Social Infobuttons: Integrating Open Health Data with Social Data Using Semantic Technology

Xiang Ji $^1$, Soon Ae Chun $^2$, James Geller $^1$

1 New Jersey Institute of Technology
2 College of Staten Island, City University of New York
OVERVIEW

1. Introduction
2. Related Work
3. Social InfoButtons
4. Conclusion and Future Work
Why Social Health Data Is Important

For doctors and nurses:
• Increase awareness of what patients are discussing
• Complement the medical knowledge
• e.g. consider the “social” side effect in treatment

For government:
• Provide intelligence for health policy
• Real-time detection of epidemics
• Public concern surveillance
Research Problems

1. Where to get the social health data

2. How to extract the different data sources

3. How to integrate resources

4. How to provide a unified querying interface
Contribution

1. Extract from social media
2. Developed web scripting scripts
3. Semantic Linking – A Matching Algorithm
4. Social InfoButtons system
RELATED WORK

InfoButtons:
Developed by Cimino et al. (1997) to meet the clinician’s information needs based on contexts.
Social InfoButtons vs. InfoButtons

• Social input (particularly, patients’ community)

• Knowledge from multiple data sources

• Visualizations (distribution map, static map, etc)
Health Data Sources

- patientslikeme®
- NYC OpenData
- BRFSS
- WebMD
- PubMed

CDC
• **BRFSS**: Behavior Risk Factor Surveillance System (CDC: Center for Disease Control and Prevention)

• **PatientsLikeMe**: Patients’ Social Network

• **PubMed**: Database of references on life sciences and biomedical topics.

• **WebMD**: Health Information Provider

• **NYC Open Data**: Public data generated by various NYC agencies and other City organizations
SOCIAL INFOBUTTONS

- PHP
- HTML
- DOM Parser
- MYSQL Database
- BRFSS
- PubMed
- UMLS
- patientslikeme®
- NYC OpenData
- WebMD

Diagram:
- PHP HTML DOM Parser
- Term Matching
- MYSQL Database
- Transform
- JENA TDB Triple Store
- SPARQL
- Social InfoButtons
Data Extraction

(1) 1,228 conditions, 911 side effects, 2,176 symptoms, 1,354 treatment purposes, and 5,608 treatments
(2) Statewide data of absolute/percentage numbers of patients
(3) The related URLs of the medical concepts
Transformation Speed Results

**Phase 1:** 16 seconds (Intel i7 1.6GHz CPU and 6G RAM)
Into 612,017 triples in an RDF file (57.1MB)

**Phase 2:** 57 seconds to transform the RDF file into a Jena TDB triple store representation (341MB)
Term Matching Algorithm

Idea: Utilize UMLS [4] Metathesaurus to search for multiple terms that refer to the same concept

Definition 3.1. Integrated data repository. \( D = \{ D_1', D_2', D_3', \ldots, D_n' \} \) is the aggregation of all data sources being integrated. \( D_i' \) represents the \( i \)th data source.

Definition 3.2. Repository of instances. \( H_i' = \{ h_{1,i'}, h_{2,i'}, h_{3,i'}, \ldots, h_{k,i'} \} \). \( H_i' \) is the set of instances from data source \( D_i' \), and \( h_{i,i'} \) is the \( i \)th instance from set \( H_i' \).

Definition 3.3. Repository of CUIs (Concept Unique Identifier) which has at least one synonym that is the same as the instance \( h_{i,i'} \). \( T_{i,i'} = \{ t \mid t \text{ is the CUI of instance } h_{i,i'} \} \).

Assumption 3.1. If the name of instance \( h_{i,i'} \) matches exactly with the label of the concept \( C_i \) in UMLS, then \( h_{i,i'} \) is a term of the UMLS concept \( C_i \), which is uniquely identified by the CUI of \( C_i \).

Assumption 3.2. If the intersection of the set \( T_{i,i'} \) and the set \( T_{j,j'} \) is not empty, and only one instance exists in the intersection, then the instances \( h_{i,i'} \) and \( h_{j,j'} \) are referring to the same concept.
Algorithm 1: Computing $S(h_{i,i'}, h_{j,j'})$

Input: instance $h_{i,i'}, h_{j,j'}, S(h_{i,i'}, h_{j,j'}) = 0$

Output: $S(h_{i,i'}, h_{j,j'})$

1. Search the UMLS with the name($h_{i,i'}$), if a str in tuple (cui, sui, str) matches the term, add cui to $T_{i,i'}$. 
2. Search the UMLS with name($h_{j,j'}$), if a match is found in UMLS, add cui to $T_{j,j'}$. 
3. for each t in $T_{i,i'}$. 
4. for each t’ in $T_{j,j'}$. 
5. if t is equal to t’ 
6. set $S(h_{i,i'}, h_{j,j'}) = 1$; 
7. break; 
8. endif 
9. end 
10. end
Information Needs

Summarized by questions, answered by SPARQL queries.

Partial Questions Answered by Social InfoButtons

<table>
<thead>
<tr>
<th>Category</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
<td>Top conditions with the most patients?</td>
</tr>
<tr>
<td>Demographics</td>
<td>Gender distribution of the patients?</td>
</tr>
<tr>
<td>Location</td>
<td>Where is the individual patient?</td>
</tr>
<tr>
<td>Condition</td>
<td>What are the symptoms of the condition?</td>
</tr>
<tr>
<td>Correlation</td>
<td>Difference between social and official data?</td>
</tr>
</tbody>
</table>
Use Case - 1

Migraine Patient:

**Information of the patients with the condition: Migraine**

How many patients? 430

who are these patients?  

how are the patients distributed in state and country level?  

where is the individual patient?  

what is the patients' gender distribution?
Treatments of the condition: Migraine

**Sumatriptan** (Prescription Drug)  [pubmed]  [webmd]
Evaluating By 49 patients

**Side Effects:**
(1) Sleepiness 28%;
(2) Nausea and vomiting 18%;
(3) Nausea 15%;
(4) Pain 13%;
(5) Tiredness 13%;
(6) Clumsiness 10%;
Use Case - 2

Government Agency: (Asthma)

A 22.4% discrepancy! Why??
Conclusions of Contributions

1) Semantic Integration Method
   - Conceptual Model
   - Term Matching Algorithm
   - Extract and Transform Structured Health Data into RDF Triples

2) Social InfoButtons
   - Comparative Visualization
   - Social Knowledge in context of Clinical Information
Future Work

(1) The scalability of the approach

(2) Utilize ontological knowledge for term matching

(3) Semantic searches to replace the embedded SPARQL query

(4) User evaluation: what kind of social information is useful

(5) investigate more open government and commercial data
REFERENCES


THANK YOU!!

QUESTIONS??