Semantic Searching

John Winder
CMSC 676
Spring 2015
Semantic Searching

searching and retrieving documents by their semantic, conceptual, and contextual meanings

Motivations:

● to do disambiguation
● to improve retrieval accuracy
  ○ precision and recall
● to unite the Semantic Web
Semantic Web

● Standardizations
  ○ data formats and schemas
    ■ XML, RDF
  ○ query languages
    ■ RDQL, SPARQL

● Key Ideas
  ○ metadata
  ○ ontologies
Ontology

- An ontology is a knowledge base
  - models hierarchies, relationships (is-a, has-a)
  - uses formal languages (inspired by databases)
- Examples:
  - WordNet (dog is-a canine is-a carnivore, etc.)
  - ConceptNet

```
PREFIX kb: <http://protege.stanford.edu/kb#>
SELECT ?painting ?painter
WHERE {
  ?painting kb:has_author ?painter
}
```

A query to retrieve a list of paintings and their painters.
Main Advancements

- **Vector Space model**
  - Boolean model
    - no partial matching
    - no clear ranking method
  - requires parallel metadata
  - rank by TF-IDF

- **Semi-structured**
  - fully structured ontological mapping has worse recall
  - keyword searching is flexible but has worse precision

Vector Space Semantic Search System by Vallet et al. [2005]
Main Advancements (cont.)

● Query Expansion
  ○ searching by meaning, beyond literal keywords
  ○ given a query, map into ontology, find new relations
  ○ returns documents even without search keywords being present in the documents
  ○ examples:
    ■ “presidents of the French government”
    ■ “reports on flooding for cities in Asia with populations under 50,000”
Main Advancements (cont.)

- Generating queries
  - search by keyword
  - parses out entity/relations

- Semantic Ranking
  - by entity
    - ReConRank in SWSE
  - by relationship
  - by document
    - annotations (Swoogle)

Ontology-Based Semantic Search System by Fernandez et al. [2011]
Mimir: Semantic Search at Scale (2015)

- Mimir, annotation-based semantic search
  - uses GATE to do NLP, extract entities/relationships
  - open source, distributed (federated) system
  - complex query parsing, indexing at three levels:
    - tokens, annotations, sub-annotations
  - applied to real world corpora (over 150 million docs)
    - immunology dataset
    - patent dataset, searching for prior art
Future Applications

- **Recommender Systems**
  - build user profiles, use history to inform results

- **Sentiment Analysis**
  - disambiguation to spot outliers in word usage

- **Reasoning (Artificial Intelligence)**
  - inference: discovering new facts
  - using ontologies to build ... more ontologies
References


Cunningham, Hamish, Maynard, Diana, Bontcheva, Kalina, and Tablan, Valentin. Gate: an architecture for development of robust hl


