Data Summarization with Hierarchical Taxonomy

Xuliang Zhu
Hong Kong Baptist University
Hong Kong, China

HKBU Database Group

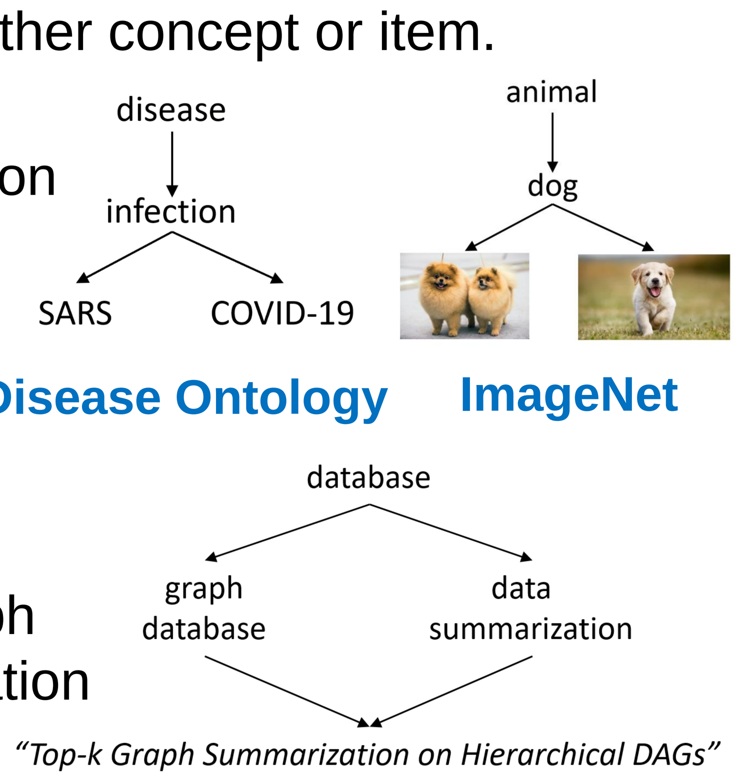
Motivations

Hierarchical DAGs (HDAG) are everywhere

Vertices are general concepts or certain items.

Directed Edges are the relationships that a general concept can summarize another concept or item.

- **Disease Ontology**
  Concepts: disease, infection
  Items: SARS, COVID-19
- **ImageNet**
  Concepts: animal, dog
  Items: images
- **ACM CCS**
  Concepts: database, graph database, data summarization
  Items: papers



HSD Problem

HSD Problem

Input: A hierarchical DAG, a query set of popular items Q, an integer k;

Output: A set S with at most k concepts;

Summary Score: $f(S) = \frac{|cov(S) \cap Q|}{|cov(S) \cup Q|}$

The larger summary score, the better selection!

S: The selection set.
cov(S): The items that set S cover.
Q: The query item set.

Objective: Find the set $S^* = \arg \max_{S \subseteq V, |S|=k} f(S)$

NP-hard: Reduction from set cover problem.

Applications: attributes filter, image set labeling, personalized recommendation

Algorithms

Transformation

Score function:

$$f(S) = \frac{|cov(S) \cap Q|}{|cov(S) \cup Q|} \geq \alpha$$
$$|cov(S) \cap Q| - \alpha \cdot |cov(S) \cup Q| \geq 0$$
$$|cov(S) \cap Q| - \alpha \cdot |cov(S) \setminus Q| \geq \alpha \cdot |Q|$$

Maximum weighted coverage:

$$g(S) = \sum_{x \in cov(S)} w(x) \geq \alpha \cdot |Q| \text{ ,where } w(x) = \begin{cases} 1 & , x \in Q \\ -\alpha & , x \notin Q \end{cases}$$

Binary search α, and then transform to the maximum weighted coverage problem.

DP on tree

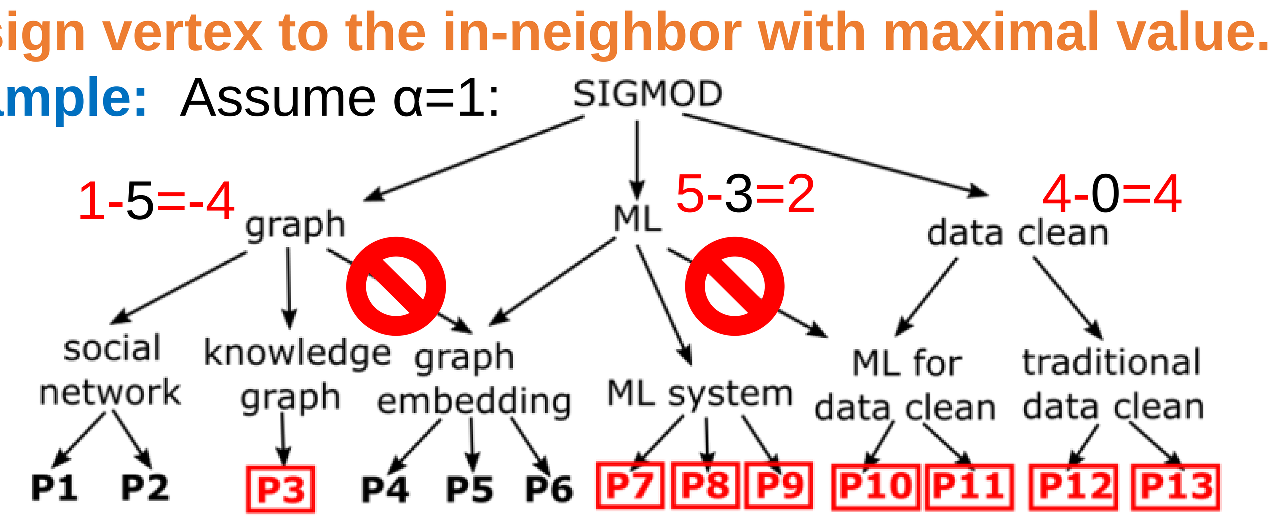
DP(v, k) the maximal weighted coverage with selecting no more than k vertices in subtree T_v.

$$DP(v,k) = \max\{DP_Y(v,k), DP_N(v,k)\}$$
$$DP_Y(v,k) = \sum_{u \in cov(v)} w(u), \text{ subject to } k \geq 1.$$
$$DP_N(v,k) = \max_{\{k_1,k_2,\ldots,k_u\}} \{\sum_{u \in N^-(v)} DP(u,k_u)\}, \text{ s.t. } \sum_{u \in N^-(v)} k_u \leq k.$$

Algorithm on HDAG

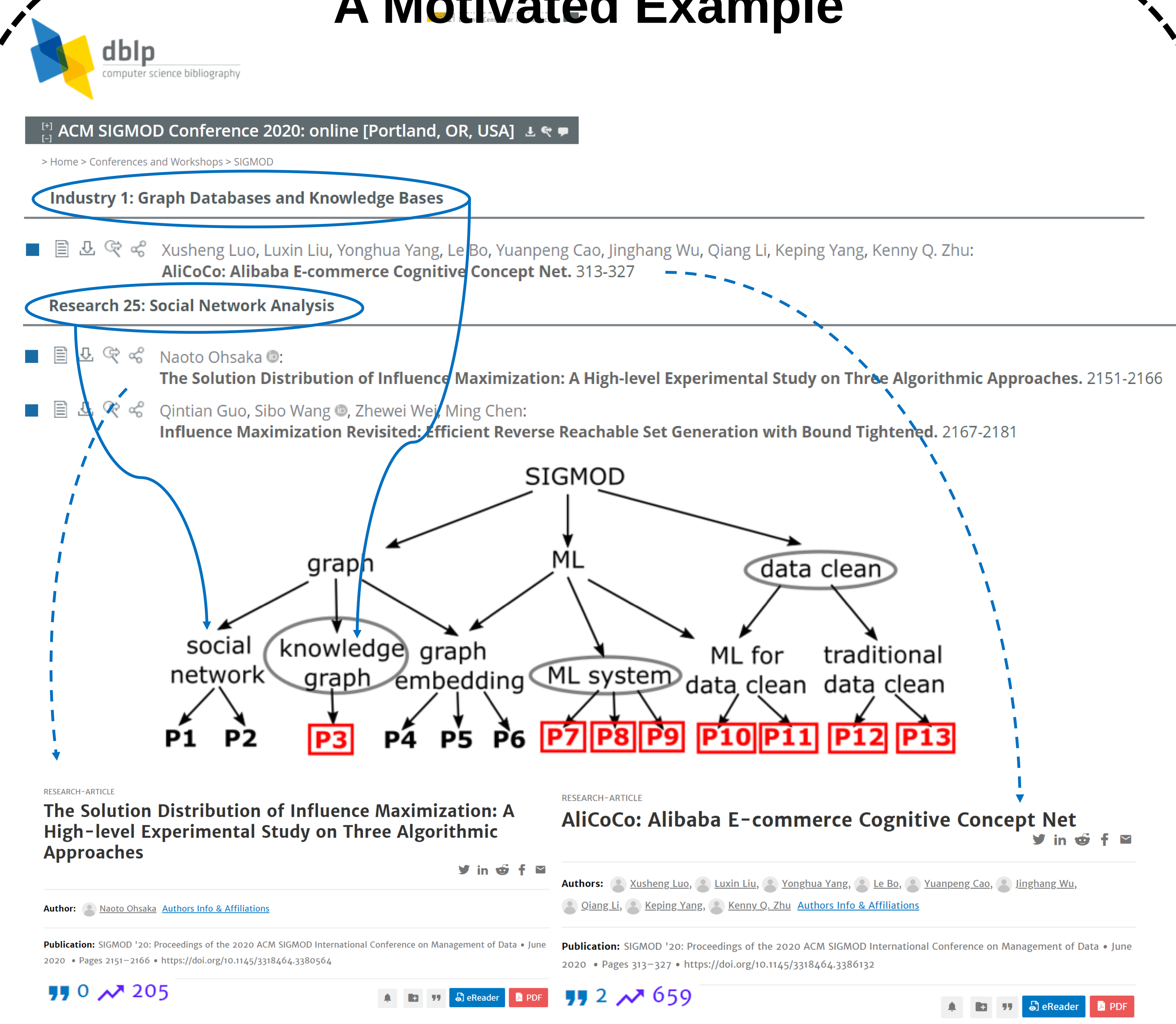
Assign vertex to the in-neighbor with maximal value.

Example: Assume α=1:



Then, apply tree algorithm.

A Motivated Example



P1: paper downloads less than 500

P3: paper downloads more than 500

Popular papers summarization in SIGMOD:

Some papers are popular and others are unpopular.
Select k topics to summarize popular papers.
Cover more popular papers, cover less unpopular papers.

An example selection (k=3):

data clean: P10, P11, P12, P13
ML system: P7, P8, P9
knowledge graph: P3

Cover all popular papers and cover no unpopular papers!

Related work

Aggregate Method [X Jing et al. 2014]

Method: Select top-k topics with maximum aggregate popular papers.
Selection (example): ML, data clean, ML system
Limitation: Lack diversity (ML & ML system)
Summary Score: $f(S) = \frac{|\{P7,P8,P9,P10,P11,P12,P13\}|}{|\{P3,P4,P5,P6,P7,P8,P9,P10,P11,P12,P13\}|} = \frac{7}{11}$

K-PCGS Method [X Zhu et al. CIKM 2020]

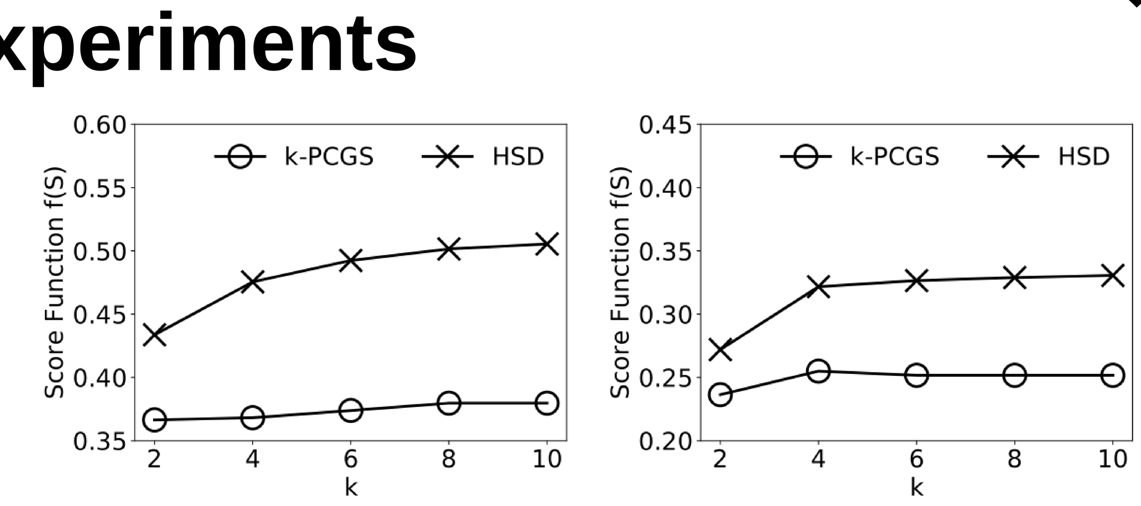
Method: Select k diverse topics with maximum summary score greedily.
Selection (example): ML, traditional data clean, knowledge graph
Limitation: Cover several unpopular papers (P4, P5, P6)
Summary Score: $f(S) = \frac{|\{P3,P7,P8,P9,P10,P11,P12,P13\}|}{|\{P3,P4,P5,P6,P7,P8,P9,P10,P11,P12,P13\}|} = \frac{8}{11}$

Our Method

Selection (example): data clean, ML system, knowledge graph
Summary Score: $f(S) = \frac{|\{P3,P7,P8,P9,P10,P11,P12,P13\}|}{|\{P3,P7,P8,P9,P10,P11,P12,P13\}|} = \frac{8}{8} = 1$

Experiments

Datasets: Disease Ontology
4,227 vertices, 8,190 edges
Evaluation metrics: f(S)
Method compare: k-PCGS



(a) Tree Dataset (b) HDAG Dataset