Building Complementary Domain Taxonomies using Query Enrichment

- Simoni S. Shah, Shraddha Bhattad, Sanket Lokegaonkar, Ganesh Ramakrishnan
Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay

Motivation: Carefully curated taxonomy by Domain Expert

Why domain taxonomy?

- When the repository is specific, generic or umbrella taxonomies do not suffice.
- Domain to be covered with higher specificity.

And taxonomies in general help

- Organizing data : textual/multimedia
- Structured content
- Effective and intuitive browsing, searching
- Recommendation

Expert-created taxonomy

- Scenario: (Growing) repository, “seeded” with content that is organised hierarchically by a domain expert.
  - Expert’s intuition is important
- Expert organises the data points of the domain based on his intuition.
- The expert’s taxonomy
  - Specific to his mental model
  - May not be comprehensive
  - New concepts may need to be included, as the repository grows
- Domain expertise expensive!
- Question: Possible to reduce the load on domain expert?

Complementary Graphs

Expert Graph  Precision Oriented

Generic (WordNet) Graph  Recall Oriented

Justification of choice:

- Multilingual (Indo) Wordnet: vast expanse and multilingual support
  - Indo-Wordnet supports 18 Indian languages
- Employ sense and synsets to map from expert graph to Wordnet.
- Main idea: Subset WordNet based on the domain,
  - Define mappings from nodes in expert-graph nodes to related Wordnet nodes.

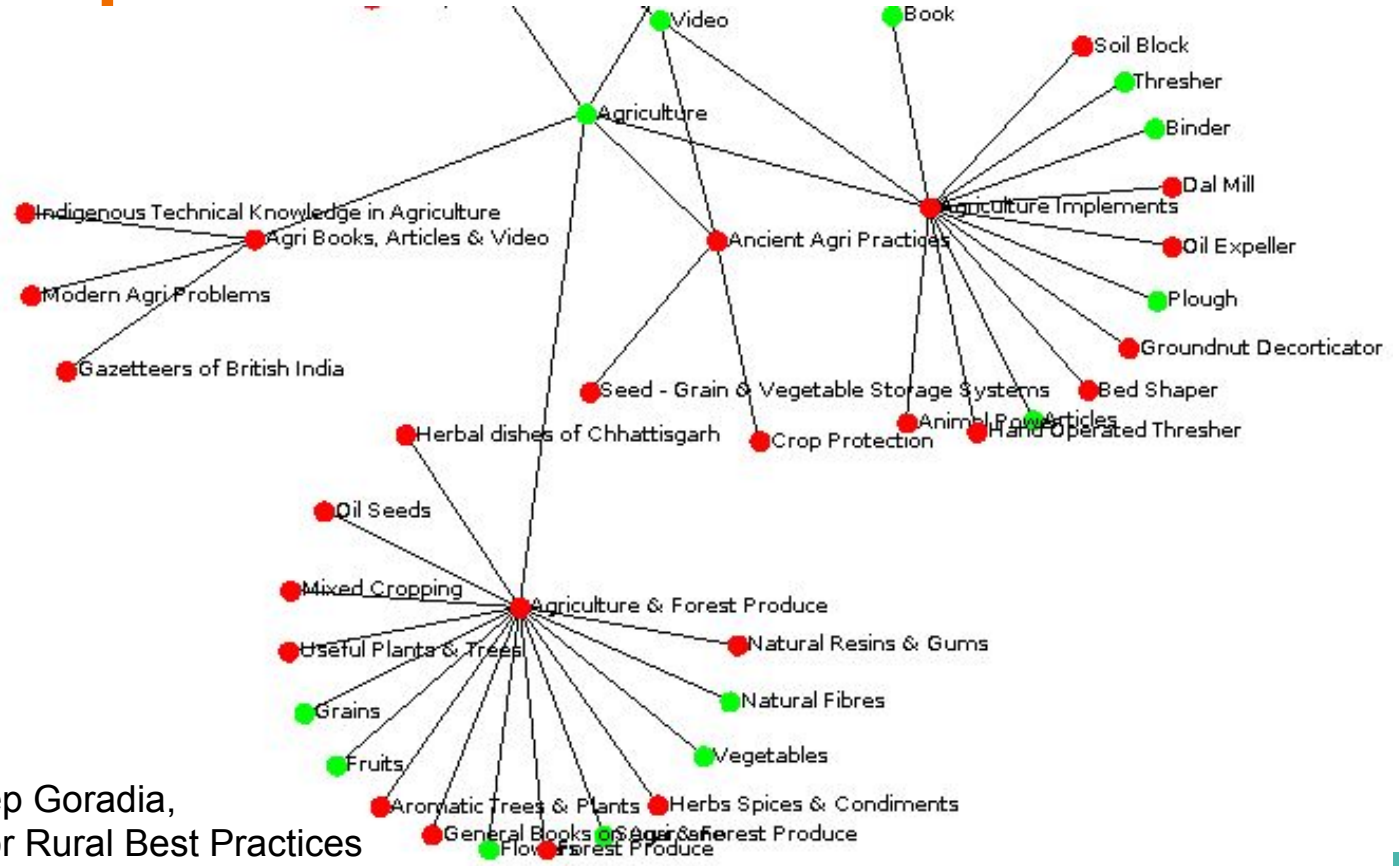
Mapping Expert Graph to WordNet

210 nodes in the Expert Graph: 119 were present in Wordnet

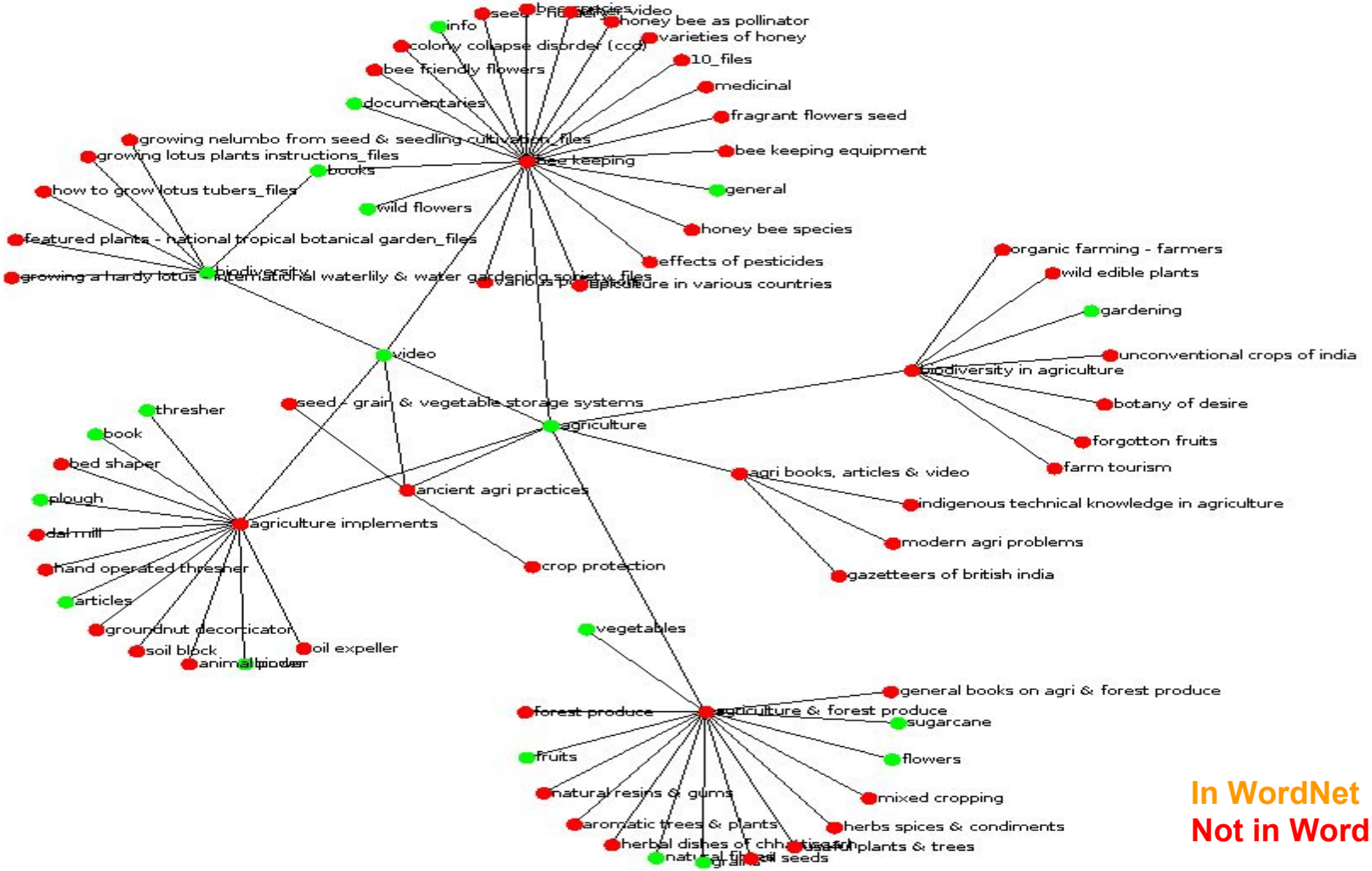
For remaining 91 nodes, we enrich the query-nodes to know further about its semantics.

1) Query Enrichment using Expert's data-organisation hierarchy
2) Query Enrichment using Wikipedia

Expert Graph



Courtesy: Sandeep Goradia,
Domain Expert for Rural Best Practices


info
colony collapse disorder (ccd)
bee friendly flowers
documentaries
growing nelumbo from seed & seedling cultivation_files
growing lotus plants instructions_files
how to grow lotus tubers_files
featured plants - national tropical botanical garden_files
growing a hardy lotus
honey bee as pollinator
varieties of honey
10_files
medicinal
fragrant flowers seed
bee keeping equipment
general
honey bee species
effects of pesticides
bee keeping
books
wild flowers
biodiversity
video
seed - grain & vegetable storage systems
agriculture
ancient agri practices
crop protection
agriculture implements
thresher
book
bed shaper
plough
dal mill
hand operated thresher
articles
groundnut decorticator
soil block
oil expeller
organic farming - farmers
wild edible plants
gardening
unconventional crops of india
biodiversity in agriculture
botany of desire
forgotton fruits
farm tourism
agri books, articles & video
indigenous technical knowledge in agriculture
modern agri problems
gazetteers of british india
vegetables
forest produce
agriculture & forest produce
general books on agri & forest produce
sugarcane
fruits
flowers
natural resins & gums
mixed cropping
aromatic trees & plants
herbs spices & condiments
seed oil
In WordNet
Not in WordNet

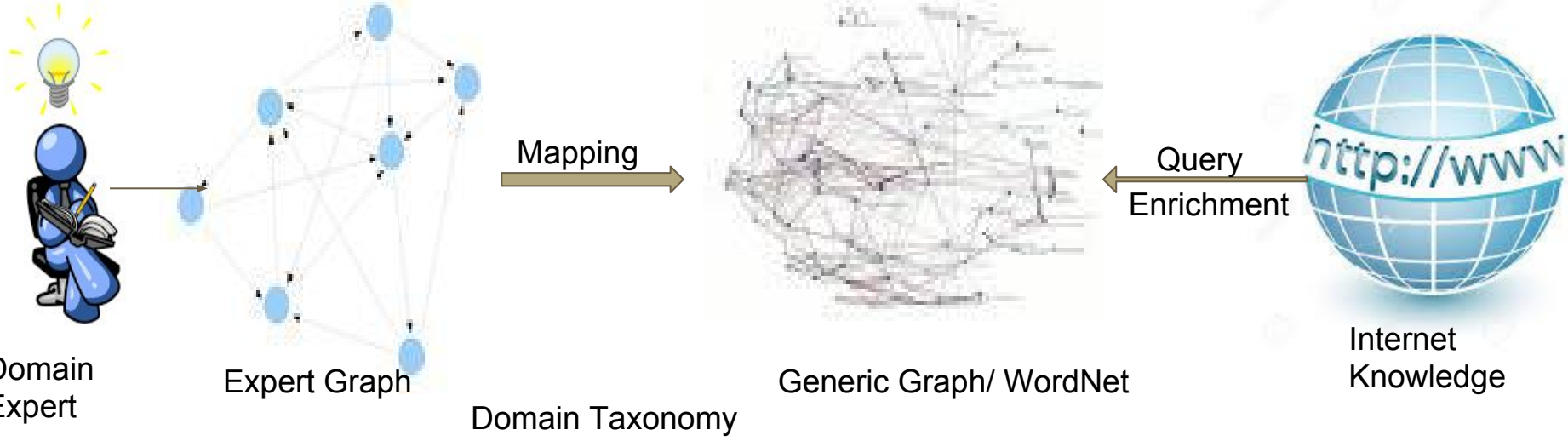
Overall Approach:

For existing and new node **n**:

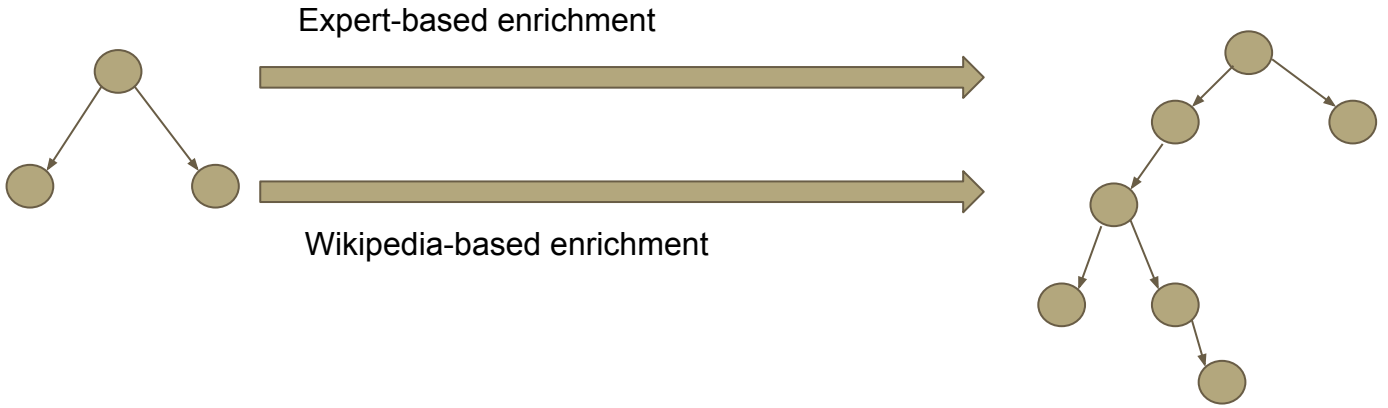
- Step 1: Build the Expert Graph using the domain hierarchy (directory structure) created by the expert
- Step 2: For every node **n** in the Expert graph, map n to a (set of) node(s) in WordNet. Here we have two possibilities:
  - Case (i): n is a green node; i.e. it exists as a node in WordNet. In this case, disambiguate the sense.
  - Case (ii): n is a red node; i.e. it does not exist in WordNet. In this case, map n to a subset of nodes in WordNet, which are semantically close to it, wrt to the given domain.
    - Use Expert-based Query enrichment and query WordNet directly.
    - Use Wikipedia- based query enrichment and find useful WordNet nodes
- Step 3: For every new node suggested by the expert, carry out the above steps.

Overall Approach

Develop a framework that use Domain Expertise complemented by Internet knowledge, in order to create and evolve a domain taxonomy.



Taxonomy subsetting using enrichment & mapping



Expert Taxonomy

Wordnet Taxonomy subsetting

Query Enrichment: [KDDCUP 2005]winners: “Building Bridges for web-query classification”

Query enrichment using Expert's Domain Hierarchy

- For every occurrence of node **n** in the expert-directory, a query is made,
  - Query consists of keywords on **path from n to the root (ancestry of n)**.
- Look for potential and semantically close mappings in Wordnet.
  - TF-IDF- weighted ranking of WordNet synsets

Examples for Expert-based Query Enrichment

For “**Silvopasture**”: Not in WordNet, present in Wikipedia

Results from Expert-based Query Enrichment: "cultivation", "production" and "irrigation".

For **“Agniastra”** : not present in WordNet or Wikipedia

Results from Expert-based Query Enrichment: pile , collection,toiletry , instrumentality.

Note that Agniastra is a pesticide created using cow-urine and neem leaves.

Query Enrichment using Wikipedia

- 76 of the 91 red nodes had Wiki-pages for them.
- Steps for mapping into WordNet using Wiki-based Query Enrichment:
  - Extract hyperlinked words from Wikipedia page of a given node.
  - Filter out words not present in WordNet
  - Word2vec-based filtering, by measuring proximity to the domain
  - For word-sense disambiguation, use related concepts.

Word2Vec pruning: the words obtained from querying Wikipedia are further examined to improve precision. Every word w of the query result is assigned a score, based on its average “semantic proximity” (cosine similarity using Word2Vec) with the nodes in the Expert graph.Only words with high score are considered.

Mapping Algorithm

- Input Query with meta-data obtained from enrichment (Expert/Wikipedia).
- Filter out stopwords, non-alphanumeric characters and stem for both query as well as WordNet
- Use TF-IDF (Term Frequency- Inverse Document Frequency) to reflect how important a word is to a document in a collection.

Examining TF-IDF's importance

- This makes scoring for specific words higher say, as for example in case of
- 'Water' : 4.66305084537
- 'Entity' : 1.0
- 'Irrigation': 8.99798587251
- 'Aquaculture': 11.1182494087
- 'Liquid': 5.17457552544
- 'Hydroponics': 10.8305673363
- 'Agriculture': 8.95876515936

Here., Aquaculture, Hydroponics, Agriculture have high scores compared to water, liquid, and entity is scored one as it is the root-node of WordNet organisation(Hence appears on every path., which we have posed as documents for TF-IDF)

Examples for Wiki-based Query Enrichment

- For **"biodynamics"**:

  Results from Wiki-based QE: "agriculture climate change horticulture bamboo processing"

  Words filtered out after Word2Vec filtering: "blanc rudolf steiner rootstock riesling red wine quartz pruning powdery mildew port wine pinot"

- For **"bee keeping":**

  Relevant words from Wiki: "agriculture, sericulture, honey, silk, beeswax, corn syrup, mealworm, cypress, burlap, pollination, apiary"

  Irrelevant words from Wiki: " ptolemy, mythology, livy, latin, iberian peninsula, eratosthenes"

Analysis of Word2vec Pruning :-

- For **"biodynamics"**,

  True positives: "agriculture horticulture floriculture compost fertilizer flower garden garden grape gardening horsetail botanical garden greenhouse composting orchard botany".

  False Negatives: "herbicide pomology" (due to a threshold on no. of words)

  For **"bee keeping"**

  True positives: "agriculture silkworm sericulture honey silk beeswax corn syrup mealworm cypress burlap pollination apiary sumac fungus ant".

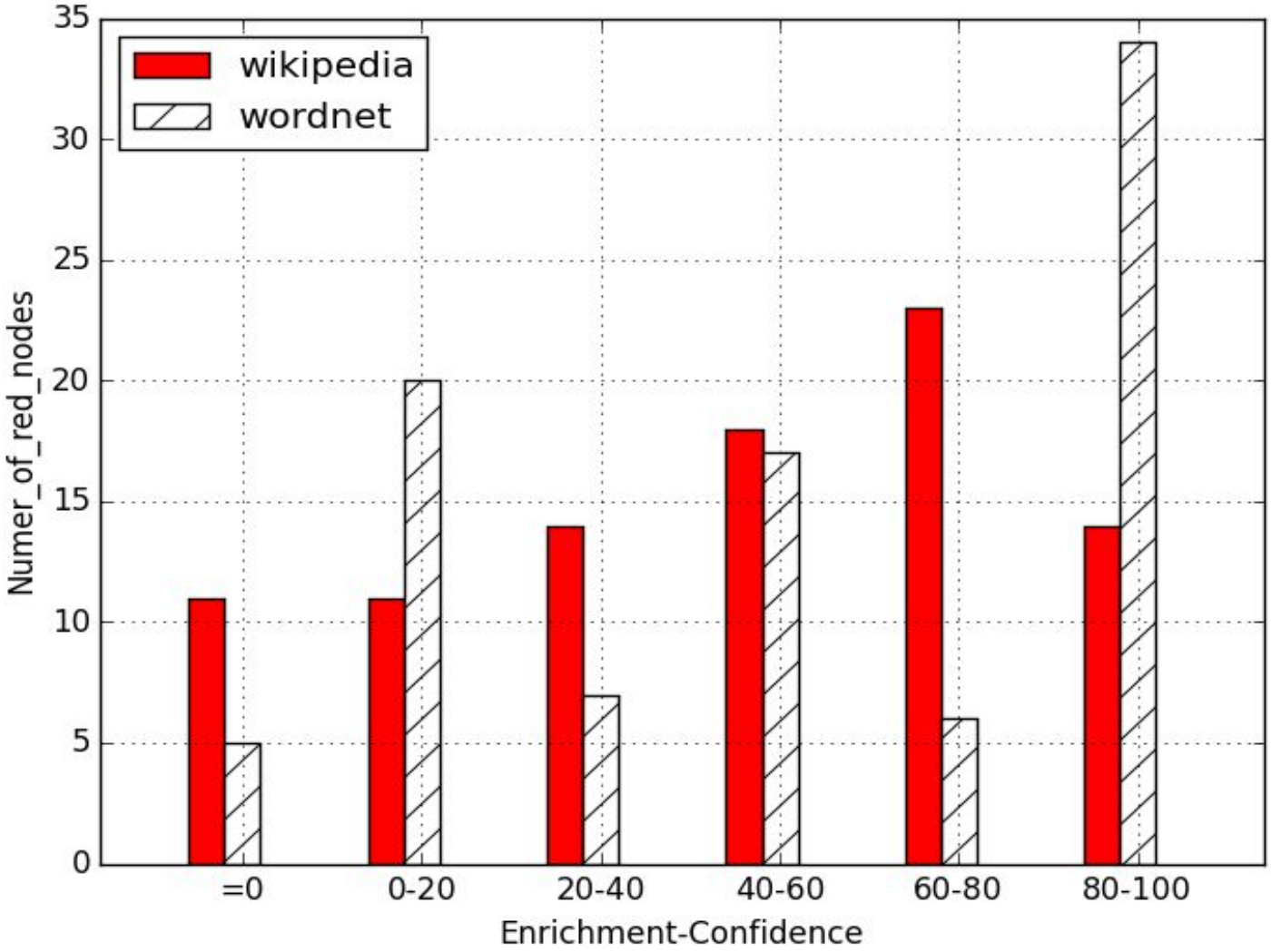


Figure depicts manual evaluations done for query enrichment statistics from ancestry as well as from Wikipedia hyperlinks

Original Query	Wordnet nodes using query ancestry	Wordnet Nodes using Word2vec filtering on Wikipedia hyperlinks
mud housing	(plaster : mixture:architecture : discipline)	(slop:chicken_manure:fertilization:mud_brick: mud_pie:mud:fertilizer:night_soil: horse_manure:cow_manure)
date palm sugar	(cane : stalk)	(coconut:palmyra:copra:coconut:jaggery:palm: carbohydrate:fan_palm:wine_palm:oil_palm)
clay refrigerator	(tray : receptacle)	(earthenware:cooler:refrigerator:terra_sigillata: mugginess:chukker:red_tide: electric_refrigerator:terra_cotta:delft)
butter churning	(butter : food)	(clarified_butter:butter:yak_butter:ghee:stick: meuniere_butter:spread:brown_butter: buttermilk:butter_cookie)
panchagavya	(toiletry : instrumentality:pile : collection: cultivation : production)	(arsenical:spray:phosphine:curd:night_soil: fertilization:horse_manure:chicken_manure: cow_manure:pesticide)
marine life	(zoology : biology:archeology :	(seaweed:seafood:cyprinid:bonefish:

Some cases to be improved upon :-

multicrop	**(melon : edible_fruit: executive_department : federal_department:class : people)**	**Miss (although there are results for multi crop)**
no electricity refrigerator #no doesn’t makes much difference	**(tray : receptacle)**	**(pump:sea_trout:rainbow_trout: stock:rosebay: coast_rhododendron: hydraulic_pump: electric_refrigerator:brown_trout: blackberry)**

Conclusion and Future Work

- Query enrichment using the paths in the data set hierarchy (that is provided by the domain expert) greatly enhances precision of WordNet querying
- More than 80% of the concepts that are not present in WordNet, have relevant Wikipedia pages.
    - Querying wiki helps in improving on recall.
- Pruning using word2vec helps us assess the relevance of wiki output, to the domain outlined by the expert.
- Both, the WordNet querying as well as Wikipedia querying intrinsically use domain expert knowledge, to provide results with better precision.
- Certainly helps relieving the domain expert from the burden of curating and maintaining a cannonical taxonomy.