KWB: An Automated Quick News System for Chinese Readers

Computer Science Department, University of Masschusetts Lowell

Yiqi Bai, Wenjing Yang, Hao Zhang, Jingwen Wang, Ming Jia, Roland Tong, Jie Wang

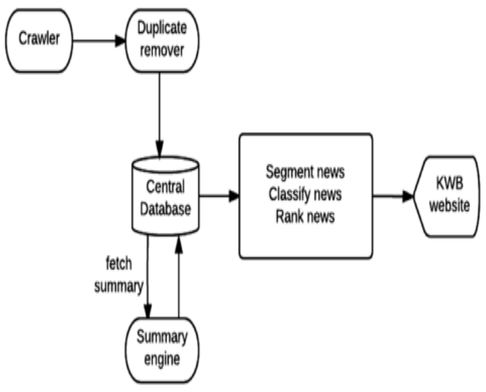
INTRODUCTION

- We are living in the era of information explosion. To help people obtain information quickly, we would want to construct an automated system that collects information and provides accurate summarization to the user in a timely fashion.
- This would be a system that integrates advanced technologies and current research results on text automation, including data collection, storage, classification, ranking, summarization, web displaying, and app development.
- **KWB** is an automated quick news system. It crawls and collects the news items from over 120 news websites in mainland China, eliminates duplicates, and retrieves a summary for each news article using a proprietary summary engine. It uses a Labeled-LDA classifier to classify the news items into 19 categories, computes popularity ranks called PopuRank of the newly collected news items in each category, and displays the summaries of news items in each category sorted according to PopuRank together with a picture.

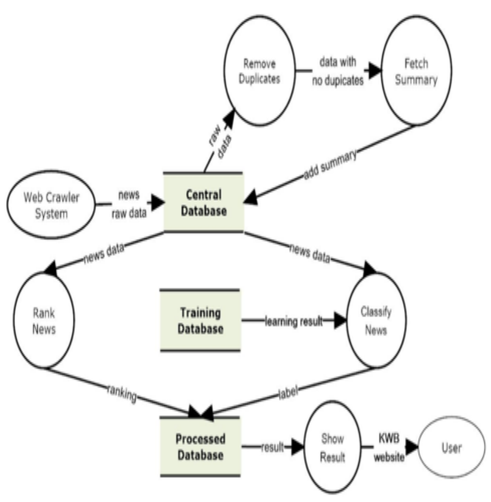
KWB ARCHITECTURE

KWB consists of five components :

- Crawlers: is responsible for collecting news items around the clock from over 120 news websites in mainland China.
- Central DB: is responsible for processing the raw data collected from the crawlers, including removing duplicated news items and fetching summaries for each news article.
- Summary engine: is responsible for returning summaries for each new article with different lengths required by applications. This is preparatory technology.
- Core processing unit: includes Chinese text fragmentation. News article classifications. Ranking each document according to PopuRank.
- Web display: is responsible for displaying on a website the news items in each category according to their PopuRanks in each day, their summaries, pictures (if there is any), and links to the original news items.



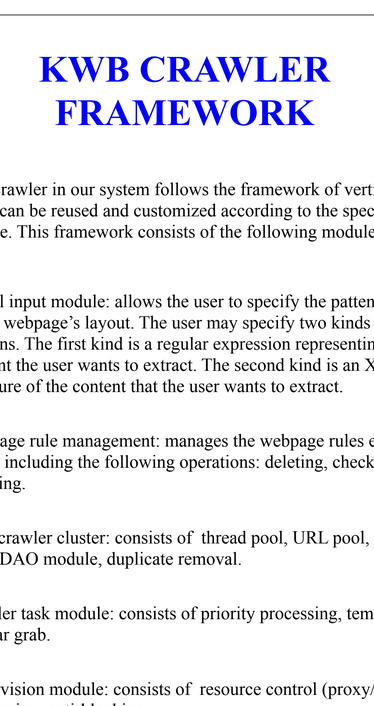
Data flow in **KWB** system in each module will operate data and save new attributes.



KWB CRAWLER FRAMEWORK

The **KWB** crawler in our system follows the framework of vertical crawling. It can be reused and customized according to the specific layout of a webpage. This framework consists of the following modules:

- Visual input module: allows the user to specify the patten of the target webpage's layout. The user may specify two kinds of patterns. The first kind is a regular expression representing what the content the user wants to extract. The second kind is an XPath structure of the content that the user wants to extract.
- Webpage rule management: manages the webpage rules entered by users, including the following operations: deleting, checking and updating.
- Core crawler cluster: consists of thread pool, URL pool, pattern pool, DAO module, duplicate removal.
- Crawler task module: consists of priority processing, temp grab, regular grab.
- Supervision module: consists of resource control (proxy/account), monitoring, anti-blocking.
- Program entrance.



CENTRAL DATABASE

Data collected from the **KWB** crawler are raw data. A new database called central DB is created to remove duplicates and retrieve summaries for raw data collected in every hour.

There are two different types of duplicates in the raw data:

- Exactly the same news items due to reposting.
- Different news items reporting the same news.

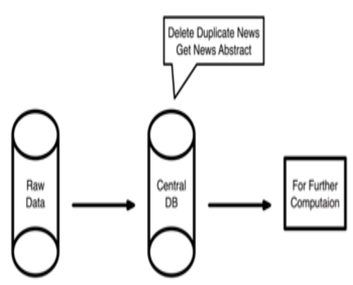
The central DB retrieves article summaries and detects duplicates in a parallel fashion. It sorts all the unprocessed raw data in increasing order according to their IDs. Starting from the first news article, repeat the following:

- Send a request to the summary engine to retrieve summaries of required lengths.
- Compute the cosine similarities of the article with the news items whose IDs fall in a small fixed time window after this article. If a duplicate is found, remove the one whose ID is in the time window.
- Move to the next news article in the shorted list.

KWB also use index to delete duplicate news. The index of the news items stored in the central DB contains four fields:

- News title.
- News URL.
- Image URL.
- First and last sentence of the news content.

News items that match any of these fields for all pairs of news items will be deleted.



POPURANK

KWB determines the popularity ranking, called PopuRank, of news items.

- Let t_v denote the current time frame.
- Let $\mathcal{D}_v = \{D_1, D_2, \cdots, D_N\}$ denote the corpus of all news items collected in this time frame with duplicates removed, where D_i is a news article and D_i contains N_i words in the model of bag of words, denoted by $D_i = (w_1, w_2, .., w_{N_i})$
- w_j appears in the current time frame t_v ,the sequence of consecutive time frames window is $T = (t_{v-u+1}, t_{v-u+2}, \cdots, t_v)$

We define the following terms:

Term frequency (TF). The term frequency of word w_j in D_i in time frame t_v, denoted by $tf(w_j, D_i, t_v)$, is the number of times it appears in D_i, denoted by N_{ij}, divided by N_i. That is,

$$tf(w_j, D_i, t_v) = \frac{N_{ij}}{N_i}.$$

Note that if $w_j \notin D_i$, then $tf(w_j, D_i, t_v) = 0$.

Term rank (TR). We define the term rank of word w_j in document D_i in time frame t_v, denoted by $tr(w_j, D_i, t_v)$, as follows:

$$tr(w_j, D_i, t_v) = \alpha \cdot tf(w_j, D_i, t_v) + \beta \cdot df(w_j, \mathcal{D}_v),$$

where $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta = 1$. For example, we may let $\alpha = 0.6$ and $\beta = 0.4$ to indicate that we place more weight on term frequency over document frequency.

Document frequency (DF). The document frequency of word w_j in the corpus $\mathcal{D}_v$, denoted by $df(w_j, \mathcal{D}_v)$, is defined as the total number of documents in $\mathcal{D}_v$ that contain w_j, denoted by N_j, divided by the total number of words in $\mathcal{D}_v$, denoted by N. That is,

$$df(w_j, t_v) = \frac{N_j}{N}.$$

Average term frequency (ATF). Let $atf(w_j, \mathcal{D}_v)$ denote the average term frequency of word w_j in corpus $\mathcal{D}_v$. That is,

$$atf(w_j, t_v) = \frac{\sum_{i=1}^{N} tf(w_j, D_i, t_v)}{N}.$$

$$avgATF(w_j, t_v) = \frac{ATF(w_j, t_v)}{\ell - 1},$$

$$avgDF(w_j, t_v) = \frac{DF(t_v)}{\ell - 1},$$

$$ATF(w_j, t_v) = \sum_{t_i \in T - \{t_v\}} atf(w_j, t_i),$$

$$DF(w_j, t_v) = \sum_{t_i \in T - \{t_v\}} df(w_j, t_i).$$

At each time frame in this window, monitor the DF and ATF values for each word. There are two cases:

- w_j is a new word, that is, it did not appear in the previous time frames in the window T , then we compute the TF-IDF values of all the new words in this time frame and mark the top d percent of the new words as popular words.
- w_j is not a new word. Compute $atf(w_j,t_v)$ and $df(w_j,t_v)$. If the ATF and DF values of word w_j at time t_v suddenly increase k_1 and k_2 times over the previous average ATF and DF values, respectively, for word w_j , denoted by $avgATF(w_j,t_v)$ and $avgDF(w_j,t_v)$.

To specify k_1 and k_2.

$$ratATF(w_j, t_v) = \frac{atf(w_j, t_v)}{avgATF(w_j, t_v)},$$

$$ratDF(w_j, t_v) = \frac{df(w_j, t_v)}{avgDF(w_j, t_v)}.$$

If

$$ratATF(w_j, t_v) > \delta,$$
$$ratDF(w_j, t_v) > \sigma,$$

where δ and σ are threshold values, then we say that word w_j is **popular** in time frame t_v.

PopuRank of news article $D_i \in D_v$ to be the sum of term rank of the popular words in D_i in time frame t_v

$$\text{PopuRank}(D_i, t_v) = \sum_{w \in H_v \cup D_i} tr(w, D_i, t_v).$$

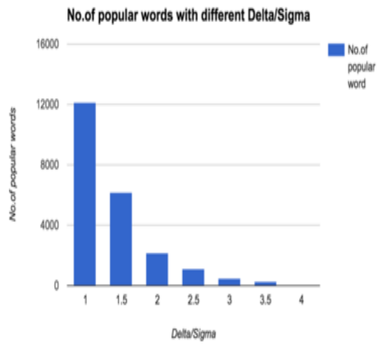
The values of parameters in following figures for **PopuRank** calculation are u = hour, l = 24, d = 20%, $\alpha = 0.6$, $\beta = 0.4$, $\delta = 1.5$, and $\sigma = 1.5$.

	A	B	C
1	Title	Hot Time (timestamp)	PopuRank
2	南宁哪里交通被刮手机数据	1430445600	579
3	广西今年投520.7亿为民办实事	1430445600	546
4	春晚福娃邓鸣贺因白血病去世	1430445600	519
5	五月起个人就乱发布天气预报	1430445600	483
6	巡视进驻中粮2.4亿流失国资没见过这么乱的企业	1430445600	478
7	京津冀协同发展2020年北京人口不超2300万	1430445600	475
8	楼购房提意政策新变化 买家迎购房"窗口期	1430445600	465
9	南宁市区小学地段划分将公布	1430917200	458
10	甘肃省人民政府关于进一步加强新时期爱国卫生工作的实施意见	1430445600	457
11	美国纺作两介石"越外债"内幕	1430445600	443
12	成都企业启动"临网"计划 民资投	1430445600	440
13	市贸促会副会长李锋亭网说	1430445600	434
14	抗战期间两介石为何布置兵却夺不住南京?	1430445600	424
15	海口养生胜地大盘点	1430445600	424
16	互联网+引领白领跨界活动	1430445600	421
17	让万隆棒棒锭造新的光彩	1430445600	421
18	习近平总书记向全国劳动群众致节日祝贺	1430445600	418
19	武长顺普通威胁提电话称某中央领导送书	1430445600	415
20	重磅！中央政治局会议释放9大信号	1430445600	413
21	揭秘贪官为何偏爱现金	1430445600	411

Parameters α and β is related to TR and **PopuRank**. The value of α and β are decided by which character, TF or DF, is regarded more important.

Title	Alpha	PopuRank
决不放弃任何一丝生的希望——"东方之星"沉船水下搜救纪实	0.9	8
	0.8	16
	0.7	15
	0.6	63
	0.5	35
	0.4	34
	0.3	35
	0.2	37
	0.1	54

Threshold δ and σ decide the numbers of popular words.



WEBDISPLAY

KWB is an automated quick news system that collects news items real-time from all major Chinese news websites, classifies the news items into 19 categories, and displays on http://www.kuaiwenbao.com news items in each category with summaries and pictures, sorted according to their PopuRank values.

We have also implemented KWB in mobile apps (Android App may be downloaded by entering http://www.kuaiwenbao.com/kuaiwenbao.apk on a web browser of an Android phone).

CONCLUSIONS

We described KWB, an automated quick news system for the Chinese reader. In particular, we described the architecture of KWB, the KWB crawler framework, the central DB, the PopuRank, and the use of KWB.

- Web crawling technologies are important mechanisms for collecting data from the Internet.
- Central DataBase filters redundant data and reduces computing and resources cost.
- PopuRank mechanism measures the popularity of a news item in certain time period.

This work was supported in part by a grant from Wantology.