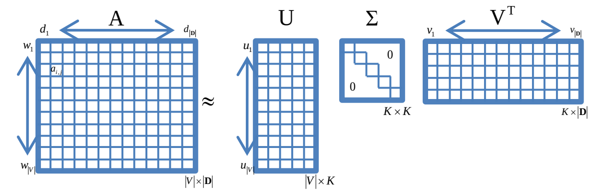
Search Results Diversification

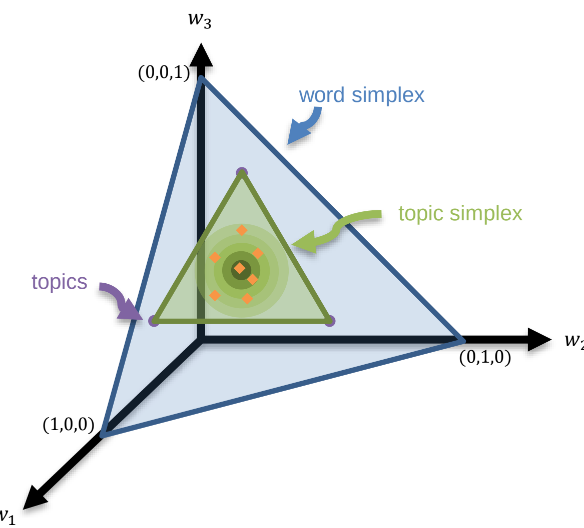
Kuan-Yu Chen (陳冠宇)

2020/11/20 @ TR-313, NTUST

Review.



- Latent Semantic Analysis
  - Vector representation for word is Σu_i^{T}
  - Vector representation for document is Σv_j^{T}



- Statistical Topic Models
  - Probabilistic Latent Semantic Analysis
    - $\mathcal{L} = \prod_{w_i \in V} \prod_{d_j \in \mathbf{D}} P(w_i, d_j)^{c(w_i, d_j)}$
  - Latent Dirichlet Allocation
    - $\mathcal{L} = \prod_{d_j \in \mathbf{D}} \int P(\theta_{d_j}|\alpha) \left(\prod_{i=1}^{|d_j|} \left(\sum_{k=1}^{K} P(w_i|T_k, \beta) P\left(T_k \middle| \theta_{d_j}\right) \right) \right) d\theta_{d_j}$

Review..

- 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

Review…

- About the M-step

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

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

Review….

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

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

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

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)`.

Homework 4.

- The evaluation measure is MAP@1000
  - The **hard** deadline is 11/26 23:59
  - You point is depended on your performance on the **private** leaderboard!
    - 50 public queries and 50 private queries

$$YourScore = 5 + \frac{YourMAP - BaselineMAP}{HighestMAP - BaselineMAP} \times 8$$

  - 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 4..

#	Team Name
1	宮澤賢治
2	康帕內魯拉
	FYI: with 32 topics
3	Test
4	M10815048_張晏銘
5	Enilesab
6	TESTT1235
7	B10615034_黃柏翰
8	Alice
9	80847002S_羅天宏
10	M10915100_郭智威
	FYI: with 8 topics
11	快樂ㄟ甘蔗man
12	hongyun
	Baseline 0.4P(w|D) 0.6P(w|BG)
13	在小世界裡畫出最耀眼的大幸福
14	trytrysee
15	騙人的吧

#	△pub	Team Name
1	▲1	康帕內魯拉
2	▲3	Enilesab
3	—	Test
4	—	M10815048_張晏銘
		FYI: with 32 topics
5	▼4	宮澤賢治
6	▲1	B10615034_黃柏翰
7	▲3	M10915100_郭智威
8	▼2	TESTT1235
9	—	80847002S_羅天宏
10	▼2	Alice
		FYI: with 8 topics
11	▲1	hongyun
12	▼1	快樂ㄟ甘蔗man
		Baseline 0.4P(w|D) 0.6P(w|BG)
13	—	在小世界裡畫出最耀眼的大幸福
14	—	trytrysee
15	—	騙人的吧

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 HW6 and/or final project

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	

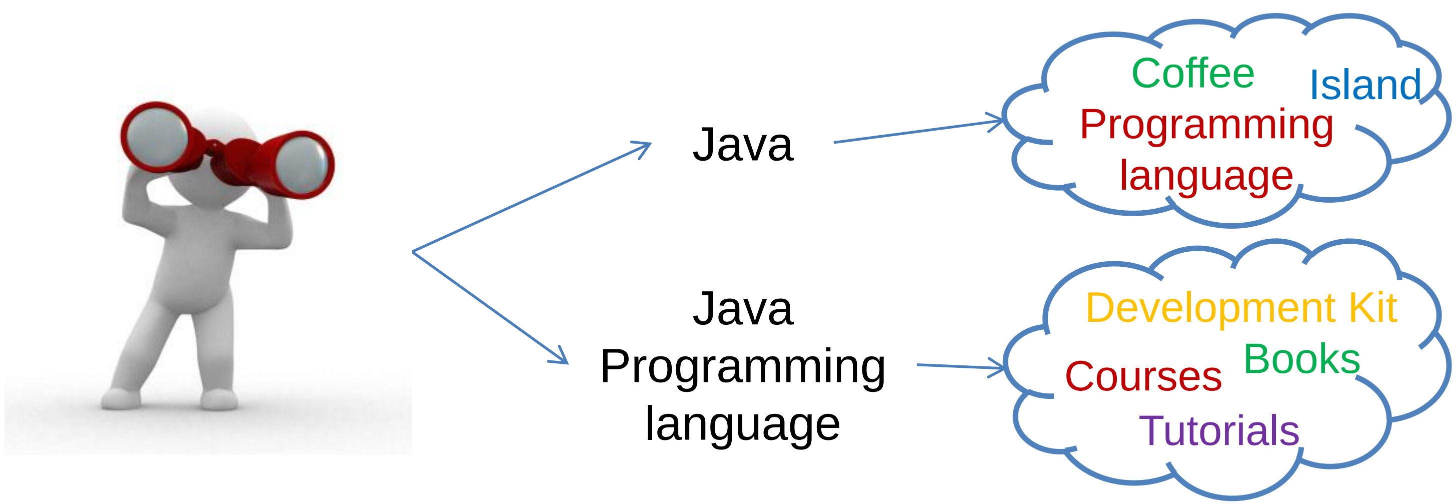
Resource

- 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

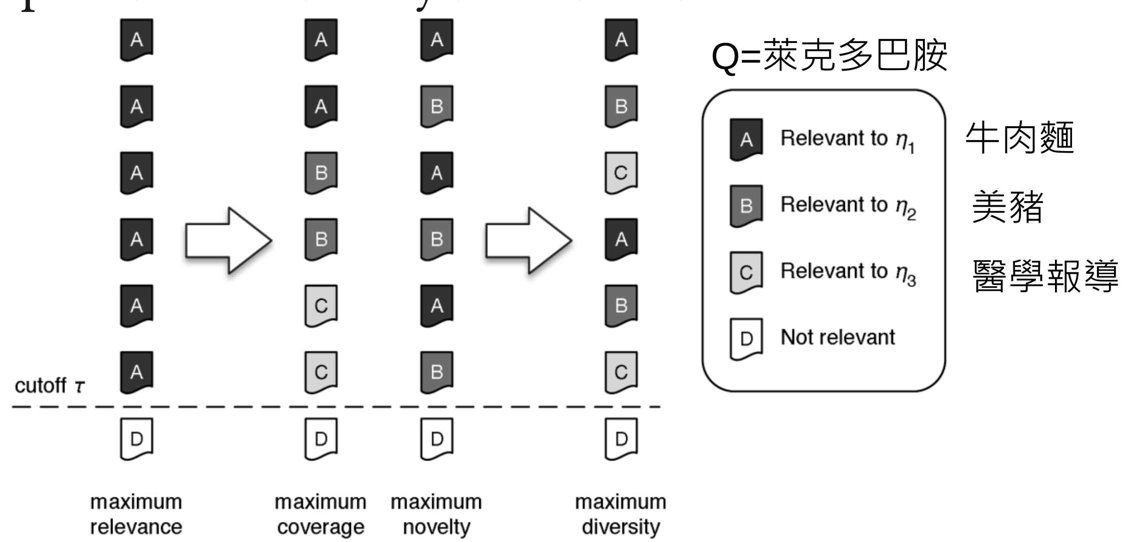


Relevance, Coverage, Novelty, & Diversity

- 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 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



Example.

Google

萊克多巴胺

全部 新聞 圖片 影片 購物 更多 設定 工具

約有 3,030,000 項結果 (搜尋時間：0.37 秒)

綜合報導

廣告 · www.twreporter.org/

美豬美牛的10大爭議 - 瘦肉精是什麼？安全嗎？

食安風險未明的萊克多巴胺，歐盟至今仍維持禁令，瘦肉精對人體有無危害？美國為何不斷要求開放？對台灣畜牧產業又會有什麼影響？《報導者》整理出10大爭議，解析全面開放美豬牛的影響。 類型: 深度在地報導, 開啟新聞革命, 非營利媒體。

化學定義

zh.wikipedia.org › zh-tw › 萊克多巴胺

萊克多巴胺- 维基百科，自由的百科全书

萊克多巴胺（Ractopamine）是一種β促效劑（β-agonist）藥物，用以助長豬、牛、火雞生出瘦肉，減少體脂肪。是瘦肉精中最常见的一種，其肉品殘留毒性遠低於 ...

化学式： $C_{18}H_{23}NO_3$ 溶解性（水）： 4100 mg/L

CAS号： 97825-25-7 SMILES： 显示▼: OC(c1ccc(O)cc1)CNC(...

瘦肉精 · 臺灣進口美國牛肉問題 · 雙盲

周刊科普

www.businessweekly.com.tw › 焦點 › 時事分析

萊克多巴胺有多可怕？別讓「不了解」蒙蔽了判斷力｜商周

2020年9月17日 — 然而萊克多巴胺究竟是什麼？這其實是歷經多年嚴謹研究，為了提升牲口肉品產量與安全使用所發展而成的飼料添加物，可減少飼料用量 ...

health.tvbs.com.tw › 營養

瘦肉精是什麼？吃了含萊克多巴胺的豬肉會怎樣？有方法排毒嗎 ...

2020年8月31日 — 政府宣布明年1月1日起，將放寬含瘦肉精的美豬美牛進口，引起國人恐慌。台北榮民總醫院臨床毒物與職業醫學科主任楊振昌提醒，萊克多巴胺是 ...

相關新聞

焦點新聞

新北率全國之先豬肉萊克多巴胺「零檢出」入午餐契約| 台灣英文新聞| 2020/11/18

Taiwan News · 1 天前

萊克多巴胺

萊克多巴胺是一種β促效劑藥物，用以助長豬、牛、火雞生出瘦肉，減少體脂肪。是瘦肉精中最常見的一種，其肉品殘留毒性遠低於具有相同功能的其他動物飼料添加物。美國謊稱在其測定的容許殘留量下合法使用，將不會對人類造成中毒或短期危害。目前的實驗數據確定會對人體產生其他副作用，人體長期攝取殘留的萊克多巴胺會造成健康問題！ 維基百科

分子式： $C_{18}H_{23}NO_3$

PubChem CID： 56052

溶解性（水）： 4100 mg/L

化學式： $C_{18}H_{23}NO_3$

CAS號： 97825-25-7

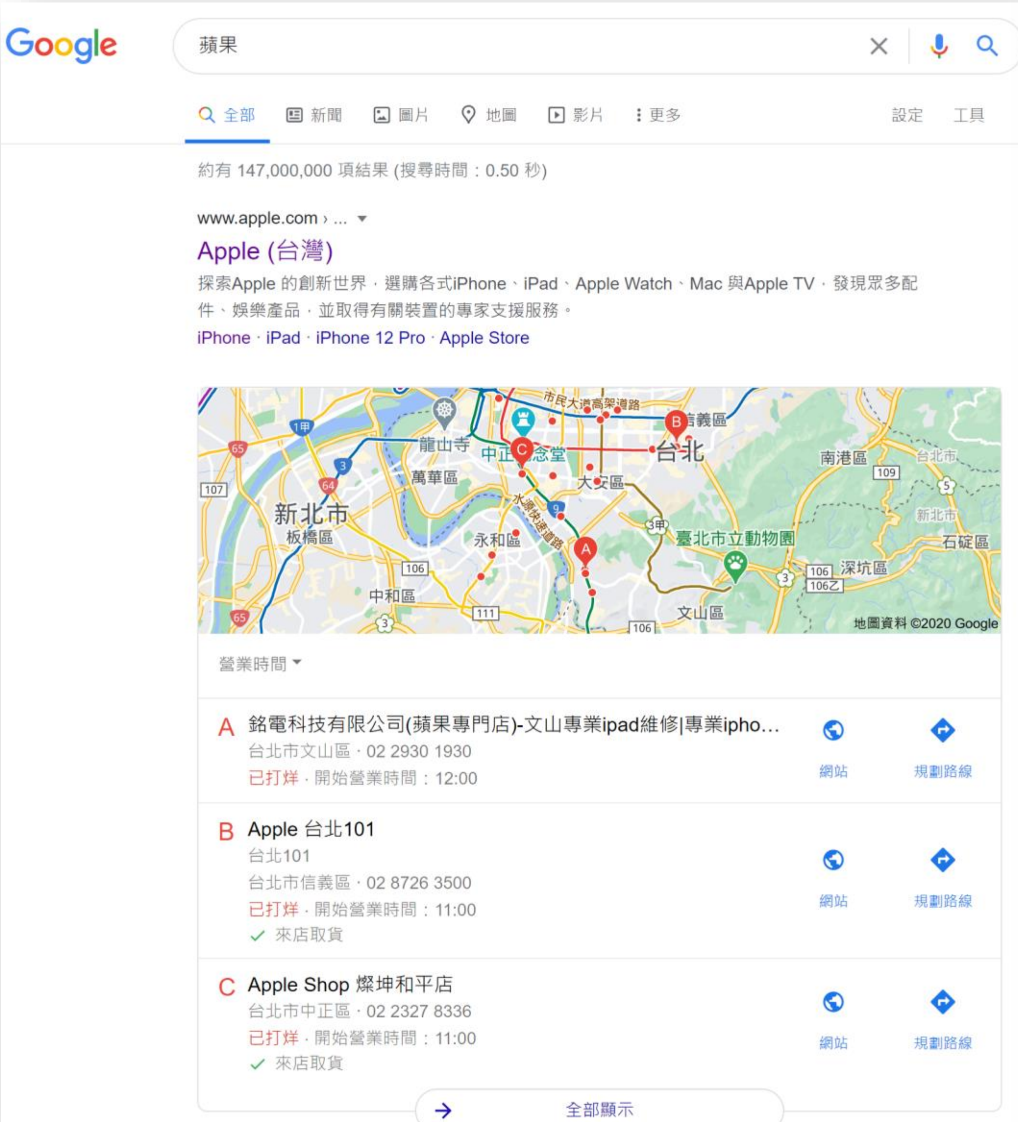
SMILES： 顯示▼: OC(c1ccc(O)cc1)CNC(C)CCc2ccc(O)cc2

其他人也搜尋了

Tretoquinol 偶氮甲醯胺 Procaterol

意見回饋

Example..



Google

蘋果

全部 新聞 圖片 地圖 影片 更多 設定 工具

約有 147,000,000 項結果 (搜尋時間：0.50 秒)

www.apple.com › ...

Apple (台灣)

探索Apple 的創新世界，選購各式iPhone、iPad、Apple Watch、Mac 與Apple TV，發現眾多配件、娛樂產品，並取得有關裝置的專家支援服務。

iPhone · iPad · iPhone 12 Pro · Apple Store

營業時間

A 銘電科技有限公司(蘋果專門店)-文山專業ipad維修|專業ipho...
台北市文山區 · 02 2930 1930
已打烊 · 開始營業時間：12:00
網站 規劃路線

B Apple 台北101
台北101
台北市信義區 · 02 8726 3500
已打烊 · 開始營業時間：11:00
來店取貨
網站 規劃路線

C Apple Shop 燦坤和平店
台北市中正區 · 02 2327 8336
已打烊 · 開始營業時間：11:00
來店取貨
網站 規劃路線

全部顯示

tw.appledaily.com › ...

首頁 | 蘋果新聞網 | 蘋果日報

提供全面新聞資訊、即時分析，全天候報道本地及全球新聞。

最新 · 娛樂蘋道 · 日報 · 要聞

appledaily.com

蘋果日報: 首頁

蘋果日報網站提供香港蘋果日報、即時新聞、動新聞、要聞港聞、娛樂、兩岸國際、體育、副刊等內容，文字、圖像、影片兼備，為你提供全面而即時的新聞資訊。

zh.wikipedia.org › wiki › 蘋果公司

蘋果公司- 維基百科，自由的百科全书

蘋果公司（英語：Apple Inc.），原稱蘋果电脑公司（英語：Apple Computer, Inc.），是总部位于美国加州庫比蒂諾的跨國科技公司，与亚马逊，谷歌、微软 ...

創辦人：史蒂夫·乔布斯; 斯蒂夫·沃兹尼亚克; ... 營業額：▼ 2601.74億美元（2019）

總資產：▼ 3385.16億美元（2019） 成立：1976年4月1日，44年前

zh.wikipedia.org › zh-tw › 苹果

苹果- 维基百科，自由的百科全书

苹果，又稱柰或林檎，是苹果树（学名：Malus pumila）的果实。苹果樹是薔薇科苹果亚科苹果属植物，為落叶乔木，果實富含矿物质和维生素，是人们最常食用的 ...

科：薔薇科 Rosaceae 界：植物界 Plantae

种：苹果 M. pumila 屬：苹果属 Malus

m.momoshop.com.tw › 品牌旗艦 › Apple

Apple,品牌旗艦- momo購物網

Apple,品牌旗艦,iPhone,iPad,Mac優惠便宜好價格！Apple 蘋果品牌推薦,並提供您完整規格種類的iPhone,iPad,Mac！

news.xfastest.com › others › apple-m1

不僅效能IBM說出蘋果拋棄Intel處理器真相| XFastest News

22 小時前 — IBM AI戰略副總裁Sumit Gupta替蘋果算了一筆帳，其假設蘋果2020年累計出貨860萬台13吋MacBook Pro及540萬台MacBook Air。一顆M1處理 ...

news.cnyes.com › 國際股 › 美股

投行提證據！iPhone 12 正啟動蘋果超級週期| Anue鉅亨- 美股

7 小時前 — 相信這將轉化為蘋果前所未有的升級週期。週四Daniel Ives 重申其觀點。蘋果十月中旬發布四款新iPhone 12，包括5.4 吋iPhone 12 Mini、6.1 吋 ...

technews.tw › 2020/11/18 › app-store-2

蘋果調降App Store 一半佣金，營收百萬美元以下小企業受惠 ...

2 天前 — 蘋果（Apple）的App Store 付費App 與App 內購佣金一向為許多開發者所詬病，因為每次Apple 都會抽佣30%；尤其今年疫情橫行，對於許多 ...

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**

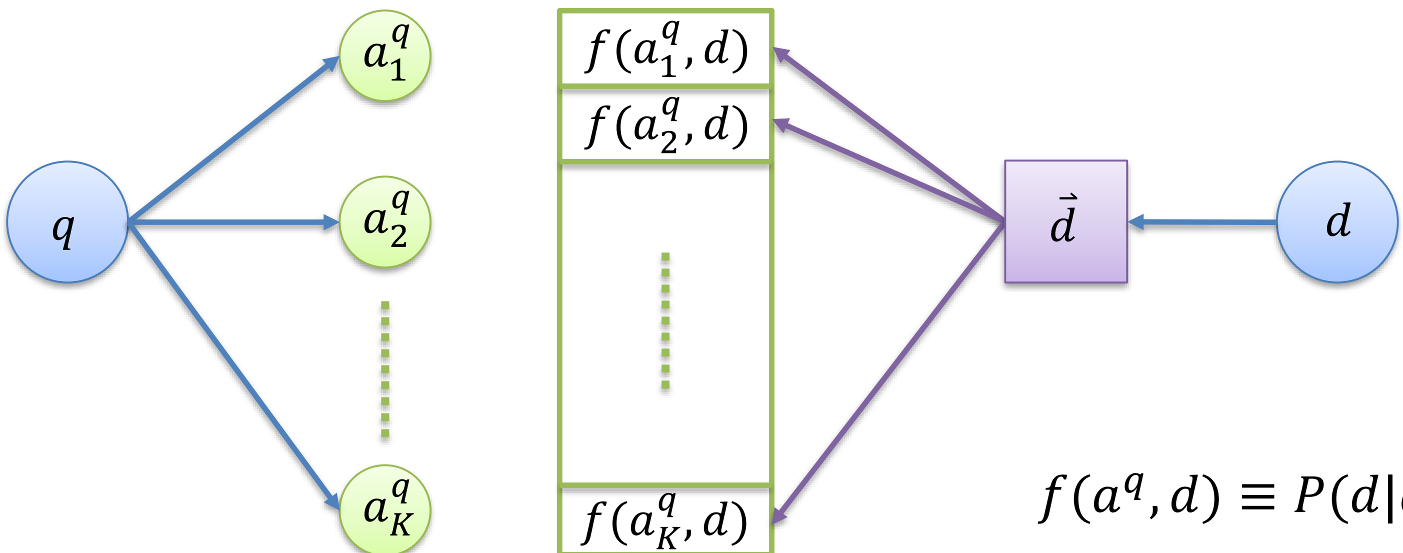
- "Marginal Relevance" cab be regarded as the metric
  - A document has high **marginal relevance** if it is both relevant to the query and contains minimal similarity to previously selected documents

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

$$score(d,q) = \lambda \cdot Rel(d,q) - (1-\lambda) \cdot \max_{\tilde{d}\in\tilde{\mathbf{D}}} sim(d,\tilde{d})$$

Explicit MMR – xMMR

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



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

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

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

$$score(d, q) = \lambda \cdot Rel(d, q) - (1 - \lambda) \cdot \max_{\tilde{d} \in \tilde{\mathbf{D}}} sim(d, \tilde{d})$$

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:

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

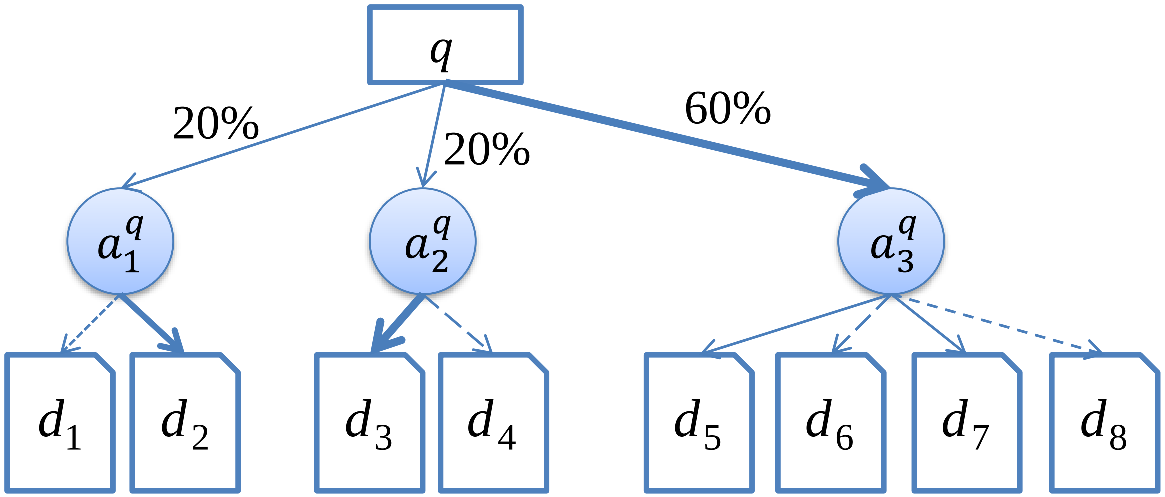
  - The mixture weight for the background model can serve as a measure of novelty or redundancy

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

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

$$score(d,q) = \lambda\cdot Rel(d,q) + (1-\lambda)\cdot\beta$$

WUME – Motivation



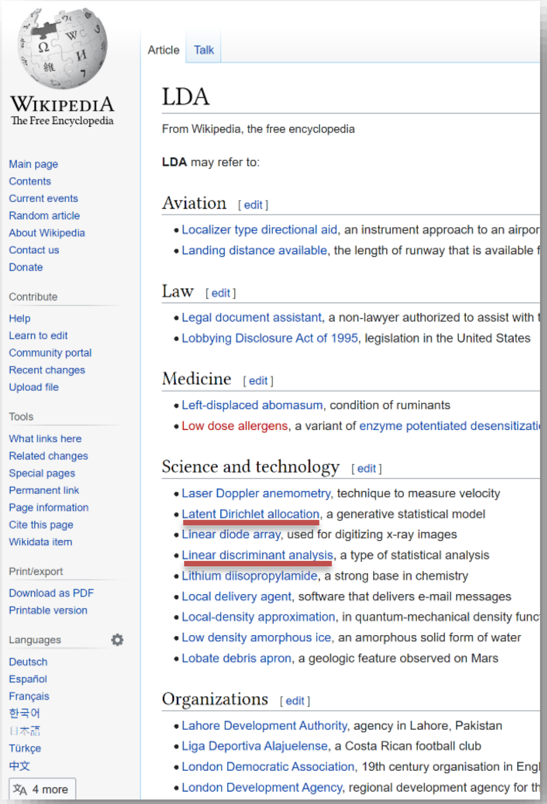
- There are three sub-queries under the given query $q = \{a_1^q, a_2^q, a_3^q\}$, and web documents $\mathbf{D} = \{d_1, \cdots, d_8\}$

- 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:

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

Google Insights for Search or Wikipedia

  - Finally, the ranking function becomes:

$$score(d,q) = \lambda \cdot Rel(d,q) + (1-\lambda) \cdot \sum_{k=1}^{K} P(d|a_k^q, q) P(a_k^q|q)$$

eXplicit Query Aspect Diversification

- xQuAD: eXplicit Query Aspect Diversification
  - When given an ambiguous query, xQuAD estimates the ranking score by:

$$score(d, q) = \lambda \cdot P(d|q) + (1 - \lambda) \cdot P\left(d, \overline{\widetilde{\mathbf{D}}}|q\right)$$

- $P(d|q)$ 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:

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

xQuAD – 2

- For $P\left(\bar{\widetilde{\mathbf{D}}}|a_k^q\right)$, xQuAD assumes that the relevance of each document in $\widetilde{\mathbf{D}}$ to a given sub-query a_k^q is independent

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

- To sum up, xQuAD suggests that:

$$\begin{aligned} P\left(d, \bar{\widetilde{\mathbf{D}}}|q\right) &= \sum_{k=1}^{K} P\left(d, \bar{\widetilde{\mathbf{D}}}|a_k^q\right) P(a_k^q|q) \\ &= \sum_{k=1}^{K} P(a_k^q|q) P\left(d|a_k^q\right) P\left(\bar{\widetilde{\mathbf{D}}}|a_k^q\right) \\ &= \sum_{k=1}^{K} P(a_k^q|q) P(d|a_k^q) \prod_{\tilde{d}_n \in \widetilde{\mathbf{D}}} (1 - P(\tilde{d}_n|a_k^q)) \end{aligned}$$

xQuAD – 3

- The final score for each document is determined by:

$$score(d,q) = \lambda \cdot P(d|q) + (1-\lambda) \cdot P\left(d, \overline{\widetilde{\mathbf{D}}}|q\right)$$

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

  - 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? Another relevance?

Questions?

kychen@mail.ntust.edu.tw