**TopN sorting algorithm**

The TopN algorithm refers to finding the largest (or smallest) first n elements from an existing array. In the healthy mom app, the TopN algorithm is mainly used in the search query, and the search results are sorted according to relevance and popularity.

TopN algorithm has three main ways to achieve, in the healthy mom APP, the team mainly uses efficient binary tree sorting, namely TreeSet and overloaded compareTo method to achieve TopN algorithm, the basic idea is as follows:

1. Loop all the elements in the HashMap. If the size of the TreeSet is less than the required number of elements n, add the structure directly to the TreeSet. The TreeSet will automatically sort according to the compareTo method of the structure and update the value of the minimum correlation.

2. If the size of TreeSet greater than the number of elements required n, then determine whether the value is greater than the minimum correlation, if greater than, update the TreeSet and the minimum correlation value.

3. After the loop is completed, all elements in the TreeSet is the largest n elements to find

**Word segmentation algorithm**

APP adopts the efficient Nlpir Chinese word segmentation system researched and developed by the Chinese Academy of Sciences. At the same time, Nlpir Chinese word segmentation system to search information segmentation, the healthy mother APP will also pass the word segmentation results of speech judgment, the separation of the words to re-integration, constitute the required keywords.

Nlpir Chinese word segmentation system has won the first prize of Qian Weichang Chinese Information Processing Science and Technology Award in 2010, the first international Sighan word segmentation contest in 2003, the first domestic comprehensive evaluation of 973 in 2002. Nlpir Chinese word segmentation system (also known as ICTCLAS2013) main functions include Chinese word segmentation, part of speech tagging, named entity recognition, user dictionary function, support GBK encoding, UTF8 encoding, BIG5 encoding; typical client has the People's Daily client phase III products, CITIC Trust Letter system, Huawei, national grid and so on.

**Data mining**

In the healthy mom APP, the frequent itemsets mining is mainly required for the relevance of the data. Therefore, the team mainly adopts the fpGrowth algorithm. The algorithm is mainly used in mining hot issues, such as: what campus seasons, which areas, which sites have an impact on the allergic constitutional groups.

The algorithm is Han Jiawei et al proposed in 2000, the correlation analysis algorithm. Compared with the Apriori algorithm, Apriori scans the candidate set to mine the frequent itemsets by continuously constructing the candidate set, which needs to scan the original data multiple times. When the original data is large, the disk I / O times are too many and the efficiency is low. However, the FPGrowth algorithm only needs to scan the original data twice, and compresses the original data through the FP-tree data structure, which is more efficient.

The basic workflow is as follows:

1. The elements in the header table are extracted from the List obtained by the scan database. All the elements in the List are the data required by the program to segment the chat records of college students and doctors and determine the result according to the word segmentation results. Starting from the bottom item of the header table, construct the conditional pattern base for each item.

2. Construct the conditional FP-tree (conditional FP-tree).

FP-Growh: Recursively tap each conditional FP-tree, accumulate the frequent itemsets until the FP-tree is found empty or the FP-tree has only one path (only one path, the combination of items on all paths Is a frequent item set).

**Automatic update**

By compiling and configuring listeners, Healthy Mom APP updates the top 100 articles in the latest articles and top articles every half a day, and updates the data mining results every three days.