Transforming Data into Knowledge

using

Knowledge Engineering &

Machine Learning

Presented by: Atif Khan

infoTrellis
Scope

Building Blocks
- ontology
- inference
- machine learning

Tools
- InfoTrellis
- ALLSight
- DBpedia
Motivation

Data Streams » Information?

.....11011111000011
.....1101100011111
.....000010010100
.....1101111111011
.....110110001100
.....011010010100

.....1101111111011
.....110110001100
.....1101110001100
.....011010010100
.....011010010100
Motivation

Information » Knowledge

Charles Smith @profcharles
Just finished registering for KDD 2012 in Beijing http://www.kdd.org/kdd2012/

Charles Smith shared a link
I am off to knowledge discovery and data mining conference in Beijing. Looking forward to Michael Jordan's keynote address
Motivation

Information » Knowledge

Charles Smith

KDD2012

URL
www.kdd2012/

Knowledge Discovery and Data Mining

Conference

Beijing

Charles Smith

Michael Jordan
Motivation

Information » Knowledge

Charles Smith
registered for
KDD2012
website
URL
www.kdd2012/
heldAt
Beijing

Charles Smith
Knowledge Discovery
and Data Mining
Conference
Beijing

Michael Jordan
Motivation

Information » Knowledge

Charles Smith
registered for
KDD2012
website
URL
www.kdd2012/
heldAt
Beijing

Charles Smith
attending
Knowledge Discovery and Data Mining
Conference
Beijing

Michael Jordan
isKeyNote
Speaker

01-24-2013
Motivation

Information » Knowledge

- Charles Smith
- KDD2012
- URL: www.kdd2012/
- Beijing

- Charles Smith
- Knowledge Discovery and Data Mining
- Conference
- Beijing

- Michael Jordan

registered for

same as

same as

isa

attending

isKeyNote Speaker

heldAt

isa

same as

heldAt
Motivation

Information » Knowledge

- Charles Smith is an academic

- KDD is a conference about *data mining* and *knowledge discovery*

- Michael Jordan is an influential academic in data mining community

- Charles Smith and Michael Jordan will both be in Beijing during KDD 2012
Building Blocks

ontology

inference

machine learning
Hybrid MDSS - Holmes**

Holmes

- a hybrid medical decision support system (MDSS)

Based on

- ontological knowledge representation
- logic-based inference
- machine learning to deal with noise

Hybrid MDSS

Line of Inquiry

- which patients can be prescribed what sleep medications?

Considerations

- patient-centric & evidence-based
- automated
- easy to explain and validate
- noise tolerant
Hybrid MDSS

Datasets

**CDC–Behavioral Risk Factor Surveillance System (BRFSS) – 2010**
Hybrid MDSS

Datasets

**multi-dimensional**

- a. demographic information
- b. medical information
- c. behavioural Information
Hybrid MDSS

Datasets

**Characteristics**

- **a.** 450K+ individuals
- **b.** 400+ attributes/record
- **c.** high “missingness”
- **d.** numeric coding
Hybrid MDSS

Datasets

- BRFSS
- MAYO CLINIC
- sleeping pill prescription protocol (HTML)

01-24-2013
Hybrid MDSS

Datasets

- BRFSS
- Mayo Clinic
- Drug Information Online

Drug-to-drug interactions (HTML)
Ontology Model

- Drug
- Patient
- Disease
- Condition
Ontology Model

- Pain pills
- Sleeping pills
- Drug
- Disease
- Condition
- Patient
- Medical Record
- BRFSS Field
- Age
- Gender
- Female
- Male
Ontology Model

Drug

Pain pills
Sleeping pills

Disease
Condition

Patient

Age
Gennder (Female, Male)

Medical Record
BRFSS Field

isa
Ontology Model

Drug

Pain pills
Sleeping pills

Disease

Condition

Patient

Age
Gender
Female
Male

Medical Record

BRFSS Field

isa

hasContraIndication
An Ontology Model

Defines

- **taxonomy**
  - a hierarchy of concepts
- **relationships**

Scope - *“domain of discourse”*

- e.g. medical decision support system
Mapping Raw Data

Patient1

Medical Record1

Gender Field hasValue 2

Age Field hasValue 66

Snore Field hasValue 1
Mapping Raw Data

Patient1

Medical Record1

Gender Field

Age Field

Snore Field

Gender Field hasValue 2 hasGender Female
Age Field hasValue 66
Snore Field hasValue 1
Mapping Raw Data

Patient1

Medical Record1

Gender Field

Age Field

Snore Field

hasValue 2 hasGender Female

hasValue 66 hasCondition Elderly

hasValue 1
Mapping BRFSS Data

- **Patient1**
  - **Gender Field**
    - **hasValue**: 2
    - **hasGender**: Female
  - **Age Field**
    - **hasValue**: 66
    - **hasCondition**: Elderly
  - **Snore Field**
    - **hasValue**: 1
    - **hasDisease**: Sleep Apnea
Mapping Expert Knowledge

Insomnia

Ramelteon

Eszopiclone

prescribedFor
Mapping Expert Knowledge

- SleepApnea
- Asthma
- Insomnia
- LungDisease
- LiverDisease
- Depression

Ramelteon

Eszopiclone

hasContraIndication

prescribedFor
Mapping Expert Knowledge

- Sleep Apnea
- Asthma
- Insomnia
- Lung Disease
- Liver Disease
- Depression

Ramelteon

- Pregnancy
- Breast Feeding
- Alcohol Abuse
- Elderly

Eszopiclone

has Contra Indication

prescribed For
Mapping Expert Knowledge

- SleepApnea
- Asthma
- Insomnia
- LungDisease
- LiverDisease
- Depression

- Ramelteon

- Pregnancy
- BreastFeeding
- AlcoholAbuse
- Elderly

- Eszopiclone

- Propoxyphene

- hasContraIndication
- prescribedFor
Mapping Expert Knowledge

:Propoxyphene a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:Wygesic a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:Trycet a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:PropoxypheneCompound65 a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:Propacet100 a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:Aspirin a :Drug; :isPrescribedFor :Pain.

:Tylenol1 a :Drug; :isPrescribedFor :Pain.

:Tylenol2 a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :SleepingMedication.

N3 representation
Mapping Expert Knowledge

N3 representation

Propoxyphene a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :Eszopiclone.

Wygesic a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :Eszopiclone.

Trycet a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :Eszopiclone.

PropoxypheneCompound65 a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :Eszopiclone.

Propacet100 a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :Eszopiclone.

Aspirin a :Drug;  
  isPrescribedFor :Pain.

Tylenol1 a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :SleepingMedication.

Tylenol2 a :Drug;  
  isPrescribedFor :Pain;  
  hasContraIndication :SleepingMedication.
Knowledge Representation-Recap

Why?
- to create, maintain and share information in a precise manner without ambiguity of meaning

How?
- ontologies

“Now! That should clear up a few things around here!”

Knowledge-Inference

a *discovery* process to find *implied* knowledge using *explicitly defined* information
Knowledge-Inference

a discovery process to find implied knowledge using explicitly defined information

logic-based

ontology
Knowledge-Inference: example

What do we **know** about Mary?

- Mary is grandmother
- Mary is a grand parent
- Mary is woman
Knowledge-Inference: example

What do we know about Mary?

- Mary is grandmother
- Mary is a grand parent
- Mary is woman

if a person has a child, and that child also has a child, then the person is a grandparent
Drug-to-Drug Interactions

**If** a patient is taking an existing drug (D1) and D1 has contraindication to another drug D2 **then** drug D2 should not be prescribed to the patient

\[
\{ \ ?P \ r :\text{Patient}. \\
\ ?D1 \ r :\text{Drug}. \\
\ ?D2 \ r :\text{Drug}. \\
\ ?P \ r :\text{isTaking} \ ?D1. \\
\ ?D1 \ r :\text{hasContraIndication} \ ?D2. \\
\} \Rightarrow \{ \ ?P \ r :\text{cannotBeGiven} \ ?D2 \}. 
\]
Inference Rules

Drug-to-Disease Interactions

- **If** a patient has a condition that has a contra indication to a drug
  **then** the *patient should not be given* the drug

\[
\{ \text{?P} \text{ a :Patient.} \\
\text{?D} \text{ a :Drug.} \\
\text{?DIS a :Disease.} \\
\text{?P :hasDisease ?DIS.} \\
\text{?D :hasContraIndication ?DIS.} \\
\} \Rightarrow \{ \text{?P :cannotBeGiven ?D}. \}
\]
Putting it all Together

- triplestore
- reasoner
- query
- inference rules
Putting it all Together

- triplestore
- reasoner
- query
- inference rules
- Result
- Proof
Putting it all Together

Logic-based, can be verified by traversing the knowledge graph.
But What about Noise?

Noise

- cripples knowledge-based solutions
- limits inference capability
But What about Noise?

Noise

- cripples knowledge-based solutions
- limits inference capability

Mary → James → Emily

hasChild

?
Data is Almost Always Noisy

use “machine learning” to deal with noise
**Machine Learning?**

**query:** given a person's age and income predict if they are **happy** or **sad**
query: given a person's age and income predict if they are happy or sad
query: is he a **happy** or **sad person** ?
Machine Learning 101

Machine Learning

- **classification**: predict class of an instance

- **regression**: prediction of a numeric value

- **clustering**: group similar items together
Machine Learning 101

Machine Learning

- **classification**: predict class of an instance
- **regression**: prediction of a numeric value
- **clustering**: group similar items together
Machine Learning 101

Machine Learning

- **classification**: predict class of an instance
Classification

0. Train
   - build a **classifier** based on *known exemplars*

1. Predict

2. Update

3. Evaluate
Dealing with *Noise*

**query:** is John sad? 

*but we don't have access to that information*
Dealing with Noise

query: is John depressed?

healthy          depressed

unknown

Training exemplars
Dealing with **Noise**

**query:** is John ☹ depressed?

**answer:**

\[
\text{sameAs}(☹, ☹) = 0.14 \\
\text{sameAs}(☹, ☹) = 0.85 \\
\text{sameAs}(☹, ☻) = 0.01
\]
Dealing with *Noise*

**query:** is John 

**answer:**

\[
\text{sameAs}(\gray{\frown}, \yellow{\frown}) = 0.14 \\
\text{sameAs}(\gray{\frown}, \red{\frown}) = 0.85 \\
\text{sameAs}(\gray{\frown}, \green{\frown}) = 0.01
\]
a hybrid decision support system with ontology-based knowledge representation and logic-based reasoning augmented with machine learning classification for noise tolerance
Recap

a hybrid decision support system with ontology-based knowledge representation and logic-based reasoning augmented with machine learning classification for noise tolerance
Taming Data in the Wild
Taming Data in the Wild

“Data-to-knowledge”
- an expensive journey

however, connecting the “data dots” is the key

“Linked Data”*
- can make this a reality

* http://linkeddata.org/
Linked Data

Brief Summary

- Tim Berners-Lee's vision
- URIs* to identify 'things'
- HTTP-based dereferencing of URIs
- structured representation (RDF**)
- hyperlink 'things' together

* URI = Universal Resource Identifier
** RDF = Resource Description Framework
DBpedia

Extract Information from Wikipedia

- unstructured information
  - articles (free text + noise)

- structured components
  - infobox templates
  - categorisation information
  - images
  - geo-coordinates,
  - external web links
Knowledge Engineering using Ontology*

- 359 classes
  - in a subsumption hierarchy

- 1,775 different properties

*http://wiki.dbpedia.org/Ontology
Knowledge Engineering using Ontology

“English version of the DBpedia knowledge base currently describes **3.77 million things**, out of which **2.35 million** are classified in a consistent Ontology, including **764,000 persons**, **573,000 places** (including **387,000 populated places**), **333,000 creative works** (including **112,000 music albums**, **72,000 films** and **18,000 video games**), **192,000 organizations** (including **45,000 companies** and **42,000 educational institutions**), **202,000 species** and **5,500 diseases**.”

http://wiki.dbpedia.org/Datasets
DBpedia – Interlinked Datasets

2007

2009
DBpedia in Action

http://www.visualdataweb.org/reelfinder/reelfinder.php

Demo/Video
infoTrellis
ALLSight Platform

know your customer
AllSight – 360 View of a Customer
AllSight – 360 View of a Customer
AllSight – 360 View of a Customer
AllSight – 360 View of a Customer
Example

[Diagram showing relationships between customer, product, location, and transactions]
Example

- customer
  - male
  - female
- teenager
- adult
- senior
- Bieber nation
- tech buff
- shopaholic
<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Acquired</strong></td>
<td>5,333,242 (8,294 avg/day)</td>
</tr>
<tr>
<td>Transaction References</td>
<td>3,182,277</td>
</tr>
<tr>
<td>Store Visit Broadcasts</td>
<td>101,233</td>
</tr>
<tr>
<td>Employee Chatter</td>
<td>58,410</td>
</tr>
<tr>
<td>Events</td>
<td>63,382</td>
</tr>
<tr>
<td>News</td>
<td>789,289</td>
</tr>
<tr>
<td>Job Ads</td>
<td>84,295</td>
</tr>
<tr>
<td>Coupons</td>
<td>771,114</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>993,242</td>
</tr>
<tr>
<td><strong>Total Qualified</strong></td>
<td>4,175,519 (5,012 avg/day)</td>
</tr>
<tr>
<td>Qualification Confidence</td>
<td>56 % avg</td>
</tr>
<tr>
<td><strong>Total Customer Matches</strong></td>
<td>184,299</td>
</tr>
<tr>
<td>Match Confidence</td>
<td>81.55 %</td>
</tr>
</tbody>
</table>
Under the Hood

Data Enrichment

- taxonomies/ontologies
- dictionaries/catalogues

Matching (decision making)

- knowledge-based rules
- machine learning
Challenges

Entity Resolution

Pre-processing

Feature Selection

Training (& retraining)

Noise
Tools of the Trade
Some Basic Tools

Knowledge Engineering

- Protégé

- Web Ontology Language – OWL

- Resource Description Framework (RDF)
  - XML
  - Notation 3 (N3)
Some Basic Tools

Semantic Reasoners

- CWM
- Euler Sharp
- Jenna
- FuXi
Some Basic Tools

Machine Learning Toolkits

- Weka
- Stanford NLP
- Apache Mahout*
- LIBLINEAR*
- LIBSVM*
- numpy, scipy (python libs)
In Summary

The collaborative work of this group has the potential to unlock data to create knowledge.

There are many
- uses cases that can benefit from this work
- tools available to facilitate this process
Thank You!!
Confidence in Your Data

“Velocity, Volume, and Variety without Veracity creates Vulnerability”*

* Adler on Data Governance
Confidence in Your Data

“Velocity, Volume, and Variety without **Veracity** creates **Vulnerability**”*

* Adler on Data Governance