IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Design and implementation of book publishing topic selection system based on collaborative filtering algorithm

To cite this article: Yuning Bian *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **563** 052018

View the article online for updates and enhancements.

This content was downloaded from IP address 221.6.32.206 on 12/10/2019 at 21:02

IOP Conf. Series: Materials Science and Engineering **563** (2019) 052018 doi:10.1088/1757-899X/563/5/052018

Design and implementation of book publishing topic selection system based on collaborative filtering algorithm

Yuning Bian[1,a], Yeli Li[1,b], Qingtao Zeng[1,c], Mengyang Liu[1,d]

[1]Beijing Institute Of Graphic Communication, China

[a]774200483@qq. com, [b]1605872754@qq. com, [c]954276545@qq. com, [d]1367274860@qq. com

Abstract: In view of the current book publishing topic planning may appear to rely on subjective experience of the problem. As well as book topic selection cannot be very good in line with the current hot vocabulary of this potential problem. Through the book sales market investigation, in line with the purpose of serving the book publishing industry. A collaborative filtering algorithm is used to design a business Intelligence Publishing topic selection system based on data mining. Follows are the process of processing data. First, we should put the crawl of online book information into the database. Second, filter to get the complete data we want. Finally, analysis the obtained data by collaborative filtering algorithm. This can be personalized to recommend the publishing house required keywords and the direction of the topic. Doing so can greatly reduce the burden on publishing house staff. It facilitates the development of the topic selection work of publishing house. Through the prediction of book publishing topic, we can understand the market law in a timely manner, cater to the customer's consumption concept. In this way, we can greatly improve the efficiency of publishing house selection and reduce the consumption of human financial and material resources.

1. Overview

1.1 The significance of subject research

The rapid development and widespread use of Internet technology has led schools, businesses and individuals as well as other groups to accumulate a lot of data. In the face of so many complex and diverse data such as simple queries in the past can only be simple additions and deletions to this data and can not be more in-depth analysis of the data, let alone automatically filter the information we enter to get streamlined and practical information [1]. In the context of such an era, a new vocabulary has emerged: data mining. The point of data mining is to extract from a large amount of noisy, incomplete, fuzzy data what people need or to find out the knowledge and information that people potentially want to acquire. In popular terms, data mining is a new information processing technology and its main task is to extract filter and transform the final output of the data to assist the business decision.

Book publishing this industry is constantly facing opportunities and challenges in the development process. Although the information in the internet age is characterized by a large number and rapid change, but once these large and complex data are properly processed, the efficiency of book publishing will be greatly increased. When publishing and computer science are combined, a new industry will be created. The book publishing industry will go to a higher stage.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

1.2 The main purpose of the project

We use Python to crawl online book information, then save the crawled information for data analysis. Through efficient and reasonable data analysis algorithm, we can realize simple book recommendation. The data of this topic depends on the authoritative website Douban books: Netizens register their own accounts on the bean flap to rate the books they have purchased or read. As long as the number of reviews for a book is 10, the Douban system will summarize all the comments. In order to avoid malicious brush points, the calculation method is to use weighted average score. That is, the final score is related to the number and average score of the book.

The main purpose of this topic is to design a book selection recommendation system for retrieving book information and its comments. This system mainly uses crawler technology to crawl book information and hot reviews from Douban book website.

2. Data Mining Overview

2.1 Introduction to Data mining technology

Simply put, data mining is about extracting the key information that users need in a lot of data or digging out potentially valuable information that users need. To make the whole process convenient and intelligent, we usually write computer code and then enter data to make the process of data mining simple. This process is also the process of filtering useful information in a database. In this filtering process, the general pattern is summed up for the next step of data mining, so that it can be summed up to make accurate predictions of future data. Of course, the results are not necessarily accurate, because there are a lot of data mining patterns that are meaningless or straightforward, or some data that is false, some of which are generated by coincidence. In real life, some situations can not be avoided, such as we may artificially tamper with some data, some data will be lost to cause breakage can not be used. What we see is not necessarily exactly right, there are surprises in any situation, everything will have general circumstances and special circumstances. This requires us to take into account all possible scenarios as much as possible when editing the algorithm.

The main task of data mining is to extract the required information for people by analyzing and processing the data stored in the data. For example, how to understand customer loyalty in the market? The key to solving this problem is to refer to the customer's personal information database and the database of the goods purchased by the customer. The buyer's previous purchase behavior can analyze what he likes and doesn't like. Once these characteristics are discovered by merchants, these customers can be used to test current actual customers, thus identifying customers whose beliefs are easy to swing and those with strong beliefs. After selecting such a customer, the merchant can take measures to treat them specifically, knowing that the cost of special treatment for all customers is very high. Similarly, the same technology can help businesses identify customers who are not satisfied with the current business offering services but are likely to be interested in other services offered by merchants, and merchants can promote their services to these potential customers. In a highly competitive society, if these potential data can be tapped by merchants, it will become a raw material to drive the development of the enterprise.

2.2 Introduction to Web Mining technology

Data mining technology in the Internet is now booming. Each search engine company gets a more reliable and authoritative consideration between the pages and the website by tapping the hyperlinks between the various linked pages. Authority is defined in the dictionary as "the high status achieved through success and influence". The word “pagerank” is used to measure the authority of a Web page, which was invented by Google's founders and used by many search engine developers in various forms. The more pages you link to your site, the more authoritative your site is, especially if the pages that link to your site themselves have high authority. This definition may sound like some circular definition, but it is perfectly feasible. Search engines Use the pagerank measure to sort pages, and then return the sorted results to the user.

Search engine can not only excavate Web content, but also excavate user's query content, such as user's query keyword. In this way, advertisements can be accurately placed according to the possible interests of customers. Search engine companies have a lot of motivation to achieve accurate advertising, because advertising providers according to the number of clicks of users to pay search engine companies corresponding remuneration. Search engine companies also tap the user's click Information, the user in the returned results of the click Information itself contains useful knowledge, this knowledge helps to further improve the user's search quality. Online booksellers get a lot of shopping recommendations by digging through the information in the book purchase database, such as "the customer who bought the book also buys the book"; Similarly, these booksellers have a lot of incentive to recommend tempting, personalized options to their customers. Movie sites recommend movies based on a user's previous choices and the choices of other users, and their level of profitability depends on the turnaround rate of their site users.

3. Introduction to the main algorithms of the system

3.1 Overview of collaborative filtering algorithms

The concept of collaborative filtering recommendation algorithm is easy to understand. The core idea of this concept is that it thinks people who have the same preferences for the same item will have a consistent interest, and those with the same interests we call them "neighbors. " When you buy something, merchants tend to recommend to you what other items your "neighbors" have purchased. And these "neighbors" we think they have a good similarity. As a result, we know that the way to divide a customer is a neighbor is to calculate each other similarities, and those with a high degree of similarity are naturally divided into groups, with people in the group being "neighbors" to each other.

For example, when you want to buy a book, you ask your friends for their opinions, and your friends are often people who are similar to your interests. The books your friends recommend to you are often very persuasive. The steps of the collaborative filtering recommendation algorithm are: First gather information to determine the user's preferences and then input all the information into the prediction part, through the filtering of the prediction algorithm we can probably recommend to the user that he may like the information, this information is recommended to the user, so that the completion of the entire system recommendation process.

Collaborative filtering recommendation system has many advantages it can automatically analyze and filter content that is difficult to use in text by machine automatic analysis and filtering to get the results we want, and he can also filter based on some complex abstract content, expanding the concept of collaborative filtering to extend the breadth of collaborative filtering [2].

3.2 Classification of collaborative filtering algorithms

Collaborative filtering algorithms can be divided into two categories: one is a user-based collaborative filtering algorithm, and the other is an item-based collaborative filtering algorithm.

The user-based collaborative filtering recommendation algorithm usually considers the similarity between two different users, as long as other users who are similar to the user are identified, and the items that the user may be interested in are predicted by something similar to the user's liking. and whether each user likes an item is considered based on the user's rating of the item. Users who rate the same item or the same type of item can be divided into a group. They have a higher degree of similarity with each other and therefore call them "neighbors". It is more reliable to recommend what "neighbors" like.

The project-based collaborative filtering algorithm is a group of items with high similarity by calculating the similarity between two items, recommending to the user items with a high similarity to their favorite items. That is, when you want to buy an item, you may also be interested in items that are similar in height to this item. Therefore, it will be highly reliable to recommend items that are similar to this item to you.

4. Network Information Crawling Technology

4.1 Introduction to Crawler Technology

There are countless web pages on the World Wide Web. They contain a great deal of information, which is all-encompassing. But many times, whether it's data analysis or product requirements, we need to extract what we're interested in and get valuable content from some websites [3]. But even humans entering the 21st century still have only two hands and one eye. It is not possible to read each page and then copy and paste it. Therefore, we need a program that automatically gets the content of the page and extracts the content according to the specified rules. This is the crawler-reptilian [4-8].

In the traditional sense, the crawler crawl process is as follows. First, the scheduler gets the Web content of the initial URL. Second, the parser parses the crawled Web page to extract the URL information for the next URL. Finally, URL Manager continues to crawl information about this page and get the URL of the next related page. This process loops until a stop condition of the system is met. Then the entire crawl process ends. The crawler process is complex, and we also need to design specific analysis algorithms based on specific Web pages to filter links that are not related to the entire topic, and to keep useful links to the URL queue waiting to be crawled. Therefore, the complete crawler typically contains the following three modules: 1. Network request Module 2. Crawl Process Control Module 3. Content Analysis Extraction module [9-12].

4.2 Simple crawler Architecture

The Simple crawler design architecture is as follows:

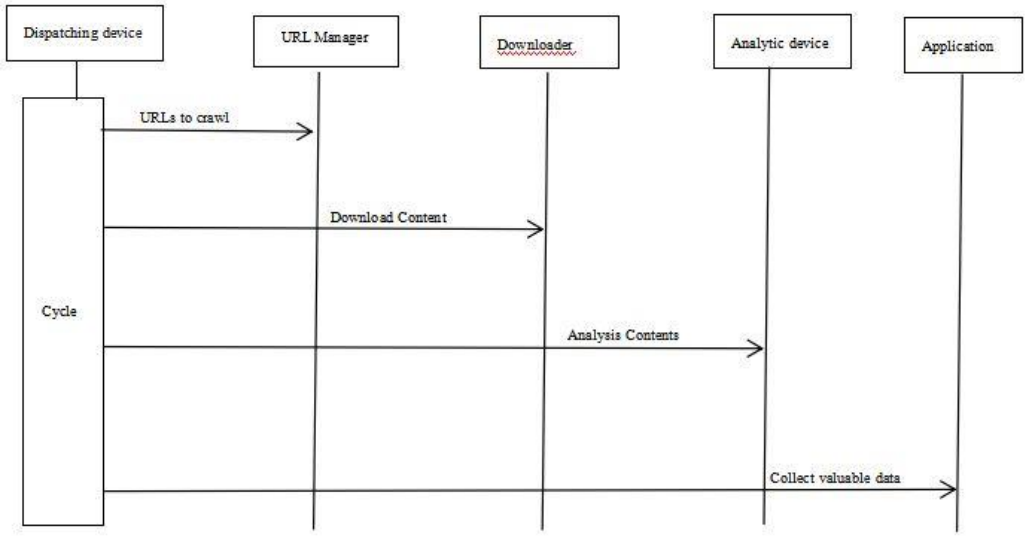


Figure 1. Simple crawler architecture

5. Design and implementation of the system

5.1 System Structure Design

Combined with the current relatively popular data mining technology and collaborative filtering algorithm, this system realizes an intelligent topic selection system of recommended books. The system structure of Intelligent Book Topic selection system based on data mining is as follows:

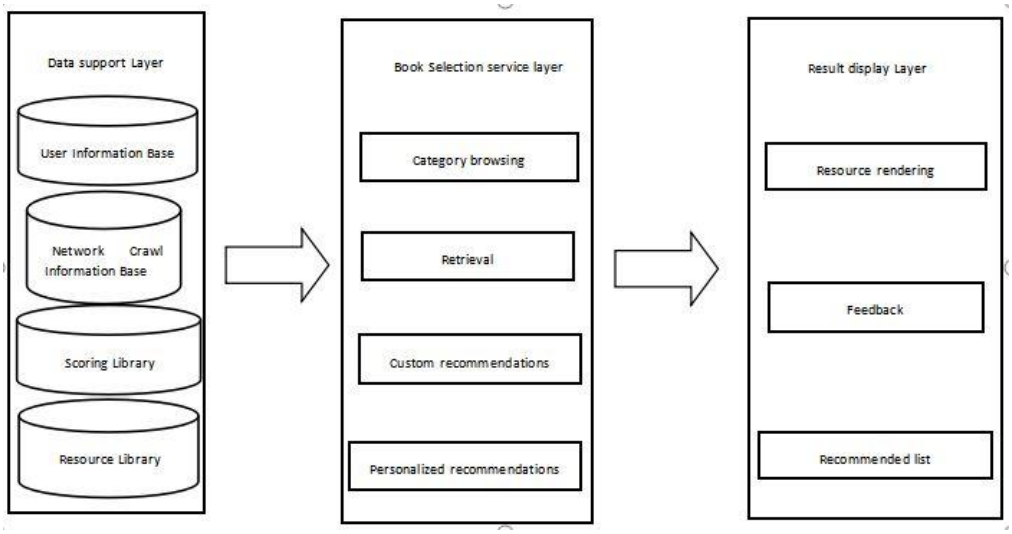


Figure 2. Architecture of Book recommendation system

Database support layer

The database support layer is mainly the information of the login account, the data information crawled on the network and the user's scoring demand information according to a certain correlation coefficient weighted data is then stored in the local database. This database support layer consists mainly of four databases: the user's login information base, the database crawled by the network, the comprehensive weighted scoring library and the resource library.

Book Selection service layer

The service layer of book topic selection is the key layer of the whole system and the layer that embodies the collaborative filtering algorithm. This layer can be based on the Publisher editor's user needs to provide the publishing house with personalized recommendations for book topic selection or processing of actual user requests, mainly including the following functions:

Custom recommendations: Data crawled from the network can know each user's score for different books, using the collaborative filtering recommendation algorithm for publishers to recommend potentially useful book keywords.

Personalized recommendation: Administrators according to different topic requirements for users to recommend some of the quality of the topic reference.

Category browsing: Because the criteria for each topic vary, all data is classified according to different criteria to facilitate the publishing house editing browsing and reference.

Retrieval: According to the publishing house search request, return the corresponding book keyword resources.

Results display Layer

The result display layer of this system is the output interface layer of the whole system, which is directly oriented to the system user, is the interface of the whole system and the publishing house selection staff direct interaction, the user can use the resources of the whole system directly through this level, and the user can interact directly with the system through this level. Enter the requirements into the corresponding processing module for the system. For example, the system to submit book Resources search requests, book resources to score, information feedback, etc. , and then, through the results of the display layer to the publishing house to edit the user to display the results of the request, render the final resources or provide a personalized list of recommendations.

5.2 System code Design

This system takes the current mainstream Douban station as the data source, the initial crawl interface is "code book", according to the URL of the initial Douban page to extract the page book name, rating, author, publishing house, publication time, Price, ISBN, Introduction, thermal review and Bean flap recommended book URL. As long as the crawl URL collection in the URL manager is not

empty or does not reach the specified number of crawls, the information for the Web page and the new URL are crawled.

The Python modules that need to be written using Python2. 7 are: URLLIB2 (advanced Web Communication module, downloading data according to supported protocols), BeautifulSoup (processing Html/xml), re (providing regular expression-related operations), Pymongo (connect the MongoDB database and manipulate it), Pyexcelerator (provides Excel related actions).

5.3 Data storage

The system uses MongoDB database to store crawling information. MongoDB is a database based on distributed file storage. It supports a very loose data structure, is similar to the JSON Bson format, the data storage format is the Key-value type. The biggest advantage of MongoDB is that the query language it supports is very powerful, and its syntax features are similar to object-oriented query languages, which support indexing of data.

The system creates a GoodBooks database in the MongoDB database and establishes the following collection: Newurls (Save the crawled URL), Oldurls (Save the Crawl URL), book (Save crawled Books information), Notfoundurls (save 404 Wrong URL), User (save user added books) [13-15].

6. Summarize

This system uses the collaborative filtering recommendation algorithm to realize the function of providing topic selection prediction to publishing house editors. The whole system design is rigorous, the idea is clear, the module completes the topic Selection prediction and the recommendation function independently, has provided the good work idea for the later internet Age book publication forecast. So that the staff can better carry out the work, greatly simplifies the staff work flow, for the publishing house editors to carry out the future selection of topics to provide a good idea.

Acknowledgments

Beijing Institute of graphic communication 2018 R&D Project(Ec201802). Beijing science and technology innovation service capability construction project(PXM2016_014223_000025)

References

[1] Liu Quanwei. Design and implementation of leader board system based on web crawler [C]. Posts and telecommunications. 2018. Ding, W. and Marchionini, G. 1997. A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.
[2] Research and application of information data acquisition method based on theme network crawler [J]. Shengya, Wei, Liu Mong. Electronic technology and software engineering. 2016 (07)
[3] Hong Junbin. Research nozzle and implementation of Web Application System vulnerability detection technology [D]. Guangdong University of Technology. 2016. .
[4] On the application of raspberry pie in students ' innovative education [J]. Wangchen, Li Ming. Information technology Education in China 2017 (05)Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.
[5] Python's application in Scientific computing [J]. Yiu Jiansheng, Li Shumei. Digital technology and Applications 2016 (11)
[6] Python based web crawler technology [J]. Chan Cheng, Yang Xiaolan, Zhu Fu. Heilongjiang Science and technology information. 2016 (36)
[7] Implementation of simple web crawler based on Python [J]. Silina, Liu Jiandong. Computer programming skills and maintenance. 2017 (08)
[8] Python based web crawler technology in Big Data Environment [J]. Checvoux. Electronic production. 2017 (09)
[9] Implementation of the combined application of Python and R language [J]. Pinkaike, Chen Ping Goose. China health Statistics 2017 (02)

[10] A probe into the Python language [J]. Shi Meng. New Communications from China 2017 (07)
[11] Design and implementation analysis of network crawler based on anonymous network [J]. Liu Shinjin. Computer products and circulation. 2017 (12)
[12] Research on the crawl strategy of theme Network crawler [J]. Liu Qing. Science and Technology Plaza. 2017 (04)
[13] Application of network crawler technology in electric power industry [J]. He Guanbo, Suyuki, Huang Yuanrong. Electronic technology and software engineering. 2017 (18)
[14] A new theme web crawler crawling strategy [J]. Song Ocean, Liu Xiaojan, Chan Haijun. Computer application and software. 2011 (11)
[15] Algorithms for web crawlers [J]. Huang Yulong. New Communications from China 2017 (23)