INTRODUCTION

- We are living in the era of information explosion. To help people obtain information quality, we need to construct an automated system that collects information and provides accurate summarization to the user in a timely manner.

- 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 200 news websites in mainland China, organizes, and retrieves a summary for each news article using a proprietary summary engine. It uses a Crawler to filter and categorize 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 according to their PopuRank values.

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

KWB CRAWLER FRAMEWORK

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

- Visual input module: allows the user to specify the pattern of the target webpage's layout. The user may specify the kinds of news that interest him most or enter a URL to download the news item the user wants to extract. The second kind is an Xpath expression for searching the content the user wants to extract. The second kind is an XPath expression for searching the content the user wants to extract. The XPath expression for searching the content the user wants to extract.

- Webpage rule management: manages the webpage rules entered by users, including the following operations: deleting, adding, and updating.

- Core crawler cluster: consists of thread pool, URL pool, pattern pool, DAG model, duplicate removal.

- Crawler task module: consists of priority processing, target path, regular path.

- Supervision module: consists of resource control (CPU, account), monitoring, anti-blocking.

- Program entrance.

CENTRAL DATABASE

Data collected from the KWB crawler are raw data. A raw database called central DB is reserved to remove duplicates and retrieve summaries for raw news 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 articles summaries and dereplicates duplicates in a parallel fashion. It stores 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 length.
- Compare the content similarities of the article with the news items whose ID is in a small fixed time window after the article. If a duplicate is found, remove the one whose ID is in the time window.
- Move to the next news article in the sorted list.

KWB uses one index to delete duplicates. The index of the news items stored in the central DB contains four fields:

- News title.
- News URL.
- Image URL.
- First and last sentences 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 \( \alpha \) denote the current time frame.
- Let \( \beta \) denote the set of all news items collected in the previous time frame.
- Let \( \gamma \) denote the news article and \( \delta \) denote the word in the model of bag of words, denoted by \( \delta \).

We define the following terms:

- \( V_i \) is the set of news items in category \( i \).
- \( D_i \) is the total number of news articles in category \( i \).
- \( \alpha \) is threshold value.
- \( \beta \) is threshold value.
- \( \gamma \) is threshold value.
- \( \delta \) is threshold value.

At each time frame in the window, retrieve the AT and DF values for each word. There are two cases:

1. \( w \) is a new word, that is, it did not appear in the previous time frames in the window \( \alpha \). Then we compute the TF-IDF value of all the new words in the time frame and mark the top \( \beta \) percent of the new words as popular words.

2. \( w \) is a new word, Compute \( \text{ATF}(w, \alpha) \) and \( \text{DF}(w, \alpha) \). If the AT and DF value of word \( w \) at time \( \alpha \) is high, then we add it to the set of popular words in each time frame, \( \delta \).

To specify \( \alpha \), we use the following formulas:

\[
\text{ATF}(w, \alpha) = \frac{\text{ATF}(w, \beta) + \text{ATF}(w, \gamma)}{2},
\]

\[
\text{DF}(w, \alpha) = \frac{\text{DF}(w, \beta) + \text{DF}(w, \gamma)}{2},
\]

where \( \alpha \) and \( \beta \) are threshold values, then we say that word \( w \) is popular in time frame \( \delta \).

The value of \( \alpha \) and \( \beta \) is decided by which character. ET or DE is required more important.

WEBDISPLAY

KWB is an automated quick news system that collects news items and ranks them according to their PopuRank values. The value of \( \alpha \) and \( \beta \) is decided by which character. ET or DE is required more important.

CONCLUSIONS

We described KWB, an automated quick news system for the Chinese readers. 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 storage cost.
- PopuRank mechanism measures the popularity of a news item in certain time period.