The THUISAM system at NTCIR-11 MINE task was designed to address a specific challenge in information retrieval and disambiguation. The framework integrates several key components and algorithms to effectively process and rank documents.

### Framework

#### Subtopic Mining

We propose a 3-step framework in Subtopic Mining: Subtask, Candidate Mining, Candidate Ranking and Hierarchy Construction.

#### Candidate Mining From Various Resources

- **Similar Queries from Query Recommendation, Random Walk on Query-URL Bigraph and Query2vec.**
- **Query2vec:** Query → words
- **Sentence:** Each query can be represented as a vector.
- **Simulate Queries with cosine similarity.**

#### Candidate Ranking With LTR Algorithms

- **Goal:** Find the high-quality subtopic candidates.
- **Rank candidates using Learning-To-Rank algorithm.**
- **Training Set:** Rank subtopics from NTCIR Intent-2 data.
- **Feature:** Similarity between query and candidate.

**Text similarity:** Length difference, Jaccard similarity, Edit Distance.

**Search Result Similarity:** number of shared results...

**Metric to optimize:** NDCG@50

#### Experimental Results

<table>
<thead>
<tr>
<th>RUNNAME</th>
<th>SYSTEM DESC</th>
<th>DFL</th>
<th>FLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUSAM-C-1A</td>
<td>[Bottom-Up] Clustering candidates, find the highest-frequency n-gram which can match one of the candidates as FLs.</td>
<td>0.2773</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-2A</td>
<td>[Bottom-Up] Cluster candidates for each cluster. Learning to Rank the n-gram, find the best ones as FLs.</td>
<td>0.2204</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-3A</td>
<td>[KB-Aided] For queries which appear in Encyclopedia, use the disambiguation items (instead of FLs and classify other candidates.</td>
<td>0.1400</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-4A</td>
<td>[Top-Down] Learning to Rank FLs candidates, use heuristic greedy select from Subtopic N-gram to classify other candidates.</td>
<td>0.2224</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-5A</td>
<td>[Top-Down] Learning to FLs candidates, use hierarchy bottom-up clustering.</td>
<td>0.4237</td>
<td></td>
</tr>
<tr>
<td>THUSAM-E-2A</td>
<td>[Top-Down] Extraction from multiple resources (all) and bottom-up hierarchical clustering.</td>
<td>0.3179</td>
<td></td>
</tr>
</tbody>
</table>

### Document Ranking

#### Probabilistic model is leveraged for document ranking, which is based on BM25 and combined with our previous proposed word pair model.

#### Result re-ranking with HITS

Top m documents sorted by either Authority or Hub Value in the search result are placed up to the front. Its new rank is determined as follows:

\[
p_{\text{renew}} = p_{\text{old}} + \alpha \times (p_{\text{authority}} + p_{\text{hub}})
\]

#### Pruned Exhaustive Search

Previous studies have demonstrated that finding the optimal solution for diversified search is NP-hard. THEOREM: Given k>l+1, if there exists a document pair \(d_l\) and \(d_d\) satisfying:

\[
(\alpha_l - \alpha_d) - (|\alpha_l - \alpha_d|) > 0
\]

The document list containing \(d_l\) in its k-th slot and \(d_d\) in its k-slot cannot be optimal diversified search result.

Notion: \(\alpha_g\) denotes the score for doc in the k-th slot.

Pruned Exhaustive Search based on the THEOREM

#### Experimental Results

<table>
<thead>
<tr>
<th>RUNNAME</th>
<th>SYSTEM DESC</th>
<th>DFL</th>
<th>FLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUSAM-C-1A</td>
<td>Exhaustive search with window size 4. The SM result is from Subtopic N-gram Learning to rank list.</td>
<td>0.6985</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-1B</td>
<td>Exhaustive search with window size 5. The SM result is from Subtopic N-gram Learning to rank list.</td>
<td>0.6943</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-2A</td>
<td>Exhaustive search with window size 4. The SM result is from heuristic greedy select from subtopics.</td>
<td>0.5982</td>
<td></td>
</tr>
<tr>
<td>THUSAM-C-2B</td>
<td>Exhaustive search with window size 5. The SM result is from heuristic greedy select from subtopics.</td>
<td>0.3087</td>
<td></td>
</tr>
</tbody>
</table>

**NDCG@50**