QIC: Incorporating Context into a User Query

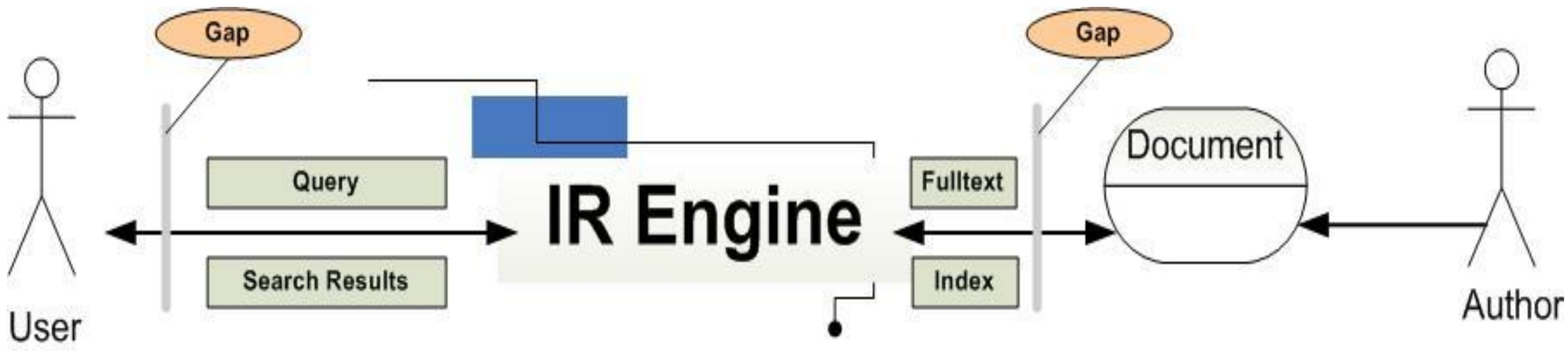
Min Song and Lori Watrous-deVersterre
Information Systems
College of Computing Sciences
New Jersey Institute of Technology

Outline

- Research Problems
- Research Goals
- QIC Overview
  - QEQIC
  - Concept Extraction
  - Learning to Rank and Dynamic Clustering
- Evaluation
- Conclusions and Future Work

Research Problems

People want search results to reflect exactly what they **meant**, **all** that they meant, and **only** what they meant, and they want it **quickly**.



- There are gaps….
  - Gap between what the user wants (information need) and the query that the user formulates
  - Gap between what the document represents and indexes that the IR engine built

Research Goals

- The purpose of the project is three-fold:
  1. **Incorporating inference of user preferences in Query Expansion**

     Our approach: QEQIC

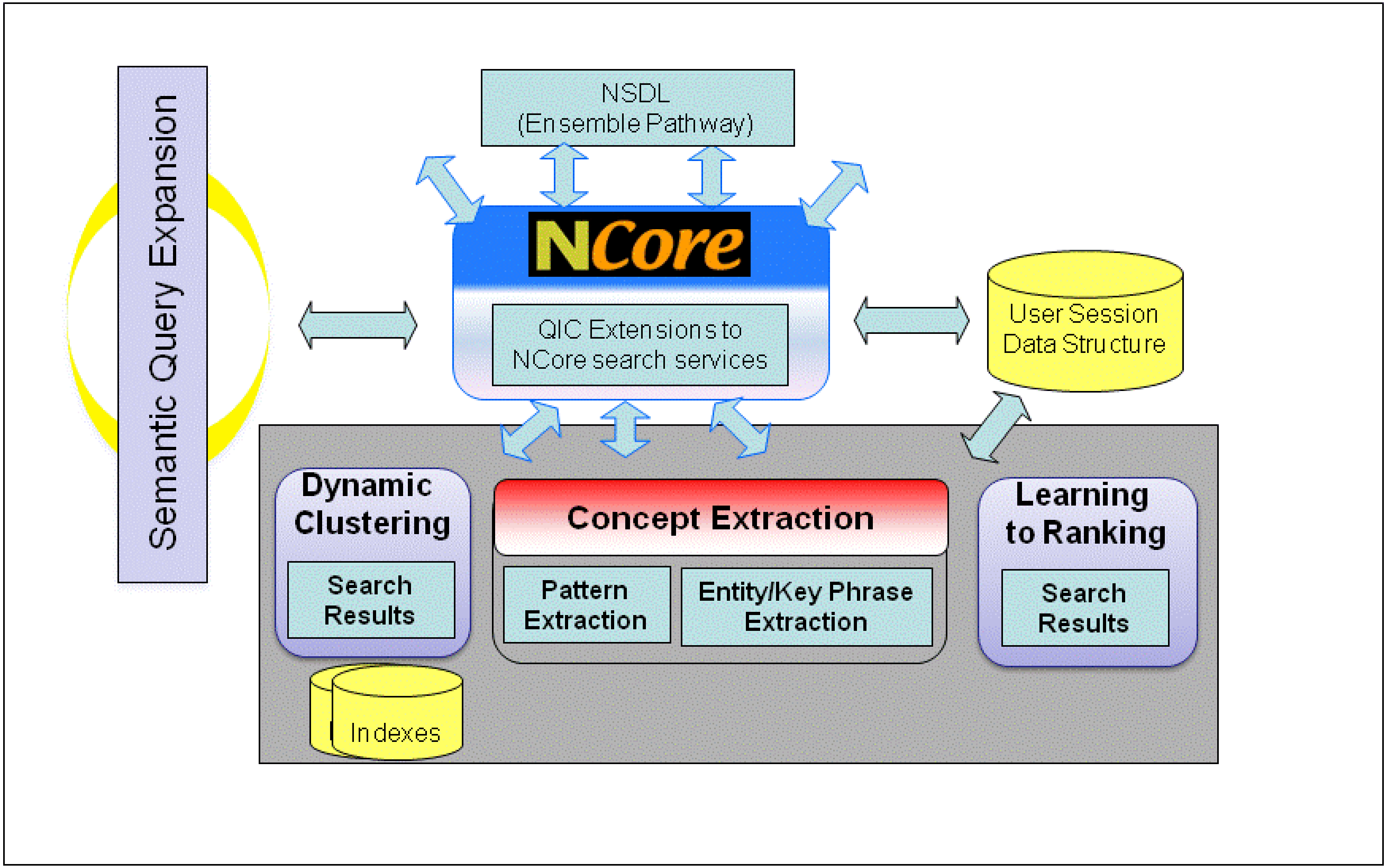
  2. **Capturing meanings** embedded in documents

     Our approach: Concept Extraction

  3. **Ranking search results with context-enriched features**

Our approach: Learning to Rank and Dynamic Clustering

QIC System Architecture



Data Collection

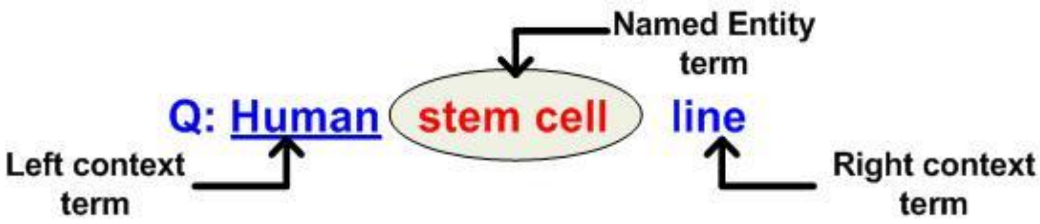
- Two different types of data:
  1. Ohsumed – biomedical data collection for proof-of-concept at the initial development phase.
  2. Ensemble Pathway – computing sciences data collections at http://www.computingportal.org/collections

jOAI, an OAI harvesting tool built in Ncore was used to crawl Ensemble.

Note: The size of data is small. This may influence the overall performance of our approach.

Query Expansion: QEQIC

- Query is initially represented as a tuple of {*context, named entities*}.



  - Named entities detected using Boosted Dictionary-based Entity Spotter (BDES).
  - A concept tuple consists of {*Computing concept, description, class*}.
    - *Computing concepts* provided by "The Free On-line Dictionary of Computing" (http://foldoc.org/).
    - *Class* assigned to a concept manually based on ACM Classification

Boosted Dictionary-based Entity Spotter

- Dictionary-based approach: tackles the problem of lack of contextual cues but:
  - too many false recognitions
  - takes too long to look up the dictionary entry.
- Our approach resolves these issues by:
  - Approximate String Distance Algorithm to retrieve candidate entries
  - Shortest-path Distance Algorithm
  - Part-Of-Speech (POS) tag
  - Syntactical properties of terms

Concept Tuple Example

- **Sentence:**

  Demonstrate the algorithm for simultaneously finding the minimum and maximum values in an array

  - **Concept:**

    Algorithm

  - **Class:**

    Theory of Computation

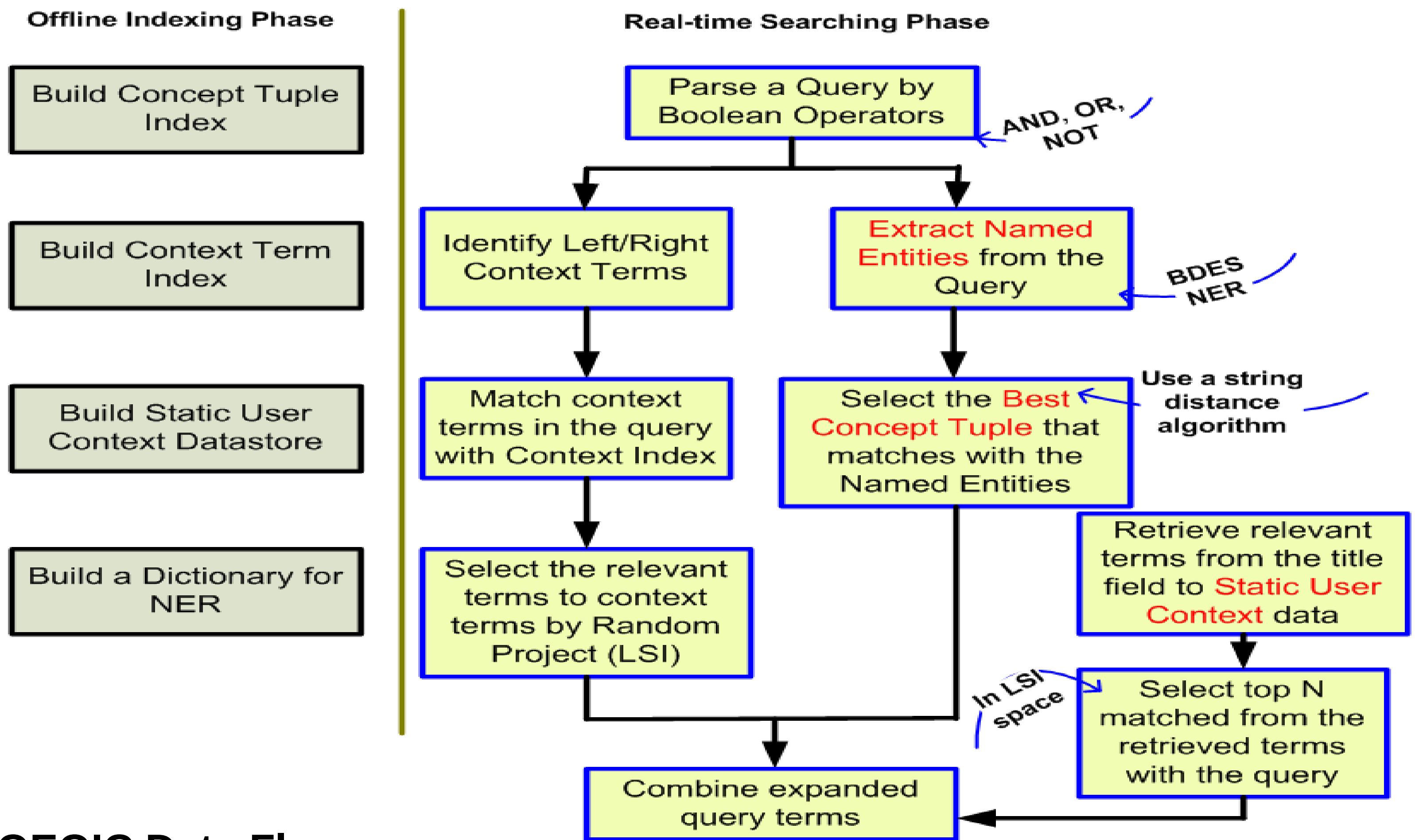
  - **Description:**

    Model of computation and algorithm

QEQIC: static user profile data

- Incorporate Static User Preference data into query expansion
  - Subject terms stored in user profile are matched with titles of data set in the Latent Semantic Index (LSI) space.
  - *N* top terms relevant to subject terms in the user profile are compared with a query.
    - If there is a good match (based on string similarity between a top term and query terms), the terms are weighted higher.

QEQIC: Semantic Query Expansion Algorithm



QEQIC Data Flow

Preliminary Results of BDES

BDES

	Genia	Geinia+MeSH	Genia+MeSH+ UMLS
Precision	0.93	0.878	0.88
Recall	0.573	0.72	0.68

BDES without POS/syntactic properties

	Genia	Geinia+MeSH	Genia+MeSH+ UMLS
Precision	0.76	0.56	0.51
Recall	0.62	0.82	0.78

Concept Extraction

- Probabilistic Combinatorial Markov Random Fields (PCMRF):
  - A supervised learning technique
  - PCMRF is a non-generative graph model.
- Training data:
  - 5000 sentences from Ensemble Pathway and other computing sciences related digital libraries.
    - These 5000 sentences are positive examples (meaning containing concepts in the sentence).
    - Combined with 5000 more sentences (negative examples), we build a concept extraction model.

Concept Extraction

- RESTful Web Services for Concept Extraction

http://localhost:8080/qic/Tagger?tag="Testing internet tagging service"


```
<tagging>
        <token>
                <name>Testing</name>
                <class>O</class>
        </token>
        <token>
                <name>internet</name>
                <class>Web</class>
        </token>
        <token>
                <name>tagging</name>
                <class>O</class>
        </token>
        <token>
                <name>service</name>
                <class>O</class>
        </token>
</tagging>
```

Dynamic Clustering of Search Results

- Clustering approach:
  - Based on a supervised learning technique - Probabilistic Combinatorial Markov Random Fields (PCMRF) technique
    - Same as our concept extraction technique
  - Requires a small set of initial training examples.
  - For performance reasons, input for clustering is a set of concepts extracted from Ensemble and stored in a database.

Rank search results with context features

- **Learning to rank** – apply supervised learning techniques to rank search results.
- Proposed technique: Mixture Support Vector Machines
  - Combines multiple models
    - Models are built with a set of features (attributes) such as TF-IDF, no. of clicks, the user's research interest, etc.
    - There are several different ways to select features:
      1) Document-driven model [11,15],
      2) Meta data-driven model [14],
      3) User static context-driven model, and
      4) User search behavior-driven model [13]

Note: The current model is based on document-related features.

Document-driven Model

- The most popular approach in learning to rank.
  - Training data is part of the LETOR package [11]
    http://research.microsoft.com/en-us/um/beijing/projects/letor/default.aspx
  - 25 features were extracted
    - 10 from title, 10 from abstract, and 5 from 'title + abstract'
    - TF, TF*IDF, BM25, Language Model ranking scores, IDF, etc

```
2 qid:1 1:3.00000000 2:2.07944154 3:0.27272727 …
25:-3.87512000 #docid = 40626
```

  - For query id "1" and document id "40626", the label is "2" (definitely relevant).

Search Behavior-driven Model

- Incorporate users' search behavior into ranking the results [13].
  - 1 - category; 2 - qid; 3 - search; 4 - abstract_text; 5 - full_text; 6 - no_visits; 7 - no_returned_citation; 8 - pos_clicked_citation; 9 - search_duration

```
0 qid:1 1:0.0 2:1.0 3:0.0 4:1 5:0 6:0 7:0.016944444
0 qid:1 1:0.0 2:2.0 3:0.0 4:2 5:0 6:1 7:4.2805557
1 qid:1 1:1.0 2:4.0 3:0.0 4:5 5:11583 6:10 7:106.29944
2 qid:1 1:8.0 2:8.0 3:11.0 4:27 5:2194 6:33 7:314.75027
0 qid:1 1:0.0 2:1.0 3:0.0 4:1 5:0 6:0 7:17.611666
0 qid:1 1:0.0 2:1.0 3:0.0 4:1 5:0 6:0 7:13.123055
1 qid:1 1:2.0 2:0.0 3:0.0 4:2 5:0 6:0 7:20.195278
```

Sample training data based on search behavior

Evaluation

- Preliminary Results with Ohsumed Data for Query Expansion
  - A set of 348,566 references from MEDLINE consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991).
  - Popular data set to apply supervised learning techniques to IR
  - Contains the 106 queries in test set, with patient and topic information, in the format:
    - .I Sequential identifier
    - .B Patient information
    - .W Information request
  - For the preliminary test, we used 12 out of 106 queries.

Evaluation

- The Approach
  - Use Recall and Interpolated Average Precision to measure the performance.
  - Investigate whether QEQIC performs better than the baseline Language Model technique.
  - Investigate whether adding concepts, semantic types, and context terms to QE improves the performance.

Preliminary Results

	QEQIC (title only)	baseline LM (title only)	QEQIC (title+abstract)	baseline LM (title+abstract)
Inter. Avg. Precision	0.135	0.108	0.172	0.139
Avg. Recall	0.359	0.256	0.407	0.323

Measure by Recall and Interpolated Average Precision

Preliminary Results

- Impact with different feature sets

	QEQIC+CON	QEQIC+CON+SEM	QEQIC+CON+SEM+ CXT
Avg. Precision	0.137	0.107	0.107
Avg. Recall	0.359	0.321	0.321

CON: Concept
SEM: Semantic Type
CXT: Context Term

Measure by Recall and Interpolated Average Precision

Conclusions and Future Work

- Conclusions
  - We developed a semantic query expansion technique, and tested it on a biomedical data collections.
  - We developed a new ranking technique for the search results with the “Learning to Rank” approach.
  - We developed a concept extraction technique and a dynamic clustering technique with Probabilistic Combinatorial Markov Random Fields.
  - We developed RESTful APIs for our techniques.
- Future Work
  - We plan to conduct a pilot study and the main experiment on Ensemble Pathway data

Acknowledgement

National Science Foundation
WHERE DISCOVERIES BEGIN

NEW JERSEY INSTITUTE OF TECHNOLOGY
Founded 1881

informatics
research lab

References

1. He, X. F., and Jhala, P. Regularized query classification using search click information. *Pattern Recognition* 41, 7 (2008), 2283–2288
2. Jansen, B. J., Booth, D. L., and Spink, A. Patterns of query modification during web searching. *Journal of the American Society for Information Science and Technology* 60, 3 (2009), 557–570.
3. Kapoor, A., Burleson, W., and Picard, R. Automatic prediction of frustration. International Journal of Human-Computer Studies, 65(8):724{736, 2007.
4. Labelle, P. R. 2007. Initiating the learning process: A model for federated searching and information literacy. *Internet Reference Services* 12(3/4), 237–52. Haworth Press database (accessed Feb. 14, 2008).
5. Pol, R. V. D. Dipe-d: a tool for knowledge-based query formulation. Information Retrieval 6 (2003), 21–47.

References

6. Park, S. Usability, User Preferences, Effectiveness, and User Behaviors When Searching Individual and Integrated Full-Text Databases: Implications for Digital Libraries, Journal of American Society for Information Science, vol 50. No. 5, pp. 456-468, 2000.

7. Tenopir, C., B. Hitchcock, and A. Pillow. 2003. Use and users of electronic library resources: An overview and analysis of recent research studies (CLIR Rep. No. 120).

8. Zhang, Y., and Moffat, A. Some observations on user search behavior. In *Proceedings of the 11th Australasian Document Computing Symposium* (2006).

9. Fan, J.-W. and Friedman, C. (2007) Semantic Classification of Biomedical Concepts Using Distributional Similarity, *Journal of the American Medical Informatics Association*,14 (4) p.467-477.

10. Bodenreider, O. and McCray, A. T. (2003) Exploring semantic groups through visual approaches, *Journal of Biomedical Informatics* 36 (2003) p. 414–432

References

11. Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y., and Hon, H.-W. Adapting ranking SVM to document retrieval. Proceedings SIGIR 2006, pp.186–193, 2006.
12. Lu, Z and Wilbur, WJ: Improving accuracy for identifying related PubMed queries by an integrated approach, Journal of Biomedical Informatics, 2008.
13. Agichtein, E., Brill, E., Dumais, S. T. and Ragno, R. Learning user interaction models for predicting web search result preferences. In SIGIR 2006, pages 3-10, 2006.
14. Learning SVM Ranking Function from User Feedback Using Document Metadata and Active Learning in the Biomedical Domain. Proceedings of the ECML/PKDD-08 Workshop on Preference Learning
15. M.-R. Amini, T.-V. Truong, and C. Goutte. A boosting algorithm for learning bipartite ranking functions with partially labeled data. In SIGIR 2008, pages 99-106, 2008.

References

16. Papadimitriou, C. H., Raghavan, P., Tamaki, H., Vempala, S. Latent Semantic Indexing: A Probabilistic Analysis, Journal of Computer and System Sciences, Volume 61, Issue 2, October 2000, Pages 217-235..

Questions?

Thanks!