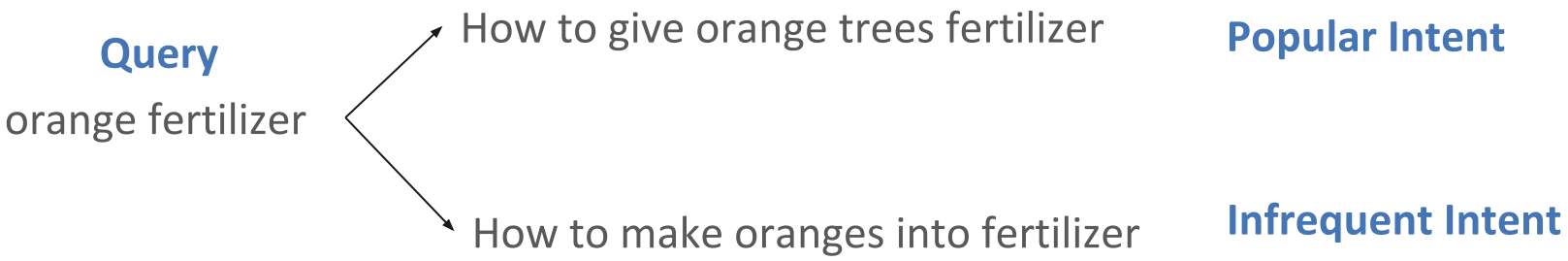
Disjunctive Sets of Phrase Queries for Diverse Query Suggestion

Ziyang Liao, Keishi Tajima
Kyoto University

WI 2019
2019.10.16

Research Background

- **Many ambiguous queries that can be interpreted in multiple ways**



- **Query results are filled with pages corresponding to popular query intents**
- **Difficult to find queries that can retrieve pages corresponding to infrequent query intents.**

Existing method 1

- **Search engines recommend queries extracted from query log data**
  - Queries for infrequent intents appear in the query log data only infrequently

    E.g., Bing query recommendation result for “Orange fertilizer”

関連キーワード

orange **trees** fertilizer

fertilizer **for** orange **trees florida**

natural fertilizer **for** orange **trees**

satsuma orange fertilizer

best shrub fertilizer

types of fertilizers for crops

alternatives to nitrogen fertilizer

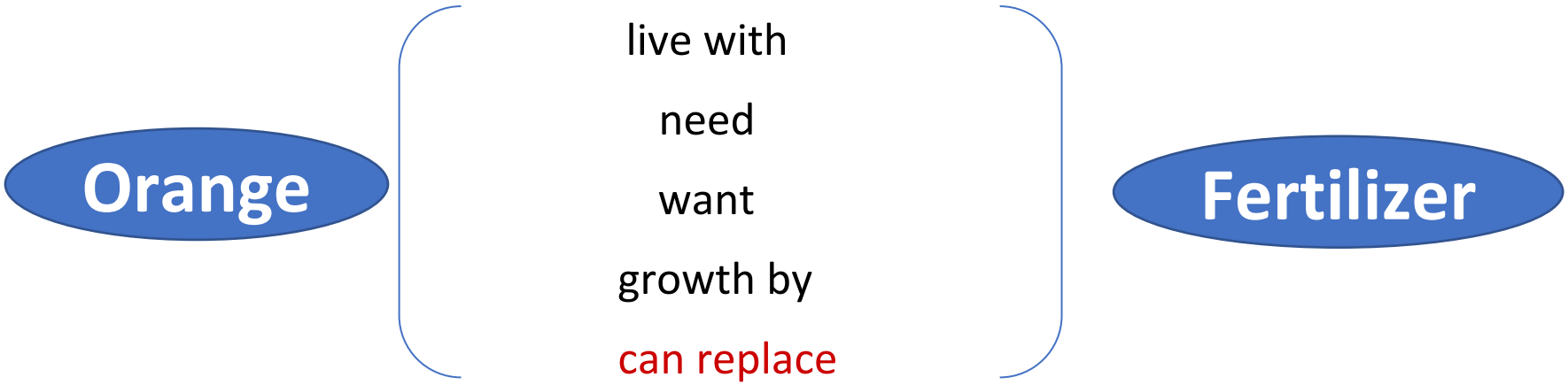
fertilizer **plant**

How to make orange into fertilizer ❌

How to give orange tree fertilizer ✅

Existing method 2

- **Extract phrases that represent relationships between query terms from a web corpus [1]**
  - Web corpus includes more information than query log data
  - It is still difficult to find phrases corresponding to infrequent query intents
  - Phrases corresponding to infrequent query intents are buried in those corresponding to popular query intents



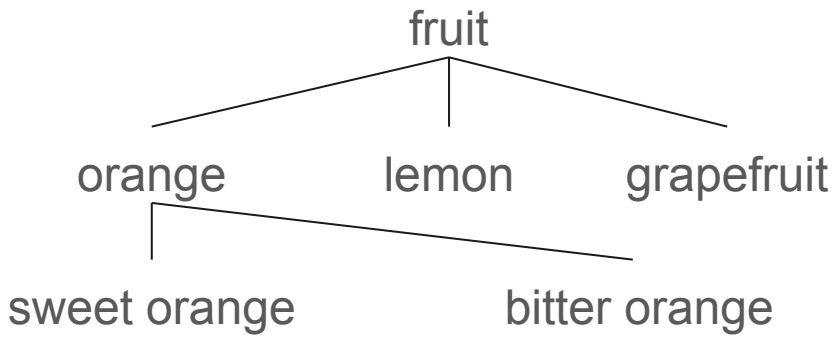
[1] Daisuke Fukuchi, Takehiro Yamamoto, Katsumi Tanaka, Query Mining Based on Term Relationship Estimation in Verbal Queries, Transactions of the Japanese Society for Artificial Intelligence, , 2017, Volume 32, Issue 1, Pages WII-J_1-15,

Our Approach 1

Expand query terms using hypernyms, sister terms, and hyponyms

To increase the chance of finding phrases corresponding to infrequent query intents, we expand each query term into the set of its hypernyms, syster terms, and hyponyms.

- hypernym
- sister term
- hyponym



Our Approach 1 example

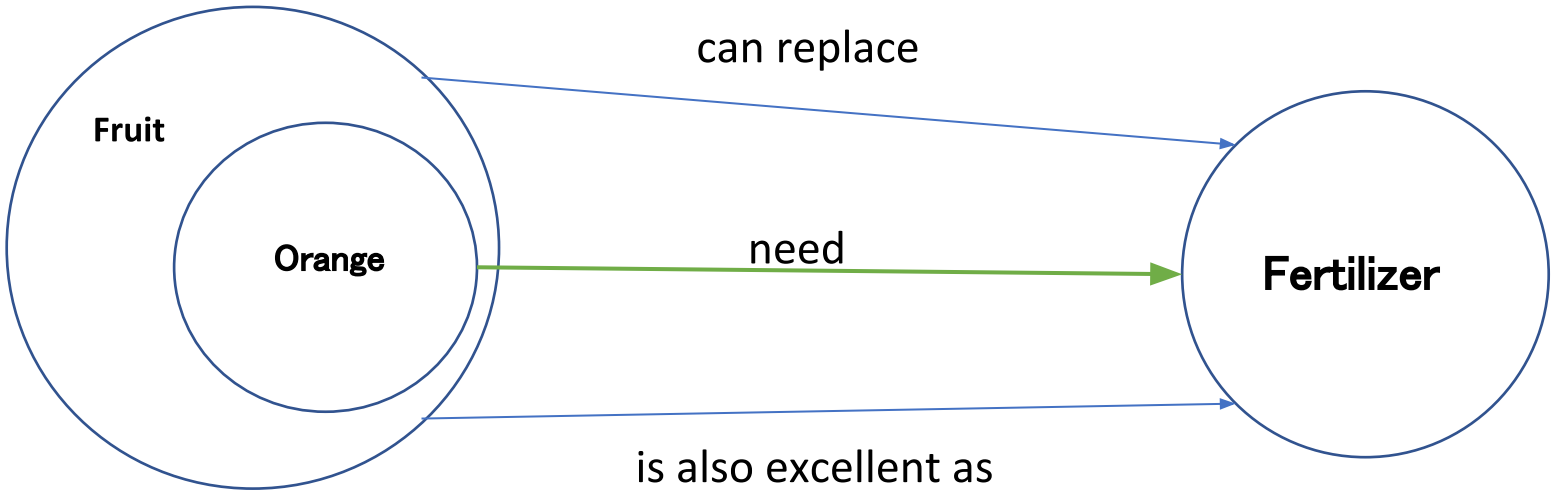
- **Expand the query term “orange”**
- **Find connecting phrases using all expanded query terms**

Query: Orange Fertilizer

Expansion: Orange → {Orange, Fruit, …}

Phrase: Fruit can replace Fertilizer

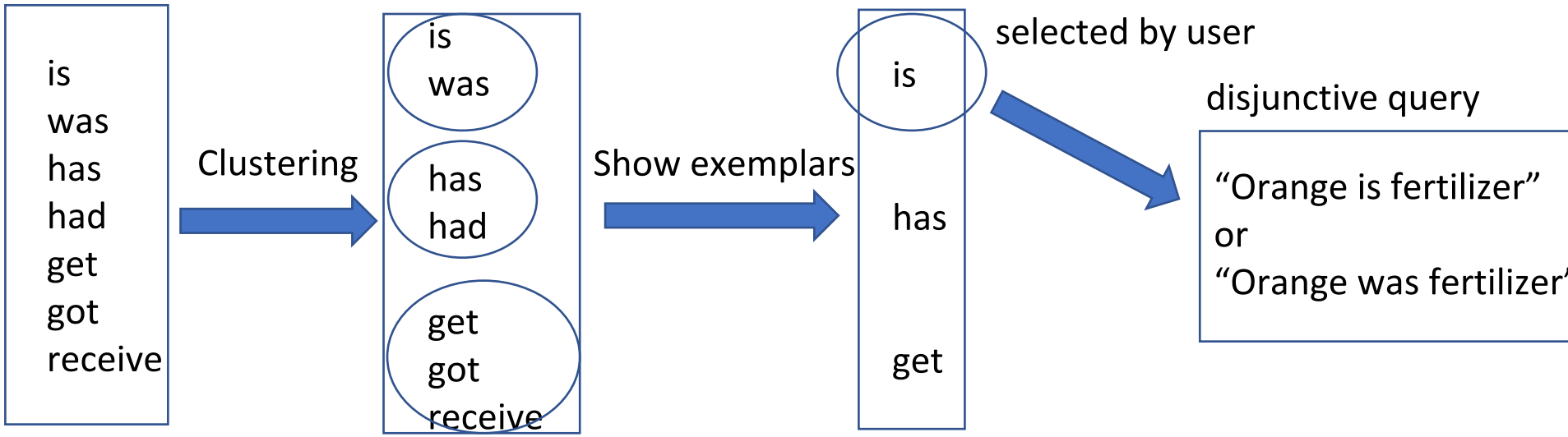
Expanded query:
“Orange can replace Fertilizer”



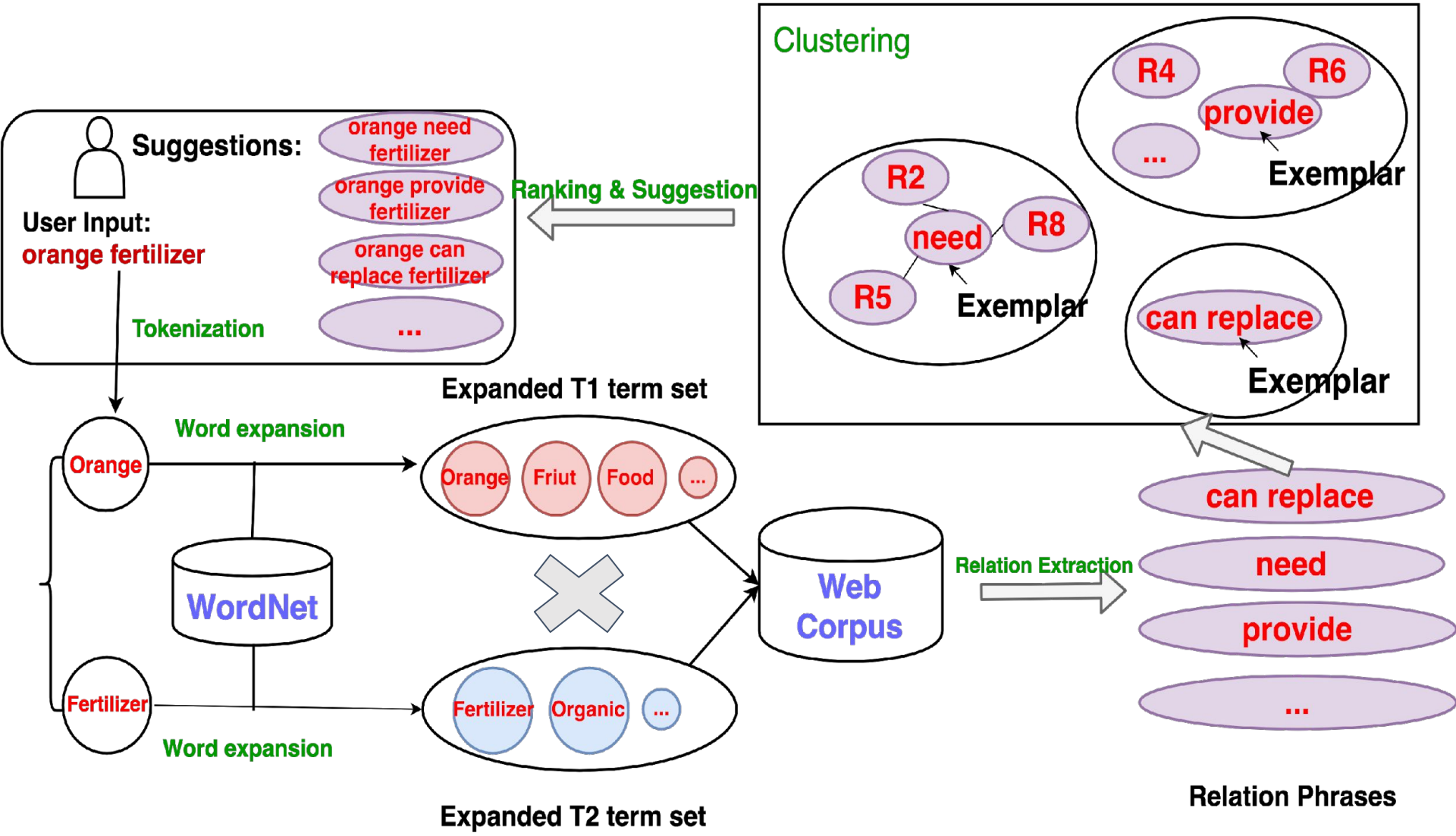
Approach 2

Clustering extracted phrases

- By query term expansion, we obtain many phrases
- Phrases for infrequent intents are buried in those for popular intents
- Cluster related phrases and only show exemplars of the clusters



Overview


Clustering
R4
R6
provide
...
Exemplar
R2
need
R8
R5
Exemplar
can replace
Exemplar
Suggestions:
orange need fertilizer
orange provide fertilizer
orange can replace fertilizer
...
Ranking & Suggestion
User Input:
orange fertilizer
Tokenization
Expanded T1 term set
Word expansion
Orange
Orange
Friut
Food
...
WordNet
Web Corpus
Relation Extraction
can replace
need
provide
...
Fertilizer
Fertilizer
Organic
...
Word expansion
Expanded T2 term set
Relation Phrases

Method details

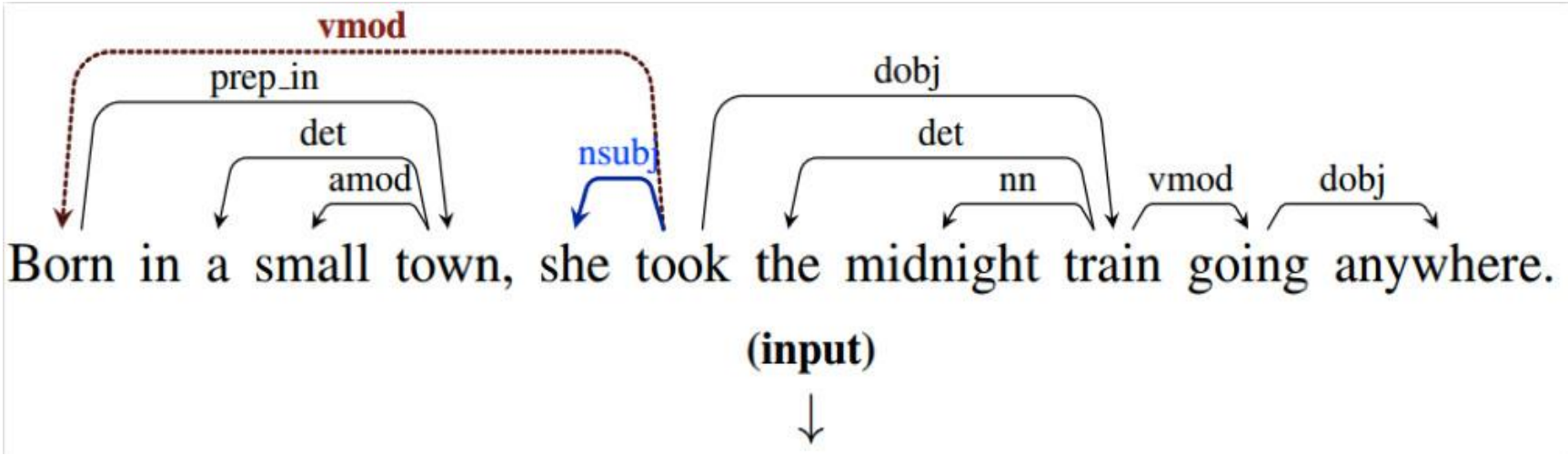
- **Query term expansion**
  - WordNet
- **Connecting phrase extraction**
  - OpenIE (Open information extraction) [2]
- **Web corpus**
  - ClueWeb
- **Clustering algorithm**
  - Affinity Propagation algorithm

[2]. Mausam, Pal, H.,&Saha, S. (2017). Bootstrapping for Numerical Open IE. ACL.

Open Information Extraction

Word extraction tool

- OpenIE (Open information extraction)



(she; took; midnight train)

Data processing procedure

Remove triples with incorrect information

- Triple with typographical errors (eg: apple ai fertilizer)
- 0 results in search engine (eg: apple is computed from fertilizer)

Similarity calculation and clustering of related phrases

- Extracting related phrases from the captured relationship triples
- Calculate similarity between related phrases
- Clustering based on similarity

Calculate the overall weight of the cluster and rank

- Score considering the frequency of elements in the cluster and the type of expansion

Clustering algorithm

Affinity Propagation algorithm

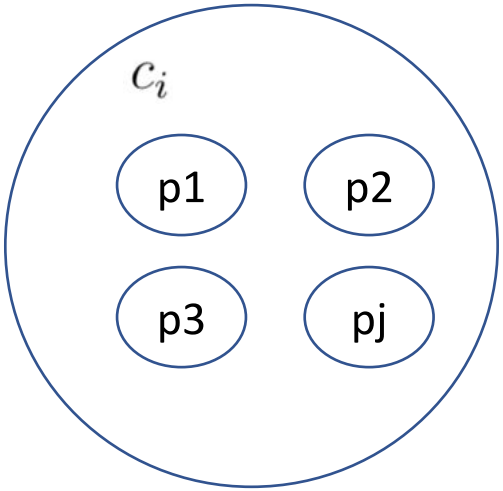
- Premise
  - Cluster target: phrases
  - Distance: Value of the semantic similarity between phrases
- Nature
  - Dynamically determines the number of clusters without the need to specify K as in K-means
  - A point to be exemplar is chosen from each cluster

Cluster ranking algorithm

Ranking score

Query: “e1 e2”

$$S(c_i) = w(e_i^1)w(e_i^2) \sum_{p_j \in c_i} f_j$$



- $S(c_i)$ cluster c_i ranking score
- $w(e_i^1)$ The weight of the expansion type of e1
- $w(e_i^2)$ The weight of the expansion type of e2
- f_j : The frequency of phrase p_j in cluster c_i

Expansion type e	**Weight value w(e)**
hypernym	1
sister terms	2
itself & hyponym	3

Ranking score calculation example

For “orange is fertilizer”

- The connecting phrase “is” is extracted by using “orange” itself and a hypernym of “fertilizer”
- The cluster size of “is” is 10

- $S(c_i)$: cluster c_i ‘s ranking score
- $w(e_i^1)$: The weight of the expansion type of e1 = 3
- $w(e_i^2)$: The weight of the expansion type of e2 = 1
- f_j : The frequency of phrase p_j in cluster c_i = 10

$$S(c_i) = w(e_i^1)w(e_i^2) \sum_{p_j \in c_i} f_j$$ = 3*1*10=30

Experiment

Procedure

1. Prepares queries that have multiple interpretations
2. Get query recommendation results
   - Query recommendation results of Google
   - Query recommendation results of Bing
   - Query recommendation results of proposed method
3. Search the web by using query recommendation results with Bing and get the top 10 results
4. Calculate the precision of the results that match the infrequent intent

Experiment

Test queries

Prepares multiple interpretable queries

Id	Query	Common search intent	Infrequent search intent
1-1	apple fertilizer	Fertilizer for apple	How to change apple into fertilizer
1-2	banana fertilizer	Fertilizer for banana	How to change banana into fertilizer
1-3	orange fertilizer	Fertilizer for orange	How to change orange into fertilizer
2-1	Kyoto bank	The Kyoto Bank	Information about the banks in Kyoto
2-2	Japan bank	The Bank of Japan	Information about the banks in Japan
2-3	China bank	The China Bank	Information about the banks in China
3-1	Steak sauce	Sauce for steak	How to make sauce using steak
3-2	Beef sauce	Sauce for beef	How to make sauce using beef
3-3	Chicken sauce	Sauce for chicken	How to make sauce using chicken

Evaluation

Get query recommendation results

Result of Google query recommendation and proposed method

Google	Proposed method
orange tree fertilizer florida	orange is fertilizer
homemade fertilizer for citrus trees	fertilizer produced orange
citrus fertilizer	fertilizer transported in orange
how often to water orange trees	fertilizer end up in orange
when to fertilize citrus trees in southern california	orange contains fertilizer
citrus fertilizer npk	fertilizer arrow orange
liquid citrus fertilizer with micronutrientsk	fertilizer pollute orange
fertilizer production	fertilizer found in orange
	orange has fertilizer
	orange made from fertilizer

Not match for infrequent intents **Match for infrequent intents**

Query recommendation for ''orange fertilizer''

Evaluation

Patterns

We calculate the result of following patterns

Patterns
1) Synonym expansion
2) Typeof expansion
3) Instanceof expansion
4) Hastypes expansion
5) Hasinstance expansion
6) Full expansion without expansion type scoring
7) Full expansion with expansion type scoring

Evaluation

Result

Calculate precision of results that match minor intentions

id	1-1	1-2	1-3	2-1	2-2	2-3	3-1	3-2	3-3
Google	0	0.14	0	0	0.08	0	0	0	0
Bing	0	0.24	0	0.03	0.04	0	0	0.03	0
Original query	0	0	0	0	0.07	0.05	0.02	0	0.03
1) Synonym expansion	0	0	0	0	0	0	0	0	0
2) Typeof expansion	0	0.38	0.05	0	0	0	0	0	0
3) Instanceof expansion	0	0	0	0	0	0	0	0	0
4) Hastypes expansion	0	0	0	0	0	0	0	0	0
5) Hasinstance expansion	0	0	0	0	0	0	0	0	0
6) Full expansion without expansion type scoring	0.05	0.41	0.06	0	0.09	0.07	0.02	0.05	0.03
7) Full expansion with expansion type scoring	0.05	0.41	0.06	0	0.1	0.09	0.03	0.1	0.03

Summary

- **Propose query expansion method for infrequent query intents**
  - To obtain more candidate queries, expand the query terms, and extract phrases that connect the two query terms.
  - A large number of phrases are extracted from the corpus and clustered to generate diverse phrases.
  - Add connecting phrases to the original query and generate query candidates.
  - The experimental results show the usefulness of the proposed method.