Mining User Intent from Search Query Logs

Md. Zia Ullah, Masaki Aono, and Md. Hanif Seddiqui
Toyohashi University of Technology, Aichi, Japan
arif@kde.cs.tut.ac.jp, aono@tut.jp, hanif@cu.ac.bd

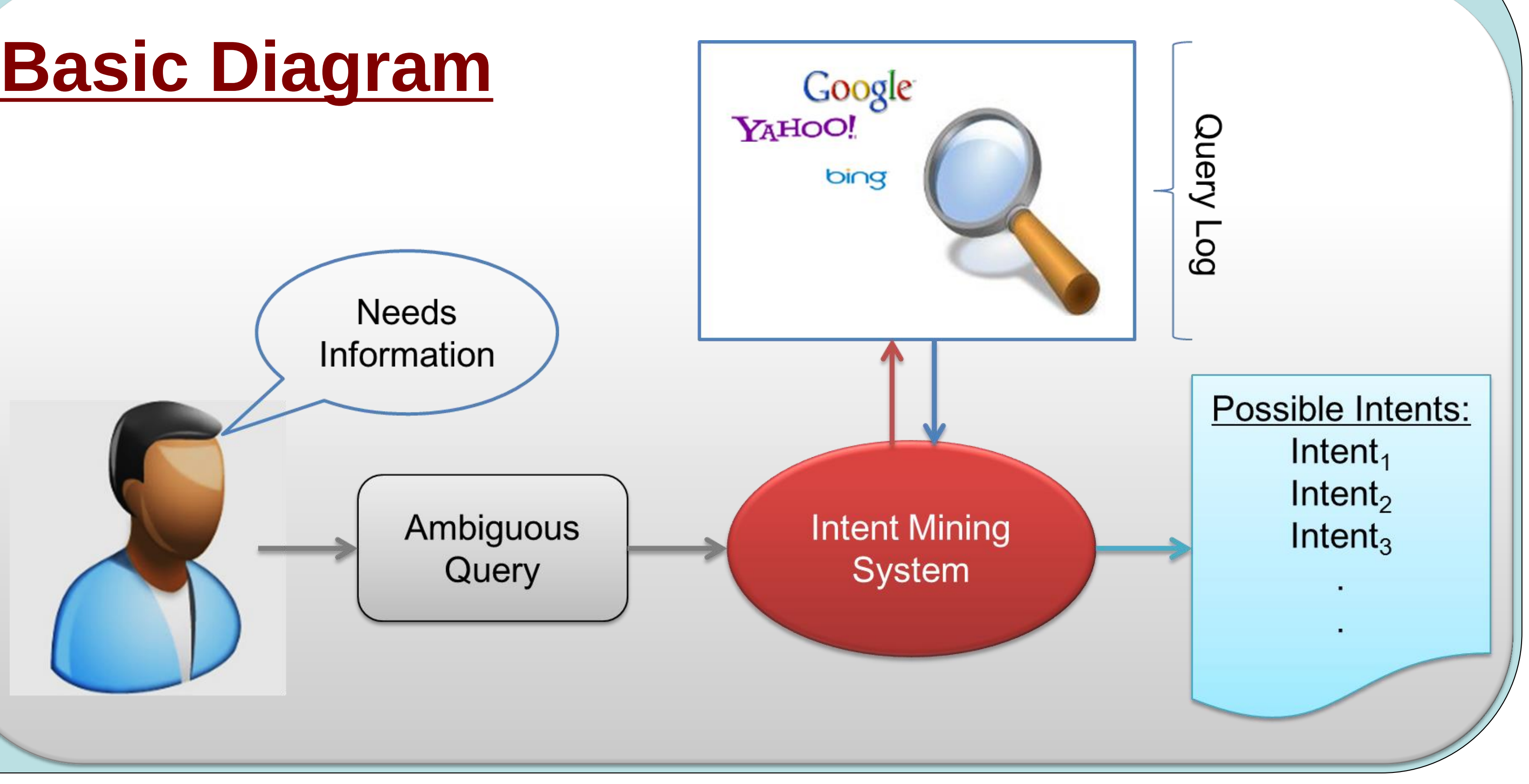
Introduction

Motivation:

- ➢ Queries are usually ambiguous and/or underspecified.
- ➢ Different users often have different intents for the same query.

To learn user's search intent, subtopic mining plays an role in information retrieval problem.

Basic Diagram



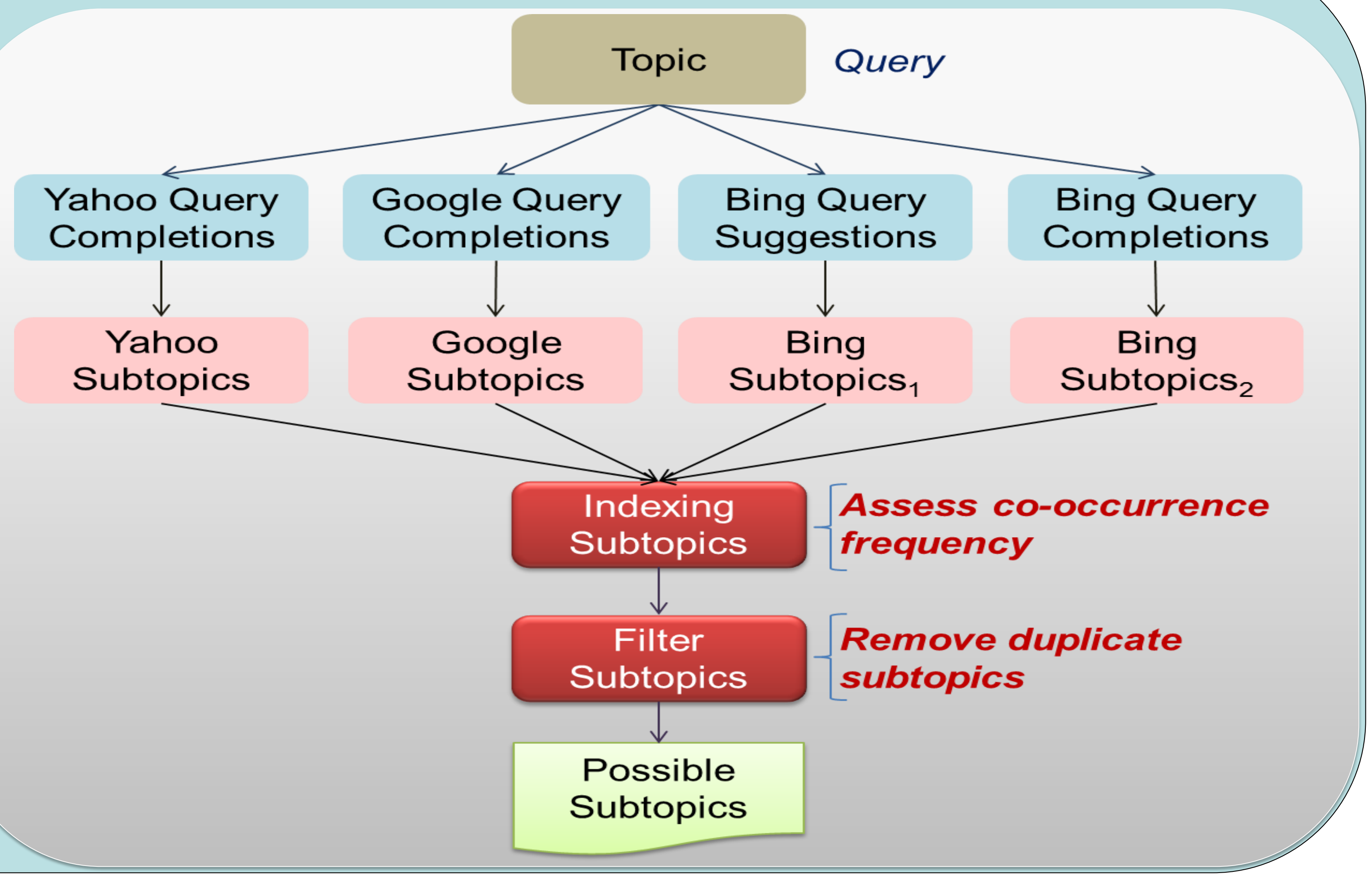
Pre-processing

Assumption:

- ➢ Subtopics are the specification or reformations of the original query.
- ➢ Some subtopics are more likely than others.

Mining Subtopics:

- ➢ Index subtopics from logs, using Lucene.
  - o Given a topic, search subtopics in across logs.
- ➢ **Estimating the co-occurrence frequency of subtopics.**
- ➢ Filtering subtopics using some rules
  - o **Removing duplicates that have similar sense.**



Main Processing

Subtopic Selection:

- ➢ Given a topic, select subtopics using rules
  - o The length of the subtopic, its Edit-Distance to the topic and some other features

Ranking:

- ➢ Estimate the rank of the subtopics
  - a. Choose the subtopics with **high frequency**,
  - b. If there is a tie, choose the subtopics with **nearest Edit-Distance** to the topic
  - c. And further, if there is also a tie, choose the subtopic with **lexicographically smaller** one.

Example:

Top 10 Subtopics for Topic "Sore Throat"
Sore Throat Infections
Sore Throat Remedies
Strep Throat Symptoms
Throat Cancer Symptoms
What Causes a Sore Throat
sore throat allergies
sore throat and cough
sore throat and ear ache
sore throat and fever

Evaluation

Primary Evaluation Metric:

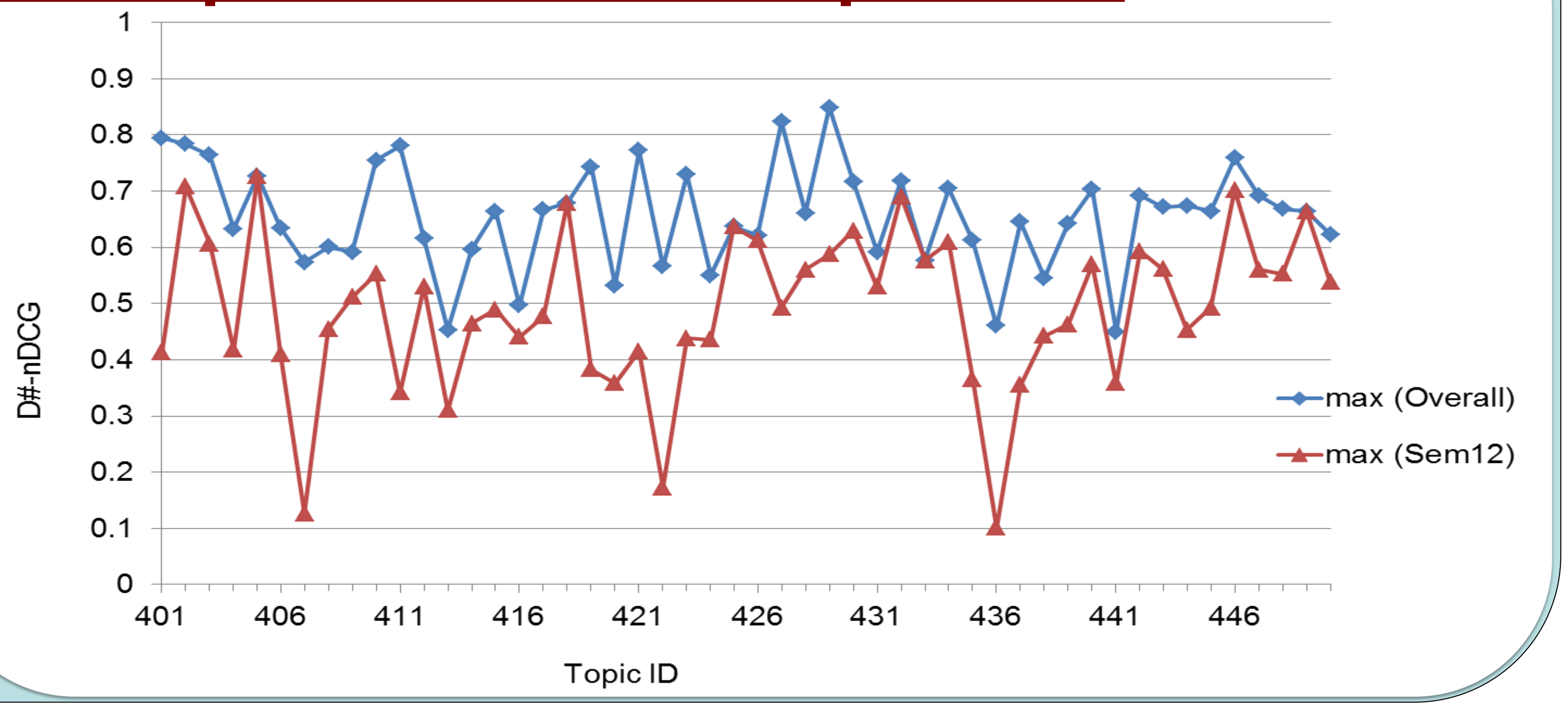
D#-nDCG:

$$D\# - nDCG@k = \gamma I - rec@k + (1-\gamma) D - nDCG@k$$

Results:

Runs	I-Rec@10	D-nDCG@10	D#-nDCG@10
SEM12-S-E-1A	**0.3780**	0.4233	0.4007
SEM12-S-E-2A	0.3777	**0.4250**	**0.4014**
SEM12-S-E-3A	0.3403	0.3573	0.3488
SEM12-S-E-4A	0.3727	0.3471	0.3599
SEM12-S-E-5A	0.3659	0.3445	0.3552

Per-topic D#-nDCG Comparison:



Conclusion

- ➢ We demonstrated that **co-occurrence** and **Edit-Distance** features achieve better result for few topics.
- ➢ Query logs are utilized only, moreover, other resources i.e. Wikipedia or **Search engine hits** might have useful features.
- ➢ Our system has lack of benefits from subtopic clustering that we leave as future work.

Discussion

Result: Needs Improvement

- ➢ utilizing Wikipedia for disambiguating some subtopics, anchor text for aggregating more subtopics
- ➢ adopting semantic similarity measures
- ➢ clustering subtopics to filter duplicating intents or extract more useful intents