CI6226 Information Retrieval System Practice

Jiang Ke
G1903408J
ke006@e.ntu.edu.sg

Liu Chengwei
G1902930A
chengwei001@e.ntu.edu.sg

Liu Ye
G1902703E
li0003ye@e.ntu.edu.sg

Yang Xuehuan
G1903559J
s190113@e.ntu.edu.sg

Zhang Cen
G1902673L

In this section, our information retrieval system is to evaluate on the measure time for creating an index and answering shorter and longer query respectively. And also the memory usage is to calculate for holding the index. We will compare the index size with document corpus size as well. Finally, we will demonstrate the correctness of information retrieval system on a small documents set.

Refine the dataset. Original Dataset contains much irrelevant information such as email comment which indicates email is unclassified and associated with U.S. Department of State. To better present information related to email body, we removed those message and store emails into XML documents amenable to quickly index the email. Table 1 shows that out of 7945 emails, we clarified 4816 emails while the others was discarded for some format problems such as incompatible to store as XML document.

	orginal	error	selected
number	7945	3129	4816

Table 1: Refine the HillaryEmails dataset

Measure the time. As shows in Table 2, we evaluated how much time was used for creating for each index and for varied sized query(len=3,5,7,10). It is clear that creating an index is fast and accounts for 1.46 millisecond. When the length of query varies, the time used varies as well. Answering query with length of 3 spent 0.73 milisecond at the lowest while that of 10 spent 3.93 millisecond at the highest. And this demonstrates that answering less length of query spent less time. Overall, both time used for creating an index and answer an query is relatively fast.

	index	query			
		len=3	len=5	len=7	len=10
time(ms)	1.46	0.73	1.92	2.52	3.93

Table 2: Time for creating index and answering query

Measure the size. Table 3 shows that size of HillaryEmails dataset. Size of original HillaryEmails dataset is 40MB. After selection, we focused on part of the dataset taking up 37MB. We created 42364 indices for this dataset. For storing the indices in text format, we have 1.5MB storage to maintain it. However, storing the indices using trie (patricia trie) reduced the storage hugely to 218KB. Thus, choosing different structure or format to store index can make a difference with respect to the memory or storage requirement accordingly.

	orginal	error	selected	index(text)	index(trie)
number	7945	3129	4816	42364	42364
size	40MB	13M	37MB	1.5MB	218KB

Table 3: Size of the index and the HillaryEmails dataset

Verify the result. To illustrate how much of correctness the information retrieval system could achieve, as shows in Table 5, we selected an email from the dataset. We construct several query string to the information retrival system to verify whether the results are correct or not in Table 4. liuye: TO DO

q_id	1
query terms	justice sensitive Muslims
relevant docs	
correctness	

Table 4: Verify the query result

6 Optimization

liuye: TO DO

doc	HillaryEmails-5898.txt
header	From: Abedin, Huma <AbedinH@state.goy> Sent: Thursday, August 19, 2010 12:42 PM To: Subject: Fw: Indian PM shows solidarity with Pakistan against floods From: Sullivan, Jacob 3 To: Abedin, Huma Sent: Thu Aug 19 11:07:01 2010 Subject: FW: Indian PM shows solidarity with Pakistan against floods Can you flag for S Singh is doing a good thing here. From: membership_services@fma.sositd.com [mailto:membership_services@fma.sositd.com] On Behalf Of News Desk Sent: Thursday, August 19, 2010 10:55 AM To: Afghanistan-Pakistan News Alerts; CENTCOM News Alerts; PACOM News Alerts Subject: Indian PM shows solidarity with Pakistan against floods
body	Indian PM shows solidarity with Pakistan against floods ISLAMABAD, Aug. 19 (Xinhua) -- Indian Prime Minister Manmohan Singh called his Pakistani counterpart Syed Yusuf Raza Gilani over phone Thursday to convey the condolences and commensurations of his government and the people of India on the loss of precious lives and damage to the properties and infrastructure caused by floods in Pakistan. The Indian prime minister said that the member States in the South Asian Association for Regional Cooperation should come together to help Pakistan in every way possible at this critical juncture, according to the Pakistani PM office. Gilani thanked his Indian counterpart for his kind gesture of calling him to express his sympathies and solidarity with Pakistan in its hour of trial. He briefed the Indian prime minister over the huge damage caused by this unprecedented natural calamity in all the provinces of the country, which, according to UN estimates, was far bigger than the effects of earthquakes and tsunami which hit Pakistan in 2005 and Haiti last year respectively. Gilani also availed the opportunity to reiterate Pakistan's keen desire to have friendly, cooperative and good neighborly relations with India. He hoped that in accordance with the agreement reached in Thimphu in Bhutan in April, the bilateral dialogue between the two countries would progress a meaningful, substantive and result- oriented way by addressing all the issues of concern to both countries including Kashmir dispute. Media Analysis and Watch Center USSTRATCOM Foreign Media Analysis Program SOS International Ltd.

References