Pseudo-relevance Feedback & Query Models

Kuan-Yu Chen (陳冠宇)

2020/11/27 @ TR-313, NTUST

HW4

#	Team Name	Score	Entries	Last
1	M10915010_盧克函	0.63447	46	6d
2	M10915045_施信宏	0.62661	26	2d
3	M10915095_薛宇翔	0.61474	49	10h
4	M10907505_游照臨	0.61284	13	9h
5	M10915027_石成峰	0.60911	31	9h
6	M10915028_陳柏勳	0.60886	29	12h
7	M10915100_郭智威	0.60248	68	9h
8	M10915201_陳牧凡	0.59516	70	3d
9	D10907005_陳昱宏	0.59213	36	14h
10	M10815048_張晏銘	0.58963	63	3d
11	B10615043_何嘉峻	0.58052	70	13h
12	B10615034_黃柏翰	0.57919	128	2d
13	B10615036_黃泰銘	0.57787	11	1d
14	B10615026_温承勳	0.57495	48	2d
15	M10915006_廖勗宏	0.57439	19	1d
16	M10915080_羅笠程	0.57409	11	12h
17	TEST	0.57201	1	15h
18	M10815036_王仁德	0.57185	2	12h
19	B10632026_吳苡瑄	0.57150	17	17h
	FYI: with 32 topics	0.56890		
20	B10615024_李韋宗	0.56758	51	16h

#	△pub	Team Name	Score	Entries	Last
1	—	M10915010_盧克函	0.52665	46	6d
2	▲5	M10915100_郭智威	0.50803	68	9h
3	▲14	TEST	0.50714	1	15h
4	—	M10907505_游照臨	0.50159	13	9h
5	▼3	M10915045_施信宏	0.50113	26	2d
6	▲3	D10907005_陳昱宏	0.49624	36	14h
7	▲5	B10615034_黃柏翰	0.49327	128	2d
8	▼2	M10915028_陳柏勳	0.49014	29	12h
9	▼1	M10915201_陳牧凡	0.48948	70	3d
10	▼5	M10915027_石成峰	0.48665	31	9h
11	—	B10615043_何嘉峻	0.48597	70	13h
12	▼9	M10915095_薛宇翔	0.48540	49	10h
13	▲6	B10632026_吳苡瑄	0.48165	17	17h
14	▼4	M10815048_張晏銘	0.48067	63	3d
15	▲7	B10615033_王璽禎	0.48044	17	1d
16	▲7	M10915012_黃偉愷	0.47906	13	6d
17	▲13	M10815013_陳思妮	0.47847	31	10h
18	▲2	B10615024_李韋宗	0.47838	51	16h
19	▼1	M10815036_王仁德	0.47793	2	12h
20	▼7	B10615036_黃泰銘	0.47630	11	1d

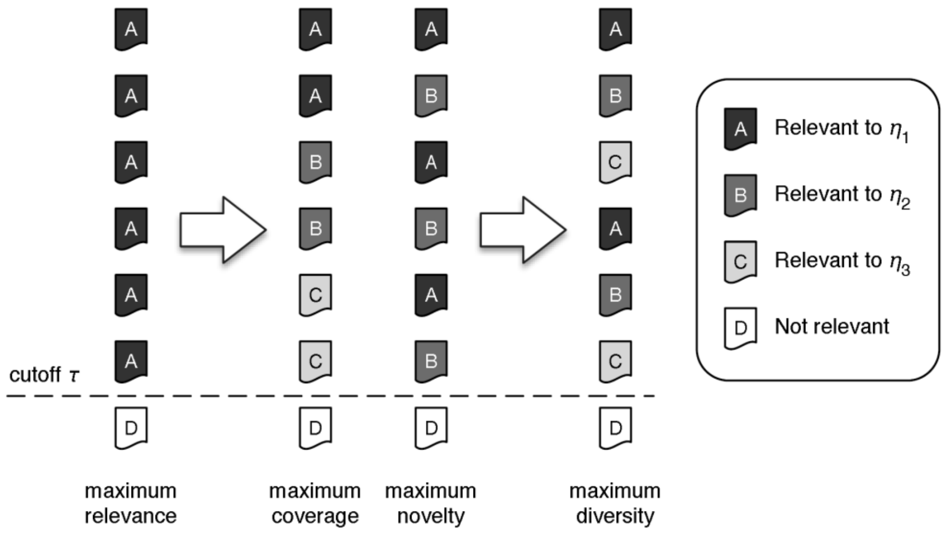
About Final Project

- Group your team!
  - 2~4 team members
  - Choose a paper

Date	Syllabus	Homework
9/18	Course Overview	
9/25	Break for Rocling2020	
10/2	Holiday for Moon Festival	
10/9	Holiday for National Day	
10/16	Classic Models	Homework-1(deadline: 10/29 23:59)
10/23	Extended Probabilistic Models	Homework-2 (deadline: 11/5 23:59)
10/30	Evaluation & Benchmark Collections	Homework-3 (deadline: 11/12 23:59)
11/6	Latent Semantic Analysis	
11/13	Statistical Topic Models	Homework-4 (deadline: 11/26 23:59)
11/20	Search Results Diversification	
11/27	Pseudo-Relevance Feedback & Query Models	Homework-5 (deadline: 12/10 23:59)
12/4	Talk	Submit Your Member List!
12/11	Representation Learning for Information Retrieval	
12/18	Supervised Retrieval Models & Information Retrieval in Practice	Homework-6 (deadline: 12/31 23:59) & Submit Your Paper Title!
12/25	Break for Your Final Project	
1/1	Holiday for Founding Anniversary	
1/8	Presentation-1	
1/15	Presentation-2	

Review

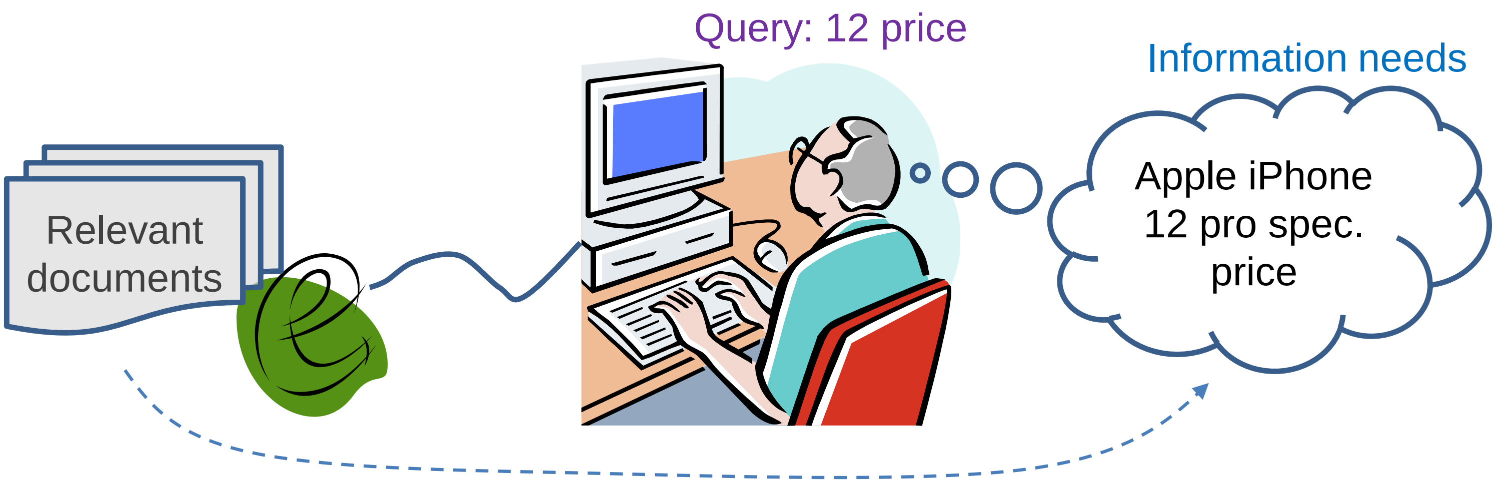
- These methods mainly differ in **diversity modeling**
  - **Implicitly**: The diversity is implicitly modeled through document similarities
    - MMR
    - SMM
  - **Explicitly**: It can be explicitly modeled through the coverage of query subtopics, and document dependency
    - xMMR
    - WUME
    - xQuAD



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

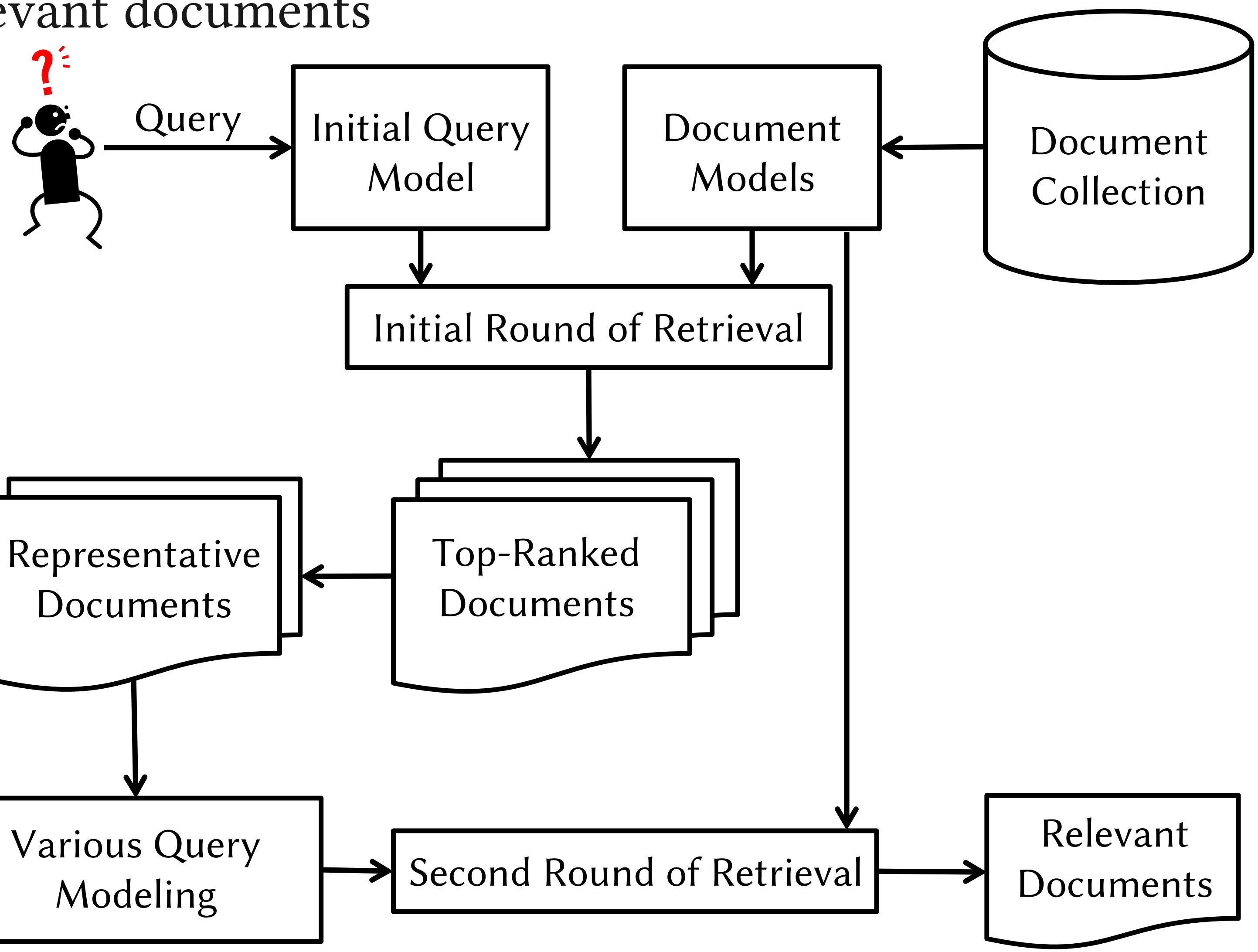
- An information need can be defined as **the reason** for which the user turns to a search engine



- 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

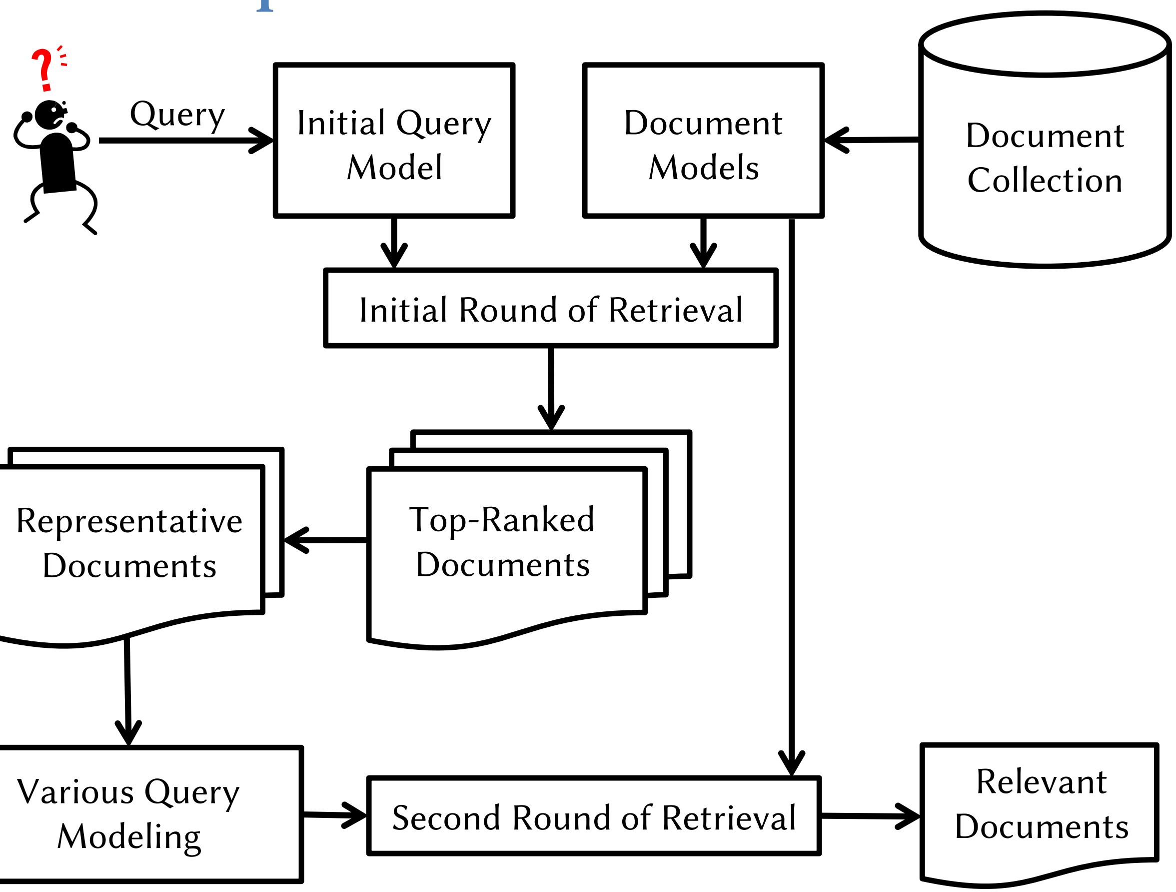
A General Flowchart of PRF

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



Research Issues

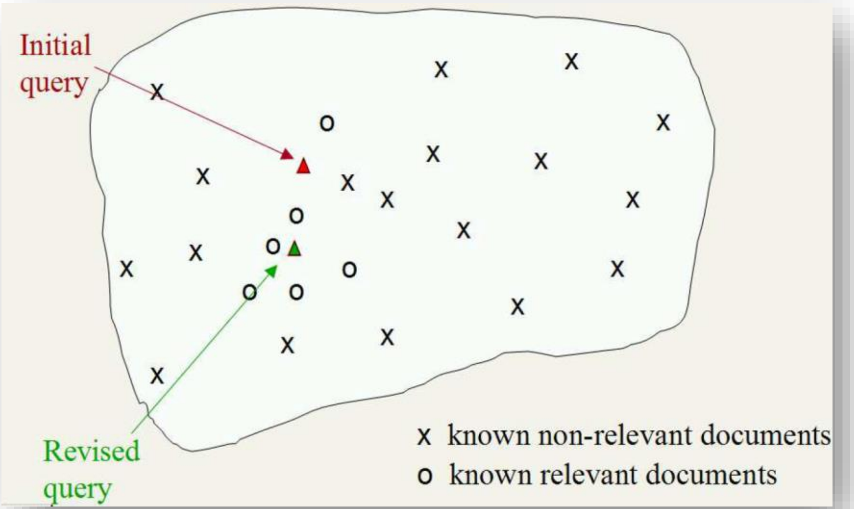
- The main issues in pseudo-relevance feedback
  - How to select relevant documents from the top-retrieved documents
  - How to **select expansion terms**



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

- 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**



The Rocchio Algorithm – 2

- The idea can be fulfilled by using the vector space model with pseudo relevant and non-relevant documents

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

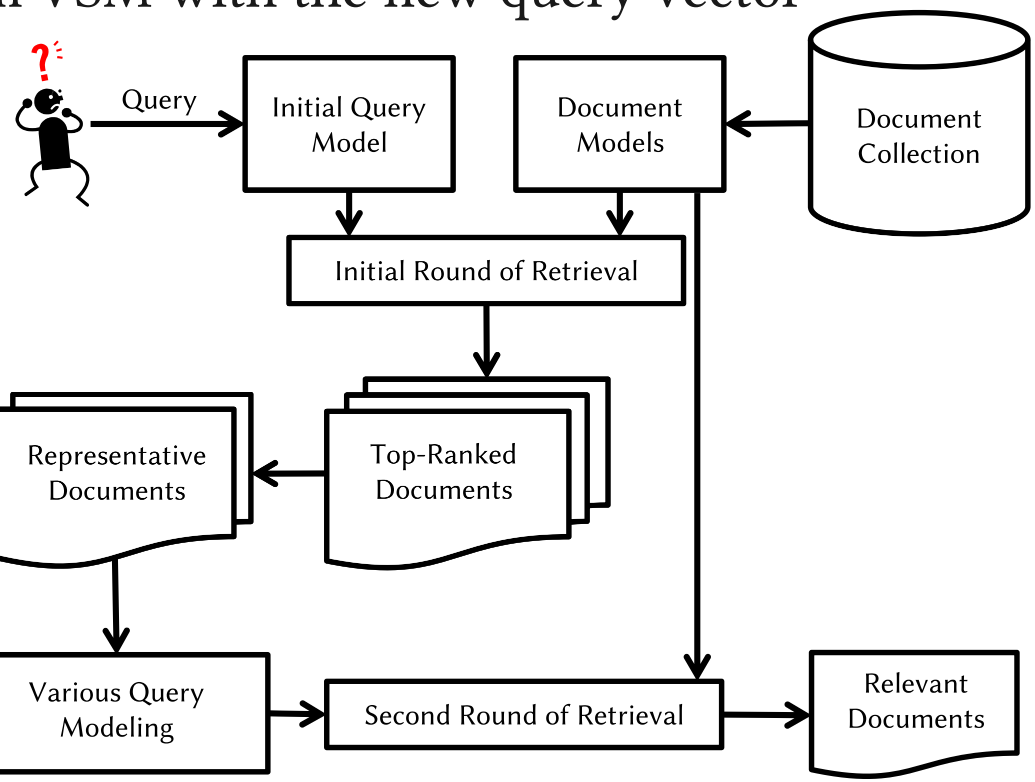
  - R_q be the set of relevant documents to a given query q
  - $\bar{R}_q$ be the set of non-relevant documents to query q
  - Each word is represented by the TFIDF score

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

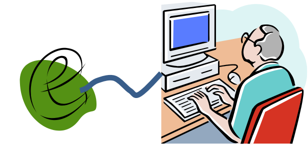
$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left(\sum_{d_j \in R_q} \overrightarrow{d_j} \right)$$

The Rocchio Algorithm – 3

- The full process will become
  1. Perform VSM
  2. Select a set of top-ranked documents
  3. Reformulate the query vector
  4. Perform VSM with the new query vector



KL-Divergence Measure



- Query likelihood measure is a classic way to employ LM to IR

$$P(d_j|q) = \frac{P(q|d_j)P(d_j)}{P(q)} \propto P(q|d_j)P(d_j)$$
$$\approx P(q|d_j) \approx \prod_{i=1}^{|q|} P(w_i|d_j)$$

- 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

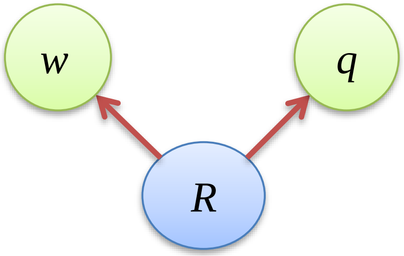
Relevance Model – 1

- 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

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

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

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



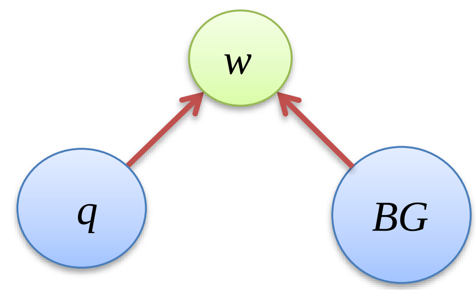
Relevance Model – 2

- Consequently, for a given pair of query and document, the relevance degree can be determined by using the new query language model
  - In order to incorporate the general information, the background model can also be integrated

$$
\begin{aligned}
KL(q||d_j) &= \sum_{w \in V} P(w|q) \log \frac{P(w|q)}{P(w|d_j)} \\
&\propto - \sum_{w \in V} P(w|q) \log P(w|d_j) \\
&= - \sum_{w \in V} [\alpha \cdot P_{ULM}(w) + \beta \cdot P_{RM}(w) + (1-\alpha-\beta) \cdot P_{BG}(w)] \log P(w|d_j)
\end{aligned}
$$

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

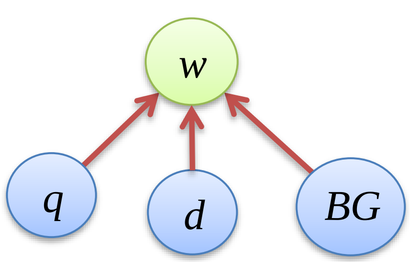
$$P_{SMM}(w) = \frac{\sum_{d_j \in R_q} c(w, d_j) P(T_{SMM}|w)}{\sum_{w' \in V} \sum_{d_{j'} \in R_q} c(w', d_{j'}) P(T_{SMM}|w')}$$

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

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

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

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

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

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

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 $\{T_1, \cdots, T_k, \cdots, T_K\}$

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

  - 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

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

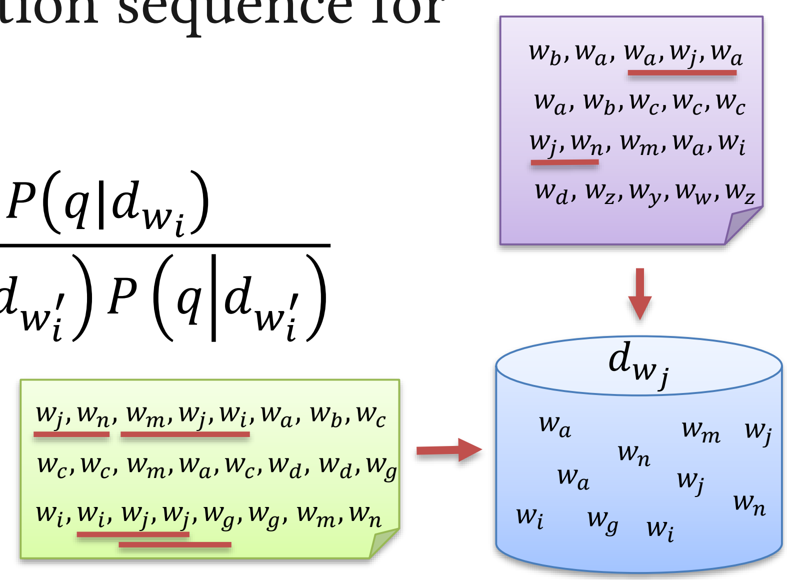
Word-based Relevance Modeling

- 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 $P(w|d_{w_i})$

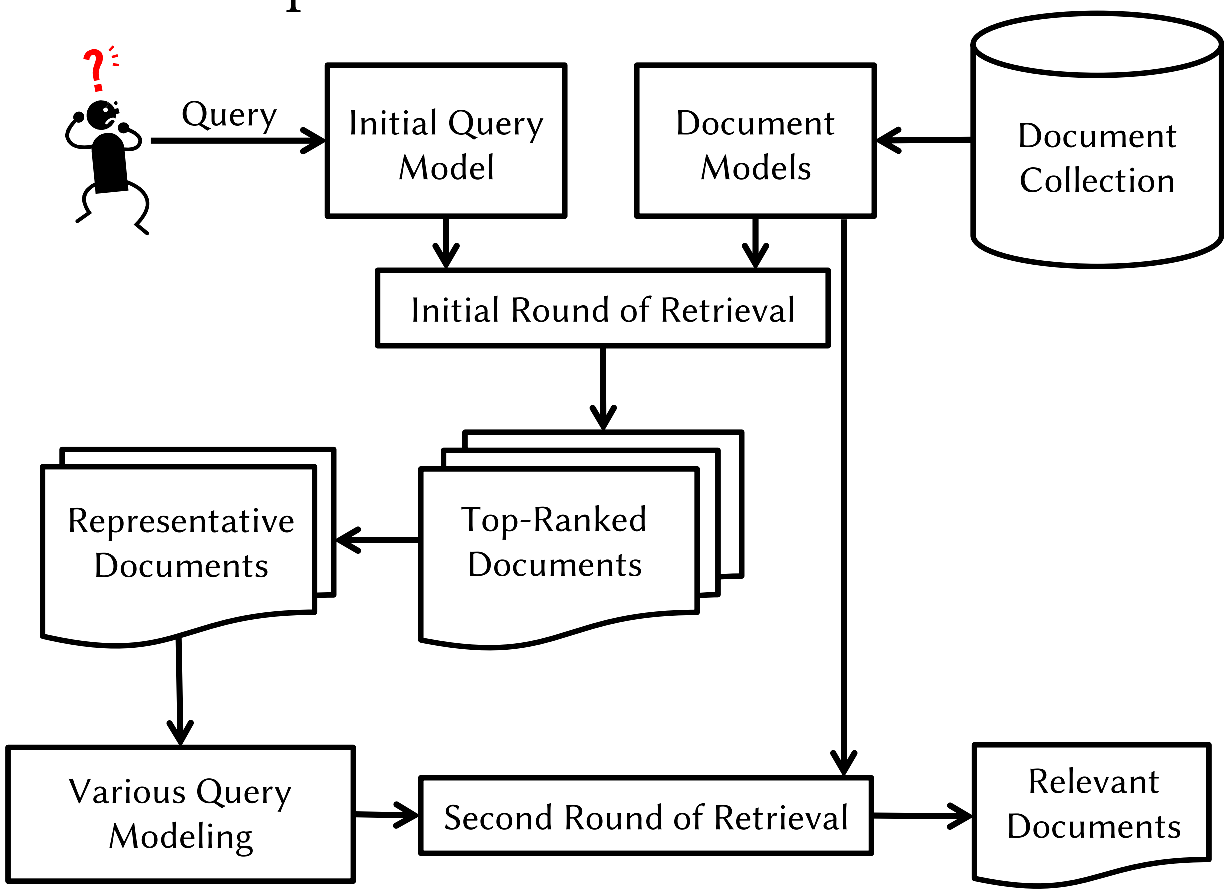
$$P_{WRM}(w) \approx \frac{\sum_{w_i \in q} P(d_{w_i}) P(w|d_{w_i}) P(q|d_{w_i})}{\sum_{w' \in V} \sum_{w'_i \in q} P\left(d_{w'_i}\right) P\left(w' \middle| d_{w'_i}\right) P\left(q \middle| d_{w'_i}\right)}$$

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



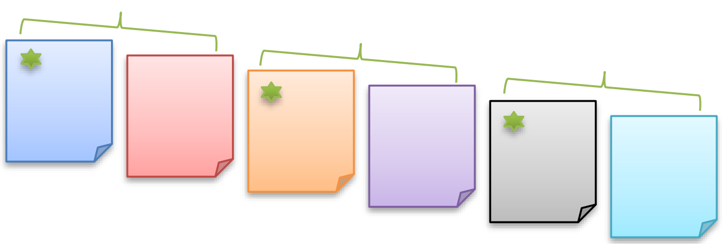
Research Issues

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

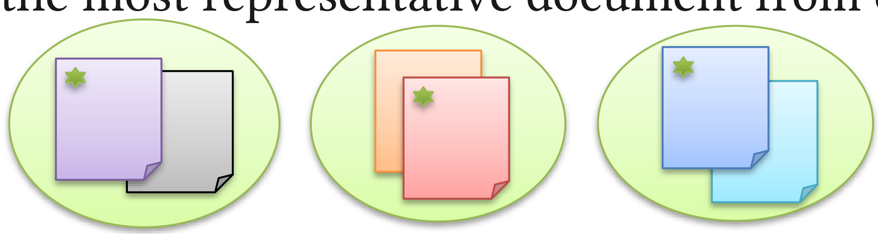


Gapped Top *K* & Cluster Centroid

- 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

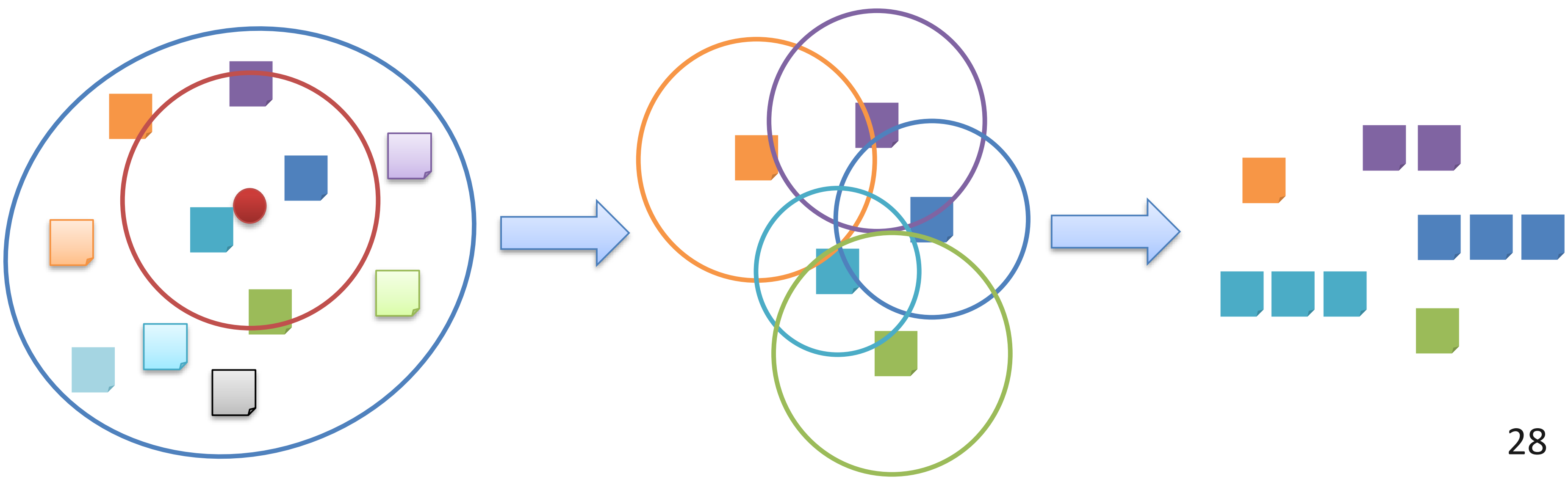
$$Diversity(d) \equiv \min_{\tilde{d} \in \tilde{\mathbf{D}}} \left(KL(\tilde{d}||d) + KL(d||\tilde{d}) \right)$$

  - 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

The Evolution

David M. Blei
Columbia University, USA

Thomas Hofmann
ETH Zurich, Switzerland

Scott Deerwester

V. Lavrenko
Edinburgh

C.X. Zhai
Illinois University

2003 Latent Dirichlet Allocation

2001 Relevance-based LM & Simple Mixture Model

1999 Probabilistic Latent Semantic Analysis

1998 Language Modeling Approaches

1994 Best Match Models (Okapi Systems)

1988 Latent Semantic Analysis

1976 Probabilistic Model

1975 Vector Space Model

1973 Boolean Model

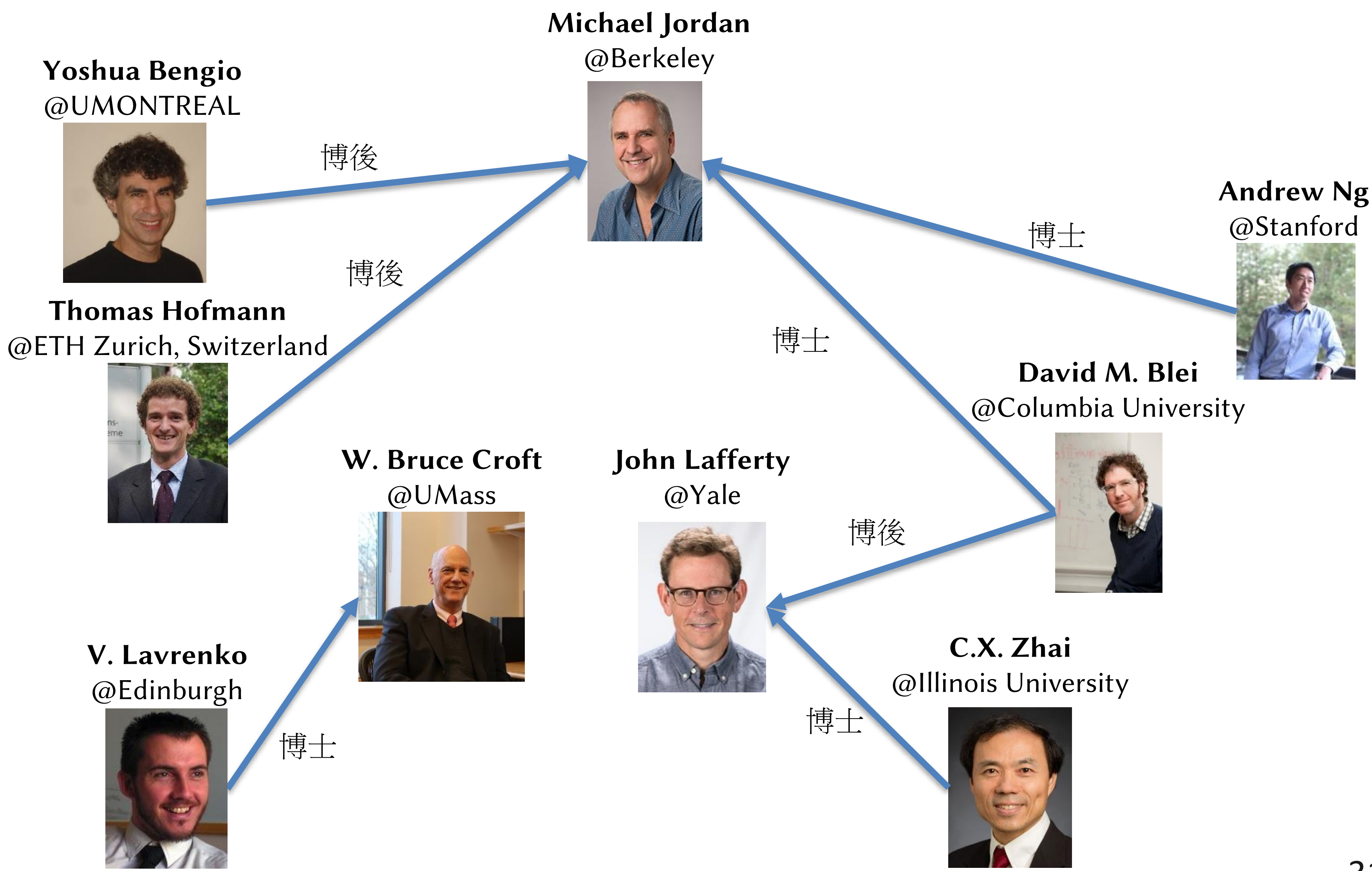
1972 Inverse Document Frequency

J. Rocchio

1965 Rocchio Algorithm

1957 Term Frequency

Relationships



Homework 5 – Description

- In this project, you will have
  - 150 Queries
    - 60% Public Queries & 40% Private Queries
  - 30,000 Documents

- Our goal is to implement a PRF algorithm for retrieval
  - **In addition to the PRF model, you can combine any models/strategies to achieve a good performance**

- Please submit a **report** and your **source codes** to the Moodle system, otherwise you will get 0 point
  - The report will be judged by TA, and the score is either 1 or 2

Homework 5 – Scoring

- Please login our competition page at Kaggle
  - https://www.kaggle.com/t/46f5a4ea8bed4a59bd1ba632226adea0
  - **Your team name is ID_Name**
    - M123456_陳冠宇
  - The evaluation measure is **MAP@5000**
  - The maximum number of daily submissions is 20
  - The **hard** deadline is 12/10 23:59am
    - You point is depended on your performance on the **private** leaderboard!
    - $YourScore = \frac{YourMAP - BaselineMAP}{HighestMAP - BaselineMAP} \times 13\%$

Homework 5 – Warning!!

#	Team Name	Notebook	Team Members	Score
	Baseline: Rocchio			0.52495
	FYI: BM25 (k1=0.8 b=0.7)			0.48874
	FYI: ULM			0.41602
	FYI: VSM			0.36969

- Please follow our rules
  - **Don’t cheat yourself, your friends, and me!**
  - **Don’t create multiple accounts!**
  - Implement the IR system by YOUSELF!
    - Enjoy the Information Retrieval Methods

Questions?

kychen@mail.ntust.edu.tw