Instituto Superior Técnico
Universidade Técnica de Lisboa

Learning To Rank Academic Experts

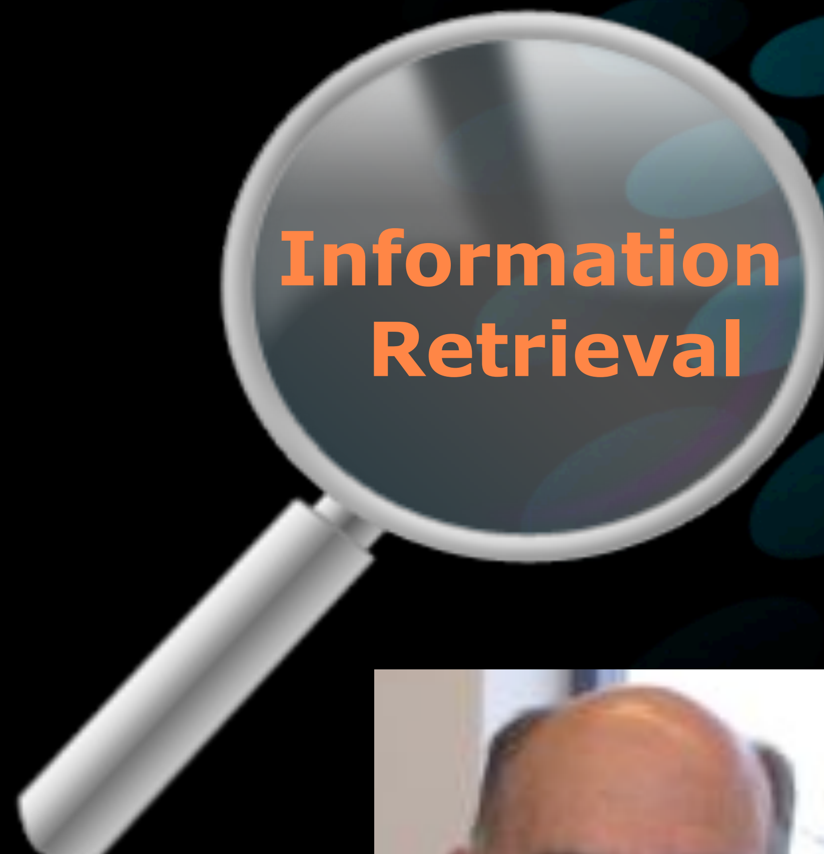
Catarina Moreira

Outline

- ✓ Introduction
- ✓ State of the Art Problems
- ✓ Features to Estimate Expertise
- ✓ Datasets
- ✓ Approaches and Results
  - ✓ Rank Aggregation Framework
  - ✓ Learning to Rank Framework
- ✓ Conclusions and Future Work

Expert Finding

Gerard Salton



Ricardo Baeza-Yates

Bruce Croft

State of the Art Problems

Usage of Generative Probabilistic Models

$$P(q|\theta_d) = \prod_{t \in q} (1 - \lambda_t) P(t|d) + \lambda_t P(t)$$

Heuristics are too simple and do not reflect expertise

Heuristics only based on the documents' textual contents

Contributions

1. Different Sets of Features to Estimate Expertise

2. Rank Aggregation Framework for Expert Finding

3. Learning to Rank (L2R) Framework for Expert Finding

Outline

- ✓ ~~Introduction~~
- ✓ ~~State of the Art Problems~~
- ✓ Features to Estimate Expertise
- ✓ Datasets
- ✓ Approaches and Results
  - ✓ Rank Aggregation Framework
  - ✓ Learning to Rank Framework
- ✓ Conclusions and Future Work

Features: Hypothesis

Multiple estimators of expertise, based on **different sources** of evidence, will enable the construction of more **accurate** and **reliable ranking models**!

Textual Similarity

Term Frequency

$$TF_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \frac{Freq(i, d_j)}{|d_j|}$$

Inverse Document Frequency

$$IDF_q = \sum_{i \in Terms(q)} \log \frac{|D|}{f_{i,D}}$$

TF.IDF

BM25

$$BM25_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \log \left(\frac{N - Freq(i) + 0.5}{Freq(i) + 0.5} \right) \times \frac{(k_1 + 1) \times \frac{Freq(i,d_j)}{|d_j|}}{\frac{Freq(i,d_j)}{|d_j|} + k_1 \times (1 - b + b \times \frac{|d_j|}{A})}$$

Profile Information

- ✓ Number of Publications with(out) query topics
- ✓ Number of Journals with(out) query topics
- ✓ Years Between Publications with(out) query topics
- ✓ Average Number of Publications per year

Graphs

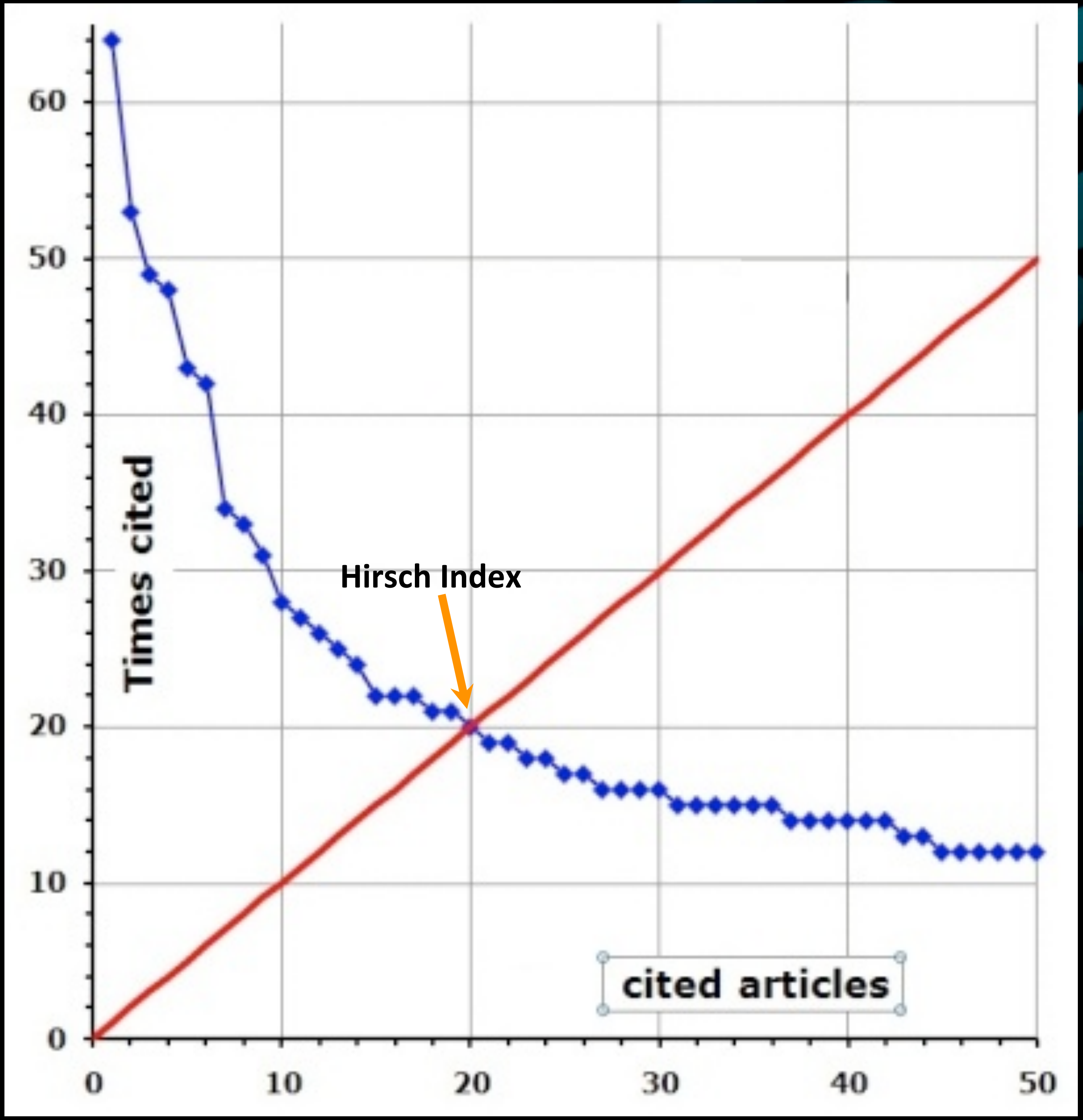
✓ Total/Max/Avg citations of the authors' papers

✓ Total Number of Unique Collaborators

✓ Publications' PageRank

✓ Academic Indexes

Hirsch Index



Other Indexes

a-Index

$$a = \text{Citations}/h^2$$

Contemporary h-Index (extension of h Index)

$$S^c(i) = \gamma * (Year(now) - Year(i) + 1)^{-\delta} * |CitationsTo(i)|$$

Trend h-Index (extension of h Index)

$$S^t(i) = \gamma * \sum_{\forall x \in C(i)|} (Year(now) - Year(x) + 1)^{-\delta}$$

Datasets

DBLP - Computer Science Dataset

- Covers journal and conference publications
- Contains abstracts and citation links
- All this information was processed and stored in a database

Property	Value
Total Authors	1 033 050
Total Publications	1 632 440
Total Publications containing Abstract	653 514
Total Papers Published in Conferences	606 953
Total Papers Published in Journals	436 065
Total Number of Citations Links	2 327 450

Datasets

Arnetminer - Validation

- Contains experts for 13 query topics
- Experts collected from important Program Committees related to the query topics

Query Topics	Rel. Authors	Query Topics	Rel. Authors
Boosting (B)	46	Natural Language (NL)	41
Computer Vision (CV)	176	Neural Networks (NN)	103
Cryptography (C)	148	Ontology (O)	47
Data Mining (DM)	318	Planning (P)	23
Information Extraction (IE)	20	Semantic Web (SW)	326
Intelligent Agents (IA)	30	Support Vector Machines (SVM)	85
Machine Learning (ML)	34		

Outline

- ✓ ~~Introduction~~
- ✓ ~~State of the Art Problems~~
- ✓ ~~Features to Estimate Expertise~~
- ✓ ~~DataSets~~
- ✓ Approaches and Results
  - ✓ Rank Aggregation Framework
  - ✓ Learning to Rank Framework
- ✓ Conclusions and Future Work

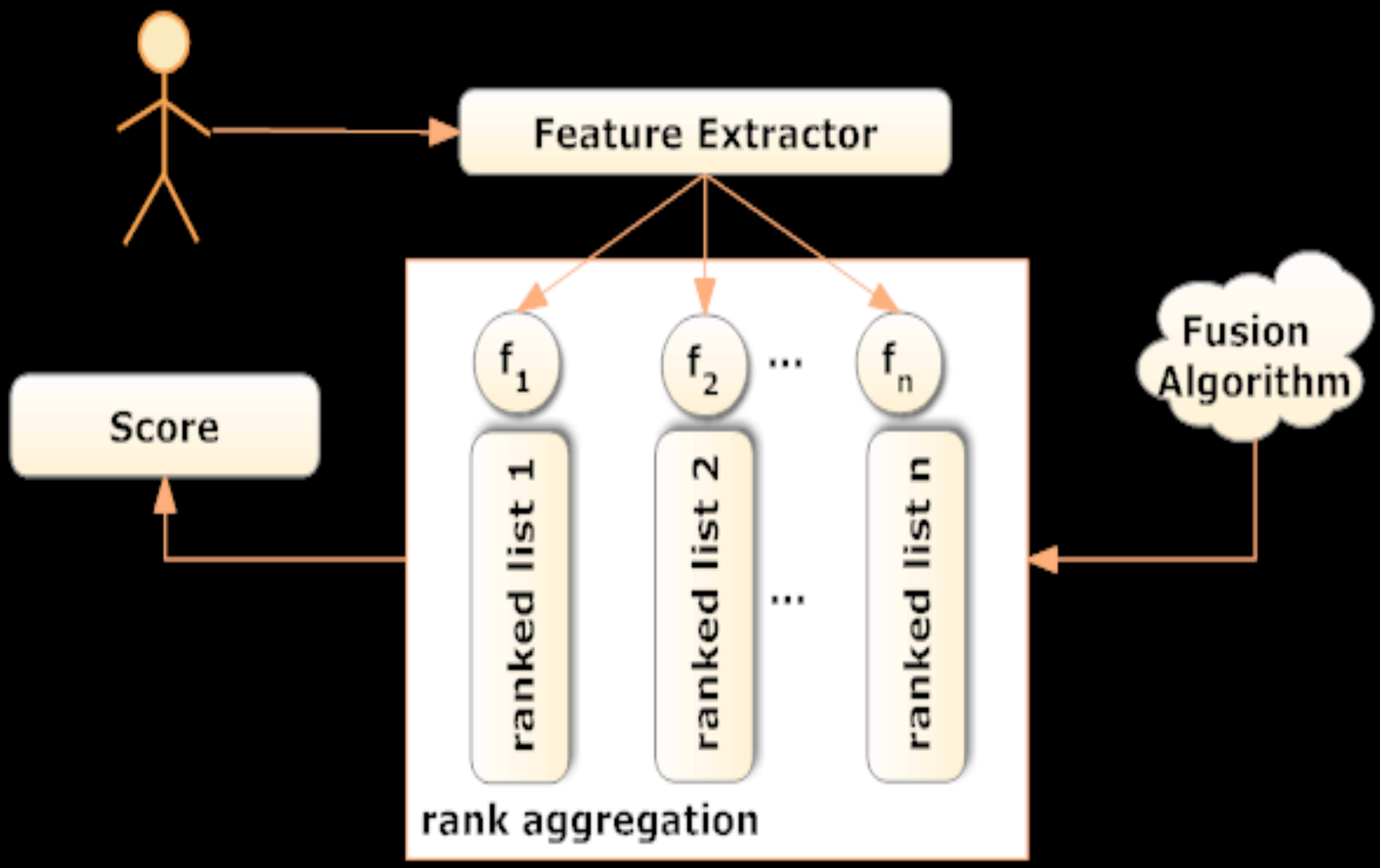
Question

How can we combine these features?

Answer

Traditional IR techniques use frameworks inspired in traditional search engines to combine different sources of evidence!

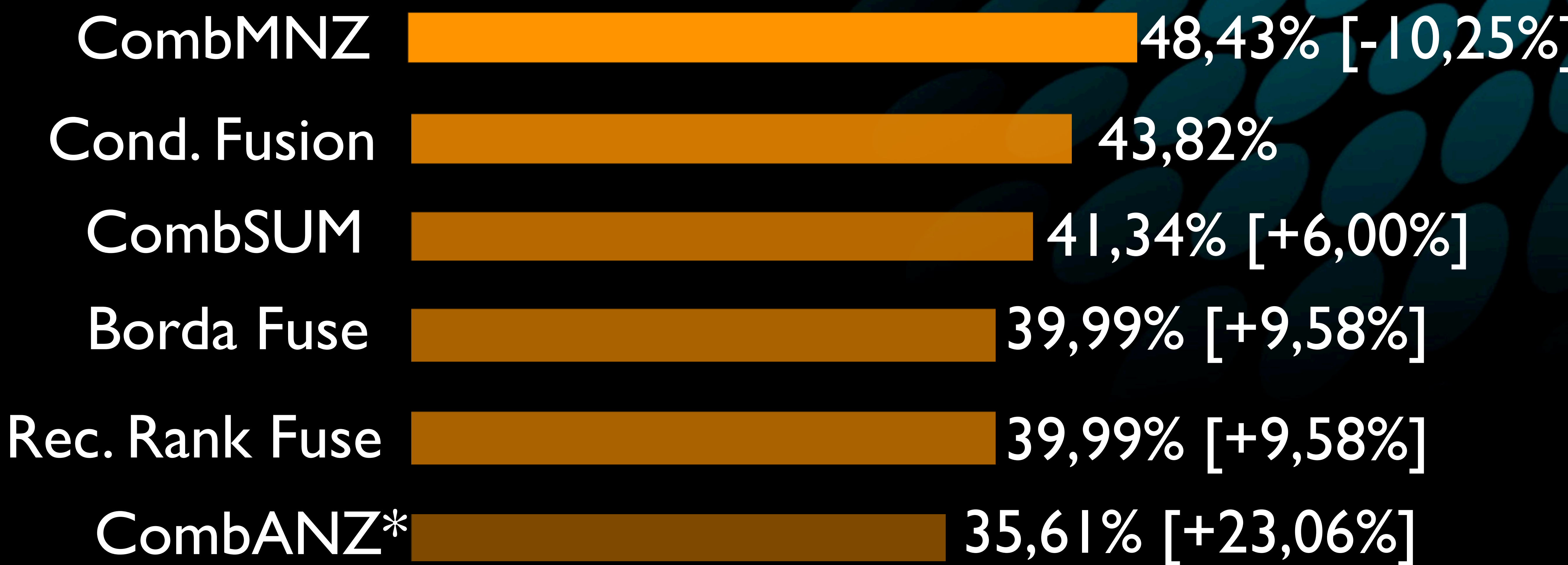
Rank Aggregation Framework for Expert Finding



Data Fusion Algorithms

- ✓ **Positional**
  - ✓ Based on the position that a candidate occupies in a ranked list
  - ✓ Algorithms: Borda Fuse and Reciprocal Rank Fuse
- ✓ **Score Aggregation**
  - ✓ Based on the score that a candidate achieved in a ranked list
  - ✓ Algorithms: CombSUM, CombMNZ and CombANZ
- ✓ **Majoritarian**
  - ✓ Based on pairwise comparisons between candidates
  - ✓ Algorithms: Condorcet Fusion

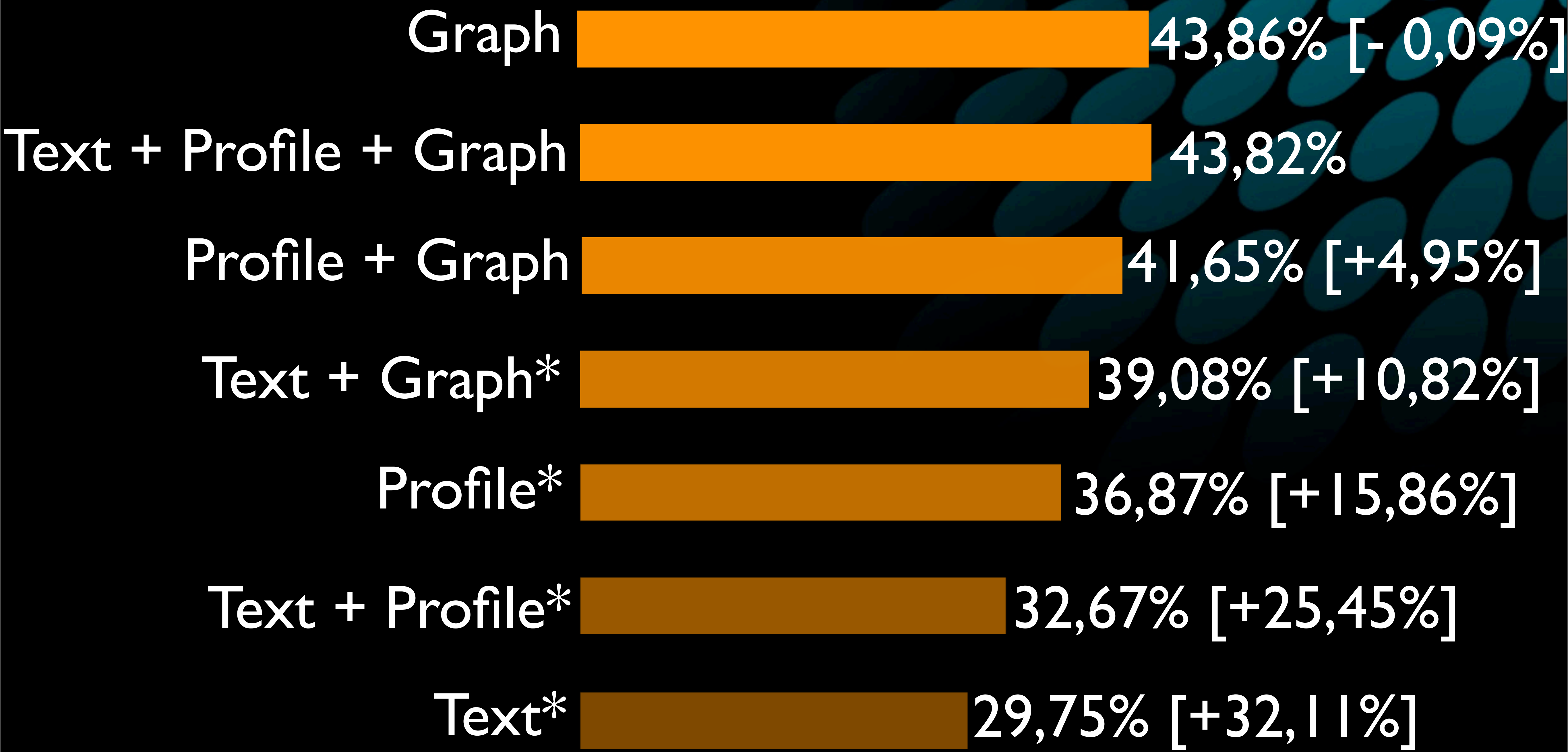
Results Rank Aggregation (MAP)*



*Sig. Tests of 0.95 conf.

*Mean Average Precision

Impact of the Features with Condorcet Fusion(MAP)*



*Sig. Tests of 0.95 conf.

*Mean Average Precision

Outline

- ✓ ~~Introduction~~
- ✓ ~~State of the Art Problems~~
- ✓ ~~Features to Estimate Expertise~~
- ✓ ~~Datasets~~
- ✓ Approaches and Results
  - ✓ ~~Rank Aggregation Framework~~
  - ✓ Learning to Rank Framework
- ✓ Conclusions and Future Work

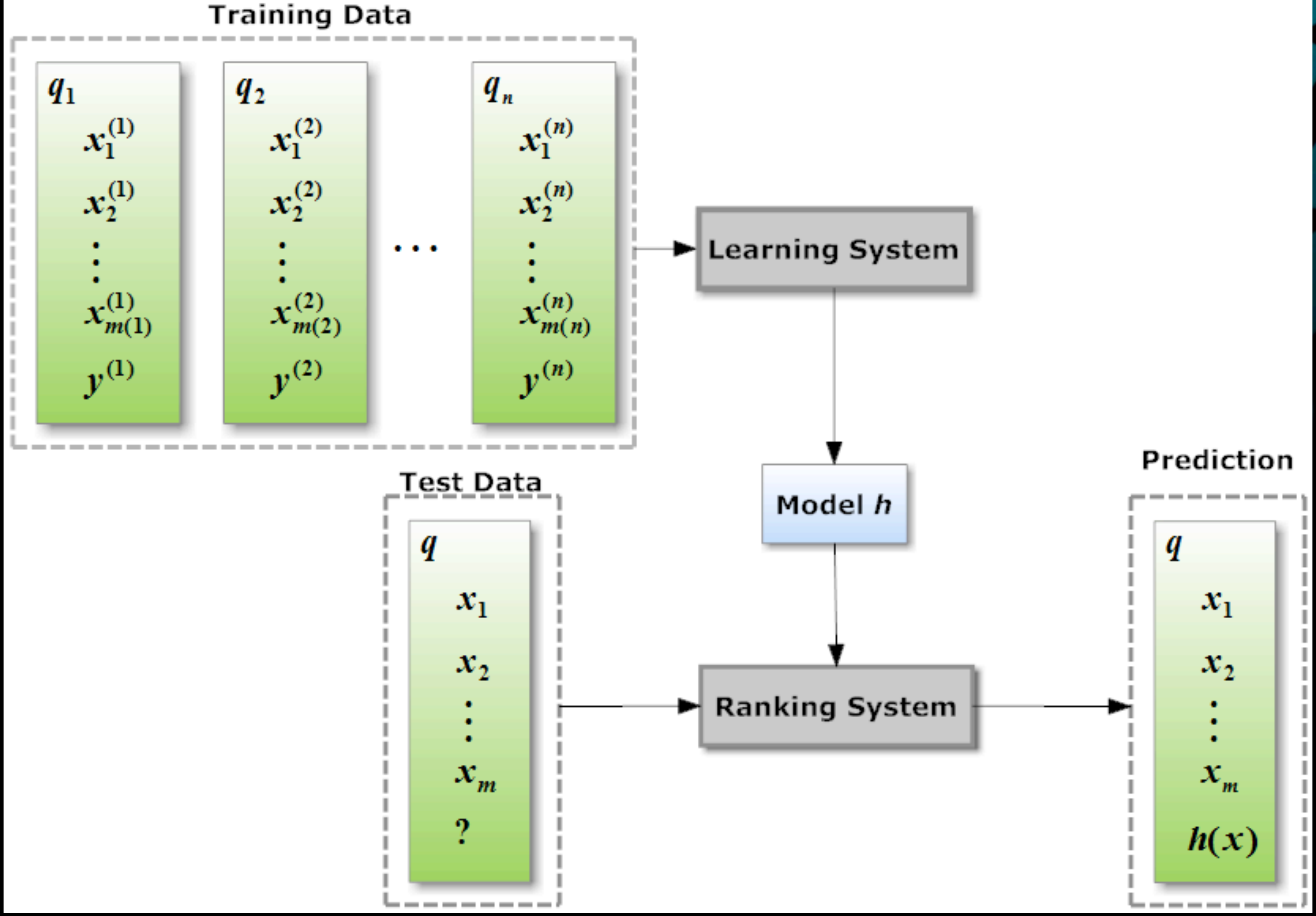
Question

How can we combine these features in an **optimal way**?

Answer

IR literature focuses on Machine learning techniques, They enable the combination of multiple estimators in an optimal way!

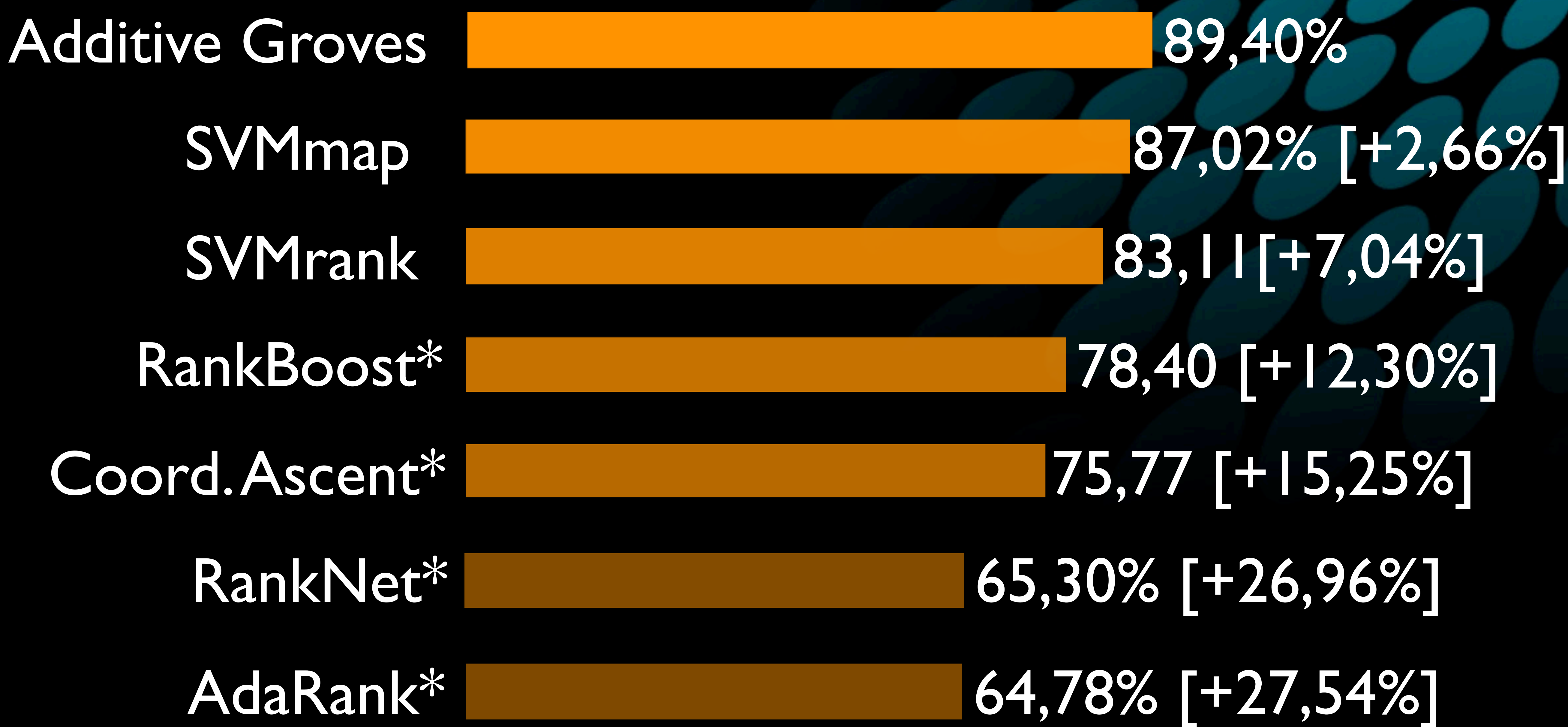
The L2R Framework For Expert Finding



L2R Algorithms

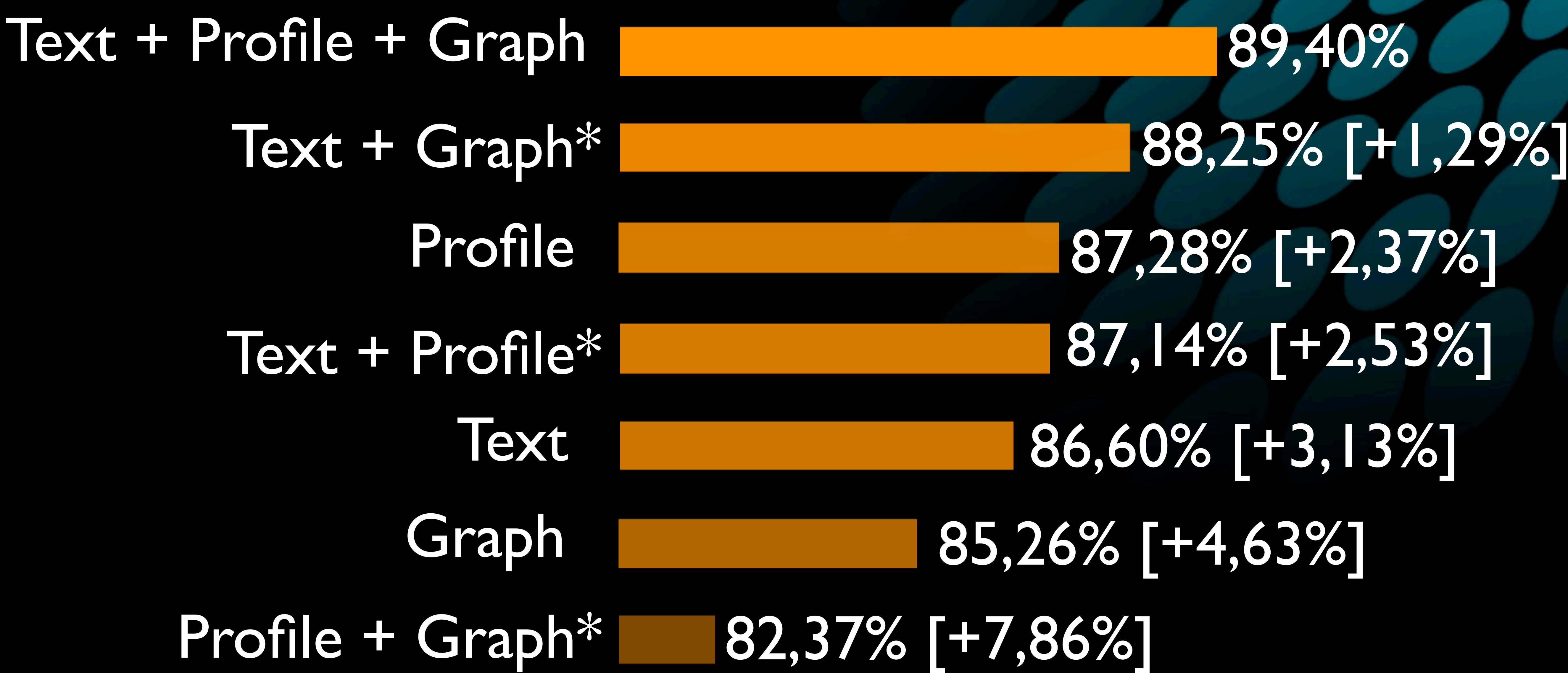
- ✓ **Pointwise**
  - ✓ **Input**: single candidate
  - ✓ **Goal**: use scoring functions to predict relevance
  - ✓ **Algorithms**: Additive Groves
- ✓ **Pairwise**
  - ✓ **Input**: **pair** of candidates
  - ✓ **Goal**: loss function to minimize number of misclassified candidate pairs
  - ✓ **Algorithms**: RankBoost, SVMrank and RankNet
- ✓ **Listwise**
  - ✓ **Input**: list of candidates
  - ✓ **Goal**: loss function which directly optimizes an IR metric
  - ✓ **Algorithm**: SVMmap, Coordinate Ascent and AdaRank

Results Learning to Rank (MAP)*



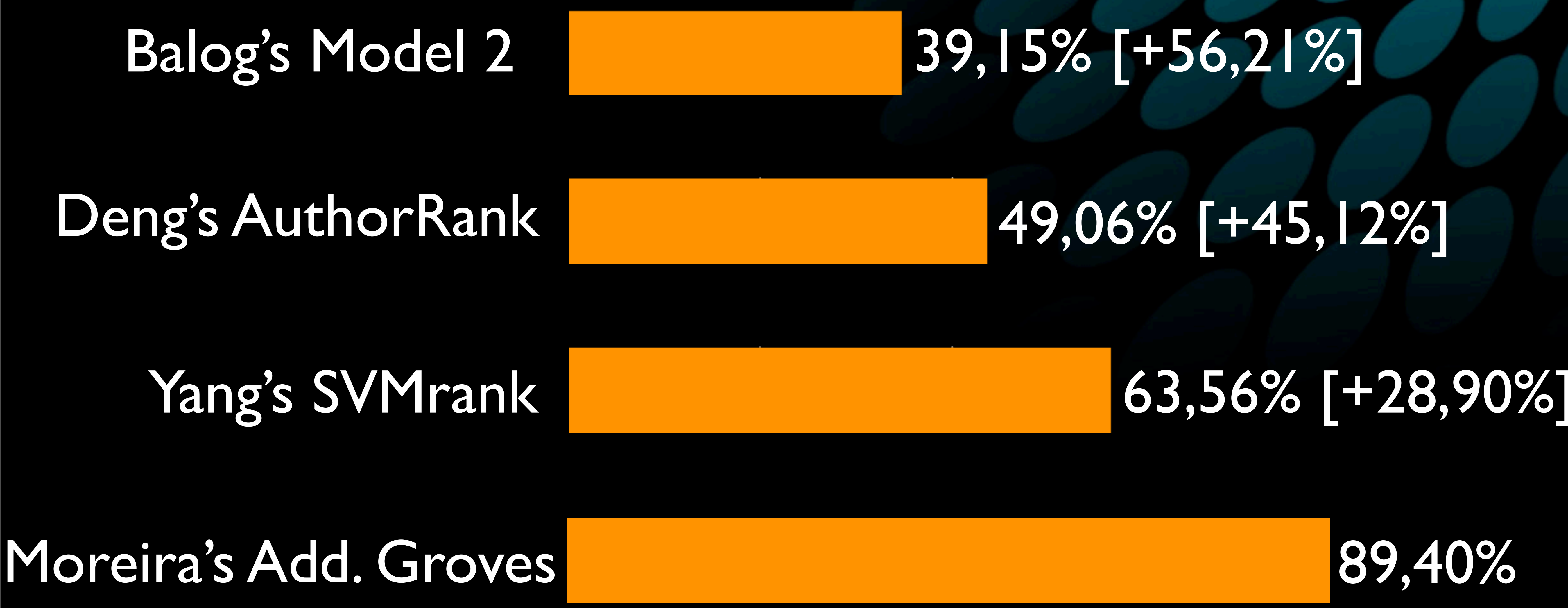
*Sig. Tests of 0.95 conf. *Mean Average Precision

Impact of the Features with Additive Groves(Map)*



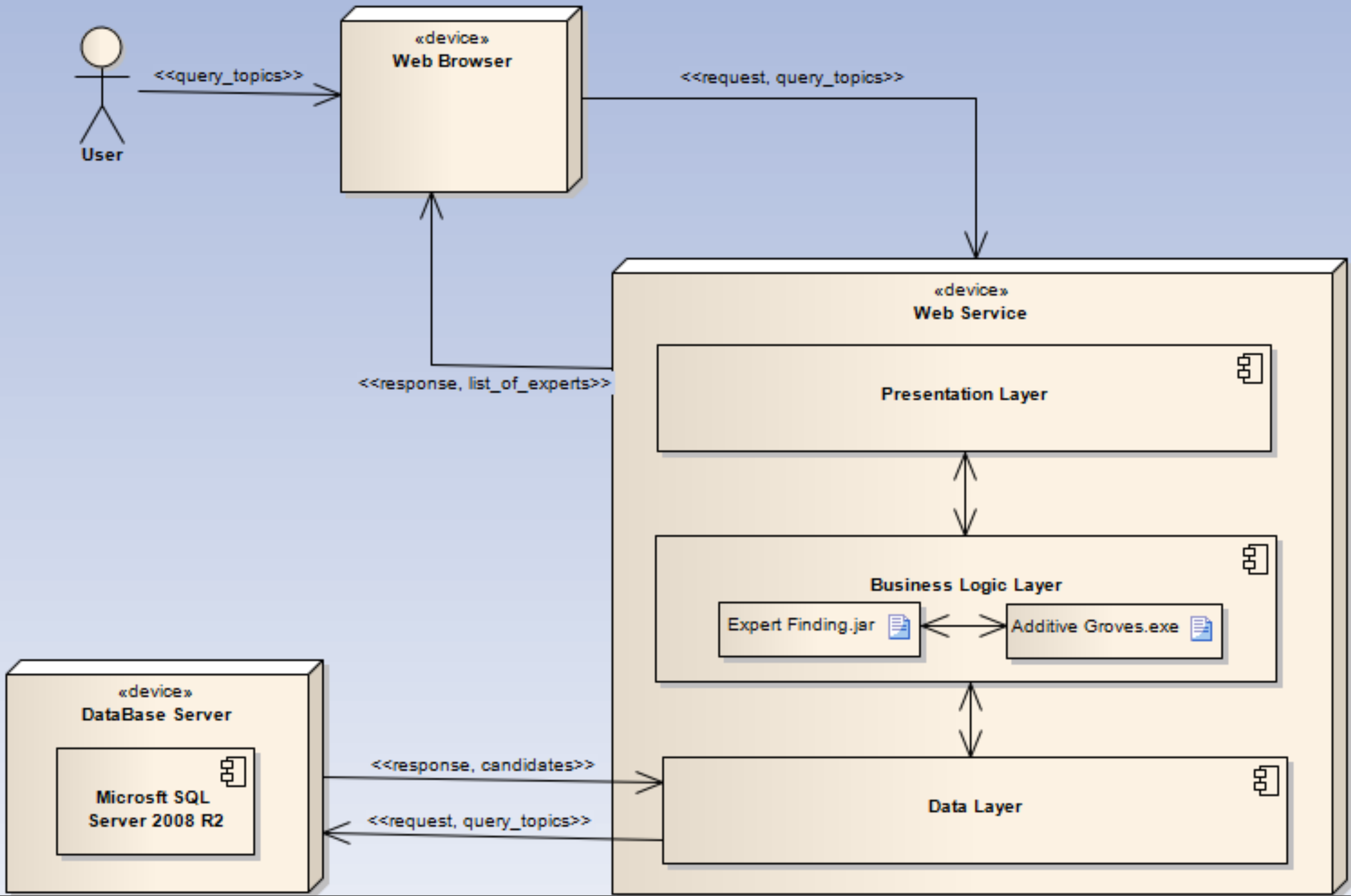
*Sig. Tests of 0.95 conf. *Mean Average Precision

Comparison with State of the Art (MAP)*



*Mean Average Precision

Prototype


User
<<query_topics>>
«device»
Web Browser
<<request, query_topics>>
<<response, list_of_experts>>
«device»
Web Service
Presentation Layer
Business Logic Layer
Expert Finding.jar
Additive Groves.exe
Data Layer
«device»
DataBase Server
Microsft SQL Server 2008 R2
<<response, candidates>>
<<request, query_topics>>

Outline

- ✓ ~~Introduction~~
- ✓ ~~State of the Art Problems~~
- ✓ ~~Features to Estimate Expertise~~
- ✓ ~~Datasets~~
- ✓ ~~Approaches and Results~~
  - ✓ ~~Rank Aggregation Framework~~
  - ✓ ~~Learning to Rank Framework~~
- ✓ Conclusions and Future Work

Conclusions

- ✓ **Effectiveness of the Learning to Rank Framework**
  - ✓ Best algorithms: Additive Groves, SVMmap and SVMrank
- ✓ **Effectiveness of the Rank Aggregation Approach**
  - ✓ Best algorithms: CombMNZ and Condorcet Fusion
- ✓ **Effectiveness of the Proposed Features**
  - ✓ Set of full features are the best

Future Work

- ✓ Feature Selection Techniques (ex: PCA)
- ✓ Expert Finding in an organizational environment (TREC dataset)
- ✓ Tasks beyond expert finding
  - ✓ Natural Language Processing
  - ✓ Geographic Information Retrieval

Publications

- ✓ **C. Moreira**, P. Calado and B. Martins, *Learning to Rank for Expert Search in Digital Libraries of Academic Publications*, In proceedings of the 15th portuguese conference on Artificial Intelligence, 2011
- ✓ **C. Moreira**, B. Martins and P. Calado, *Using Rank Aggregation for Expert Search in Academic Digital Libraries*, In Simpósio de Informática, INFORUM, 2011
- ✓ **C. Moreira**, A. Mendes, L. Coheur and B. Martins, *Towards the Rapid Development of a Natural Language Understanding Module*, In proceedings of the 11th conference on intelligent virtual agents, 2011