PTT Article Recomender

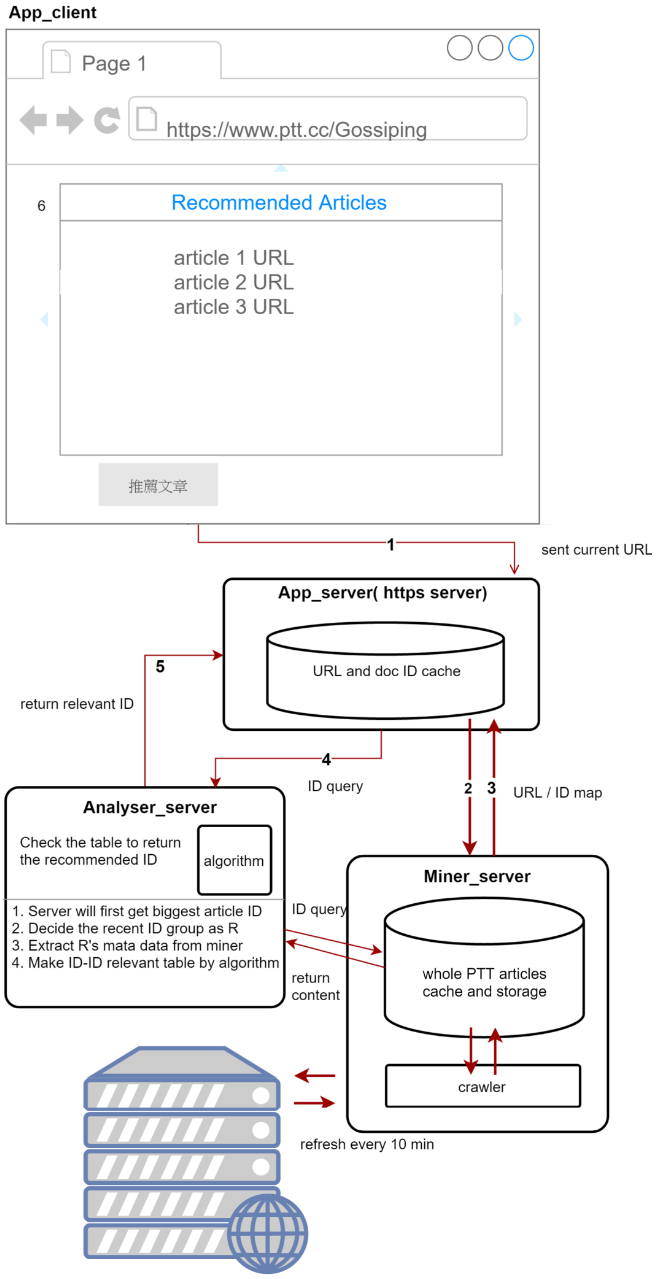
B01902065,B01902109,B01902013,B01902103

Abstract

When surfing the gossiping board on web ptt. You may find some articles that are appealing to you. You may then want to find some articles that are related to the interesting one. Unfortunately, it will be a tough task since you can not simply type "/" on the website. You may have to open another page so as to directly google it or check the board by yourself. Our project is aim to provide you with a convenient way to do this job. After installing our chrome extension, you will have a new button on the buttom-right corner. When clicking, PTT article recommender system will give you a list of URLS that are relevant to the articles that you are interested in.

Keywords: PTT, BM25, Recommender

1 System Structure


App_client
Page 1
https://www.ptt.cc/Gossiping
6
Recommended Articles
article 1 URL
article 2 URL
article 3 URL
推薦文章
1
sent current URL
App_server(https server)
URL and doc ID cache
5
return relevant ID
4
ID query
2
3
URL / ID map
Analyser_server
Check the table to return the recommended ID
algorithm
1. Server will first get biggest article ID
2. Decide the recent ID group as R
3. Extract R's mata data from miner
4. Make ID-ID relevant table by algorithm
ID query
return content
Miner_server
whole PTT articles cache and storage
crawler
refresh every 10 min

In this report, we will first demonstrate how we design the recommender system in brief and then give the work flow of our recommender. Second, we will show details of how to find relevant articles that our system would recommend. Finally, we will show experiment of our system and give the distribution table of work.

1.1 Component Design

To recommend articles from articles on the web ptt, we have to first crawl all the articles from ptt to our Miner_server. The Miner_server will use L2 cache structure to store the data and update it periodically. The Miner will also give a special ID number based on articles' URL. We will have a Analyser_server that will first read all articles from Miner based on ID and predict the relevant articles of each ID so as to make a table and keep it in the memory. After building the environments, our App_server will serve as the bridge from users to our system. It will give recommended articles based on the URL that the user is reading now. The work flow of our system is given below and summarized as the graph on the left:

1 The App_client will sent URL of this page and ask server to output the relevant articles when user opens a new article.

2∼3 Server asks Miner the corresponding ID of this URL and get return from Miner.

4 Server asks Analyser to return the relevant articles' ID.

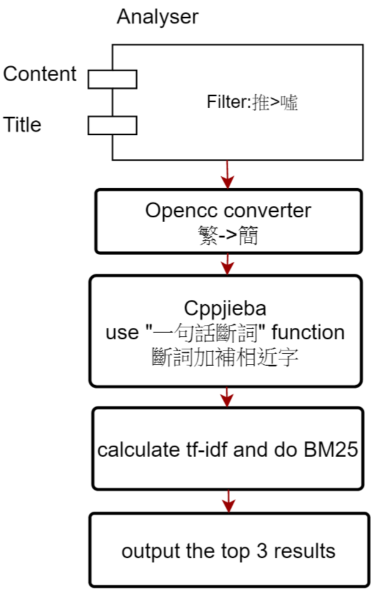
5 Analyser checks the pre-build table to answer the request and App_server will also ask the corresponding URL of Analyser's return and finally return to the client.

6 Client will make a pop-up window containing the hyper-links of recommended articles when user clicks the " 推薦文章" bottom.

1.2 Details of Retrieval

In our analyser, we use some open source on github such as cppjieba [1] and opencc [2] to help us divide both content and title of the article into chunks and then use BM25 model to calculate retrieval status value so as to predict the relevant articles, note that cppjieba is designed for simplified Chinese so we first use opencc to translate the articles so as to avoid some encoding bugs and increase precision, the procedure can be summarized as the graph below:

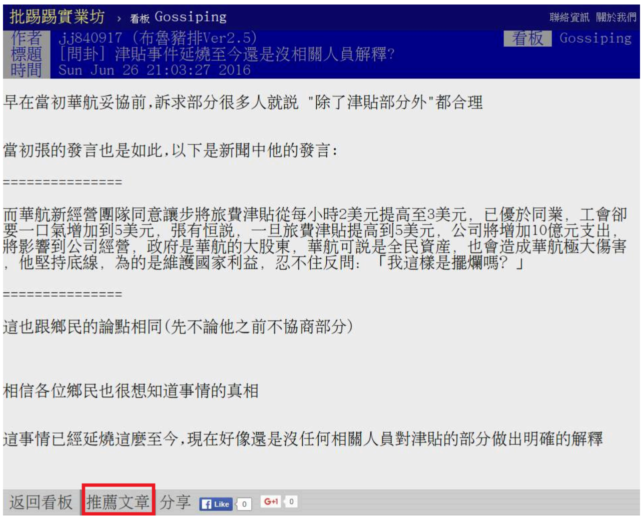
[1] https://github.com/yanyiwu/cppjieba
[2] https://github.com/BYVoid/OpenCC



In our BM25 model, we set k_1=1.5,b=0.75 and use the following formula:

$$RSV_d = \sum_{t\in q} \log(\frac{N - df_t + 0.5}{df_t + 0.5}) \cdot \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b*(L_d/L_{ave}))+tf_{t,d}}$$

2 Experiment



作者 jj840917 (布魯豬排Ver2.5) 看板 Gossiping
標題 [問卦] 津貼事件延燒至今還是沒相關人員解釋?
時間 Sun Jun 26 21:03:27 2016

早在當初華航妥協前,訴求部分很多人就說 "除了津貼部分外"都合理

當初張的發言也是如此,以下是新聞中他的發言:

===============

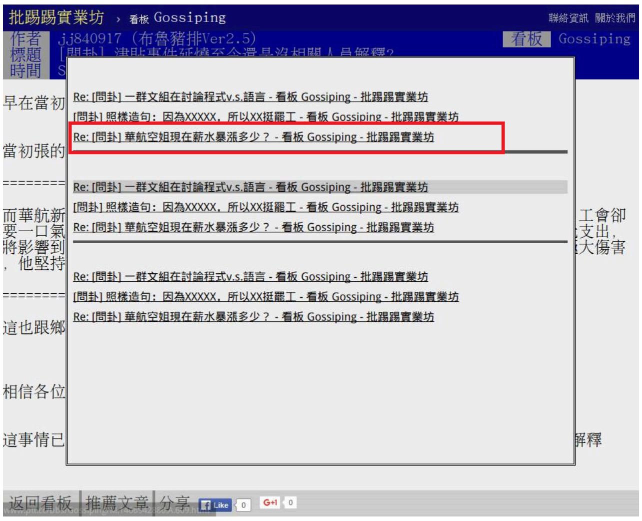
而華航新經營團隊同意讓步將旅費津貼從每小時2美元提高至3美元，已優於同業，工會卻要一口氣增加到5美元，張有恒說，一旦旅費津貼提高到5美元，公司將增加10億元支出，將影響到公司經營，政府是華航的大股東，華航可說是全民資產，也會造成華航極大傷害，他堅持底線，為的是維護國家利益，忍不住反問：「我這樣是擺爛嗎？」

===============

這也跟鄉民的論點相同(先不論他之前不協商部分)

相信各位鄉民也很想知道事情的真相

這事情已經延燒這麼至今,現在好像還是沒任何相關人員對津貼的部分做出明確的解釋

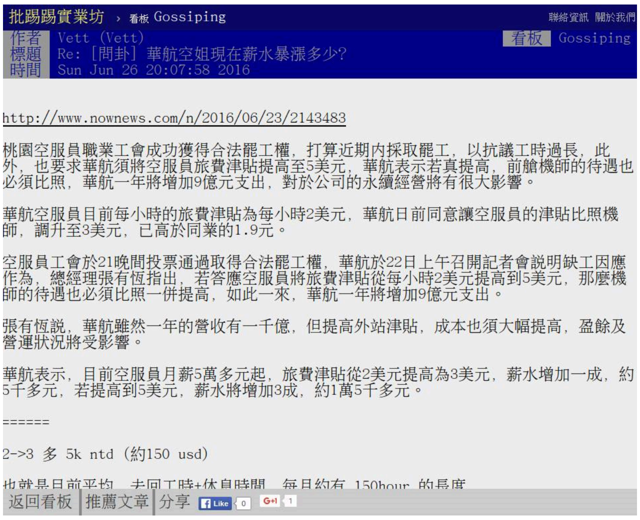


作者 jj840917 (布魯豬排Ver2.5) 看板 Gossiping

Re: [問卦] 一群文組在討論程式v.s.語言 - 看板 Gossiping - 批踢踢實業坊
[問卦] 照樣造句：因為XXXXX，所以XX挺罷工 - 看板 Gossiping - 批踢踢實業坊
Re: [問卦] 華航空姐現在薪水暴漲多少？ - 看板 Gossiping - 批踢踢實業坊

Re: [問卦] 一群文組在討論程式v.s.語言 - 看板 Gossiping - 批踢踢實業坊
[問卦] 照樣造句：因為XXXXX，所以XX挺罷工 - 看板 Gossiping - 批踢踢實業坊
Re: [問卦] 華航空姐現在薪水暴漲多少？ - 看板 Gossiping - 批踢踢實業坊

Re: [問卦] 一群文組在討論程式v.s.語言 - 看板 Gossiping - 批踢踢實業坊
[問卦] 照樣造句：因為XXXXX，所以XX挺罷工 - 看板 Gossiping - 批踢踢實業坊
Re: [問卦] 華航空姐現在薪水暴漲多少？ - 看板 Gossiping - 批踢踢實業坊



作者 Vett (Vett) 看板 Gossiping
標題 Re: [問卦] 華航空姐現在薪水暴漲多少?
時間 Sun Jun 26 20:07:58 2016

http://www.nownews.com/n/2016/06/23/2143483

桃園空服員職業工會成功獲得合法罷工權，打算近期內採取罷工，以抗議工時過長，此外，也要求華航須將空服員旅費津貼提高至5美元，華航表示若真提高，前艙機師的待遇也必須比照，華航一年將增加9億元支出，對於公司的永續經營將有很大影響。

華航空服員目前每小時的旅費津貼為每小時2美元，華航日前同意讓空服員的津貼比照機師，調升至3美元，已高於同業的1.9元。

空服員工會於21晚間投票通過取得合法罷工權，華航於22日上午召開記者會說明缺工因應作為，總經理張有恆指出，若答應空服員將旅費津貼從每小時2美元提高到5美元，那麼機師的待遇也必須比照一併提高，如此一來，華航一年將增加9億元支出。

張有恆說，華航雖然一年的營收有一千億，但提高外站津貼，成本也須大幅提高，盈餘及營運狀況將受影響。

華航表示，目前空服員月薪5萬多元起，旅費津貼從2美元提高為3美元，薪水增加一成，約5千多元，若提高到5美元，薪水將增加3成，約1萬5千多元。

======

2->3 多 5k ntd (約150 usd)

也就是目前平均 去回工時+休息時間 每月約有 150hour 的長度

3 Distribution of work

ID	work
b01902013	Analyser
b01902065	Survey and report
b01902109	Code manager and everything
b01902103	Miner

4 Future work

To be honest, we try to do sentiment analysis on web PTT at first. We follow the professor's advice and then survey Bing-Liu's paper[3] and tutorial[4]. In this paper,professor Liu says that he will first divide the content of article into chunks(noun) and use the modifier of chunks (adjective) to be the opinion word. The major problem is how to learn the sentiment of each opinion words. Professor Liu will first label 30 adjectives as the seed and then iteratively search synonyms and antonyms of labelled word on WorldNet to expend the labelled opinion words. We try to mimic his work but find out that the Chinese Worldnet provided by LOPE Lab is somewhat broken and very hard to use so we narrow ourself to just recommend articles on web ptt without doing sentiment analysis. In the future, we wish we can know how to use the Chinese Worldnet to do sentiment analysis on PTT and we will try to publish our report to the chrome extension market.

[3] https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf

[4] https://www.cs.uic.edu/~liub/FBS/Sentiment-Analysis-tutorial-AAAI-2011.pdf