Presentation of :

Knowledge discovery by automated identification and ranking of implicit relationships

Jonathon D. Wren, Raffi Bekeredjian, Jelena A. Stewart, Ralph V. Shohet, Harold R. Garner

Bioinformatics, 2004

Presented by Nancy Baker January 25, 2005 ILS 279

What is knowledge discovery?

- Knowledge Discovery is
  "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad,Piatetsky-Shapiro, and Smyth 1996).

Background: Literature Mining for knowledge discovery

- Information overload
- Millions of journal articles recording scientific findings
- No one can read them all: need automated approaches

Contribution of Swanson

- "undiscovered public knowledge"
- "non-interactive literatures"
- A-B-C model

A ------ B ------ C

Past work: Swanson and others who use co-occurence

- Swanson
  - based on keywords in titles
- Others
  - MeSH terms
  - Mapping text to UMLS concepts

In both cases the size of the domain is a problem.

Approach in this paper

- Use the A – B – C model as a basis
  - Choose the A terms
  - Literature mining to find associations to A terms (B entities)
  - Query B objects to find relationships to other objects – C
  - Look at implicit (not explicit) A-C relationships
  - Rank relationships to find the statistically exceptional ones

The A Set

- Data Sources:
  - OMIM - diseases and clinical phenotypes
  - HGNC - genes
  - LocusLink - genes
  - MeSH - chemical compounds and drugs
- 33,539 unique objects (85,234 including synonyms)

Identify Relationships

- Co-occurrence in MEDLINE record
  - Abstract
  - sentence
- Caveat – co-occurrence may not always the existence of a biologically meaningful relationship
- Need a way to estimate the importance of co-occurrence

Importance of Co-occurrence

- Fuzzy logic – not 0 or 1, somewhere in between
- Score based on frequency
- Calculate expected value based on relative connectivity
  - Assume a random network
  - How far does this relationship deviate

Implementation: estimate of precision and recall

- (Why are they putting this section here?)
- In general, precision and recall measurements are difficult in text-mining
  - Gold standard
  - Test corpus

Precision

- Manual estimation based on sample
- Looked at 25 randomly selected MEDLINE records
  - Found that 2 objects co-mentioned within the same sentence were more likely related (83%) than objects mentioned in abstract (53%)
  - Sentence co-mentions alone misses relationships (43%)

Trivial vs. non-trivial relationships

- Found non-persistent relationships
  - In first half of MEDLINE but not in second half
  - Assumed false or not interesting relationships
- Rates similar to power decay function
- Decided OK to use as error probability

Recall

- Studied recall rates using abstracts vs. full text articles
  - Chose 4 objects, one of each type, had to have 2 review articles in last 3 yrs
  - Compared objects
- Results
  - 30 objects in the literature not in database, for various reasons
  - 141 of 181 objects in database (78%)
  - 98% could have been because terms were in abstracts (spelling errors)

Processing MEDLINE records

- 12,037,763 MEDLINE records
- Created a network of 3,482,204 unique relationships between objects
- Many objects had a high number of connections – unwieldy number
- Needed to rank potential significance
- Obs/Exp calculated

Cardiac Hypertrophy

- An example – why?
- Looked at compounds with implicit relationships to cardiac hypertrophy
- Cholopromazine ranked high
- Mouse trials showed CPZ lowers hypertrophy
- A relationship between cardiac hypertropy/CPZ not mentioned previously in literature

Discussion

- A new relationship was found
- Shortcomings of method
  - Finding uninteresting relationships
  - Time consuming to find nature of relationship
  - Comparison to random network model assumes text is non-random. Is it?
  - Method has utility as information increases

Comments

- Confusing paper
  - Structure
  - Formulas
  - GDB and Genome Ontology – were they used?
- Why cardiac hypertrophy? Did it come to the top in results or was is originally a disease of interest?

Text Mining Issues

- Evaluation of methods – precision/recall
- The human component – someone must decide whether connection is interesting and potentially useful
- Collaboration