Search Results Diversification

Kuan-Yu Chen (陳冠宇)

2019/11/01 @ TR-514, NTUST

Review

- Latent Semantic Analysis
- Topic Models
  - Probabilistic Latent Semantic Analysis
  - Latent Dirichlet Allocation

About HW3.

- The Expectation-Maximization algorithm
  - E-step

$$P(T_k|w_i,d_j)=\frac{P(w_i|T_k)P(T_k|d_j)}{\sum_{k=1}^{K}P(w_i|T_k)P(T_k|d_j)}$$

  - M-step

$$P(w_i|T_k)=\frac{\sum_{d_j\in\mathbf{D}}c(w_i,d_j)P(T_k|w_i,d_j)}{\sum_{i'=1}^{|V|}\sum_{d_j\in\mathbf{D}}c(w_{i'},d_j)P(T_k|w_{i'},d_j)}$$

$$P(T_k|d_j)=\frac{\sum_{i=1}^{|V|}c(w_i,d_j)P(T_k|w_i,d_j)}{\sum_{i'=1}^{|V|}c(w_{i'},d_j)}=\frac{\sum_{i=1}^{|V|}c(w_i,d_j)P(T_k|w_i,d_j)}{|d_j|}$$

  - $P(w_i|T_k)$ and $P(T_k|d_j)$ are random initial with two constrains
    - $\sum_{w_i\in V}P(w_i|T_k)=1$, for every topic T_k
    - $\sum_{k=1}^{K}P(T_k|d_j)=1$, for every document d_j

About HW3..

- The probability $P(q|d_j)$ should be calculated in log domain

$$P(q|d_j)=\prod_{i=1}^{|q|}\left[\alpha\cdot P(w_i|d_j)+\beta\cdot\left(\sum_{k=1}^{K}P(w_i|T_k)P(T_k|d_j)\right)+(1-\alpha-\beta)\cdot P_{BG}(w_i)\right]$$

$$\log P(q|d_j)=\sum_{i=1}^{|q|}\log\left[\alpha\cdot P(w_i|d_j)+\beta\cdot\left(\sum_{k=1}^{K}P(w_i|T_k)P(T_k|d_j)\right)+(1-\alpha-\beta)\cdot P_{BG}(w_i)\right]$$

$$=\sum_{i=1}^{|q|}\left\{[\log\alpha+\log P(w_i|d_j)]\oplus\left[\log\beta+\log\left(\sum_{k=1}^{K}P(w_i|T_k)P(T_k|d_j)\right)\right]\oplus[\log(1-\alpha-\beta)+\log P_{BG}(w_i)]\right\}$$

numpy.logaddexp

`numpy.logaddexp`(*x1, x2, /, out=None, *, where=True, casting='same_kind', order='K', dtype=None, subok=True[, signature, extobj]*) = <ufunc 'logaddexp'>

Logarithm of the sum of exponentiations of the inputs.

Calculates `log(exp(x1) + exp(x2))`. This function is useful in statistics where the calculated probabilities of events may be so small as to exceed the range of normal floating point numbers. In such cases the logarithm of the calculated probability is stored. This function allows adding probabilities stored in such a fashion.

Parameters:

x1, x2 : *array_like*
Input values.

out : *ndarray, None, or tuple of ndarray and None, optional*
A location into which the result is stored. If provided, it must have a shape that the inputs broadcast to. If not provided or *None*, a freshly-allocated array is returned. A tuple (possible only as a keyword argument) must have length equal to the number of outputs.

where : *array_like, optional*
Values of True indicate to calculate the ufunc at that position, values of False indicate to leave the value in the output alone.

****kwargs**
For other keyword-only arguments, see the ufunc docs.

Returns:

result : *ndarray*
Logarithm of `exp(x1) + exp(x2)`.

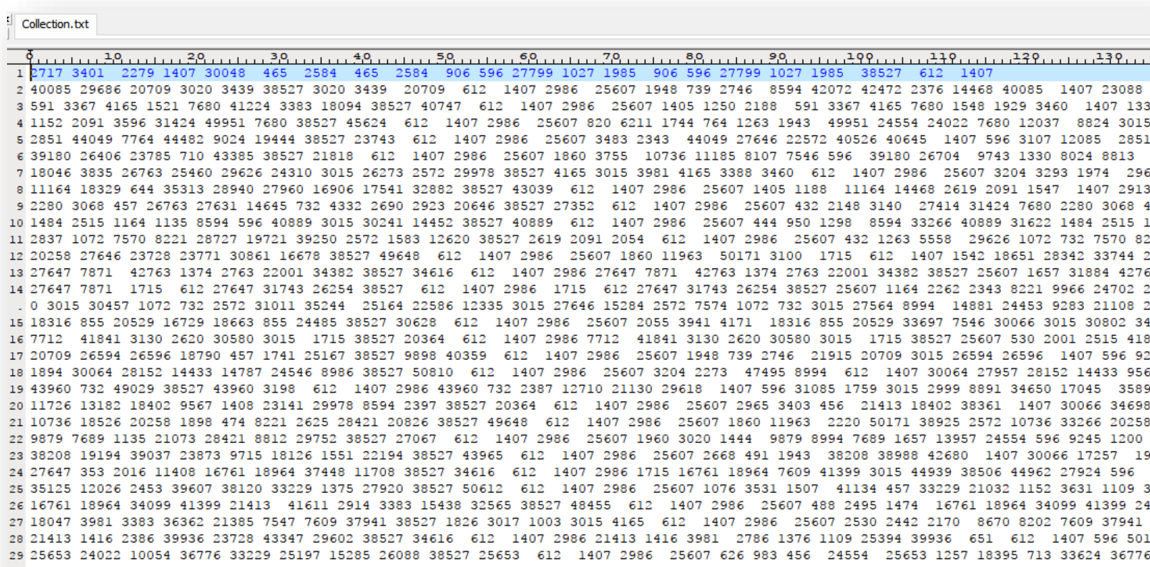
About HW3...

#	Team Name	Notebook	Team Members	Score	Entries
1	B			0.52528	55
2	ye			0.51220	43
3	M			0.50242	55
	ULM			0.50229	
4	te			0.49764	9
5	B			0.49679	47
6	M			0.49538	27
7	Pe			0.49281	20
8	M			0.49250	1
9	B			0.49096	16
10	M			0.45864	75
11	M			0.44721	61
	VSM			0.26967	
12	M			0.22473	15
13	M			0.22344	7
14	M			0.19460	6
15	te			0.10610	3
16	M			0.05606	8
17	al			0.03981	13
18	M			0.02019	1

About HW3…

- A Collection.txt is released
  - Each line is a document
  - About 18 thousand documents in total

- Train a PLSA model on the corpus with 128 topics, we can achieve 0.53x by tuning the parameters easily
  - But the fold-in strategy is needed!



```
Collection.txt
2717 3401  2279 1407 30048  465  2584  465  2584  906 596 27799 1027 1985  906 596 27799 1027 1985  38527  612  1407
40085 29686 20709 3020 3439 38527 3020 3439  20709  612  1407 2986  25607 1948 739 2746  8594 42072 42472 2376 14468 40085  1407 23088 2
591 3367 4165 1521 7680 41224 3383 18094 38527 40747  612  1407 2986  25607 1405 1250 2188  591 3367 4165 7680 1548 1929 3460  1407 1332
1152 2091 3596 31424 49951 7680 38527 45624  612  1407 2986  25607 820 6211 1744 764 1263 1943  49951 24554 24022 7680 12037  8824 3015
2851 44049 7764 44482 9024 19444 38527 23743  612  1407 2986  25607 3483 2343  44049 27646 22572 40526 40645  1407 596 3107 12085  2851
39180 26406 23785 710 43385 38527 21818  612  1407 2986  25607 1860 3755  10736 11185 8107 7546 596  39180 26704  9743 1330 8024 8813  2
18046 3835 26763 25460 29626 24310 3015 26273 2572 29978 38527 4165 3015 3981 4165 3388 3460  612  1407 2986  25607 3204 3293 1974  2962
11164 18329 644 35313 28940 27960 16906 17541 32882 38527 43039  612  1407 2986  25607 1405 1188  11164 14468 2619 2091 1547  1407 2913
2280 3068 457 26763 27631 14645 732 4332 2690 2923 20646 38527 27352  612  1407 2986  25607 432 2148 3140  27414 31424 7680 2280 3068 45
1484 2515 1164 1135 8594 596 40889 3015 30241 14452 38527 40889  612  1407 2986  25607 444 950 1298  8594 33266 40889 31622 1484 2515 11
2837 1072 7570 8221 28727 19721 39250 2572 1583 12620 38527 2619 2091 2054  612  1407 2986  25607 432 1263 5558  29626 1072 732 7570 822
20258 27646 23728 23771 30861 16678 38527 49648  612  1407 2986  25607 1860 11963  50171 3100  1715  612  1407 1542 18651 28342 33744 29
27647 7871  42763 1374 2763 22001 34382 38527 34616  612  1407 2986 27647 7871  42763 1374 2763 22001 34382 38527 25607 1657 31884 42763
27647 7871  1715  612 27647 31743 26254 38527  612  1407 2986  1715  612 27647 31743 26254 38527 25607 1164 2262 2343 8221 9966 24702 29
0 3015 30457 1072 732 2572 31011 35244  25164 22586 12335 3015 27646 15284 2572 7574 1072 732 3015 27564 8994  14881 24453 9283 21108 25
18316 855 20529 16729 18663 855 24485 38527 30628  612  1407 2986  25607 2055 3941 4171  18316 855 20529 33697 7546 30066 3015 30802 346
7712  41841 3130 2620 30580 3015  1715 38527 20364  612  1407 2986 7712  41841 3130 2620 30580 3015  1715 38527 25607 530 2001 2515 4184
20709 26594 26596 18790 457 1741 25167 38527 9898 40359  612  1407 2986  25607 1948 739 2746  21915 20709 3015 26594 26596  1407 596 924
1894 30064 28152 14433 14787 24546 8986 38527 50810  612  1407 2986  25607 3204 2273  47495 8994  612  1407 30064 27957 28152 14433 9561
43960 732 49029 38527 43960 3198  612  1407 2986 43960 732 2387 12710 21130 29618  1407 596 31085 1759 3015 2999 8891 34650 17045  35898
11726 13182 18402 9567 1408 23141 29978 8594 2397 38527 20364  612  1407 2986  25607 2965 3403 456  21413 18402 38361  1407 30066 34698
10736 18526 20258 1898 474 8221 2625 28421 20826 38527 49648  612  1407 2986  25607 1860 11963  2220 50171 38925 2572 10736 33266 20258
9879 7689 1135 21073 28421 8812 29752 38527 27067  612  1407 2986  25607 1960 3020 1444  9879 8994 7689 1657 13957 24554 596 9245 1200
38208 19194 39037 23873 9715 18126 1551 22194 38527 43965  612  1407 2986  25607 2668 491 1943  38208 38988 42680  1407 30066 17257  191
27647 353 2016 11408 16761 18964 37448 11708 38527 34616  612  1407 2986 1715 16761 18964 7609 41399 3015 44939 38506 44962 27924 596  6
35125 12026 2453 39607 38120 33229 1375 27920 38527 50612  612  1407 2986  25607 1076 3531 1507  41134 457 33229 21032 1152 3631 1109 33
16761 18964 34099 41399 21413  41611 2914 3383 15438 32565 38527 48455  612  1407 2986  25607 488 2495 1474  16761 18964 34099 41399 245
18047 3981 3383 36362 21385 7547 7609 37941 38527 1826 3017 1003 3015 4165  612  1407 2986  25607 2530 2442 2170  8670 8202 7609 37941 1
21413 1416 2386 39936 23728 43347 29602 38527 34616  612  1407 2986 21413 1416 3981  2786 1376 1109 25394 39936  651  612  1407 596 5016
25653 24022 10054 36776 33229 25197 15285 26088 38527 25653  612  1407 2986  25607 626 983 456  24554  25653 1257 18395 713 33624 36776
15912 17639 16665 24022 10054 18586 25197 33229 26088 38527 17639  612  1407 2986  25607 2626 22715  24554 18896  1278 820  662  1788 17
```

About Final Project

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

- Do you have GPU units?
  - We have to make sure you can do HW5 and final project

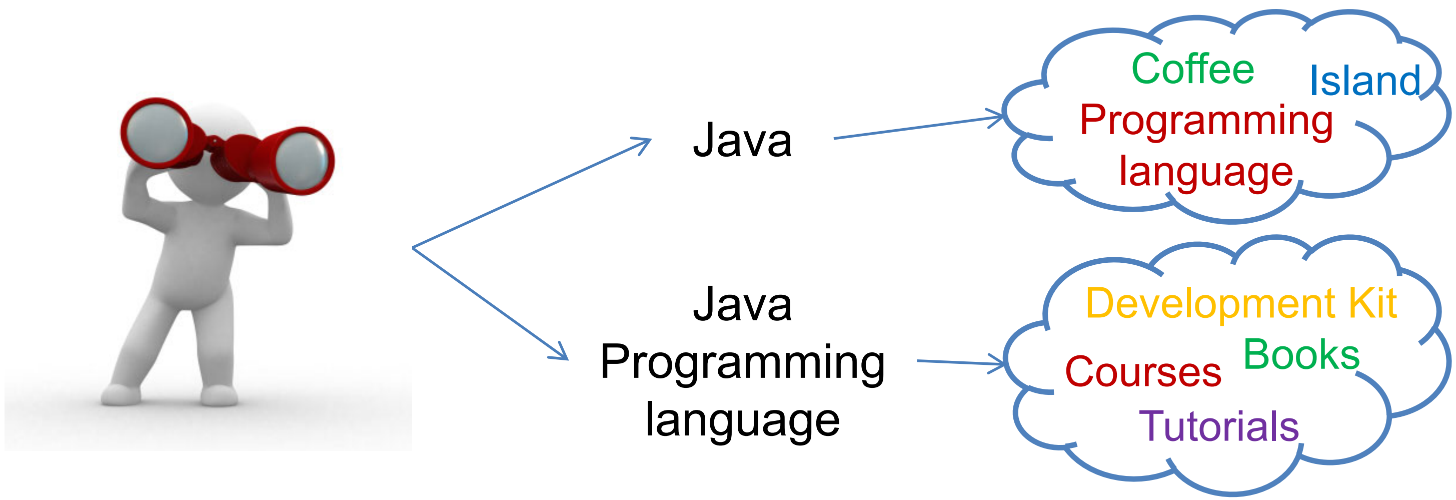
Resource

- Conferences
  - ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR)
  - International Joint Conferences on Artificial Intelligence (IJCAI)
  - ACM Conference on Information Knowledge Management (CIKM)
  - Annual Meeting of the Association for Computational Linguistics (ACL)
  - International Conference on Learning Representations (ICLR)

- Journals
  - Journal of the American Society for Information Science (JASIS)
  - ACM Transactions on Information Systems (TOIS)
  - Information Processing and Management (IP&M)
  - ACM Transactions on Asian Language Information Processing (TALIP)
  - Information Retrieval Journal (IRJ)

Introduction – What's going on?

- Traditional retrieval functions ignore the relations among returned documents
  - Top ranked documents may contain relevant yet **redundant information**
  - In order to maximize the satisfaction of different search users, it is necessary to diversify search results
  - Search results diversification can play an initial step for many search system

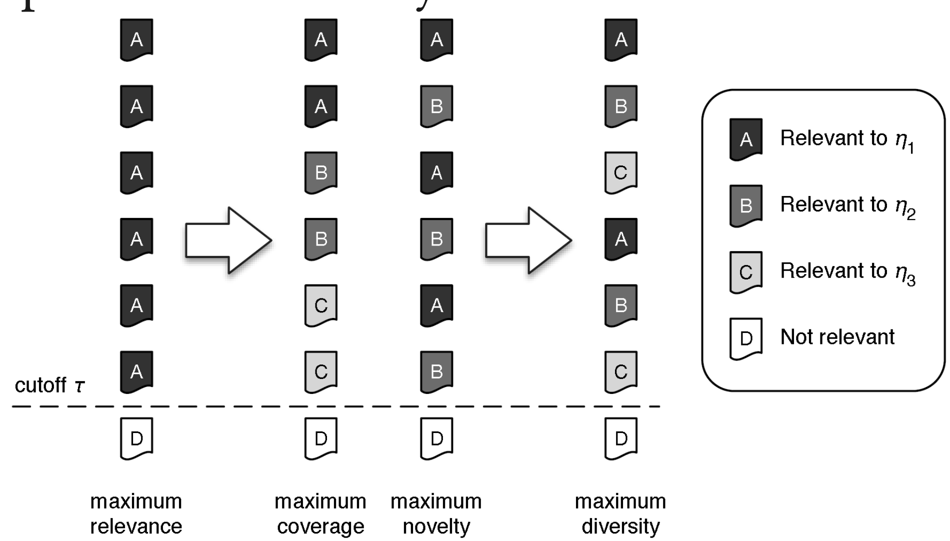


Relevance, Coverage, Novelty, & Diversity

- Most of the retrieval models assume that the relevance of a document can be estimated with **certainty** and **independently** of the estimation of the other retrieved documents
  - Ambiguous queries
    - Ensuring a high **coverage** of the possible information needs
  - Redundancy results
    - Ensuring the retrieved documents provide a high **novelty**

Relevance, Coverage, Novelty, & Diversity

- Coverage and novelty can also be conflicting objectives
  - A ranking with maximum coverage may not attain maximum novelty
    - Although covering all information needs, the ranking may place all documents covering a particular need ahead than others
  - A ranking with maximum novelty may not attain maximum coverage
    - Although covering each need as early as possible in the ranking, not all possible needs may be covered



Introduction – Various Modeling

- Many diversification methods have been proposed
  - balance the relevance and the redundancy: MMR
  - distinguish previous topics and new coming: SMM
  - language modeling approach: WUME
  - probabilistic framework: xQuAD

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

Introduction – Notations

Symbol	Description
q	A given query
a_k^q	Sub-queries (aspect), $q = \{a_1^q, \cdots, a_K^q\}$
K	Number of sub-queries
R	The user's information need
$\mathbf{D}$	A set of documents, $\mathbf{D} = \{d_1, \cdots, d_{\lvert\mathbf{D}\rvert}\}$
$\widetilde{\mathbf{D}}$	A subset of documents which already selected by new method, $\widetilde{\mathbf{D}} = \{\tilde{d}_1, \cdots, \tilde{d}_{\lvert\widetilde{\mathbf{D}}\rvert}\}$

Maximal Marginal Relevance – MMR

- MMR motivated the need for "relevant novelty" as a potentially superior criterion
  - An approximation to measuring relevant novelty is to **measure relevance and novelty independently**

- "Marginal Relevance" cab be regarded as the metric
  - A document has high **marginal relevance** if it is both relevant to the query and contains minimal similarity to previously selected documents

$$Div_{MMR}(d,q) = -\max_{\tilde{d}\in\tilde{\mathbf{D}}} sim(d,\tilde{d})$$

$$d^* = \operatorname*{argmax}_{d\in\mathbf{D}} \lambda \cdot Rel(d,q) - (1-\lambda)\cdot \max_{\tilde{d}\in\tilde{\mathbf{D}}} sim(d,\tilde{d})$$

Simple Mixture Model – SMM

- Given the observed new document, we estimate the mixing weight for the background model θ_{BG} and the previous topic model θ_T
  - The simplest previous topic model can be modeled as:

$$P(w|\theta_T) = \sum_{\tilde{d} \in \tilde{\mathbf{D}}} \frac{1}{N} P(w|\tilde{d})$$

  - The mixture weight for the background model can serve as a measure of novelty or redundancy

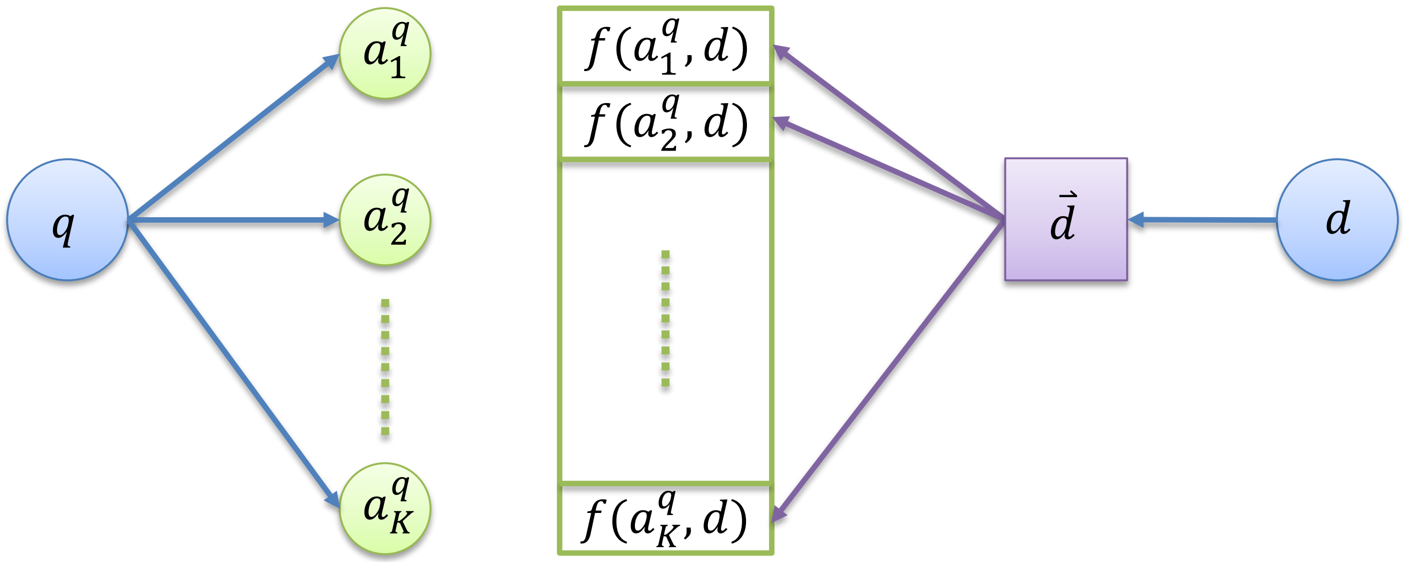
$$L(\beta|d, \theta_{BG}, \theta_T) = \prod_{w \in V} \left((1-\beta) \cdot P(w|\theta_T) + \beta \cdot P(w|\theta_{BG})\right)^{c(w,d)}$$

$$d^* = \underset{d \in \mathbf{D}}{\operatorname{argmax}} \, \lambda \cdot Rel(d,q) + (1-\lambda) \cdot \underset{\beta}{\operatorname{argmax}} \, L(\beta|d, \theta_{BG}, \theta_T)$$

$1 - \beta = P(\theta_T|d)$
$\beta = P(\theta_{BG}|d)$

Explicit MMR – xMMR

- For a given query with its sub-queries, each document can be represented by a K-dimensional vector over sub-queries

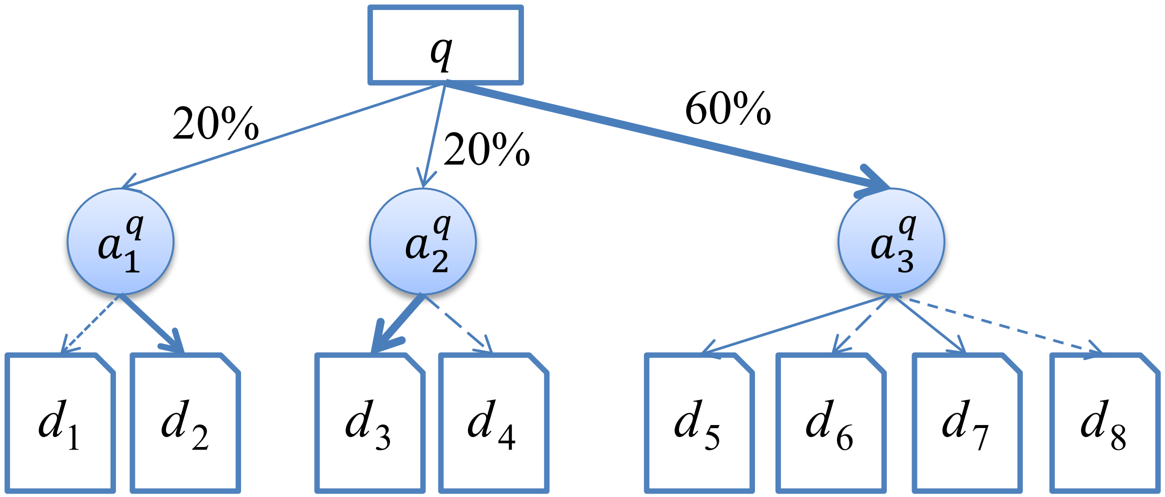


$$f(a^q, d) \equiv P(d|a^q) \quad f(a^q, d) \equiv cos(a^q, d)$$

- By doing so, the redundancy score can be defined by considering sub-queries

$$d^* = \underset{d \in \mathbf{D}}{\operatorname{argmax}} \lambda \cdot Rel(d, q) - (1 - \lambda) \cdot \max_{\tilde{d} \in \tilde{\mathbf{D}}} sim(\vec{d}, \vec{\tilde{d}})$$

WUME – Motivation



- There are three sub-queries under the given query $q = \{a_1^q, a_2^q, a_3^q\}$, and web documents $\mathbf{D} = \{d_1, \cdots, d_8\}$

- Although d_3 is more relevant to one of the sub-query a_2^q than d_5 to a_3^q, given that a_2^q attracts less user interest than a_3^q, d_3 **should** still be ranked lower than d_5

WUME

- WUME formalize the diversification method as:
  - Given a query Q, the probability that a retrieved document meets user's information need R can be written as:

$$P(R|d) = \frac{P(R)P(d|R)}{P(d)} \propto P(d|R)$$

  - Take sub-query information into consideration:

$$P(d|R) \approx P(d|q) = \sum_{k=1}^{K} P\left(d\middle|a_k^q, q\right)P(a_k^q|q)$$

Google Insights for Search or Wikipedia

  - Finally, the ranking function becomes:

$$d^* = \underset{d \in \mathbf{D}}{\operatorname{argmax}}\, \lambda \cdot Rel(d, q) + (1 - \lambda) \cdot \sum_{k=1}^{K} P\left(d\middle|a_k^q, q\right)P(a_k^q|q)$$

eXplicit Query Aspect Diversification

- xQuAD: eXplicit Query Aspect Diversification
  - When given an ambiguous query, xQuAD builds a new ranked list by:

$$d^* = \underset{d \in \mathbf{D}}{\operatorname{argmax}} \lambda \cdot P(d|q) + (1-\lambda) \cdot P\left(d, \overline{\widetilde{\mathbf{D}}}|q\right)$$

- $P(d|q)$ is the likelihood of document d being observed given the initial query

  The probability can be regarded as modeling ***relevance***

- $P\left(d, \overline{\widetilde{\mathbf{D}}}|q\right)$ is the likelihood of observing this document but not the documents already in $\widetilde{\mathbf{D}}$

  The probability can be regarded as modeling ***diversity***

xQuAD – 1

- In order to derive $P\left(d,\bar{\bar{\mathbf{D}}}|q\right)$, xQuAD explicitly consider the possibly several aspects underlying the initial query as a set of sub-queries

  - By assuming $\sum_{k=1}^{K} P(a_k^q|q) = 1$, xQuAD calculates $P\left(d,\bar{\bar{\mathbf{D}}}|q\right)$ by considering sub-queries:

$$P\left(d,\bar{\bar{\mathbf{D}}}|q\right) = \sum_{k=1}^{K} P\left(d,\bar{\bar{\mathbf{D}}}|a_k^q\right)P(a_k^q|q)$$

  - Further, $P\left(d,\bar{\bar{\mathbf{D}}}|a_k^q\right)$ can be broken down by independent assumption:

$$P\left(d,\bar{\bar{\mathbf{D}}}|a_k^q\right) = \overset{\text{coverage}}{P\left(d|a_k^q\right)}\overset{\text{novelty}}{P\left(\bar{\bar{\mathbf{D}}}|a_k^q\right)}$$

xQuAD – 2

- xQuAD also assumes that the relevance of each document in $\widetilde{\mathbf{D}}$ to a given sub-query a_k^q is independent

$$P(\bar{\widetilde{\mathbf{D}}}|a_k^q) = P\left(\bar{\tilde{d}}_1, \cdots, \bar{\tilde{d}}_{|\widetilde{\mathbf{D}}|} \middle| a_k^q\right) = \prod_{\tilde{d}_n \in \widetilde{\mathbf{D}}} P(\bar{\tilde{d}}_n | a_k^q) = \prod_{\tilde{d}_n \in \widetilde{\mathbf{D}}} (1 - P(\tilde{d}_n | a_k^q))$$

- To sum up, xQuAD suggests that:

$$Div_{xQuAD}(d, q) = \sum_{k=1}^{K} \underbrace{P(a_k^q|q)}_{\text{the importance of } a_k^q} \underbrace{P(d|a_k^q)}_{\text{the relevance of } d \text{ to } a_k^q} \prod_{\tilde{d}_n \in \widetilde{\mathbf{D}}} \underbrace{(1 - P(\tilde{d}_n|a_k^q))}_{\text{the satisfaction degree } a_k^q}$$

  - Instead of comparing a document d to all documents already selected in $\widetilde{\mathbf{D}}$, xQuAD estimates the utility of any document satisfying the sub-query a_k^q, given how well it is already satisfied by the documents in $\widetilde{\mathbf{D}}$

Analytical Comparisons

- Diversity Modeling:
  - MMR and SMM **implicitly** model the diversity through document similarities
  - xMMR, WUME and xQuAD **explicitly** model the diversity through the coverage of query subtopics

- Document Dependency:
  - WUME assumes that the diversity score of a document is independent of other documents
  - The other three methods assume that the diversity score depends on the previously selected documents

General Framework

- Most of these methods **iteratively select** the document that is not only **relevant** to the query but also **diversified** to cover more query subtopics, explicitly or implicitly

- All of methods fit into a general framework that iteratively selects with the highest relevance and diversity scores:

$$d^* = \underset{d \in \mathbf{D}}{\operatorname{argmax}}\, \lambda \cdot Rel(d, q) + (1 - \lambda) \cdot Div(d, q)$$

Experimental Results

	TREC09 result		TREC10 result	
	α-nDCG@20	α-nDCG@100	α-nDCG@20	α-nDCG@100
MMR^*	0.365	0.427	0.344	0.415
$WUME^*$	0.479	0.546	0.579	0.630
$xQuAD^*$	0.482	0.550	0.588	0.636

- All the parameters in each method are set to the optimum values
  - Both xQuAD and WUME perform significantly better than MMR
    - Using explicit sub-queries in diversification is better
    - The performances of xQuADand WUME are not significantly different

Conclusions

- The experiment result shows that the explicit sub-query modeling and sub-query importance penalization strategies perform better

- It is interesting to find that how the sub-queries affect the overall performance

- Finally, we can think about that what's the difference between sub-queries and latent topics?
  - Supervised v.s. Unsupervised?

$$P(d|R) \approx P(d|q) = \sum_{k=1}^{K} P\left(d \middle| a_k^q, q\right) P(a_k^q|q)$$

- Beyond relevance or another relevance?

Questions?

kychen@mail.ntust.edu.tw