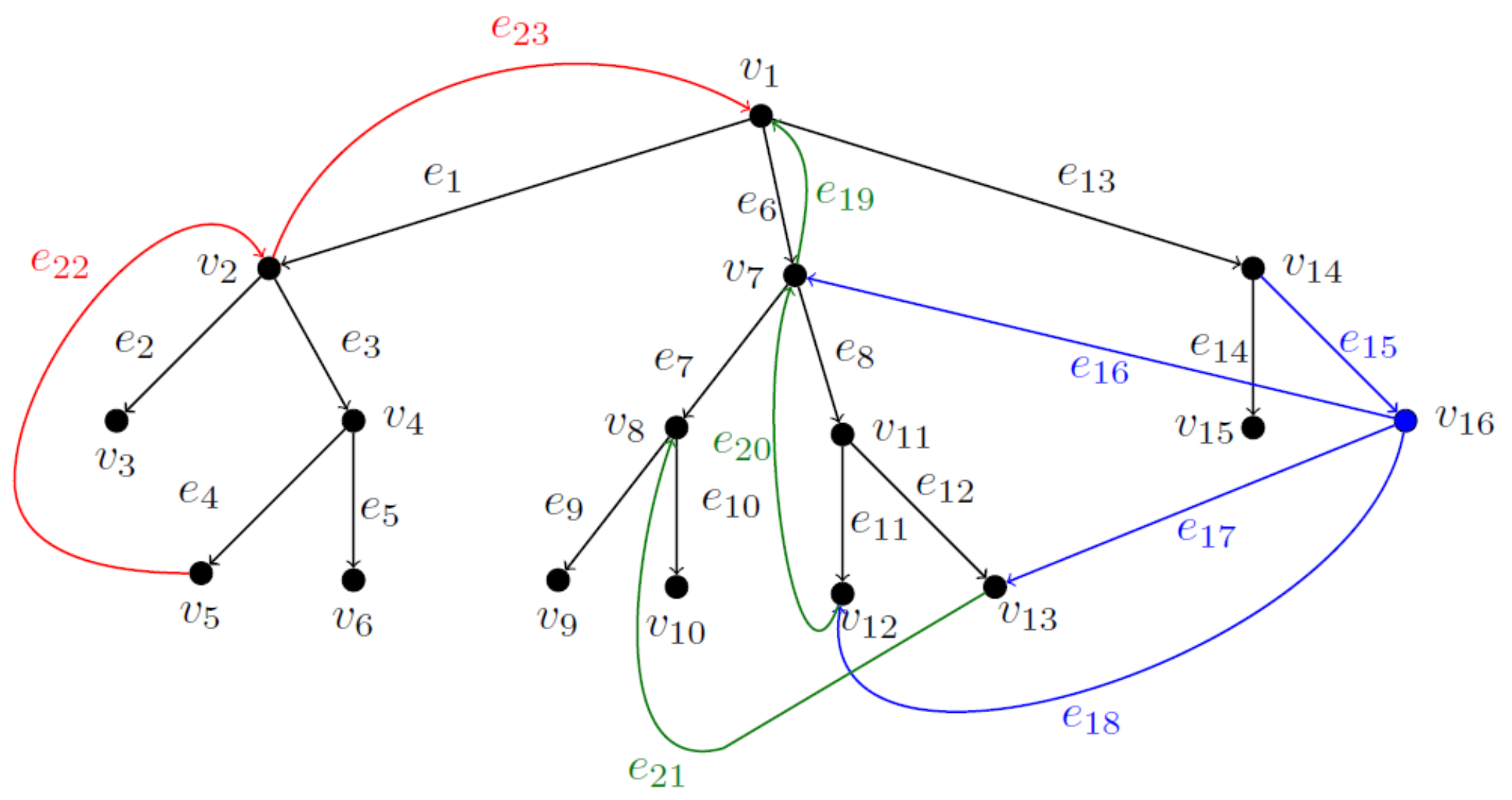
- $sSim(\mathrm{G}_i,\mathrm{G}')$: Structure similarity of G_i to G'
- $\mathrm{G}_1 << \mathrm{G}_2$: G_1 appears before G_2 in the ranked list

$\mathrm{G}_1 << \mathrm{G}_2$ if and only if [$sSim(\mathrm{G}_1,\mathrm{G}')$)> $sSim(\mathrm{G}_2,\mathrm{G}')$]

$\mathrm{G}_1 \ll \mathrm{G}_2$ or $\mathrm{G}_2 \ll \mathrm{G}_1$if [$sSim(\mathrm{G}_1,\mathrm{G}')$ = $sSim(\mathrm{G}_2,\mathrm{G}')$]

Computational Experiments

- Generalized tree dataset: $\mathbf{G}$ ={G_1, G_2,…, G_4}
  - G_1contains $v_{i=\{1,2,...15\}}$ and $e_{j=\{1,2,...14\}}$
  - G_2 contains $v_{i=\{1,2,...16\}}$ and $e_{j=\{1,2,...18\}}$
  - G_3 contains $v_{i=\{1,2,...16\}}$ and $e_{j=\{1,2,...21\}}$
  - G_4 contains $v_{i=\{1,2,...16\}}$ and $e_{j=\{1,2,...23\}}$
- Given query: G'
  - G' contains $v_{i=\{1,2,...15\}}$ and $e_{j=\{1,2,...14\}}$



Metadata of Electronic Health Records

Vertex Labels

$l(v_1)$: Patient	$l(v_9)$: With Minimal Activity
$l(v_2)$: Rheumatoid Arthritis	$l(v_{10})$: Morphine
$l(v_3)$: Cardiomegaly	$l(v_{11})$: Set
$l(v_4)$: Set	$l(v_{12})$: PCI
$l(v_5)$: Diclofenac	$l(v_{13})$: Aspirin
$l(v_6)$: Omeprazole	$l(v_{14})$: Set
$l(v_7)$: Coronary Disease MI	$l(v_{15})$: Senior
$l(v_8)$: Chest Pain	$l(v_{16})$: Set

Edge Labels

$l(e_1)$: chronic illness	$l(e_{13})$: personal characteristics
$l(e_2)$: permanent damage	$l(e_{14})$: age group
$l(e_3)$: treatments	$l(e_{15})$: health history
$l(e_4)$: nsaid	$l(e_{16})$: illness
$l(e_5)$: support therapy	$l(e_{17})$: nsaid
$l(e_6)$: diagnosis-recent	$l(e_{18})$: endovascular procedure
$l(e_7)$: chief complaint	$l(e_{19})$: controlled in
$l(e_8)$: treatments	$l(e_{20})$: effective on
$l(e_9)$: setting of occurance	$l(e_{21})$: had benefit for
$l(e_{10})$: treatments	$l(e_{22})$: effective on
$l(e_{11})$: endovascular procedure	$l(e_{23})$: controlled in
$l(e_{12})$: medication	

Similarity values and Ranked List

Generalized Tree	Structure Similarity	Rank
G_1	1.0	1
G_2	0.8439	2
G_3	0.5131	3
G_4	0.4609	4

Conclusion

Using our attributed generalized tree representation, metadata is able to express complex relations between objects in e-health domain. In addition, semantic and pragmatic information can be represented using label and weight attributes. Our similarity algorithm for generalized trees leads to more precise ranked results by integrating the similarity of corresponding attributes with the structure similarity.

The authors would like to acknowledge the valuable contribution of Mehrdad Kiani, MD. in this research as the expert in health domain.

Publication:

Kiani, M. and Bhavsar, V.C. and Boley, H. 2014. Structure similarity of attributed generalized trees, in: 8th IEEE International Conference on Semantic Computing (ICSC2014), Newport Beach, California, USA, pp. 100-107.

References:

[1] H. Boley, B. Grosof, M. Kifer, M. Sintek, S. Tabet, and G. Wagner, “Object-oriented ruleml”, http://www.ruleml.org/indoo/indoo.html, 2004.