Pseudo-relevance Feedback & Query Models

Kuan-Yu Chen (陳冠宇)

2018/11/09 @ TR-514, NTUST

HW3

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	M10715004			0.50982	42	2d
2	—	M10715052			0.48138	9	5h
3	—	M10715062			0.47324	42	1d
4	—	B10415018_沈政一			0.47312	50	21h
5	—	B10415004_楊晉復			0.45749	48	13h
6	—	B10415045_施泓仰			0.44680	14	16h
7	—	B10332015_羅子原			0.44400	8	2h
8	—	M10615110			0.43939	13	2h
9	▲1	M10715025_廖傑明			0.42423	27	1d
10	▲1	M10715010			0.42025	58	2d

Progress

Date	Syllabus	Homework
9/14	Course Overview	
9/21	Classic Models	Homework-1
9/28	Extended Probabilistic Models	
10/5	Break @ Rocling 2018	
10/12	Evaluation & Benchmark Collections	Homework-2 & HW2 Description
10/19	Latent Semantic Analysis and Topic Models	Homework-3
10/26	Search Results Diversification	
11/2	Midterm Exam	
11/9	Pseudo-Relevance Feedback & Query Models	Homework-4
11/16	Invited Talk	
11/23	Introduction to Deep Learning	Submit Your Member List and Paper Title!
11/30	Representation Learning for Information Retrieval	Homework-5
12/7	Supervised Retrieval Models	
12/14	Presentations	
12/21	Presentations	
12/28	Presentations	
1/4	Competition	

Review

- Topic Models
  - PLSA
  - LDA

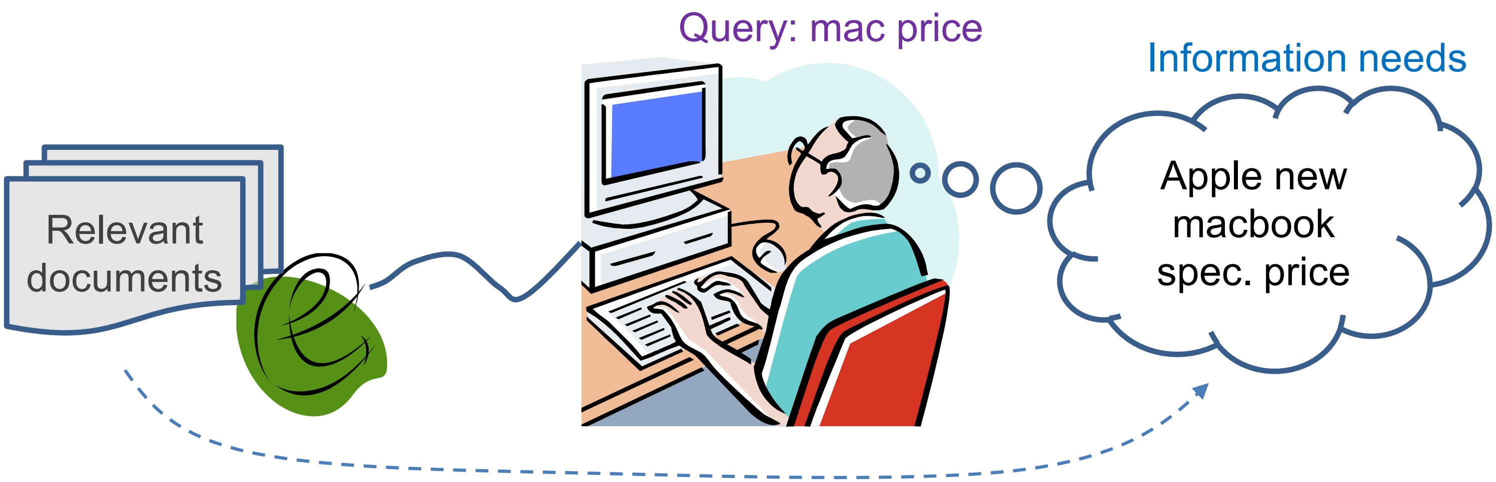
- Search Results Diversification
  - MMR
  - SMM
  - xMMR
  - WUME
  - xQuAD

- Clarity

M. Larson and G. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235-422, 2012.

Introduction

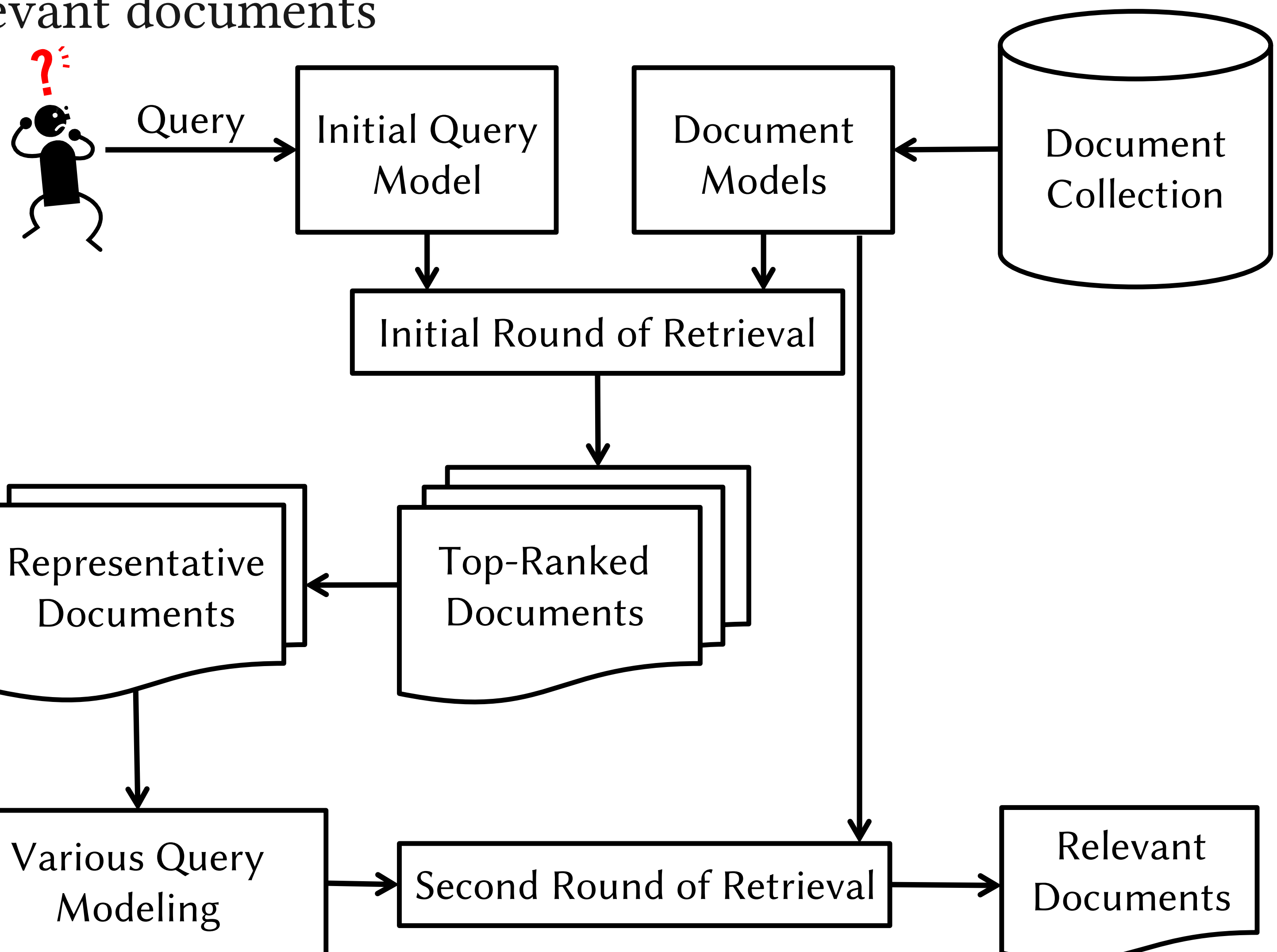
- An information need can be defined as **the reason** for which the user turns to a search engine



- Each query usually consists of only a few words, the corresponding representation might not be appropriately estimated
  - Several effective formulations to enhance the query representation by **pseudo-relevance feedback** process

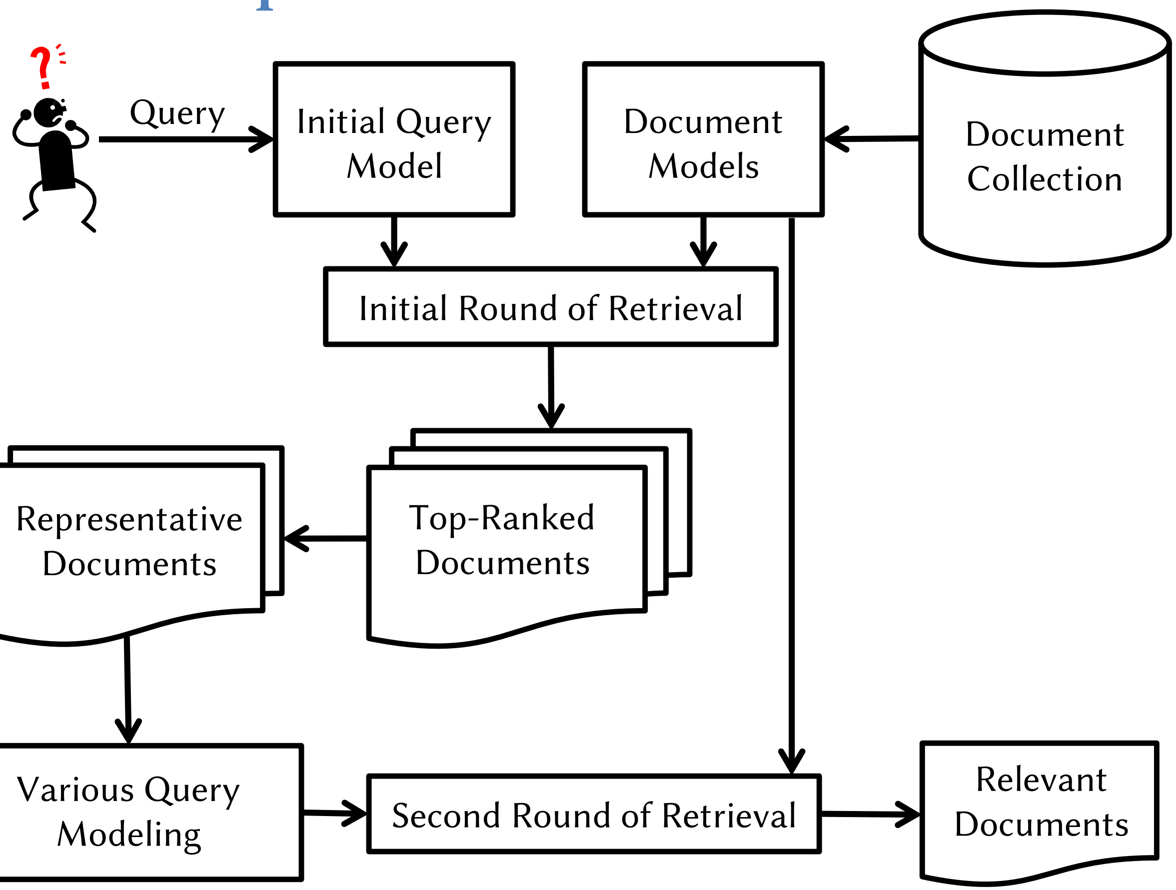
A General Flowchart of PRF

- "Pseudo" means that we assume top-ranked document are relevant documents



Research Issues

- The main issues in pseudo-relevance feedback
  - How to select relevant documents from the top-retrieved documents
  - How to **select expansion terms**



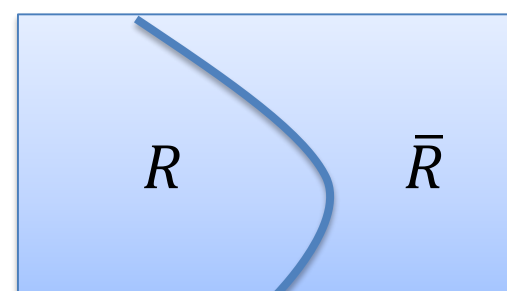
The Rocchio Algorithm – 1

- Rocchio's relevance feedback model is a classic query expansion method and it has been shown to be effective in boosting information retrieval performance
  - It is a way of incorporating pseudo relevance feedback information into the vector space model

$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left(\sum_{d_j \in R_q} \overrightarrow{d_j} \right) - \gamma \cdot \frac{1}{|\bar{R}_q|} \cdot \left(\sum_{d_{j'} \in \bar{R}_q} \overrightarrow{d_{j'}} \right)$$

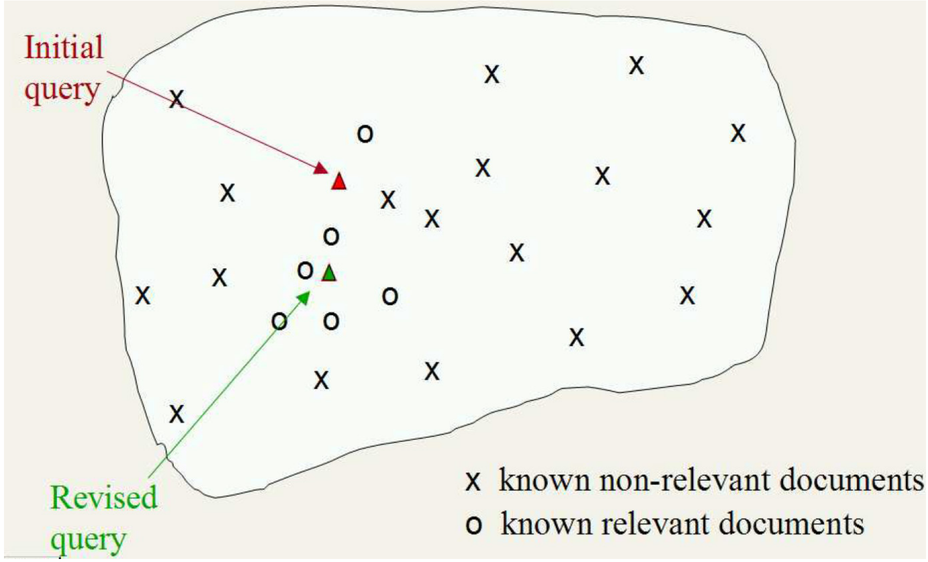
  - R_q be the set of relevant documents to a given query q
  - $\bar{R}_q$ be the set of non-relevant documents to query q
  - Each word is represented by the TFIDF score

Document Collection



The Rocchio Algorithm – 2

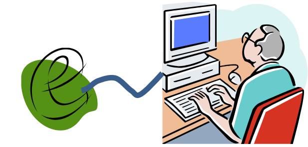
- Starting from the original query $\vec{q}$, the new query moves you some distance toward the centroid of the relevant documents and some distance away from the centroid of the non-relevant documents



- A simplified variant is to consider the positive feedback documents only

$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left(\sum_{d_j \in R_q} \overrightarrow{d_j} \right)$$

KL-Divergence Measure



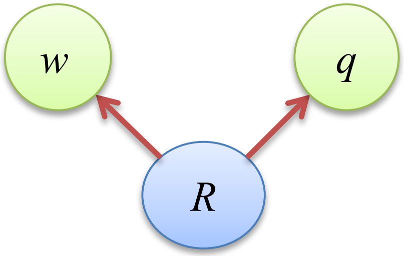
- Another basic formulation of LM for IR is the Kullback-Leibler (KL)-Divergence measure

$$KL(q||d_j) = \sum_{w \in V} P(w|q) \log \frac{P(w|q)}{P(w|d_j)} \propto -\sum_{w \in V} P(w|q) \log P(w|d_j)$$

  - A query is treated as a **probabilistic model** rather than simply an **observation**
  - KL-divergence supports us to achieve a better result by considering **both** query and document models

Relevance Model

- The relevance modeling (RM) is a well-practiced approach
  - Each query is assumed to be associated with a concept R (or relevance class/information need)
    - Both the query and relevant documents are drawn from the concept R
  - The RM model assumes that words w that **co-occur** with the query in the concept will have higher probabilities

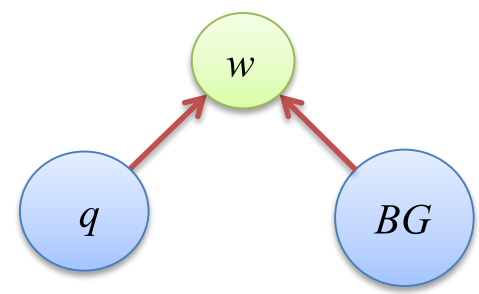


$$P_{RM}(w) \equiv \frac{P(w,q|R)}{\sum_{w' \in V} P(w',q|R)} \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w,q|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j) P(w',q|d'_j)}$$

$$= \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) \prod_{i=1}^{|q|} P(w_i|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j) P(w'|d'_j) \prod_{i=1}^{|q|} P(w_i|d'_j)}$$

Simple Mixture Model – 1

- An alternative formulation to extract relevance cues is simple mixture model (SMM)
  - It assumes that words in the set of pseudo-relevance feedback documents are drawn from two-component mixture model:
    - One component is the query model
    - The other is a background model



- The SMM model $P_{SMM}(w)$ is estimated by maximizing the log-likelihood of the set of top-ranked documents R_q expressed as follows:

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left((1-\alpha) \cdot P_{SMM}(w) + \alpha \cdot P(w|BG)\right)^{c(w,d_j)}$$

Simple Mixture Model – 2

- Estimate the parameters
  - E-step

$$P(T_{SMM}|w) = \frac{(1-\alpha)\cdot P_{SMM}(w)}{(1-\alpha)\cdot P_{SMM}(w) + \alpha\cdot P(w|BG)}$$

  - M-step

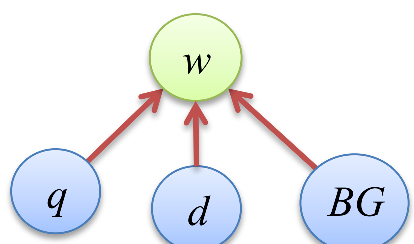
$$P_{SMM}(w) = \frac{\sum_{d_j \in R_q} c(w,d_j) P(T_{SMM}|w)}{\sum_{w' \in V}\sum_{d_{j'} \in R_q} c(w',d_{j'}) P(T_{SMM}|w')}$$

$$\begin{aligned}\mathcal{L} &= \prod_{d_j \in R_q}\prod_{w \in V}\left((1-\alpha)\cdot P_{SMM}(w) + \alpha\cdot P(w|BG)\right)^{c(w,d_j)} \\ &= \prod_{d_j \in R_q}\prod_{w \in V}\left(P_{SMM}(w|T_{SMM})P(T_{SMM}) + P(w|BG)P(BG)\right)^{c(w,d_j)}\end{aligned}$$

Tri-Mixture Model – 1

- The TriMM model $P_{TMM}(w)$ is estimated by maximizing the log-likelihood of the set of top-ranked documents
  - It assumes that words in the set of pseudo-relevance feedback documents are drawn from three-component mixture model:
    - One component is the query model
    - Another component is the document-specific model
    - The other is a background model

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \Big((1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG) \Big)^{c(w,d_j)}$$



Tri-Mixture Model – 2

- Estimate the parameters
  - E-step

$$P\left(T_{TMM}\middle|w,d_j\right)=\frac{(1-\alpha-\beta)\cdot P_{TMM}(w)}{(1-\alpha-\beta)\cdot P_{TMM}(w)+\alpha\cdot P(w|d_j)+\beta\cdot P(w|BG)}$$

$$P\left(T_{d_j}\middle|w,d_j\right)=\frac{\alpha\cdot P(w|d_j)}{(1-\alpha-\beta)\cdot P_{TMM}(w)+\alpha\cdot P(w|d_j)+\beta\cdot P(w|BG)}$$

  - M-step

$$P_{TMM}(w)=\frac{\sum_{d_j\in R_q}c(w,d_j)P\left(T_{TMM}\middle|w,d_j\right)}{\sum_{w'\in V}\sum_{d_{j'}\in R_q}c(w',d_{j'})P\left(T_{TMM}\middle|w',d_{j'}\right)}$$

$$P\left(w\middle|d_j\right)=\frac{c(w,d_j)P\left(T_{d_j}\middle|w,d_j\right)}{\sum_{w'\in V}c(w',d_j)P\left(T_{d_j}\middle|w',d_j\right)}$$

A Unified Framework – 1

- It is obvious that the major difference among the representative models mentioned above is how to capitalize on the set of documents and the original query

- A principled framework can be obtained to unify all of these models (and their extensions) by using a generalized objective likelihood function:

$$\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left(\sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)}$$

A Unified Framework – 2

$$\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left(\sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)}$$

- **Relevance modeling (RM)**: when E only consists of the user query, M consists of a set of document models corresponding to the top-ranked (pseudo-relevant) documents, and we assume the document models are known, then it can be deduced to the RM model

$$P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j)P(w|d_j) \prod_{i=1}^{|q|} P(w_i|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j)P(w'|d'_j) \prod_{i=1}^{|q|} P(w_i|d'_j)}$$

$$= \frac{\sum_{d_j \in R_q} P(d_j)P(w|d_j)P(q|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j)P(w'|d'_j)P(q|d'_j)}$$

$$= \sum_{d_j \in R_q} P(w|d_j) \frac{P(d_j)P(q|d_j)}{\sum_{d'_j \in R_q} P(d'_j)P(q|d'_j)}$$

$$\sum_{w' \in V} P(w'|d'_j) = 1$$

A Unified Framework – 3

$$\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left(\sum_{m \in M} P(w|m) P(m) \right)^{c(w,e)}$$

- **Simple mixture modeling (SMM)**: if we hypothesize that M consists of two components: one component is a generic background model and the other is an unknown query-specific topic model, the weight of each component is presumably fixed in advance, and the observations are those top-ranked documents

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left((1-\alpha) \cdot P_{SMM}(w) + \alpha \cdot P(w|BG) \right)^{c(w,d_j)}$$

A Unified Framework – 4

$$\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left(\sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)}$$

- **Tri-Mixture modeling (TMM)**: if we hypothesize that M consists of three components: the first component is a generic background model, the second model is a document-specific model, and the last one is an unknown query-specific topic model, the weight of each component is presumably fixed in advance, and the observations are those top-ranked documents

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \Big((1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG) \Big)^{c(w,d_j)}$$

A Unified Framework – 5

$$\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left(\sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)}$$

- **Others**: without loss of generality, some other state-of-the-art language models also can be deduced from the proposed general objective function, such as the **positional relevance model**, the **cluster-based methods**, the **topic models**, and among others

$$\mathcal{L} = \prod_{w_i \in V} \prod_{d_j \in \mathbf{D}} P(w_i, d_j)^{c(w_i, d_j)} = \prod_{d_j \in \mathbf{D}} \prod_{i=1}^{|d_j|} P(w_i, d_j)$$

$$= \prod_{d_j \in \mathbf{D}} \prod_{i=1}^{|d_j|} \left(P(d_j) \sum_{k=1}^{K} P(w_i|T_k) P(T_k|d_j) \right)$$

Topic-based Relevance Modeling

- TRM assumes that the additional cues of how words are distributed across a set of latent topics can carry useful global topic structure for relevance modeling
  - The pseudo-relevant documents are assumed to share a set of pre-defined latent topic variables $\{T_1, \cdots, T_k, \cdots, T_K\}$

$$P_{TRM}(w) \approx \frac{\sum_{d_j \in R_q} \sum_{k=1}^{K} P(d_j) P(T_k|d_j) P(w|T_k) P(q|T_k)}{\sum_{w' \in V} \sum_{d'_j \in R_q} \sum_{k'=1}^{K} P(d'_j) P(T_{k'}|d'_j) P(w|T_{k'}) P(q|T_{k'})}$$

  - As with PLSA and LDA, the probabilities $P(w|T_k)$ and $P(T_k|d_j)$ can be estimated using inference algorithms like EM or VB-EM algorithms on the whole document collection

$$P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) P(q|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j) P(w'|d'_j) P(q|d'_j)}$$

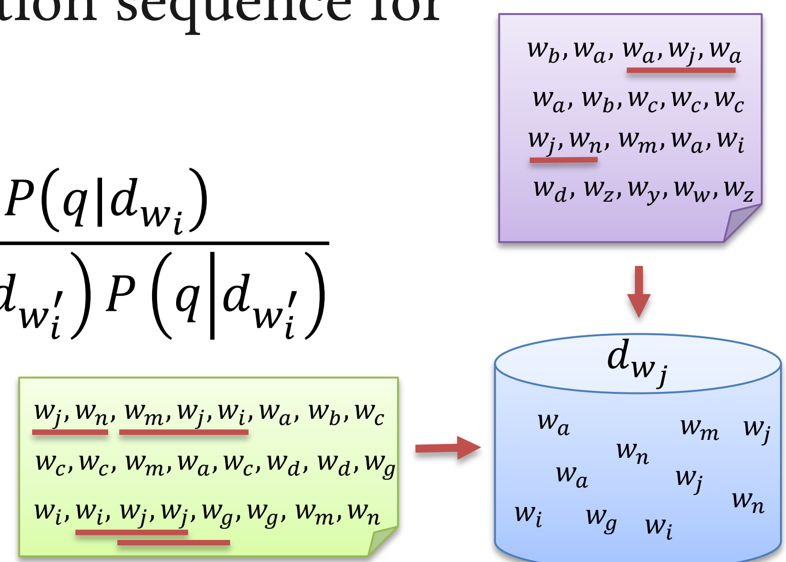
Word-based Relevance Modeling

- The most challenging aspect facing RM is how to efficiently infer the relevance class
  - The relevance class of a given query is commonly approximated by the top-ranked documents returned by an IR system

- The WRM model of each word in the language can be trained by concatenating those words occurring within a context window to form a relevant observation sequence for estimating $P(w|d_{w_i})$

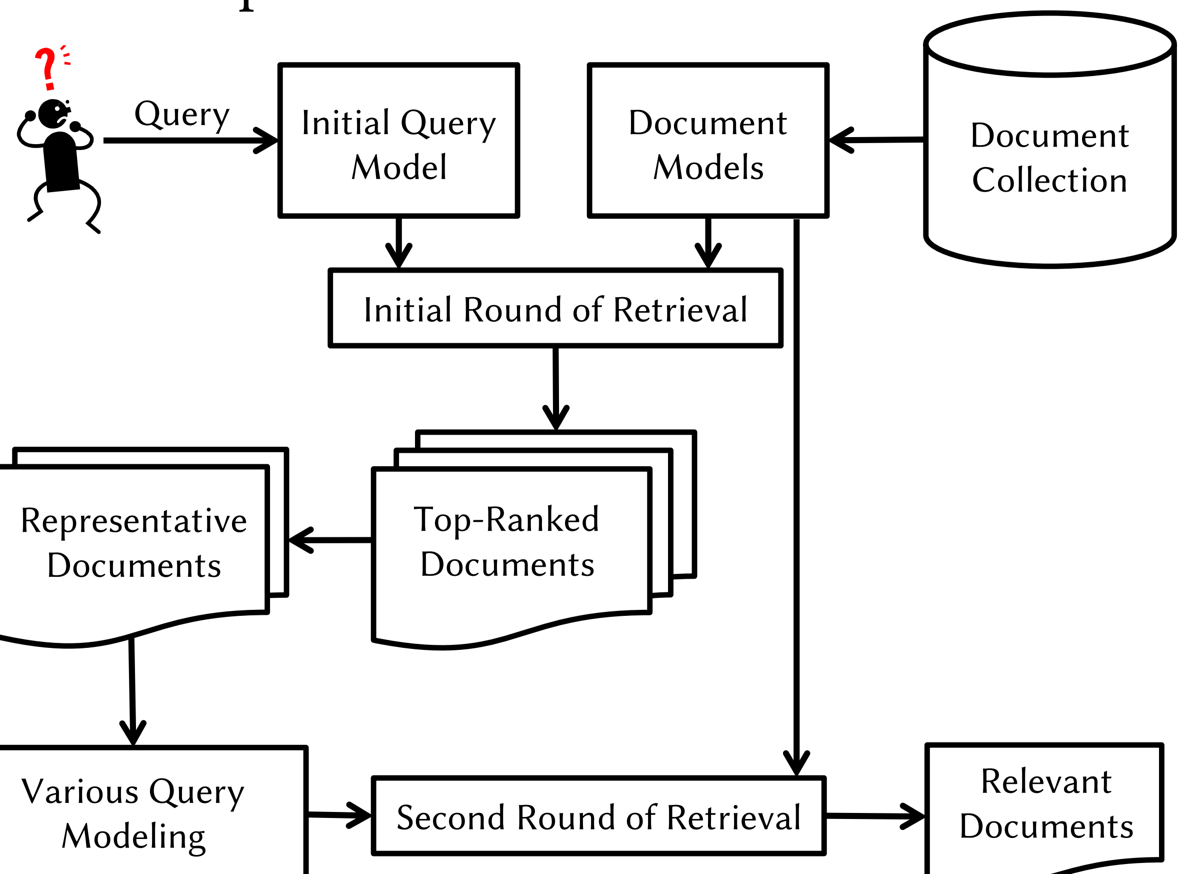
$$P_{WRM}(w) \approx \frac{\sum_{w_i \in q} P(d_{w_i}) P(w|d_{w_i}) P(q|d_{w_i})}{\sum_{w' \in V} \sum_{w'_i \in q} P\left(d_{w'_i}\right) P\left(w' \middle| d_{w'_i}\right) P\left(q \middle| d_{w'_i}\right)}$$

$$P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) P(q|d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j) P(w'|d'_j) P(q|d'_j)}$$



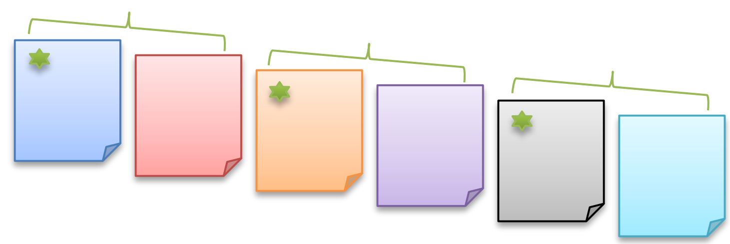
Research Issues

- The main issues in pseudo-relevance feedback
  - How to **select relevant documents** from the top-retrieved documents
  - How to select expansion terms

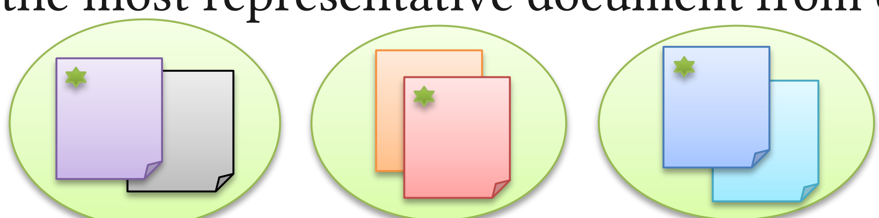


Gapped Top *K* & Cluster Centroid

- In order to select a set of pseudo-relevant documents, which can cover most of the possible aspects of the query, a few selecting methods have been proposed
  - **Gapped Top *K***
    - partition the documents into K clusters based solely on the relevance scores
    - select documents with the highest relevance score in each cluster to form the feedback document set



  - **Cluster Centroid**
    - partition top-ranked documents into K clusters
    - select the most representative document from each cluster



Active Relevance, Density, & Diversity

- Active-RDD algorithm extends the MMR algorithm by adding an extra term, which reflects the document density
  - Relevance

$$Rel(d) \equiv KL(q||d) = \sum_{w \in V} P(w|q) log \frac{P(w|q)}{P(w|d)}$$

  - Density
    - Jeffreys divergence

$$Density(d) \equiv \frac{-1}{|\mathbf{D}|} \sum_{d_j \in \mathbf{D}} \left(KL(d_j||d) + KL(d||d_j) \right)$$

  - Diversity

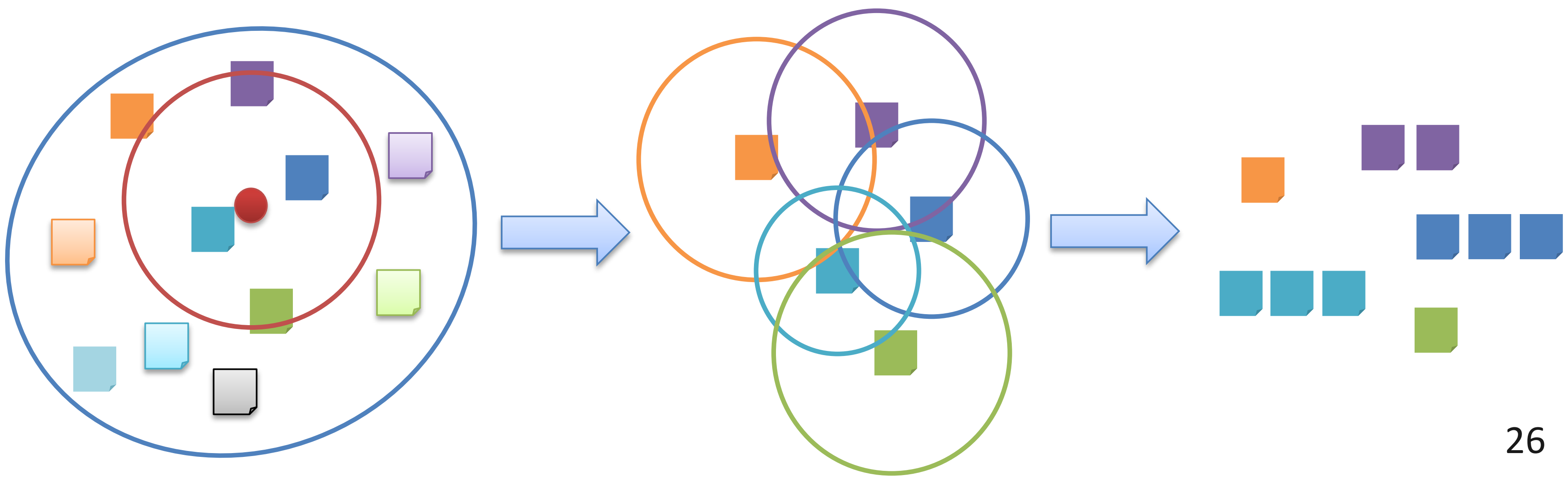
$$Diversity(d) \equiv \min_{\tilde{d} \in \tilde{\mathbf{D}}} \left(KL(\tilde{d}||d) + KL(d||\tilde{d}) \right)$$

  - Active-RDD

$$d^* = \underset{d \in \{\mathbf{D} - \tilde{\mathbf{D}}\}}{\operatorname{argmax}} \alpha \cdot Rel(d) + \beta \cdot Density(d) + (1 - \alpha - \beta) \cdot Diversity(d)$$

Resampling Method

- The essential idea is that a document that appears in multiple highly-ranked clusters will contribute more to the query terms than other documents
  - The **dominate documents** in the sampled clusters are used for feedback **with redundancy**
  - The overlapping cluster method is used to identify **dominant documents** for the query to emphasize good representative terms in dominant documents



Conclusions

- The methods for tackling the fundamental problem can be classified into **global** methods and **local** methods
  - Global methods are techniques for expanding or reformulating query terms independent of the query and initial search results
    - Thesaurus or WordNet
    - automatic thesaurus generation
    - spelling correction
  - Local methods adjust a query relative to the documents that initially appear to match the query
    - Relevance feedback
    - Pseudo relevance feedback (Blind relevance feedback)
    - (Global) indirect relevance feedback

Homework 4 – Description

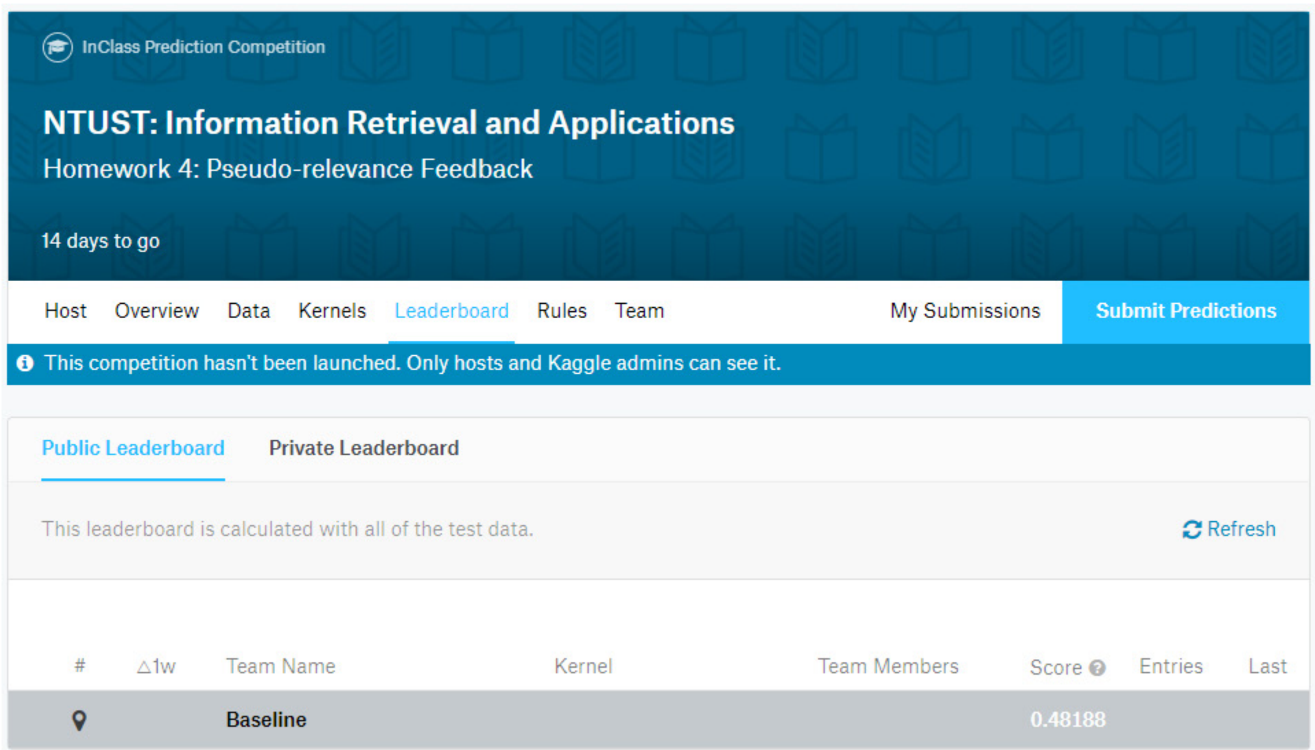
- In this project, you will have
  - 800 Queries
  - 2265 Documents

- Our goal is to implement the Rocchio algorithm (or LM-based methods) for retrieval

Homework 4 – Kaggle

- Please login our competition page at Kaggle
  - https://www.kaggle.com/t/aed5e4d90570477d8fb1745552cd904e

- **Your team name is ID_Name**
  - M123456_陳冠宇



Homework 4 – Submission Format

submission.txt

```
1 Query,RetrievedDocuments
2 20001.query,VOM19980619.0700.0347 VOM19980225.0700.0510 VOM19980317.0900.0192 VOM19980317.0900.0330 VOM19980
. 00.0173 VOM19980302.0700.0241 VOM19980303.0700.2287 VOM19980530.0730.0166 VOM19980404.0700.2088 VOM19980616.0
. 216 VOM19980614.0700.0357 VOM19980626.0700.0409 VOM19980403.0700.0489 VOM19980523.0730.0220 VOM19980524.0730.
. VOM19980624.0900.0077 VOM19980625.0700.0363 VOM19980605.0730.0152 VOM19980602.0730.0102 VOM19980603.0730.0280
. 9980522.0730.0037 VOM19980228.0700.0327 VOM19980414.0900.0260 VOM19980223.0700.0765 VOM19980505.0700.0529 VOM
. 503.0730.0136 VOM19980319.0900.3416 VOM19980620.0730.0034 VOM19980302.0700.0209 VOM19980302.0900.2091 VOM1998
. 0900.0207 VOM19980305.0900.1926 VOM19980521.0730.0029 VOM19980504.0700.0376 VOM19980314.0700.0239 VOM19980619.
. .0137 VOM19980611.0700.0150 VOM19980326.0700.2112 VOM19980522.0900.0269 VOM19980503.0700.0412 VOM19980428.090
. 4 VOM19980422.0900.0021 VOM19980605.0700.0194 VOM19980611.0700.0046 VOM19980223.0700.2728 VOM19980614.0730.02
. M19980303.0700.0696 VOM19980326.0900.0149 VOM19980505.0700.0481 VOM19980614.0730.0034 VOM19980226.0900.1964 V
. 80523.0730.0083 VOM19980316.0700.0356 VOM19980609.0900.0009 VOM19980314.0700.2300 VOM19980302.0700.2137 VOM19
. 4.0700.0458 VOM19980319.0900.2169 VOM19980305.0700.2126 VOM19980515.0700.0472 VOM19980403.0700.0129 VOM19980
. 30.0142 VOM19980618.0700.0234 VOM19980319.0900.0647 VOM19980527.0700.0528 VOM19980607.0730.0033 VOM19980305.0
3 20002.query,VOM19980530.0730.0101 VOM19980611.0900.0216 VOM19980506.0900.0089 VOM19980624.0700.0434 VOM19980
. 00.2021 VOM19980604.0900.0246 VOM19980606.0700.0562 VOM19980303.0900.2085 VOM19980225.0700.0999 VOM19980312.0
. 171 VOM19980220.0900.1979 VOM19980305.0700.0763 VOM19980627.0700.0360 VOM19980225.0700.0302 VOM19980529.0700.
. VOM19980612.0730.0192 VOM19980319.0700.2737 VOM19980630.0700.0071 VOM19980526.0730.0131 VOM19980403.0700.0489
. 9980430.0900.0192 VOM19980502.0700.0307 VOM19980616.0700.0420 VOM19980319.0900.3468 VOM19980303.0900.1926 VOM
```

Homework 4 – Scoring

- The evaluation measure is **MAP@50**

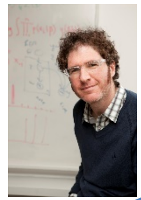
- The maximum number of daily submissions is 20

- The **hard** deadline is 11/23 11:00am
  - $YourScore = 4 + \frac{YourMAP - BaselineMAP}{HighestMAP - BaselineMAP} \times 6\%$

- You should submit source codes and a mini report onto the Moodle system
  - TA will ask you to demo your program
  - In this HW, you can **ONLY** leverage PRF models to do retrieval

The Evolution

David M. Blei
Columbia University, USA

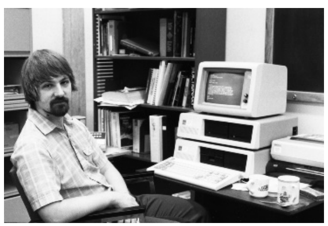


Thomas Hofmann
ETH Zurich, Switzerland

2003 Latent Dirichlet Allocation

2001 Relevance-based LM & Simple Mixture Model

Scott Deerwester



1999 Probabilistic Latent Semantic Analysis

V. Lavrenko
Edinburgh

C.X. Zhai
Illinois University

1998 Language Modeling Approaches

1994 Best Match Models (Okapi Systems)

1988 Latent Semantic Analysis

1976 Probabilistic Model

1975 Vector Space Model

1973 Boolean Model

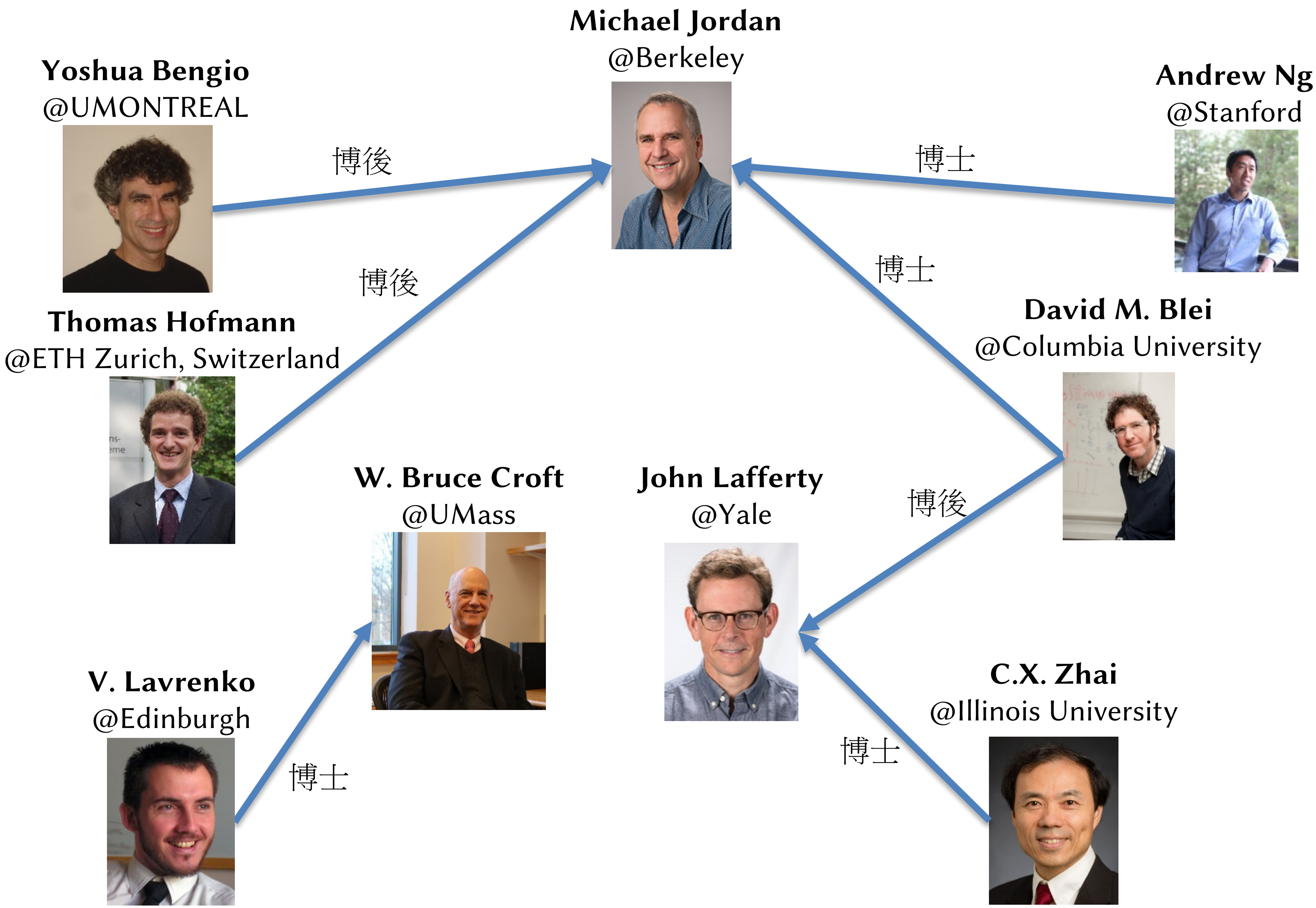
1972 Inverse Document Frequency

J. Rocchio

1965 Rocchio Algorithm

1957 Term Frequency

Gossiping


Michael Jordan
@Berkeley
Yoshua Bengio
@UMONTREAL
博後
Andrew Ng
@Stanford
博士
博後
博士
Thomas Hofmann
@ETH Zurich, Switzerland
David M. Blei
@Columbia University
W. Bruce Croft
@UMass
John Lafferty
@Yale
博後
V. Lavrenko
@Edinburgh
博士
C.X. Zhai
@Illinois University
博士

Questions?

kychen@mail.ntust.edu.tw