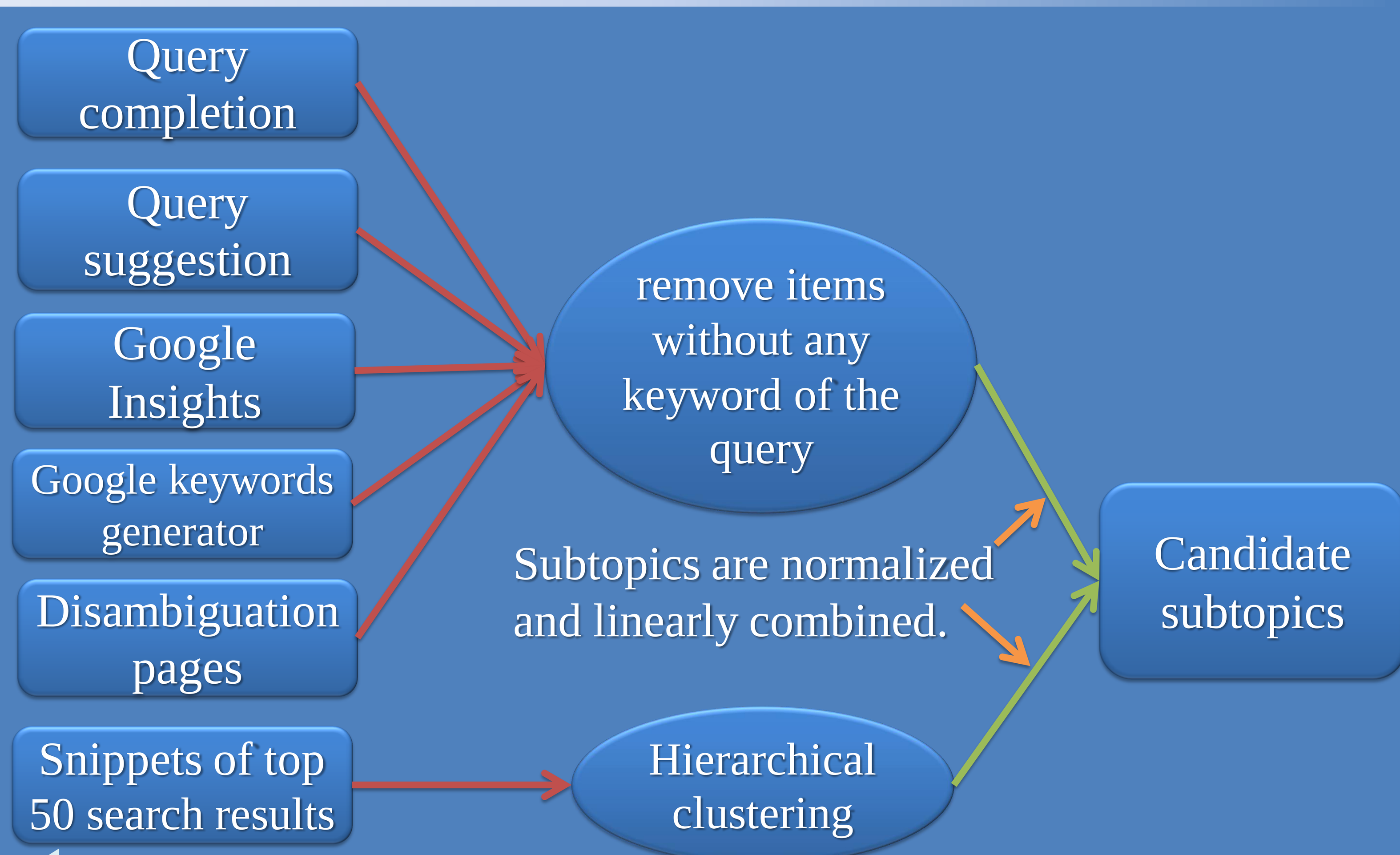
THUIR at NTCIR-10 INTENT-2 Task

Yufei Xue, *Fei Chen*, Aymeric Damien, Cheng Luo, Shuai Huo, Min Zhang, Yiqun Liu, Shaoping Ma
Department of Computer Science and Technology, Tsinghua University, Beijing, China
chenfei27@gmail.com

English Subtopic Mining

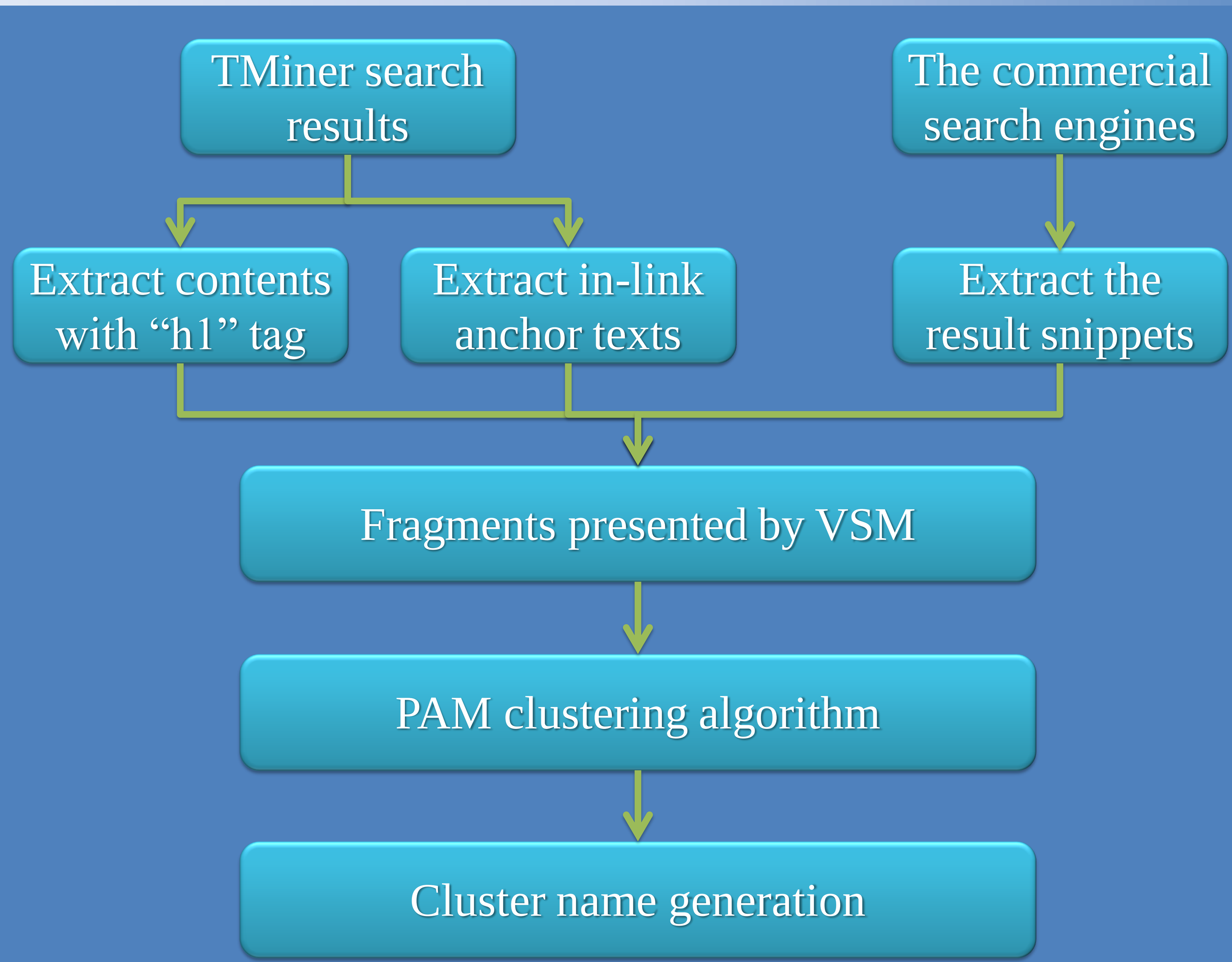
External Resource Based Subtopic Mining



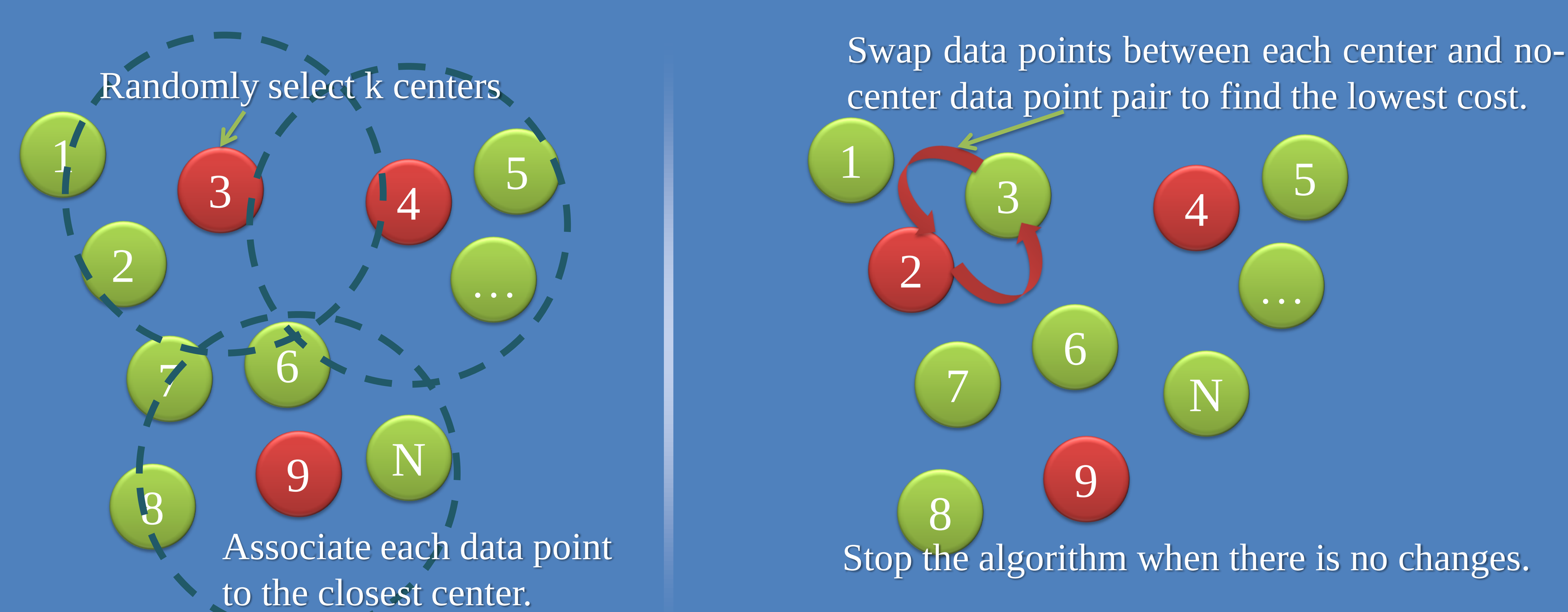
- The weight of different resources.

Resource	Weight	Resource	Weight
Jaccard similarity	0.05	Google Insights	0.15
Google Keywords generator	0.75	Query suggestion/completion	0.05

Top Results Based Subtopic Mining



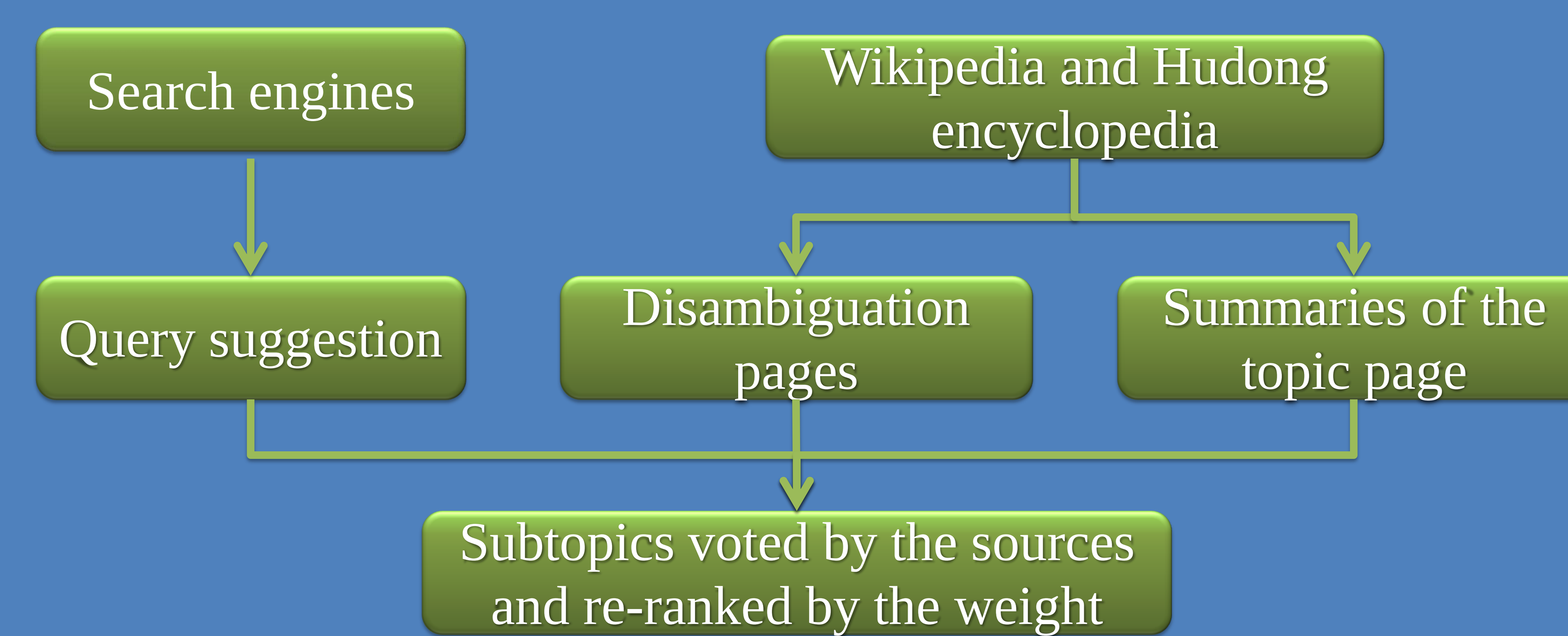
- PAM clustering algorithm



- Subtopics mined in these two ways are linearly combined.
- The duplicated subtopics are removed according to the WordNet-based semantic similarity.

Chinese Subtopic Mining

Extract Candidate Subtopics



$$weight_{new} = votes + 0.05 \times (coverage\ rate) + 0.005/(intent\ length)$$

- We combine the title and the snippet of the top 10 search results to form a snippet document and give every term in this document a score.

$$TermScore(t) = \sum_{i=0}^{10} (freq_{snippet}(t) + \lambda \times freq_{title}(t))$$

$$where\ \ freq_{snippet}(t) = \sum_{i=0}^{10} (freq_{snippet}(t) \times CT_i)\ \text{and}\ freq_{title}(t) = \sum_{i=0}^{10} (freq_{title}(t) \times CT_i)$$

- Only the top k terms in the snippet document are considered, and the term score are normalized.

$$NormScore(t_i) = 1.0 - (\alpha - \beta) \times \frac{i}{k} \Rightarrow SnippetScore(s,k) = \sum_{i=0}^{i<k} NormScore(t_i) \times I_{t_i\ in\ s}$$

$$Score(s,k) = \lambda \times OrigScore(s) + (1-\lambda) \times SnippetScore(s,k)$$

LDA on Snippet Click Document

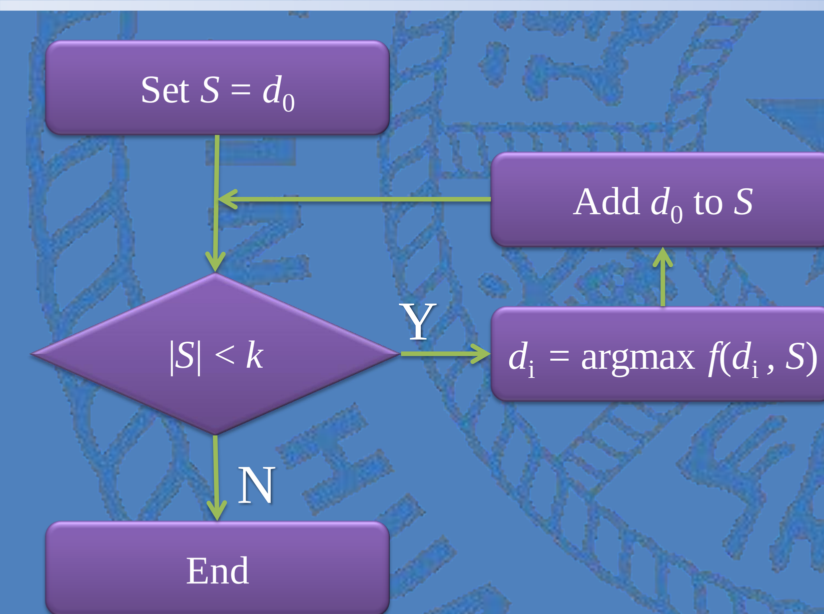
- Remove all the appearances of given query from *d*, and get a new document *d'*.
- Estimate the latent topics $t_1, t_2, \ldots, t_n$ of *d'*.
- Get two words with the largest probabilities within each topic, denoted by w_{k1} and w_{k2}.
- Connect up *q* to w_{k1} and w_{k2}, and get 4 different phrases.
- If any of the phrases has appeared in the snippet click document *d*, add the phrase into the intent candidate list with weight 0.4.

Document Ranking

Selective Diversification

- We only diversify the search result when a query is informational.
- To identify whether a query is informational or navigational, we leverage C4.5 algorithm to learn a decision tree.
- The features used in this algorithm are as follows:
  - nCS(*q*)=(Sessions of *q* that involves less than *n* clicks) / (session of *q*)
  - nRS(*q*)=(Session of *q* that involves clicks only on top *n* results) / (Session of *q*)
  - CD(*q*)=(Click on the most popular result of *q*) / (Click on all results of *q*)

Result Diversification Based on Novelty



$$f(d_i, S) = \sum_{d_j \in S} (N - r_j) + \lambda \times \cos(w_i, w_j)$$