Extracting, representing and mining Semantic Metadata from text: Facilitating Knowledge Discovery in Biomedicine

Cartic Ramakrishnan

**Advisor:**
Dr. Amit Sheth

**Committee Members:**
Dr. Michael Raymer  
Dr. Guozhu Dong  
Dr. Thaddeus Tarpey  
Dr. Vasant Honavar  
Dr. Shaojun Wang
Overview

• Define Knowledge Discovery

• Contributions
  - Text Mining
  - Knowledge Discovery

• Past work

• Future Work
What Knowledge Discovery is NOT

Search
- Keyword-in-document-out
- Keywords are fully specified features of expected outcome
- Searching for prospective mining sites

Mining
- Know where to look
- Underspecified characteristics of what is sought are available
- Patterns
What is knowledge discovery?

“knowledge discovery is more like sifting through a warehouse filled with small gears, levers, etc., none of which is particularly valuable by itself. After appropriate assembly, however, a Rolex watch emerges from the disparate parts.” – James Caruther

“discovery is often described as more opportunistic search in a less well-defined space, leading to a psychological element of surprise” – James Buchanan

Opportunistic search over an ill-defined space leading to surprising but useful emergent knowledge
Element of surprise – Swanson’s discoveries

- Magnesium
- Migraine
- Calcium Channel Blockers
- Spreading Cortical Depression
- Stress
- Swanson’s Discoveries

Associations discovered based on keyword searches followed by manually analysis of text to establish possible relevant relationships.
Knowledge Discovery in AI
The robot scientist

Planned search over a well-defined (axiomatic) space leading to knowledge discovery.

Knowledge discovery by humans is done in non-axiomatic ill-defined spaces over multi-modal data.

Scientific literature is ill-defined and loosely structured source of data used in scientific investigations.

Assigning structure and interpretation to text (Semantics)

implemented robotic system that applies techniques from artificial intelligence to carry out cycles of scientific experimentation. The system automatically generates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, reports the results, and then generates new hypotheses inconsistent with the data, and then repeats the cycle. Here we apply the system to the determination of gene function using deletion mutants in yeast (Saccharomyces cerevisiae) and an experiment with experiments. We built and tested a detailed logical model (involving genes, proteins and metabolites) of the aromatic amino acid synthesis pathway. In biological experiments that automatically reconstruct parts of this model, we show that an intelligent experiment selection strategy is competitive with human performance and significantly outperforms, with a cost decrease of 3-fold and 100-fold (respectively), both cheapest and random-experiment selection.
Knowledge Discovery = Extraction + Heuristic Aggregation

Nicolas Flamel (traditionally c. 1330 – 1418) was a successful alchemist and manuscript-seller who developed a posthumous reputation as an alchemist due to his reputed work in the field of alchemy.

An alchemical book, published in Paris in 1612 as L'oeuvre des figures hieroglyphiques and in London in 1624 as Exposition de l'Art de la Connaissance des Figures, was attributed to Flamel. The book was supposedly commissioned by Flamel for a Symposium of the Innocents in Paris, long perished before the time the work was published. In the publisher's introduction Flamel's search for the Philosopher's Stone was described. According to that introduction, Flamel had made his life's work to understand the text of a lost manuscript which he had purchased; the introduction claims that, around 1380, he traveled to Spain for assistance with translation. On the way back, he reported that he met a wise, who identified Flamel's book as being a copy of the original book of Abruham. With this knowledge, over the next few years Flamel and his wife allegedly decoded enough of this book to successfully replicate its recipe for the Philosopher's Stone, producing first silver in 1392, then gold.

According to the introduction to his work and additional details that have accrued since its publication, Flamel was the most accomplished of the European alchemists, and had learned his art from a Jewish convert on the road to Santiago de Compostela. Others thought Flamel was the creation of seventeenth-century editors and publishers desperate to produce modern printed editions of supposedly ancient alchemical treatises then circulating in manuscript for an avid reading public; Deborah Harkness put it succinctly. The modern assertion that many references to him or his writings appear in alchemical texts of the 1500s, however, has not been linked to any particular source. The essence of his reputation is that he succeeded at the two magical goals of alchemy – that he made the Philosopher's Stone which turns lead into gold, and that he and his wife Perenelle achieved immortality.

Flamel's house still stands in Paris, and is now the oldest house in the city. The ground floor contains a restaurant.

### Contents (hide)
1. Life
2. In popular culture
3. Notes
4. References
5. External links

#### Life

During his lifetime, Flamel and his wife provided lodging and meals for the poor in their home, in exchange for prayer; they were devout Catholics. Later in life they were noted for their wealth and philanthropy.

Flamel lived into his 80s, and in 1410 designed his own tombstone, which was carved with arcane alchemical signs and symbols. Some believe that he died shortly after the tombstone was created. Later after that, a local criminal, who wished to acquire Flamel's reputed gold, went to Flamel's residence. Finding nothing, but undeterred, he was said to have then gone to the grave with only a spade and a lantern, and dug up the grave. Upon opening the coffin, he was disappointed to find an absence of gold, but shocked to find in its place the corpse of Nicolas Flamel. Some modern sources claim that it was just the grave of the wrong person who was not dead at the time, while others claim that he faked his own death, citing as proof the fact that long after 1410, several books were published in his name. The tombstone is presently at the Musée de Cluny in Paris.

Expanded accounts of his life are taken as legendary. In addition to the mysterious book of twenty-one pages filled with encoded alchemical symbols and arcane writing, he may also have studied some texts in Hebrew. Interest in Flamel revived in the nineteenth century: Victor Hugo mentioned him in The Hunchback of Notre Dame. Eric Saba was intrigued by Flamel. Flamel is often referred to in late-twentieth-century fictional works such as the Harry Potter books and movies as well as The Da Vinci Code.

#### In popular culture

- Nicolas Flamel's story is alluded to in J.K. Rowling's first Harry Potter book, Harry Potter and the Philosopher's Stone, in which he is something of a MacGuffin, though he is the clue to the whole mystery of the book, he never actually makes an appearance. He was friends with Albus Dumbledore and is said to have lived for hundreds of years until the Philosopher's Stone was destroyed following the events of the book. He was 866 years old.

- Flamel has been alleged to be the eighth "Grand Master of the Prexy of Scoor" (1398-1418) as part of a 1900s where his name was printed in the French National Library in the "Diaries Secrets". This resulted in him being mentioned in the 1980s pseudonymous book The Holy Blood and the Holy Grail, Umberto Eco's 1989 novel Faccettoria's Pendulum, and Dan Brown's 2003 novel, The Da Vinci Code. Many of the names of "Grand Masters" were evidently chosen for some sort of connection with alchemy.

- Klauss and his wife Perenelle Flamel were described as characters created by the Italian author Umberto Eco, who also wrote the Italian novel The Name of the Rose (1980), and an Italian novel named "Niccolò and
This MEK dependency was observed in BRAF mutant cells regardless of tissue lineage, and correlated with both downregulation of cyclin D1 protein expression and the induction of G1 arrest.

*MEK dependency ISA Dependency_on_an_Organic_chemical
*BRAF mutant cells ISA Cell_type
*downregulation of cyclin D1 protein expression ISA Biological_process
*tissue lineage ISA Biological_concept
*induction of G1 arrest ISA Biological_process
Overview

• Define Knowledge Discovery

• Contributions
  – Text Mining
    • Compound Entities
    • Complex relationships
  – Knowledge Discovery
    • Subgraph discovery

• Past work

• Future Work
Knowledge Discovery over text

Assigning interpretation to text

Text → Extraction of Semantics from text → Semantic Metadata Guided Knowledge Explorations → Semantic Metadata Guided Knowledge Discovery

→ Triple-based Semantic Search → Semantic browser → Subgraph discovery

Semantic metadata in the form of semi-structured data
Ontology-enabled Information Extraction

Comparison with standard IE

Standard IE

- Simple entities
- Typically supervised pattern-based
- Output not structured
- Type specific patterns to train
- No focus on semantics

Our approach

- Compound and modified entities
- Unsupervised
- Output structured to support knowledge discovery
- Not restricted to specific entity types
- Assign semantic interpretations to sentence
Information Extraction via Ontology assisted text mining – Relationship extraction

Biologically active substance

Lipid

Disease or Syndrome affects causes complicates

Fish Oils Raynaud's Disease

instance_of instance_of

UMLS Semantic Network

MeSH PubMed 9284 documents 4733 documents 5 documents 13
Background knowledge and Data used

UMLS – A high level schema of the biomedical domain
- 136 classes and 49 relationships
- Synonyms of all relationship – using variant lookup (tools from NLM)
- 49 relationship + their synonyms = ~350 verbs

MeSH
- 22,000+ topics organized as a forest of 16 trees
- Used to query PubMed

PubMed
- Over 16 million abstract
- Abstracts annotated with one or more MeSH terms
An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.

- Entities (MeSH terms) in sentences occur in modified forms.
  - "adenomatous hyperplasia" modifies "estrogen".
  - "An excessive endogenous or exogenous stimulation" modifies "estrogen".

- Entities can also occur as composites of 2 or more other entities.
  - "adenomatous hyperplasia" and "endometrium" occur as "adenomatous hyperplasia of the endometrium".
Method – Identify entities and relationships in Parse Tree

VBZ induces

An excessive endogenous or exogenous stimulation by estrogen

Modifiers
Modified entities
Composite Entities
An excessive endogenous or exogenous stimulation modifies:
- Estrogen

Adenomatous hyperplasia is a composite entity that includes:
- Modified entity 1
- Modified entity 2

Modified entities include:
- Modified entity 1
- Modified entity 2

Composite entities include:
- Composite entity 1
- Composite entity 2

Endometrium is a part of:
- Composite entity 1
- Composite entity 2

The diagram illustrates the relationships and modifications in the context of hormonal stimulation.
Swanson's discoveries - Associations between Migraine and Magnesium [Hearst99]

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high levels of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability

Data sets generated using these entities (marked red above) as boolean keyword queries against pubmed

Bidirectional breadth-first search used to find paths in resulting RDF
Paths between Migraine and Magnesium

Table 2. Paths between Migraine and Magnesium

<table>
<thead>
<tr>
<th>Path length</th>
<th>Total Number of paths found</th>
<th># of interesting paths</th>
<th>Max. # of named relationships in any path</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>260</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>4103</td>
<td>1864</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>106450</td>
<td>33403</td>
<td>5</td>
</tr>
</tbody>
</table>

 Paths are considered interesting if they have one or more named relationship Other than hasPart or hasModifiers in them
CONCLUSION

- Rules over parse trees are able to extract structure from sentences
- Our definition of compound and modified entities are critical for identifying both implicit and explicit relationships
- Swanson’s discovery can be automated – if recall can be improved – what hurts recall?
Interesting Observations from this preliminary work
Observations – Sentence characteristics and Parsing

Sentence characteristics

Univariate analysis identified positive inotropic drug requirement (p = 0.011, odds ratio [OR] = 3.41), ejection fraction (EF) (p = 0.001, OR = 0.92), cross-clamp time (p = 0.046, OR = 0.97), left internal mammary artery (p = 0.023, OR = 0.49), chronic obstructive pulmonary disease (COPD) (forced expiratory volume in 1 second <75% of predicted value (p = 0.009, OR = 2.02), intra-aortic balloon pump (p = 0.045, OR = 1.23), body mass index (p = 0.035 OR = 5.60), and CII (p < 0.001, OR = 0.36) as predictors of SVT.

Parser issues

- Dearth of features – parsers that provide richer features needed

(TOP (S (NP (NP (DT An) (JJ excessive) (ADJP (JJ endogenous) (CC or) (JJ exogenous) ) (NN stimulation) ) (PP (IN by) (NP (NN estrogen) ) (VP (VBZ induces) (NP (NP (JJ adenomatous) (NN hyperplasia) ) (PP (IN of) (NP (DT the) (NN endometrium) ) ) ) ) ) ) ) ) ) )

hyperplasia of the endometrium.
Observations – Complex entities with nesting and overlapping structure

Complex Entities

- Chevy Chase, Chevy Chase bank building on 5th and 3rd
- Sentential forms of entities are often quite complex
  - (e.g. Reactive oxygen intermediate-dependent NF-kappaB activation)
- Structurally and semantically complex nested entities
  - Human Immunodeficiency Virus Type-2 Enhancer Activity
    [[[Human Immunodeficiency Virus_{disease}] Type-2_{disease}] Enhancer Activity_{biological_process}]
  - CD28 surface receptor
    [[[CD28_{protein_molecule}] surface receptor_{protein_family_or_group}]


General strategy for dealing with complex sentences

- Identify and extract complex entities across a given corpus
- Replace occurrences of all complex entities with single word place holders
- Re-parse the sentence to extract relationships

Tactic used

- Use a feature rich parse like a dependency parse to segment sentences into Subj→Pred→Object
- Subjects and Objects represent compound entities
- Use corpus statistics to predict constituents of compound entities
Unsupervised Joint Extraction of Compound Entities and Relationship

Cartic Ramakrishnan, Pablo N. Mendes, Shaojun Wang and Amit P. Sheth
"Unsupervised Discovery of Compound Entities for Relationship Extraction"
EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns
Joint Extraction approach

Dependency parse – Stanford Parser

Anti-Ro(SSA) autoantibodies are associated with T cell receptor beta genes in systemic lupus erythematosus patients.

(B)
Stanford Dependency Hierarchy

dep - dependent
  aux - auxiliary
    auxpass - passive auxiliary
cop - copula
conj - conjunct
cc - coordination
arg - argument
  subj - subject
    nsubj - nominal subject
      nsubjpass - passive nominal subject
csubj - clausal subject
  comp - complement
    obj - object
      dobj - direct object
      iobj - indirect object
      pobj - object of preposition
  attr - attributive
  ccomp - clausal complement with internal subject
  xcomp - clausal complement with external subject
compl - complementizer
mark - marker (word introducing an advcl)
rel - relative (word introducing a remod)
  acomp - adjectival complement
agent - agent
ref - referent
expl - expletive (expletive there)

mod - modifier
  advcl - adverbial clause modifier
  purpcl - purpose clause modifier
tmod - temporal modifier
rcmod - relative clause modifier
amod - adjectival modifier
inflmod - infinitival modifier
partmod - participial modifier
num - numeric modifier
number - element of compound number
appos - appositional modifier
nn - noun compound modifier
abbrev - abbreviation modifier
advmod - adverbial modifier
  neg - negation modifier
poss - possession modifier
posse - possessive modifier ('s)
prt - phrasal verb particle
det - determiner
prep - prepositional modifier
  sdep - semantic dependent
  xsubj - controlling subject

Figure 2: The grammatical relation hierarchy.
Hierarchy used to generalize the rules

Small set of rules over dependency types dealing with
- modifiers (a.mod, nn) etc. subjects, objects (nsubj, nsubjpass) etc.

Since dependency types are arranged in a hierarchy
- We use this hierarchy to generalize the more specific rules
- There are only 4 rules in our current implementation


<table>
<thead>
<tr>
<th>Relation</th>
<th># occurrences</th>
<th>% occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>dependent</td>
<td>4690</td>
<td>100.0</td>
</tr>
<tr>
<td>mod</td>
<td>2710</td>
<td>57.8</td>
</tr>
<tr>
<td>ncmod</td>
<td>2377</td>
<td>50.7</td>
</tr>
<tr>
<td>xmod</td>
<td>170</td>
<td>3.6</td>
</tr>
<tr>
<td>cmod</td>
<td>163</td>
<td>3.5</td>
</tr>
<tr>
<td>arg_mod</td>
<td>39</td>
<td>0.8</td>
</tr>
<tr>
<td>arg</td>
<td>1941</td>
<td>41.4</td>
</tr>
<tr>
<td>subj</td>
<td>993</td>
<td>21.2</td>
</tr>
<tr>
<td>ncssubj</td>
<td>984</td>
<td>21.0</td>
</tr>
<tr>
<td>xssubj</td>
<td>5</td>
<td>0.1</td>
</tr>
<tr>
<td>csubj</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>subj or dobj</td>
<td>1339</td>
<td>28.6</td>
</tr>
<tr>
<td>comp</td>
<td>948</td>
<td>20.2</td>
</tr>
<tr>
<td>obj</td>
<td>559</td>
<td>11.9</td>
</tr>
<tr>
<td>dobj</td>
<td>396</td>
<td>8.4</td>
</tr>
<tr>
<td>obj2</td>
<td>19</td>
<td>0.4</td>
</tr>
<tr>
<td>ioj</td>
<td>144</td>
<td>3.1</td>
</tr>
<tr>
<td>clausal</td>
<td>389</td>
<td>8.3</td>
</tr>
<tr>
<td>xcomp</td>
<td>323</td>
<td>6.9</td>
</tr>
<tr>
<td>ccomp</td>
<td>66</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table 2: Frequency of each type of GR (inclusive of subsumed relations) in the 10K-word corpus.
Anti-Ro(SSA) autoantibodies are associated with T cell receptor beta genes in systemic lupus erythematosus patients.
Preliminary results

[Genetic_Function The cardiac myosin heavy chain Arg-403->Gln mutation]
[Disease_or_Syndrome hypertrophic cardiomyopathy]
[Amino_Acid_Peptide_or_Protein The cardiac ventricular myosin]
[Organism_Attribute heavy chain phenotype]
[Amino_Acid_Peptide_or_Protein the fibronectin receptor, talin, vinculin and actin]
[Gene_or_Genome the RAD51 and RAD52 genes]
[Tissue the plasma membrane]
[Body_Part_Organ_or_Organ_Component the ligated kidneys]
[Body_Part_Organ_or_Organ_Component the developing kidney]
[Amino_Acid_Peptide_or_Protein other cytoskeletal proteins such as tau or actin]
[Amino_Acid_Peptide_or_Protein a ubiquitous actin-binding protein]
[Amino_Acid_Peptide_or_Protein the ubiquitous actin]
[assay The centrifugation assay and the DNase I inhibition assay]
[Social_Behavior a role for synaptotagmin family members in actin function]
<table>
<thead>
<tr>
<th>Relationship</th>
<th>Sentence Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>increased</td>
<td>A pre-treatment of cells with SGE from partially fed ticks in amounts salivary glands → increased → the level of both viral nucleocapsid N protein phosphoprotein P in a dose-dependent manner</td>
</tr>
<tr>
<td>inhibits</td>
<td>alpha-catenin → inhibits → beta-catenin signaling</td>
</tr>
<tr>
<td>inhibits</td>
<td>MgCl2 → inhibits → these effects of profilin, most likely</td>
</tr>
<tr>
<td>causes</td>
<td>The cardiac myosin heavy chain Arg-403 Gln mutation → causes → hypertrophic cardiomyopathy</td>
</tr>
<tr>
<td>causes</td>
<td>Moreover, addition of profilin to steady-state actin filaments → causes → slow depolymerization</td>
</tr>
<tr>
<td>causes</td>
<td>(11-22 microM) into infected PtK2 cells → causes → a marked slowing of actin tail elongation and bacterial migration</td>
</tr>
<tr>
<td>binds</td>
<td>the cytoplasmic domain of E-cadherin → binds → either beta-catenin or plakoglobin</td>
</tr>
<tr>
<td>binds</td>
<td>a constituent → binds → RBC alpha-spectrin antibody plus the presence of significant quantities of actin</td>
</tr>
</tbody>
</table>
Analysis of compound entities

Errors – some sources
- 4 rules therefore compound entities composed of other compound connected by
  - Prepositions
  - Punctuations
- Verbs interpreted as nouns by the parser, wind up as part of entities

A Fix – Corpus statistics
- Mutual Information
  - Human Immunodeficiency Virus Type-2 Enhancer Activity
    [[[Human Immunodeficiency Virus_disease] Type-2_disease] Enhancer Activity_biological_process]
Predicting the constituents to compound entities

Given compound entities
- Predict which token subsequences are entities
- Identify their semantic type in UMLS

Central idea in constituent prediction
- A token sequence is likely to form an entity if that sequence occurs often across a given corpus
- But mere co-occurrence does not work
  - Illeal Neoplasm vs. Neoplasm of the Illeum
- Instead we use dependency co-occurrence
What is Mutual information?

A measure for discovering interesting word collocations

- information that two random variables share: it measures how much knowing one of these variables reduces our uncertainty about the other

\[
I(w_i, w_j) = p(w_i, w_j) \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \text{ where } p(w_i, w_j) = p(w_i)p(w_j | w_i)
\]

\[
p(w_j | w_i) = \frac{\text{count}(w_i, w_j)}{\text{count}(w_i)}
\]
Dependency-based mutual information

\[ p_d(w_j|w_i) = \frac{\text{count}_d(w_i = \text{dep} \land w_j = \text{gov}) + \text{count}_d(w_j = \text{dep} \land w_i = \text{gov})}{\text{count}_d(w_i = \text{dep} \lor w_i = \text{gov})} \]

Collecting dependencies
- We parse 800,000 sentences using the Stanford parser
- Index all dependencies using a Lucene index

Advantage
- Capture long range dependencies between words
- Adjacency not required
Predicting constituents

Greedy mutual information based word grouping used to predict constituents

- Given a sequence of tokens as input
- Compute dependency based mutual information for all token pairs
- Starting at each token in turn attach all other tokens to it that increase the average mutual information of the token group so far

Variants of this algorithm

- Compute the average dependency-based mutual information across the corpus - use that as the threshold
Results

Results from the BioInfer corpus
- Compound entities found
- Constituent entities predicted
- Triples found

Results from the OMIM corpus
- Compound entities found
- Triples found
Evaluations

Problems with automatic evaluations using gold standard
- Meant for specific types of entities
- Mark entity mentions not compound concepts

Manual Evaluations
- Expensive
- Requires domain expert

Our Solution
- Build a tool to compare S-P-O triple with the original sentence
- Allow evaluator to assess “correctness”
Manual Evaluation

- Test if the RDF conveys same “meaning” as the sentence
- Juxtapose the triple with the sentence
- Allow user to assess correctness/incorrectness of the subject, object and triple

□ Is this triple ok?

http://knoesis.org#ce_70273 --> dominant-negative|forms|of|dmp53@en □ Is this subject ok?
http://knoesis.org#inhibited_
http://knoesis.org#ce_70274 --> cultured|cells|and|radiation-induced|apoptosis|in|developing|tissues@en □ Is this object ok?

Dominant-negative forms of Dmp53 inhibited transactivation in cultured cells and radiation-induced apoptosis in developing tissues.
Demo of Evaluation tool

Dataset & Result characteristics

- OMIM 90,000 sentences obtained from 1248 records returned for keyword “renal”

- 328 Mb of RDF generated
  - Containing 155K triples
  - 4045 entities are related to UMLS classes
  - 126 of the 136 classes in UMLS are instantiated

- Demo
Evaluation conducted using this tool

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Number of triples per relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>encodes</td>
<td>397</td>
</tr>
<tr>
<td>is expressed</td>
<td>356</td>
</tr>
<tr>
<td>induced</td>
<td>305</td>
</tr>
<tr>
<td>produced</td>
<td>221</td>
</tr>
<tr>
<td>inhibited</td>
<td>181</td>
</tr>
<tr>
<td>derived</td>
<td>172</td>
</tr>
<tr>
<td>affect</td>
<td>166</td>
</tr>
<tr>
<td>binds</td>
<td>140</td>
</tr>
</tbody>
</table>

Evaluation on OMIM RDF
- 1938 triple-sentence comparisons
- 2 evaluators (not domain experts)
- Currently system in use by domain experts at
  - CCHMC (2 experts)
  - Awaiting results
Results of Manual Evaluation

Triple accuracy

- is expressed
- encodes
- induced
- inhibited
- binds
- produced
- derived
- affect
Applications

Triple-based semantic search
Semantic Browser
Semantic Metadata Guided Knowledge Explorations and Discovery
An excessive endogenous or exogenous stimulation modifies the endometrium, which induces adenomatous hyperplasia. This process is further modified by estrogen, resulting in hyperplasia.
Discovering informative subgraphs (Harry Potter)

- Given a pair of end-points (entities)
- Produce a subgraph with relationships connecting them such that
  - The subgraph is small enough to be visualized
  - And contains relevant “interesting” connections

We defined an interestingness measure based on the ontology schema

- In future biomedical domain scientists will control this with the help of a browsable ontology
- Our interestingness measure takes into account
  - Specificity of the relationships and entity classes involved
  - Rarity of relationships etc.

Schema-driven edge weight assignment

- Entertainment company
- Manufacturing company
- Oil company
- Automotive company
- Electronics company
- Sporting goods company
- Ford Motors
- Cartic’s Company

Weights:
- 1.0
- 0.67
- 0.33
- <0.5

Legend:
- rdfs:subClassOf
- rdfs:subPropertyOf
- p ⊃ rdf:Property

Instances:
- Entertainment_Company
- Manufacturing_Company
- Oil_Company
- Automotive_Company
- Electronics_Company
- Sporting_Goods_Company
- Cartic’s Company
- Executive
- Supervisor
- Employee
- Manages
- Leads
- Owned_by
- Endorses
- Competes_with
Heuristics used to bias edge weights

Two factor influencing interestingness
Algorithm

• Bidirectional lock-step growth from S and T
• Choice of next node based on interestingness measure
• Stop when there are enough connections between the frontiers
• This is treated as the candidate graph
Algorithm

Model the Candidate graph as an electrical circuit

- S is the source and T the sink
- Edge weight are treated as conductance values
- Using Ohm’s and Kirchoff’s laws
  - find maximum current flow paths through the candidate graph from S to T
- At each step adding this path to the output graph to be displayed we repeat this process till a certain number of predefined nodes is reached
Results
Overview

- Define Knowledge Discovery
- Contributions
  - Text Mining
  - Knowledge Discovery
- Past work
- Future Work
Past work

Automatic hierarchy creation from Text
- Input text
- Output topic hierarchy

TaxaMiner: an experimentation framework for automated taxonomy bootstrapping - [pdf] - uga.edu
V Kashyap, C Ramakrishnan, C Thomas, A Sheth - International Journal of Web and Grid Services, 2005 - Inderscience
... Catic Ramakrishnan, Christopher Thomas and A. Sheth ... For the SW vision to Page
3. 242 V. Kashyap, C. Ramakrishnan, C. Thomas and A. Sheth ...
Cited by 16 - Related articles - Web Search - All 11 versions

Ranking complex relationships in RDF graphs
- Ranking paths in RDF graphs

Ranking Complex Relationships on the Semantic Web - [pdf] - uga.edu
... Wiener, IB Arpinar, C. Ramakrishnan, AP Sheth - IEEE INTERNET COMPUTING, 2005 - doi.ieeecomputersociety.org
... Halaschek-Wiener, University of Georgia I. B. Arpinar, University of Georgia
Catic Ramakrishnan, University of Georgia Amit P. Sheth, University of Georgia ...
Cited by 40 - Related articles - Web Search - All 10 versions

Conflict-of-interest detection

Semantic analytics on social networks: experiences in addressing the problem of conflict of interest ... - [pdf] - umbc.edu
... C Ramakrishnan, L Ding, P Kolari, AP Sheth, IB ... - Proceedings of the 15th international conference on World ..., 2006 - portal.acm.org
... Boonergs Aloman-Meza 1, Moonakshi Nagarajan 1, Catic Ramakrishnan 1, Li Ding 2, Pranam Kolari 2, Amit P. Sheth 1, I. B. Arpinar 1, Anupam Joshi 2 ...
Cited by 53 - Related articles - Web Search - All 7 versions
Other major papers influencing my work

Position papers

- Three types of semantics

Semantics for the Semantic Web: The Implicit, the Formal and the Powerful - [PDF] ▶ hdomi.gr
A Sheth, C Ramakrishnan, C Thomas - International Journal on Semantic Web & Information Systems, 2005 - IGI Global
... Building practical Semantic Web applications (e.g., see TopQuadrant, 2004; Sheth & Ramakrishnan, 2003; Kachyap & Shklar, 2002) require certain core capabilities ...
Cited by 75 - Related articles - Web Search - All 13 versions

- Survey of Semantic Web technologies

[PDF] ▶ Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and ...
... Integration and Analysis Amit Sheth 1,2 and Carty Ramakrishnan 2 1 Semagix and 2 LSDIS lab, University of Georgia Abstract Semantics ...
Cited by 41 - Related articles - View as HTML - Web Search - All 14 versions

- Relationship Web

Relationship Web Blazing Semantic Trails between Web Resources
AP Sheth, C Ramakrishnan - IEEE INTERNET COMPUTING, 2007 - doi.ieeecomputersociety.org
The memory extender (or memex) vision that Vannevar Bush outlined in 1945 1 pointed out the limitations of a topic-hierarchy-centric document organization mechanism and proposed a contrasting view. Describing how the human brain ...
Cited by 1 - Related articles - Web Search - OhioLINK OLinks - BL Direct - All 4 versions
Overview

- Define Knowledge Discovery
- Contributions
  - Text Mining
  - Knowledge Discovery
- Past work
  - Taxonomy construction
  - Ranking complex relationships in RDF graphs
  - Conflict of Interest detection
- Current & Future Work
Hypothesis Driven retrieval of Scientific Literature

Keyword query: Migraine[MH] + Magnesium[MH]
Strength of a connection

Associating a confidence value with extracted relationships
- Information loss in the extraction process
- Temporal aspects
- Bibliometrics
  - Venue impact
  - Author expertise
- Multiple schemas
Mechanistic Models
Publications

• Conference Publications:
  - “Ontology Learning via Unsupervised Joint Entity and Relationship Extraction” Manuscript under preparation WWW2009
  - “Semantic Search over Biomedical Literature” Manuscript under preparation WWW2009
  - “Joint Extraction of Compound Entities and Relationships to support Semantic Browsing over Biomedical Literature” Cartic Ramakrishnan, Pablo N. Mendes, Rodrigo A.T.S da Gama, Guilherme C. N. Ferreira & Amit P. Sheth Under Review
  - "Unsupervised Discovery of Compound Entities for Relationship Extraction" Cartic Ramakrishnan, Pablo N. Mendes, Shaojun Wang and Amit P. Sheth EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns
Publications

- **Journal Publications:**
• **Workshop Publications**
  - Matthew Perry, Maciej Janik, **Cartic Ramakrishnan**, Conrad Ibañez, Ismailcem Budak Arpınar, Amit P. Sheth: Peer-to-Peer Discovery of Semantic Associations. P2PKM 2005

• **Book Chapters**

• **Posters**
  - V. Kashyap, **C. Ramakrishnan**, T.C. Rindflesch Towards (Semi-)automatic Generation of Bio-medical ontologies. In AMIA 2003 Annual Symposium on Biomedical and Health Informatics
  - **Cartic Ramakrishnan**, Pablo N. Mendes and Amit P. Sheth “Ontology-driven data capture, representation, retrieval, and mining - Facilitating Knowledge Discovery in Biomedicine” Ohio Collaborative Conference on Bioinformatics (OCCBIO) 2008

• **Tutorials**
  - [Text Analytics for Semantic Computing - the good, the bad and the ugly](#)
    - Instructors: **Cartic Ramakrishnan**, Meenakshi Nagarajan and Amit Sheth
Experiences

Teaching & Mentoring
- TA for Semantic Web Fall 2003 and Semantic Enterprise Fall 2004 @UGA
  - Helped Dr. Sheth with Course design and grading
  - Guided graduate students in open-ended course projects
- Mentored interns @ Kno.e.sis Summer 2007 and Summer 2008

Collaborations & Internships
- Interned @
  - NLM, NIH Summer 2002, Summer 2004 – Dr. Vipul Kashyap
  - IBM Almaden Summer 2006 – Dr. Tanveer Syeda-Mahmood
- Research collaborations with
  - CCHMC – Dr. Bruce Aronow
  - AFRL – Dr. Victor Chan

Grant Writing
- NSF CISE grant December 2006 with Dr. Sheth & Dr. Dong
- NSF Cyber-discovery Initiative December 2007 Dr. Sheth & Dr. Bruce Aronow
Acknowledgements

Special Thanks to
Pablo N. Mendes, Christopher Thomas, Meena Nagarajan, Karthik Gomadam, Ajith Ranabahu, Dr. Shaojun Wang, Dr. Raymer

Thanks to all other Kno.e.sis members and our Summer 2008 interns Rodrigo Gama, Guilherme De Napoli, Kamal Baid, Hemant Purohit

Special thanks to Dr. Amit Sheth and my committee members
On a lighter note