Mining MEDLINE for Implicit Links between Dietary Substances and Diseases

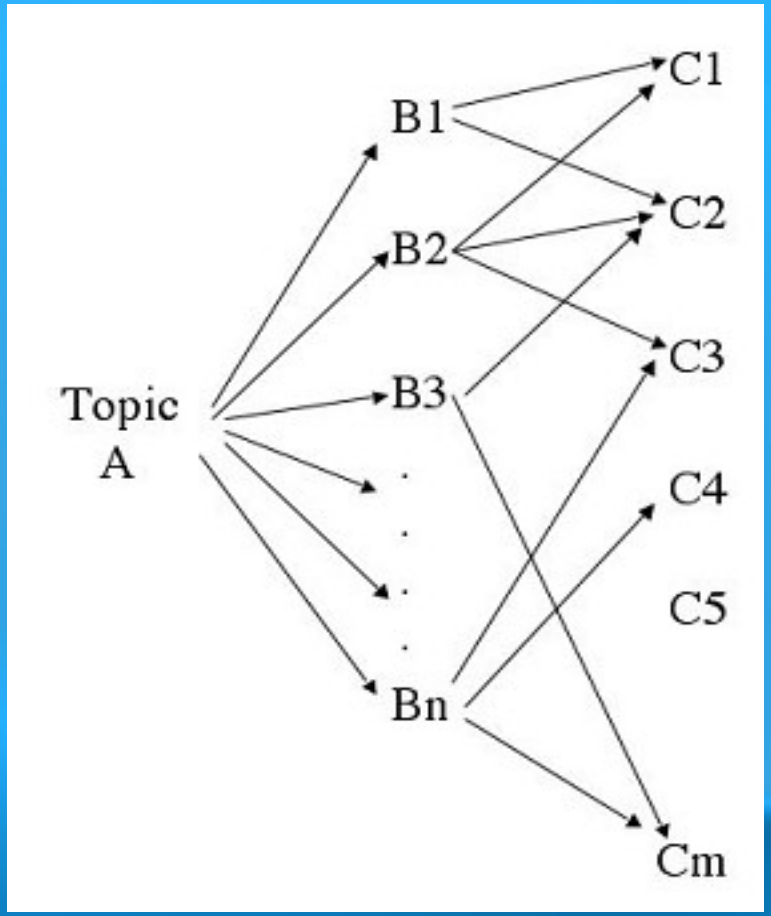
Padmini Srinivasan & Bisharah Libbus

Presented by Fernando Gutierrez

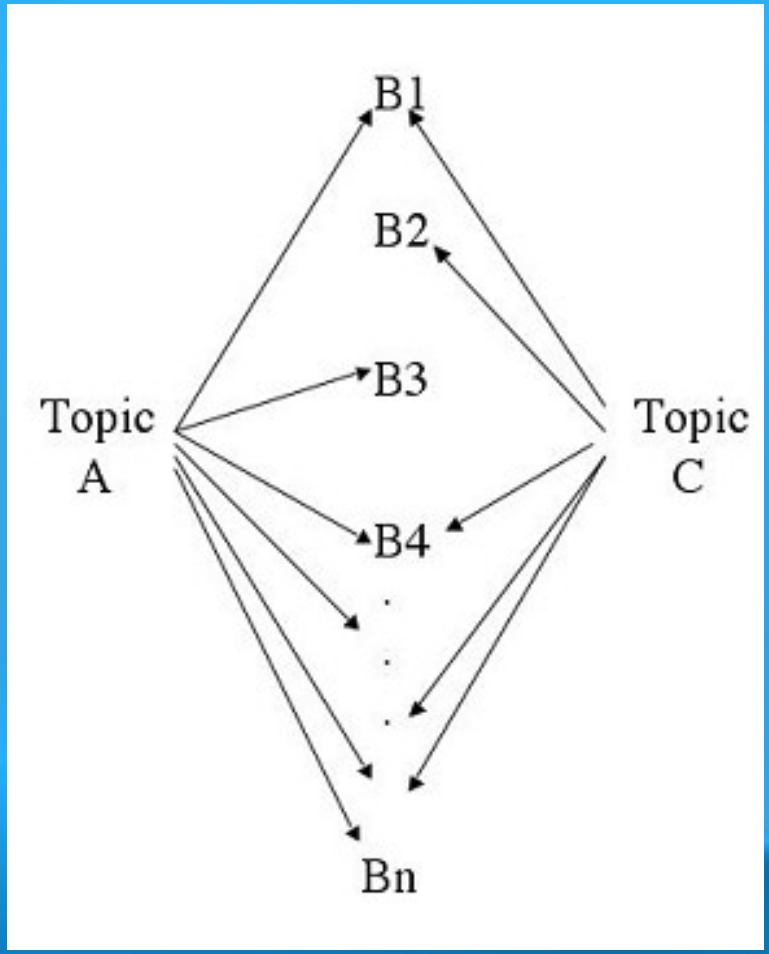
Hypothesis Generation

- Open discovery
  - Starts with single topic (A), goes through intermediate topics (B), finds terminal topic (C).
- Closed discovery
  - Starts with two topics (A,C), finds connecting topic (B).

Hypothesis Generation:
Open Discovery



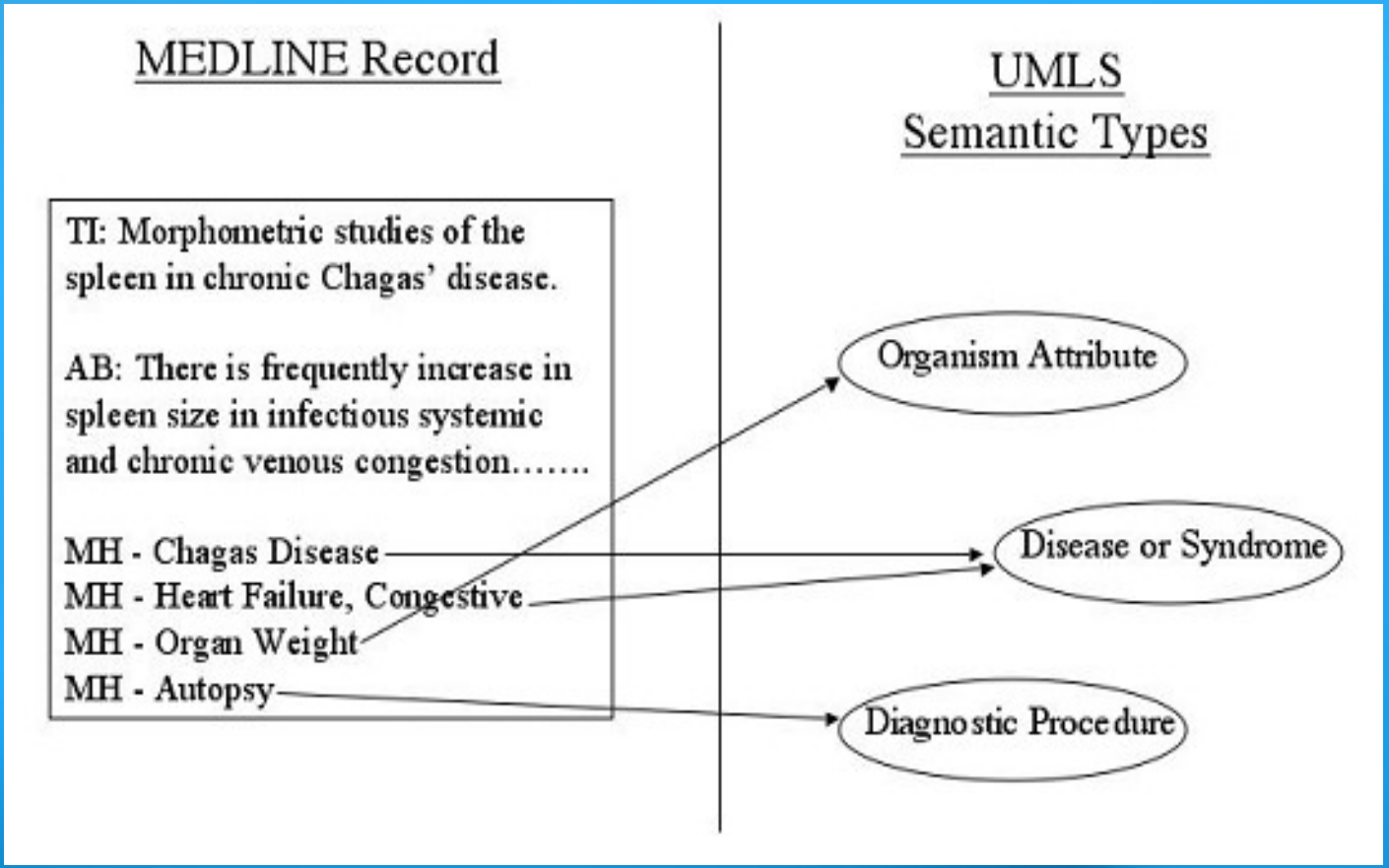
Hypothesis Generation:
Closed Discovery



Open Dicovery

- Term are extracted from MEDLINE
  - MeSH and UMLS
  - Ex: *intrerferon type II*
    - *Immunologic Factor*
    - *Pharmacologic Substance*

Term Extraction:
Using MeSH & UMLS



Topic Profile

Terms Normalized Weight:

- TF*IDF:
  - TF: number of times the MeSH term occurs in the retrieved documents
  - IDF: inverse document frequency
- Normalized:

$$weight(t_i) = v_i / \sqrt{v_1^2 + v_1^2 + ... + v_r^2}$$

Topic: Raynauds - limited to publications before 1986
PubMed Search: Raynaud AND human AND 1960[DP]:1985[DP]
Number of documents retrieved: 2,733
Number of MeSH term instances in the document set: 52,271
Number of unique MeSH terms in the document set: 2,972
Profile: (top 5 terms for a few semantic types are shown below)

Semantic Type: *Body Space or Junction:*
finger joint (1.0), wrist joint (0.81), elbow joint (0.55), esophagogastric junction (0.33)

Semantic Type: *Cell:*
neutrophils (1.0), blood platelets (0.78), erythrocytes (0.71), eosinophils (0.53), lymphocytes (0.5)

Semantic Type: *Cell Function:*
platelet aggregation (1.0), platelet adhesiveness (0.56), neural conduction (0.5), erythrocyte aggregation (0.44)

Semantic Type: *Organ or Tissue Function*
regional blood flow (1.0), microcirculation (0.41), vasoconstriction (0.41), blood flow velocity (0.41), hemodynamics (0.31)

Semantic Type: *Disease or Syndrome*
raynaud's disease (1.0), scleroderma, systemic (0.23), vascular diseases (0.09), occupational diseases (0.077), cold (0.074)

Semantic Type: *Eicosanoid:*
epoprostenol (1.0), prostaglandins e (0.65), prostaglandins (0.52), alprostadil (0.51), prostaglandins e, synthetic (0.15)

Semantic Type: *Organism Function:*
aged (1.0), blood pressure (0.29), exertion (0.1), body temperature regulation (0.09), pregnancy (0.07), menstruation (0.04)

Semantic Type: *Physiologic Function:*
blood viscosity (1.0), blood circulation (0.63), pulse (0.38), vascular resistance (0.33), collateral circulation (0.13)

Number of unique MeSH terms in profile: 2,972
Total number of MeSH term entries in profile: 4,419 (a term can be in multiple semantic types)
Top 5 Semantic types ranked by number of terms: *Disease or Syndrome (686), Pharmacologic Substance (359), Organic Chemical (291), Laboratory Procedure (224), Body Part, Organ, or Organ Component (198)*
Number of semantic types with at least 1 term in profile: 114 (out of 134 possible)

Open Discovery Algorithm

Input: (1) an A topic, (2) ST-B and ST-C: two sets of UMLS semantic types and (3) M

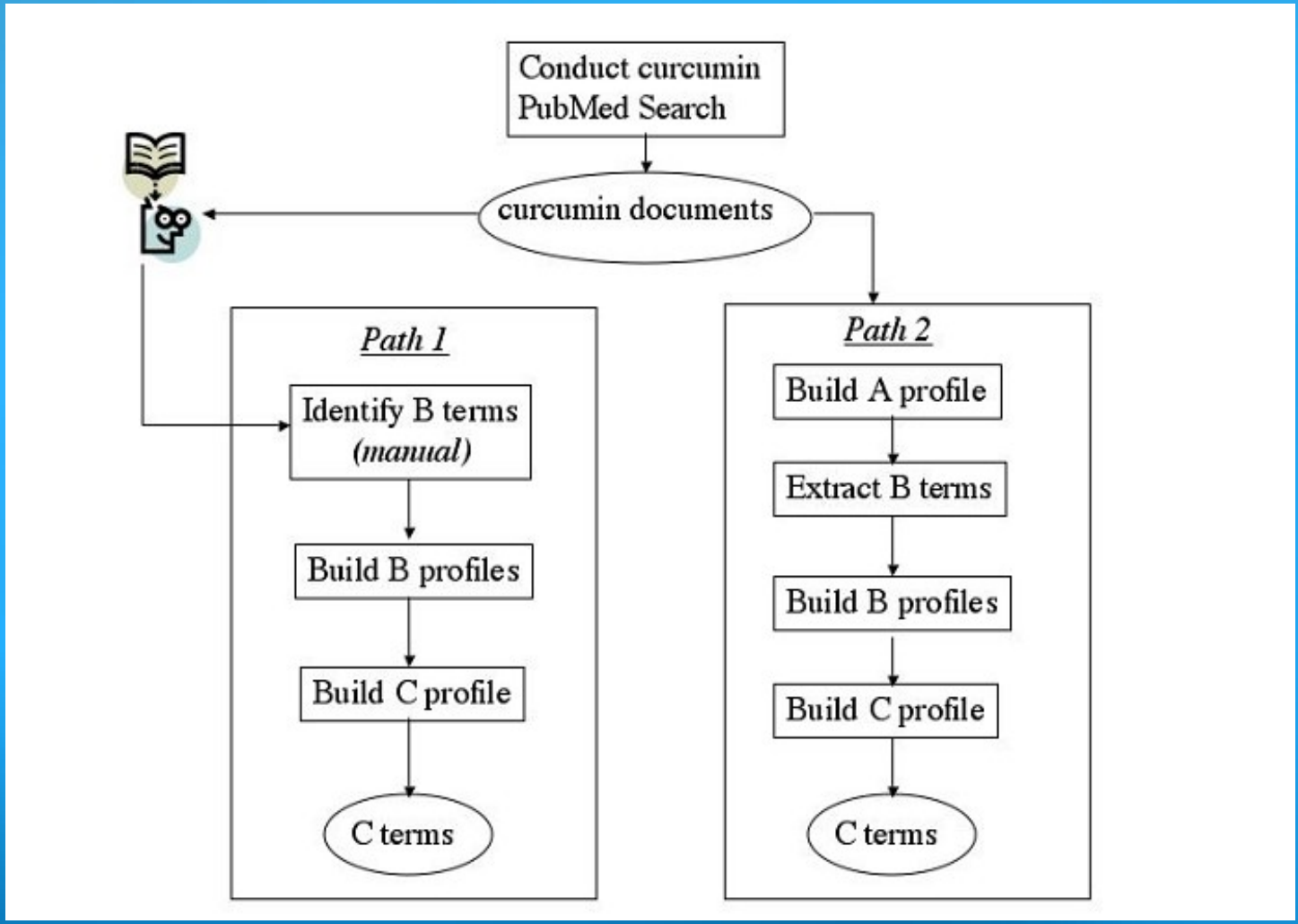
1. Search PubMed for A, and build its topic profile (AP).
2. For each semantic type in ST-B, select the M top ranking MeSH terms from AP. Remove duplicates. Call these (B1, B2, B3, etc.).
3. Search PubMed for terms B1, B2, B3, etc. (independently) and build their profiles (BP1, BP2, BP3, etc.).
4. Build a combined profile limited to ST-C semantic types where the combined weight of a MeSH term is the sum of its weights in BP1, BP2, BP3, etc. (CP).
5. Eliminate term t in CP if a PubMed search on A AND t retrieves documents.

Output: For each semantic type in ST-C, output MeSH terms in CP ranked by the combined weight.

Experiment

- Term: Turmeric/curcumin
  - *Turmeric OR Curcumin OR Curcuma*
- Semantic type:
  - *Gene or Genome*
  - *Enzyme*
  - *Amino Acid, Peptide or Protein*

Experiments



Results

- PubMed document retrieved: 1,175
- Manual refined
  - Synonyms

  Overlap 43% top 10 C terms between methods
  - Retinal Disease
  - Crohn´s Disease
  - Spinal Cord

Results

Disease	A10	M	Disease	A10	M
Retina	1 (1)	1 (1)	Cystic Fibrosis	8 (33)	
Spinal Cord	2(2)	2 (8)	Epilepsy	9 (35)	
Cytomegalovirus	3 (18)	3 (10)	Uremia	10 (36)	
Amyotrophic Lateral Sclerosis	4 (25)		Choriocarcinoma		6 (22)
Crohn Disease	5 (26)	7 (32)	Sarcoma Kaposi		8 (34)
Lupus Erythematosus Systemic	6 (27)	5 (19)	Graves Disease		9 (39)
Hodgkin Disease	7 (29)	4 (17)	Sjorgens Syndrome		10 (42)

Conclusion

- Each of the cases found
  - Plausible connections between disease and curcumin
  - B terms not so good performance
  - C terms good performance
  - Manually refine C.