Web Retrieval and Mining Programming HW1

B02902029 楊子由

Vector Space Model Implement

以 VSM 和 Rocchio Feedback 實作一個小型的 IR System。

Database

使用Sqlite作為資料庫，存有三個 Table 。初始化大概需要75分鐘。

- 紀錄 unigram、bigram：Termid, id1, id2, df (doc_freq for short)
- 紀錄在文件中的出現：Termid, Docid, Time
- 紀錄文件長度：Docid, Len

Query處理

在每筆詢問裡面都有一個很容易切的子元素：concepts，都是以「、」作為分隔，並且切下來的每個詞都在該筆詢問佔有一定的全重。所以我是這樣做的：

- 先根據「、」切好每個名詞。
- 是 Unigram 就直接放進 Query Term Set。
- 是 Bigram 就放進 Query Term Set ，並且也把兩個字分別做成Unigram放進 Query Term Set ，因為兩個字的 Term 有時切成單詞也會有意思。
- N-gram，N是二以上的就切成 Bigram 放進 Query Term Set 。
- 最後 Query Term Set 裡面的 Term 就是要挖出來的資料。
- 把出現在<narrative>裡面帶有「不相關」句子的 Bigram 權重提高1.5倍

TF-IDF

TF-IDF 是一個常用於 IR 裡面的加權方式，在 VSM 裡面可以把每個文件轉成一個有線維度的向量。實作上是用 Okapi BM25 ：

$$w_i = \left(\log \frac{N}{n_i + 0.5} \right) \frac{f(k+1)}{f + k(1 - b + b\frac{D}{D_{avg}})}$$

其中，f 和 k 是可調整參數，D和 D_{avg} 分別是文件長度和平均文件長度，這兩個可以在建立資料庫時預處理好，文件長度定為Term的個數。

相似度

$$sim(q, d) \equiv \frac{q \cdot v}{|q||v|}$$

Rocchio Feedback

Rocchio Feedback 是一個基於 Relevance Feedback 所做出來的方式，藉由認為據有關聯的文件來修改查詢向量。

$$Q' = aQ + \frac{b}{|C_r|} \sum_{q \in C_r} q_i - \frac{c}{|C_{nr}|} \sum_{q \in C_{nr}} q_i$$

其中， C_r 和 C_{nr} 分別代表了「關聯」和「不關聯」的文件。實作上，定義「關聯」集合是舊的排名的前 R 名， R 是參數，至於「不關聯」的文件，是很難去定義的，原先用 TF-IDF 所產生的向量維度必然是正的，不會出現「負」的相似性，最後幾名又通常是0，即使被當作「不關聯」的文件也是沒有用處，所以並沒有定義「不關聯」的文件。

Query在和關聯文件合併向量時，可能會出現新的維度，由於過多過小的維度會影響到準確度，所以有設定一個閥值，在文件向量 normalize 後的值必須超過 g ，才會被納入 Query 向量。

結果

評估以 MAP 來算。

Before 0.74

Public Test Score	Query Gram	TF	IDF
0.40511	Only bigram/trigram From <concepts>	$\frac{f}{1+f}$	$\log \frac{N}{\max(k, 0.000001)}$
0.56555	bigram/trigram/ 4-gram	$\frac{f}{1+f}$	$\log \frac{N}{\max(k, 0.000001)}$
0.64611	Full <concepts> to bigram	$\frac{f}{1+f}$	$\log \frac{N}{\max(k, 0.000001)}$
0.74395	Full <concepts> to bigram	$\frac{(k+1)f}{f+k(1-b+b\frac{L}{L_{avg}})}$	$\log \frac{N}{\max(k, 0.000001)}$

Okapi-BM25

參數	Train Score	Public Test Score
b = 0.75, k = 1.5	0.76377	0.75133
b = 0.75, k = 2	0.76749	0.75414
b = 0.75, k = 2.5	0.77241	0.75521
b = 0.75, k = 3	0.77055	0.75707
b = 0.75, k = 4	0.77476	0.75362

Rocchio Feedback

- Before Rocchio Feedback(Train) : 0.77055
- After Rocchio Feedback :

前 R 名	b	g (閥值)	Train Score	Public Test Score
20	1	0.1	0.65331	0.58172

10	0.5	0.1	0.72537	--
10	0.1	0.1	0.69535	--
10	1	0.1	0.72520	--
10	0.7	0.01(almost every term)	0.77316	0.70255
20	0.1	1(no new term)	0.77876	0.75726
20	0.05	1	0.77736	0.75781
20	0.01	1	0.77038	0.75638

參考資料

- Wiki : Okapi BM25
- Wiki : tf-idf
- Wiki : Rocchio algorithm