The Search Engine Weighted on Resource

Quality for Education and Learning

Youwei Huang

*School of Science, Software Engineering, Monmouth University, West Long Branch, NJ US 07764*

S1296394

6/1/2020

# ABSTRACT

Nowadays the most common and convenient way to look for information is searching on Internet. The top Internet search engines are handful, such Google, Bing, Yahoo, Baidu, and so on. They use web crawler technologies to sniff the whole visible network and then provide users simple indexes and links as the results to help users find the source of resources. Most of those Internet search engines have good performance in terms of speed and precise keyword search ability, but wide-range search engines have their cons, such as massive potential results with uncertain quality. The goal of this project is to propose a new search engine, targeting on education, to provide “high quality” learning resources to users. High quality could be an subjective judgement and limited to the resources available for searching. In this project, it refers to means “user-trusted” or “user-liked”, based on common practices in education and learning. A few new search engine technologies are proposed to support efficient storage and enhanced searching for high-quality learning resources, in particular, to address two main issues: (1) how to build such a high-performance search engine; (2) how to define the quality of resources. The details include the strategies designed to optimize general information querying, storage, ranking, and most importantly, finding the relevant learning resources that are “high-quality” for users. Prototyping and experimental study are conducted to conceptually prove this research.

Keywords: search engine, big data, education, learning resources, DLRV, RD

# CATALOG

[ABSTRACT 0](#_Toc1593359264)

[CATALOG 1](#_Toc1664586772)

[1 INTRODUCTION 1](#_Toc1837386636)

[2 RELATED WORK 4](#_Toc365914712)

[2.1 Web Crawler 4](#_Toc1878760200)

[2.2 Distributed Storage System 5](#_Toc1987669764)

[2.3 Query and Rank 5](#_Toc871075735)

[2.4 Value of resources 6](#_Toc1058677322)

[3 CHALLENGES AND DIRECTIONS 7](#_Toc583290909)

[3.1 Issues of Collecting Contents 7](#_Toc778827014)

[3.1.1 Resource Crawler 7](#_Toc636241093)

[3.1.2 Data Storage and Query 7](#_Toc1054631435)

[3.2 Issues of Defining Resource Value 7](#_Toc34379072)

[3.2.1 Value related to resources 8](#_Toc234788219)

[3.2.2 Value related to user types 8](#_Toc569226509)

[4 SYSTEM DESIGN 11](#_Toc1989280121)

[4.1 Requirements 11](#_Toc1100279158)

[4.1.1 Hardware requirements 11](#_Toc1790051280)

[4.1.2 Software requirements 11](#_Toc753363337)

[4.1.3 Non-functional requirements 12](#_Toc943363225)

[4.2 Software Engineering Process 13](#_Toc1403318007)

[4.2.1 System deployment structure design 14](#_Toc1886545961)

[4.2.2 Database and Entity&Relationship 16](#_Toc1685994500)

[4.2.3 Use Cases 17](#_Toc767878631)

[4.2.4 Package Diagram & Class Diagram 17](#_Toc930163131)

[4.2.5 Sequence Diagram 21](#_Toc991945882)

[5 ALGORITHMS 23](#_Toc1009032649)

[5.1 RD Way 23](#_Toc1875584576)

[5.1.1 Storage and Query 23](#_Toc1654227794)

[5.1.2 Rank of Results 25](#_Toc913035186)

[5.2 DLRV Way 26](#_Toc1047718823)

[5.2.1 Disadvantages of traditional RD way 26](#_Toc1100103410)

[5.2.2 Storage and Query 28](#_Toc430138311)

[5.2.3 Relational Database 28](#_Toc737621811)

[5.2.4 Non-Relational Database On Disk 29](#_Toc1466018122)

[5.2.5 Non-Relational Database Cache 32](#_Toc161414863)

[5.2.6 Three Layered Search 33](#_Toc577807927)

[5.2.7 Cache Switch Method 36](#_Toc189610209)

[5.2.8 Rank based on DLRV 39](#_Toc1220092185)

[5.2.9 Static and Dynamic 39](#_Toc1161098837)

[5.2.10 Resource Evaluation [you can give a better name] 40](#_Toc968437223)

[A. Suitability 40](#_Toc1856333278)

[B. Reliability 40](#_Toc68246624)

[C. Practicability 41](#_Toc1002816296)

[D. Feedback 41](#_Toc2091121497)

[E. Popularity 41](#_Toc637473133)

[F. Cost 42](#_Toc844612769)

[G. Static Value 42](#_Toc1043917007)

[H. Dynamic Value 42](#_Toc280040766)

[I. General formula 42](#_Toc1597976106)

[6 EXPERIMENTAL STUDY 43](#_Toc1987280233)

[6.1 Purposes 43](#_Toc1683358773)

[6.2 Test Methods 43](#_Toc1337038420)

[6.2.1 Black-box Testing 44](#_Toc1525791085)

[6.2.2 White-box and Control Group 44](#_Toc303753756)

[6.3 Test Conditions 45](#_Toc119717903)

[6.3.1 Storage and Query Condition 45](#_Toc370253320)

[6.3 Results and Analysis 46](#_Toc1312786405)

[7 COCLUSION 47](#_Toc1995302479)

[CITATIONS 49](#_Toc2024481114)

# 1 INTRODUCTION

There are many search engines available for online users. Some support wide range Internet search, such as Google, Bing, and Yahoo. Some are built within systems, such as YouTube, which has its own search engine; There are also search engines tailored to meet certain needs, like better privacy, copyright protection, and so on. For many years, and currently, Google has been dominating over 80% of market share. For this reason, Google is the main search engine considered for comparison in our research. Imagine there is someone interested in learning Java. One might Google “learn java”, and then a bunch of results are displayed: some labeled as ‘Ad’; some are videos; and many more other links, while you can keep clicking ‘see more’ to get more results. There are huge collection of results of different kinds, different sources, and certainly different qualities. It could be quite intimidating if learning java is something new for this user. Making a choice alone could become the big time-consuming step before actually starting to learn. We propose a new and unique search engine that targets educational resources, to promote and support self-learning. Providing “high quality” learning resources to users is the essential for this search engine. Here “High quality” is an subjective judgement from users, but based on common practices in education and learning. The more users are for it, the higher quality it is proved to be. Therefore, the score of quality is indeed limited to the resources that can be verified or tested by users. It will be an improving process over the time. High-quality can be understood as “user-trusted” or “user-liked”.

In this project, a few new search engine technologies are proposed to support efficient storage and enhanced searching for high-quality learning resources, in particular, to address two main issues:

1. How to build such a high-performance search engine.
2. How to define the quality of resources.

Before starting the detailed discussion, Table 1 lists the terminologies that are used this thesis.

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **DLRV** | Degree of Learning Resource Value. This system calculates the value of each resource, as the criteria for ranking and recommendation by the search engine. |
| **Suitability** | Resource title, content and tags match search keywords. |
| **Cost** | The cost of finding and using a resource, both time and money cost. |
| **Popularity** | The trend of searches, clicks, and comments. |
| **Feedback** | User feedback, positive or negative comments, mark on a resource. |
| **Practicability** | Usage of a resource, being visited, referred and shared. |
| **Reliability** | Resource reliability refers to whether the source of resources is reliable, e.g. whether it has been verified by authority or professionals. |

Table 1. The terminologies

Figure 0 illustrates the overview of the major components prototype in the system. The processes are divided into four parts: collecting data (collecting), resources search (recommendation or rank), data analysis (improvement), verification of valuable resources (verification or test). Normally, search engines are based on Web 2.0 technology today. Here DLRV (degree of learning resource value) is the method to define and improve the definition of resources value (quality), which will be explained and discussed in the later chapters.



Figure 0 The general framework of the whole system and key technologies adopted by the search engine.

Finally, experimental and test study are conducted to conceptually prove this research. The details will be presented in the later chapters.

# 2 RELATED WORK

Search engines have been studied for many years. All vast majority of search engine providers try to design superior algorithms to rank the quality of links and strategies to improve storage-query speed. The following is some work related to most popular search engines.

## 2.1 Web Crawler

Web crawler is a core component of most search engines. Web crawler provides the function of data collection, which can update the database behind the search engine to ensure that the data for user query is up-to-date.

The result of crawling is a collection of websites at a central or distributed location.[[1]](#endnote-0) There are a few web crawlers behind current mainstream search engines. Two of them are listed as followed.

1. **Googlebot**

The most famous search engine as we know, Google, uses a kind of web crawler named Googlebot. Googlebot collects web pages and build a searchable index for Google Search engine. This name is used to refer to two different types of web crawlers: a desktop crawler (to simulate desktop users) and a mobile crawler (to simulate a mobile user)[[2]](#endnote-1).

1. **Bingbot**

Bingbot is a web-crawling robot (type of [internet bot](https://en.wikipedia.org/wiki/Internet_bot" \o "Internet bot)), deployed by [Microsoft](https://en.wikipedia.org/wiki/Microsoft" \o "Microsoft) October 2010 to supply [Bing](https://en.wikipedia.org/wiki/Bing_(search_engine)" \o "Bing (search engine)).[[3]](#endnote-2) Bingbot has the same principle and tasks with Googlebot. Bingbot collects web page information from Internet nodes and stores it in distributed system.

The work of web crawlers is very similar. They crawl the information of web pages from Internet nodes as the resource library content of search engines. The data obtained by these crawlers is provided by the meta information of HTML pages (always are the title, description, keywords). There are some characteristics for these search engine bots: (1) It is impossible to crawl all the data from the Internet; (2) Crawlers do not consider data correctness or quality; (3) Crawler is a kind of automated script. (4) Both of the bots have one thing in common, that is allowing users to block crawlers.

## 2.2 RD and RDS

**RD** (Resource Discovery), this is a process of searching valuable information on the Internet.[[4]](#endnote-3) The IETF-**RD** argues that resource discovery should provide the user consistent, organized view of information.[[5]](#endnote-4) **RDS** (Resource Discovery Server) can return a set of resources, these resources are links or indexes of web pages from the Internet. Various of search engines support **RD**, such as: Google, Bing, Baidu, etc.

When I am writing this paper, I type ‘English Learning’ keywords on Google and it returns 10,010,000,000 results within only 0.48 seconds. How does search engine do such a fast query?

Users only concern with a few useful results of all the resources results. How does search engine rank them and give a matched list of valuable results?

The above two questions describe the two core tasks of traditional **RD** search engines, storage-query and results-rank. The following sections are going to bring you some existed common solutions used by common search engines.

## 2.3 Distributed Storage System

Almost all search engines that need to use distributed storage system to store a large quantity of resources, Google is used as an example in this section.

Google has its own file system named **GFS** (Google file system). **GFS** is a scalable and classical distributed file system for large distributed data-intensive applications[[6]](#endnote-5). **GFS** has been used in Google since 2003, and it is not open source. However, its basic storage techniques are generally made public.

The following Figure 1 is the basic structure and work principle of **GFS**.



Figure 1 Google File System (GFS) storage and query

When the search engine application query keywords, the processed keywords first go to a master server. The master server doesn’t store any chunks or files, it only stores the file system namespace and mappings to the chunk locations. A large number of resources are stored on the chunk servers, files are divided into multiple chunks. The main server can return the corresponding address of the chunk, then the search engine can directly find chunks through chunk servers.

The reason why Google does this is because Google gets billions of resources from the Internet crawler every day, which can not be stored by a single server, or in a database or a data table. These resources, web pages and web addresses can only be divided into several small chunks and stored in the distributed file system. Moreover, one same chunk may exist on different servers for many times. The chunk server itself can copy these chunks to prevent loss. Chunks may be stored in any corner of the **GFS** servers, but since the master server records all the chunks mapping relationships, chunks are very easy to find and search engines get the data sets from the file system in a very short time.

In the distributed system, it is not very difficult to store and query millions of data. The specific search algorithm in the distributed system is no longer redundant in this paper. You can refer to the related papers of distributed system or **GFS**. What we need to discuss here is, in a traditional search engine system or a resource management system, distributed storage is the only or a general solution.

## 2.4 Query and Rank

The main purpose of efficient storage is for efficient query. Fast query speed is one of important user experiences expected for all search engines. Google, Bing, Baidu and other major search engines commonly use the following cycles to process a user’s search request:

1. Accept a user query
2. Parse query strings
3. Figure out the keyword order
4. Look up the information in databases
5. Rank the results
6. Send back the results

In order to search for related resources among thousands or more data, traditional search engine systems use cache, pre-fetching results, memory indexes and other methods to shorten or speed up the search life cycle.

For the step (5), one basic ranking method that search engine systems use is Vector-Space[[7]](#endnote-6). The Vector-Space ranking principle follows a score of significance. First we give the main function, Rank().

Function **Rank** includes a set of algorithms (will be discussed in detail below), it will return a list of sorted valuable resources after measuring the different aspects of resources. In this method, the search keywords of users’ query(Queries) are essential in this function which are used to match the resources.

Traditional search engine uses Vector-Space model to compute the results of rank function. In Rank() function, there are two main parameters, ‘Queries’ and ‘Resources’. Let’s simplify the two parameters to Q and R in the following context. Q is a vector, in computer language, it can be an array like *[ Q(1), Q(2) ... Q(i) ]* and R is another array *[ R(1), R(2) ... R(j) ]*. It comes to ‘many *Q* to many *R*’, the **Rank** function needs to calculate the count of *Q(1)* appears in *R(1)*, the count of *Q(2)* appears in *R(1)*, to summarize as the following formula.

*SR(j)* is the score of *R(j)*, **count()** is another function which is used to count the number of times that *Q(i)* appears in *R(j)*.

*V* is the final value of a resource, *SR* has been explained, Freq is the frequency of the resource shows in the whole internet(in the range of search engine statistics). In the example, there are *j* resources that the search engine need to sort. When the *SR* (score of a resource) is higher, the more valuable of the resource (here is web page) for the user is. The more frequency of the resource appearing on the network, the more multiples, leading to higher value, he more ahead the resources are ranked in the final results. In order to compare and sort all the resources, we have to calculate from *V(1)* to *V(j)* which means to calculate all the values of resources.

The final *V* is used to rank the resource list and display to the users. The order is as follows.

The first parameter is all the final values of all the resources, second parameter means it uses descending sort.

Modern search engines also collect users' behaviors, which nowadays have a lot of compliance to follow for privacy and security reasons, so search results returned by search engines are highly related to users' interests and habits, unless such feature is chosen to be disabled.

The above two sections 2.3, 2.4 discuss about the storage-query model and results-rank which are used by main current search engines. These methods will also be used in our learning resources search engine, we will also expand some algorithms which are suitable for the current learning resource system, refer to chapter 5.

## 2.5 Value of resources

In the paper of Identifying Valuable Resources, which is published in 2007 on European Management Journal, value of recourses was discussed, from the point of business and management. It is difficult to identify resources in a firm if there is no agreed definition of what ‘valuable’ means.[[8]](#endnote-7) A valuable resource can be rare, inimitable, and non-substitutable to be a source of sustainable competitive advantage.[[9]](#endnote-8) The problem is known by RBV (resource-based view) advocates, but the value of resources was never clearly defined. It is said, to identify the value of resources in business and management area, there are some questions to ask: What is the source of resources? Valuing one resource or many resources? The past, present, future value of a resource? Objective or subjective valuations? What’s the cost of resources? The conclusion is that valuable resources can generate three types of competitive advantage: cost advantage, the ability to premium price, and volume-based advantage.[[10]](#endnote-9)

This idea of using multiple characteristics of a resource to determine the degree of value is inspiring. Similarly, the normalized resources attributes and evaluation methods can be also adopted to determine the values of learning resources, with careful design. The general evaluation method is borrowed from the business field. First, put forward the problems of learning resources; then, give some resource attributes according to these problems. One resource is “divided” into several pieces by several attributes; and these attributes can easily be evaluated; finally, we combine all the values of the attributes into a final value of the resource. However, because learning resources are quite different for those in business and management, all characteristics need to be redesigned. For example, in business and management area, cost is an important element, mostly in monetary, but the cost advantage for learning resource is generally in time. Moreover, in our design, learning resources are endowed with more attributes, which will be discussed in Chapter 3.

# 3 CHALLENGES AND DIRECTIONS

## 3.1 Issues of Collecting Contents

An all-purpose search engine like Google requires super computing power and storage capacity. The search engine proposed in this thesis is to be used for education specifically, aiming to support users to find useful learning resources. Take the massive amount of educational videos alone, the burden on storage system will be huge. Moreover, copyright protection definitely cannot be neglected.

### 3.1.1 Resource Crawler

Because of these concerns, different from the traditional search engine, this learning resource search engine does not use crawlers to obtain web page contents. The system only obtains and stores the information about the learning resources. To show the difference, this solution is named as “**Resource Crawler”**, instead of web crawler. As a result, this search engine will be a significantly light system to achieve high performance. The copyright issues and system over-storage problems can be both avoided.

The “Resource Crawler” collects the information of resource, in another words, the meta-data of resources, such as titles, locations, tags, publishers, descriptions, comments, and etc., many of which obviously require input from users. Meta-data are organized in system database, supporting inquires. Through these information, the value (quality) of resources can be defined or estimated, and users can search and locate the original resource content. It helps users find useful resources and study.

### 3.1.2 Data Storage and Query

Even though only resource information are stored, the database needs to handle the potentially very high volume of growing data. Users need quick searches; and they depend on a high performance design of the search engine system, in particular, an efficient storage structure to support information access. Every query is filtered from such a huge amount of data. The system built for this thesis uses a 3-layered storage structure and an encoded keyword mapping method to improve the search efficiency. For the application in the actual production environment, the storage needs to adopt the distributed system design.

## 3.2 Issues of Defining Resource Value

There are many challenges to address to collect valuable resources. First of all, how to determine whether a recourse is valuable? The Internet Crawler “spiders” are not something smart like human beings, the only thing they do is collecting and bringing the copies of the information back. Much meaningless and even fake or unhealthy information is obtained too. The search engine should not only filter bad impropriety or dangerous information, but also recommend the content according to the user's interest; however, that is still far from the definition of being valuable resources. Some issues can be magnified to different areas and user groups.

### 3.2.1 Value related to resources

In the field of education, there are many issues exposed by traditional search engines. It could be directly regarding resource quality. Below are some of the practical questions related to resources:

Is the resource reliable? (Does it provide correct information? Is the provider reliable?)

How relevant is the resource to the user's search? (e.g. matching keywords.)

Is the resource up-to-date? (When was the resource last updated? )

Is information appropriate? (Is the resource safe?)

Is the information redundant? (Are they repeated copies of the content?)

How much time is needed to go through the resource? (e.g. length of a video.)

What is the monetary cost? (Is it free? How is the price?)

How many positive feedback? (Do other users find the resource useful?)

To solve this problem, the new search engine is designed to meet educational purpose, to support different types of users, returning valuable search result and maintain query efficiency in term of speed. Several techniques are created. We summarize them into two problems to address:

How to collect and manages education resources?

How to define a valuable learning resource?

How to quickly process queries from users ?

### 3.2.2 Value related to user types

Even if the resources are reliable and safe, for different user groups, they can be still be good or bad, because users of different ages, level of educations, skill experiences, and even learning preferences have different needs and expectation of what they look for. Ideally, search results should be close to the ability of different users. Below are some questions more from user perspective.

What age group is the recourse suitable for?

What grade level is the resource for?

What type of resource is it? (e.g. recording, cartoon, for user preference)

Is it for professionals in the field?

What experience level is the resource for? (e.g. entry-level, intermediate, senior)

Is it made for users of special ability? (e.g. language, disability)

Inspired by the advantages of valuable resources in business field as discussed in Section 2.4, we define the value of learning resources, based on the following characteristics:

Reliability

Practicability

Suitability

Popularity

Feedback

Cost

These 6 characteristics are used to determine whether a learning resource is reliable and with high-quality learning resource. As shown in Fig 2, the 6 characteristics are used to evaluate the value of resources, of which light blue is the dynamic characteristic and gray ones are static characteristics. ‘Dynamic Characteristics’ are generated dynamically during searching and ranking, based on input keywords, while the ‘Static characteristics are stored in the database statically. “Static” is not “constant”. It means a characteristic is only changed when the data of resource is changed, such as feedback, cost, and etc. Dynamic characteristics are generated in real time.



Figure. 2 The characteristics for DLRV

# 4 SYSTEM DESIGN

## 4.1 Requirements

### 4.1.1 Hardware requirements

The whole search engine platform requires several high-performance servers which can potentially undertake billions of requests from users. Search results are listed in the browser. We call these web applications or B/S architecture, B/S is browser and server, a kind of application allows users use browser access to server. Generally speaking, the requirement on server performance and system configuration is to meet the demand of user number and resource volume. Therefore, in the early stage, there is no detailed requirement report for the whole set of server-side hardware, but we lay out some basic requirements presumably.

The dual core processor E3 based on X86 system

More than 16GB RAM

80GB disk storage

Independent database server

The distributed deployment server based on Linux is prepared with container and virtualization technology, but will not be used in the experimental time

Data analysis server and other micro services

For user-end or test-end, we require PC and mobile devices to test all the web pages and functions to work well on various browsers.

### 4.1.2 Software requirements

The complexity of software requirements is much higher than that of hardware. All algorithms, technical details and functional requirements are implemented and verified by software programming. We can use the normal web development environment, tools, languages and related SDK to implement.

1. **MVC design pattern**

MVC pattern is a very classic design pattern in software engineering. It was first proposed by Trygve Reenskaug in 1978[[11]](#endnote-10), and later became popular in web development area. A framework based on this design pattern can be called MVC framework. All the development and implementation described in this thesis are based on this design pattern. In another words, the search engine system in this project adopts MVC framework. MVC separates model, view and controller. In actual development, model is data level, view is front-end, and controller is the part of business logic. MVC design pattern can achieve high cohesion and low coupling, and it separates data, view and business logic. MVC improves the development efficiency, code cleanliness, and has higher scalability. The purpose of using this mode is to make the search engine easy to optimize and expand the function in the experiment.

1. **Related application software**

MVC framework based on Node and express.

The view layer is based on VUE[[12]](#footnote-0), also called front end, which is used by users.

Reverse proxy server and HTTP server are based on Nginx.

MySQL database, a kind of relational database, stores a large number of data generated for the search engine system.

Redis, a kind of non-relational database, which stores data in memory, used as cache in our design.

### 4.1.3 Non-functional requirements

To build a real search engine system, there are more to consider as requirements, such as of those that are non-functional. While we don’t take all of them in our prototyping, they are listed here for completeness.

1. **Performance**

Google answers 100 billion searches per month[[13]](#endnote-11). That means the average of a day is at least 3 billion, based on the statistics in 2012. Our learning resource search engine doesn’t need such high search performance because we are targeting at one special area, instead of all the users and resources on the Internet.

Around 2017, there are more than 30 million children use Google education apps[[14]](#endnote-12), this is not including college students and other adults, so our system needs at least double of this amount (children) for users’ requests, assumedly to accommodate 60 million users per day. More formally, DAU (Daily Active User) is at least 60 million. To accommodate extra requests from the increasing DAU, we raise the performance bottleneck to 100 million DAU in our development plan.

1. **Reliability**

The operation of all the services are 24 hours. To ensure users get search results within 1 second after starting the search, the response speed should be less than 1 second for each query.

1. **Security**

System layer security:

Firewall between server nodes, access control on blacklist, white-list and iptables technologies.

Data backup to prevent the data loss disaster.

When main servers crash, use the reserved servers instead.

Quick recovery plan for crashed servers.

1. **Business layer security:**

The security points of business logic are listed bellow

User verification

API requests security

User behavior logs

Cookie or cache security

User privacy

Encrypt and decrypt data

## 4.2 Software Engineering Process

In real system design, there are engineering design, system architecture deployment design, and unified modeling language. Online system and theoretical verification of the system should follow all of the design principles. In Figure 3, the design of software process follows the life cycle of software engineering and adopts agile model, which approaches development requirements and solutions through the collaborative effort of [self-organizing](https://en.wikipedia.org/wiki/Self-organizing_communities" \o "Self-organizing communities) and [cross-functional](https://en.wikipedia.org/wiki/Cross-functional_team" \o "Cross-functional team) teams and their [customer(s)](https://en.wikipedia.org/wiki/Customer" \o "Customer)/[end user(s)](https://en.wikipedia.org/wiki/End_user" \o "End user).[[15]](#endnote-13) It advocates adaptive planning, evolutionary development, early delivery, and [continual improvement](https://en.wikipedia.org/wiki/Continual_improvement_process" \o "Continual improvement process), and it encourages flexible responses to change.[[16]](#endnote-14)



Figure 3 System Development Process, based on Agile

The whole implementation process is divided into 6 parts in this paper. Full software requirements will not be described in this thesis. Only some requirements are documented as to be explained next, including Framework design and UML. In addition, system development will not be fully described as well, except the key algorithms and methods specially created for this project. Testing and verification will be presented in later chapter 6.

### 4.2.1 System deployment structure design

The deployment of the whole search engine system follows the normal web deployment mode as illustrated in Figure 4.



Figure 4 System deployment architecture

If server deployment is distributed in multiple servers in the same Intranet or multiple networks, each server has its own work task and provides API or open port to other connections. This is a very popular deployment method of Web services, as it can handle large concurrent requests, reduce the coupling between services, and improve security. Multiple servers can be managed by different teams or individuals, making it easier and more efficient to cooperate with other each other. In Figure.4, from top to bottom, from left to right, there are user clients (PC or mobile with browsers), CDN, reverse proxy server, firewall, business logic server group, business server and database connection. There is firewall between them to control access, for the security of data center. Part of database server data is stored in high-speed non-relational database, such as Redis or MongoDB, to deal with some high-frequency search engine requests. High-performance storage structure will be detailed in Chapter 5 three layered search.

The purpose of CDN is to speed up the existence of static files. Here, CDN stands for a content delivery network, or content distribution network[[17]](#endnote-15), static files can be distributed on multiple nodes of the Internet.

Commonly applied in distributed system, when users access static data, the nearest fastest server is tried first, to improve the user experience.

The reverse proxy server distributes user requests to upstream servers, which can effectively reduce the possibility of congestion. At the same time, no server downtime will affect user requests.

There must be a firewall between the reverse proxy server and the cluster server to control the access list, which can be a white list and prohibit illegal users from directly accessing the cluster.[[18]](#footnote-1)

There are many servers in the server cluster, most of them are controllers dealing with business logic, There are also some servers specialized in processing big data. DLRV’s calculation is included. These servers, which are responsible for data processing, work without rest, sort and classify the resources and tags from the database, score the resources based on DLRV algorithms (ref: Chapter 5) and provide the core business for users to search the valuable resources.

Non-relational database, such as Redis and MongoDB, are very important parts of the whole system. Search engines have high requirements for the speed of search and data acquisition, and the structure of relational database can be very complicated. For some simple tag search, non-relational database and even cache database based on memory can provide search engine with more efficient results. Redis can save high-frequency search keywords in memory based on some page switching algorithms, which can effectively improve the search speed. More discussions will be provided Chapter 5 and 6.

### 4.2.2 Database Design

ER model (Entity-relationship model) is used to present to the logic of the entities and relationships among them. Modern web application development is typically database driven, and the design of relational database follows ER model design. Figure 5 is the complete database (relational database only) design diagram of the search engine system.



Figure 5 ER model in UML of the relational database in the system

In Figure 5, we can read: user table is used to store user’s information; resources table stores resource information; users are the owners of the resources. One user can add many resources. A user can pick multiple resources to organize a course. One course can include many resources. If a resource is used in a course by a user, “usage” will increase by 1 for that resource. ‘ctypes’ and ‘rtypes’ are tables to store the types of courses and resources respectively.

‘Users are the owners of the resources’, this become a extremely important point which makes this search engine different from others. Users have the right to monitor the quality of the resources. This is the key point of reliability (one of six characteristics) evaluation in DLRV system.

### 4.2.3 Use Cases



Figure 6 Use case diagram for the search engine functions

Figure 6 is the use case diagram of the search engine system which covers 7 main use cases. Most are for just the basic functions, and the core use case is for the search function. Other small or trivial functional details are omitted here. Users can generally considered as potential learners, though they can search for other users, e.g. parents for their children. Users can search resources, upload resources information manually, set up courses by group resources (resource information indeed) into a list, and evaluate courses or resources. When a resource is cited by a course, system will add 1 to usage value of this resource.

These use cases, including “review resources”, “cite resources”, “review course” and “upload”, will affect the DLRV system in defining the values of the six characteristics. For example, they can change the value of usage, feedback, reliability, and so on. These are called “user behavior effects”.

### 4.2.4 Package and Class

Package is a [namespace](https://www.uml-diagrams.org/namespace.html) used to group elements together that are semantically relevant or might change together. It is a general purpose mechanism to organize elements into groups to provide better structure for system model.[[19]](#endnote-16) For the server side of the search engine system as designed for this project, Figure 7 shows the main packages, mapped as different folders or collections. Package have dependencies among them.

* **Controller**. This is a package includes all the controllers. It is designed in MVC mode. Controllers deal with all the business logic, take responsibilities for connecting data and views, and accept users’ requests and responses. Controller depends on private libs and public modules.
* **Server.** It is the entrance of the whole system. This package manages the files to work as a web server. The data from user-side enter this package first. It depends on controller because the request and data from user-side need controller to serve them. It depends on public modules.
* **Config.** A package manages the connection configuration of various servers, like mail server, database server, OSS server and all other servers needed in the search engine system. At the same time, this package contains some configuration of the system itself. Config package doesn’t depend on any other packages.
* **Private libs.** The package contains all the private modules, plugins used only in this system. It depends on public modules and Model package.
* **Models.** Models as a unit is the central component of MVC. It is the application's dynamic data structure, independent of the user interface.[[20]](#endnote-17) It can manage the data, logic and rules in the system.
* **Modules**. Modules are the public modules. Public modules are some free software, some of which are public plugins for various developing language. They are from the Internet, generally maintained by the community or individuals. Most of them are open-sources.



Figure 7 System package diagram

In Figure 7, **Config** and **Modules** packages have the most dependence from other resources. **Config** contains all the configuration of the whole system. **Modules** here are public plugins downloaded from the Internet for this project.

**A. Controller Package**



Figure 8 Class diagram of the package Controller

The controller package is the core package of the whole system. It contains classes which are used to deal with the user's business logic. The **Search** class solves the search requests from all users. The **Config** class returns the configuration of the site to the front end. The **Resource** class is responsible for data collection, resource classification, deletion and other functions. Through the **Resource** class, users can also edit courses and get course lists.

**B. Private Package**



Figure 9 Class diagram of the package Private

The core class of private package is the class **Public**, which means public library in a private project. It is a public library specially developed for only this system (search engine), it includes functions such as get random numbers, format date, check user info, and etc, which are used with high frequency and commonly. It can be imported and used by controllers to reduce code redundancy and coupling. It is a very common design idea in software engineering. High cohesion, loose coupling[[21]](#endnote-18). Class DB provides a set of methods to operate database, such as inset, query and delete.

**C. Server Package**



Figure 10 Class diagram of the package Server

The server package is the web server start-up entry. It listens and distributes the user's request to the controllers. The mapping mode used between the server package and the controllers is called **Convention Routing**. Figure 11 explains what is the convention routing mode.



Figure 11 Router to Controllers

The solution is to automatically map the user's router to the same class name and action name under the controller. The characteristic and advantage of the convention routing is that it does not need to configure the route files, which reduces the development time and the writing of method documents. It can reduce the possible misunderstanding between the front-end and back-end communication.

In this system, the static file has its own unique mapping way, different from the controller, so it can isolate the access of code and media files, for security guarantee. Static files, like images, CSS files, fonts and other files, can be used by the website visitors or browsers, but not the source code and executable files.

### 4.2.5 Sequence Diagram

The main function of search engine system is to search valuable resources, so in many sequences series, this part mainly studies the logic of search sequence. The following is the UML design of search sequence.



Figure 12 sequence diagram of search process

The whole search process of the system: after getting the user's request from the server, the keywords are sent to the Search controller to screen the data in resource database. When the matching and approximate data results are found, the Search controller calculates and verifies the value of the resources through the DLRV module service, and then returns the organized results to the controller. Finally, the controller returns a sorted list of results to the user browser through the HTTP server. Users get valuable resources that match their requests.

# 5 ALGORITHMS

This chapter describes key algorithms and methods proposed for this learning resource search engine, to achieve the goal of “good user experience”. It means: (1) Fast access to search results, and (2) Valuable learning resources, which is in line with the theme of the thesis. From programming point-of-view, they are ‘storage and search efficiency’ and ‘resource sorting and filtering’.

The definitions and methods in traditional search engines is about **RD**, which stands for Resource Discovery[[22]](#endnote-19). **RD** has been explained in chapter 2.2. The improved innovative algorithms used in the learning resource search engine system is named **DLRV**, which stands for “Degree of Learning Resource Value”. Like the approach of **RD**, **DLRV** also contains two parts: storage-query and results-rank.

## 5.1 Disadvantages of RD

In the chapter 2, the related work, we have discussed:

1. Distributed file system to store and query the resources
2. Vector-Space algorithm to compute and rank the value of results.

But in the special search engine of learning resources, these methods have some disadvantages. It needs to be pointed out that the disadvantages of **RD** approach listed here is only regarding the case of a special learning resource search engine in this thesis.

The distributed system takes a large place of storage and needs hundreds of distributed servers. In this system, we do not save the original files or resources, we only save the links and main information of resources. This logic is like the master server used in **GFS**. Therefore, if we distribute resources to multiple servers, it will waste a lot of physical resources, and in terms of software design, it is also extremely complicated to create such a system.

The bigger difference between learning resource search engine and traditional **RDS** is the second part, results ranking methods. Go back to chapter 3 about challenges, we have listed a number of search engine problems in this particular area, learning resources. If we use the basic ideas given in the section 2.4: to count the number of times keywords appear in a document, then multiply the count number and frequency of appearance on the network, finally we get a score to reflect the value, or significance of a resource. This score is too simple for learning resources. Besides the degree of matching keywords and appearing frequency, learning resources have more important attributes, which determine the quality of the resources. The recommendation ranking of the search results of learning resources must combine the quality of the resources themselves, and also consider the matching degree and frequency in traditional **RD**. In the section 3.2, the figure 2, displays the 6 vital attributes of a learning resource: Suitability, Cost, Popularity, Reliability, Practicability, Feedback. Among above attributes, **RD** can measure two, suitability and popularity. High matching can be involved in suitability, frequency of appearance on the internet can be included in popularity, that means high suitability and popularity are used commonly in **RDS**.

Therefore, in the learning resource system, the remaining missing attributes become the defects of traditional search engines, and only sorting the values of resources based on the above two points (Suitability and Popularity) does not work well in learning resource search engine system. Compare to **RD** search engines, DLRV search engine contains more attributes according to table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Suitability** | **Popularity** | **Reliability** | **Practicability** | **Feedback** | **Cost** |
| **RD** | Yes | Yes | No | No | No | Not sure |
| **DLRV** | Yes | Yes | Yes | Yes | Yes | Yes |

Table 2 Characteristics in RD and DLRV

In addition to the methods in resource storage and ranking aspects, there are some other features of the shortcomings compared to common search engines, here to do a simple list.

The common search engines can not allow users to comment, reply, like on the resources.

The common search engines can not allow users to reference resources.

These search engines don’t bring user feedback into DR, this approach leads to the lack of human action values when search engine do the query and sort. Starting from these shortcomings of DR, in the next chapter we will show how DLRV improves all these attributes of learning resources in storage-query and results-rank two main aspects.

## 5.2 Storage and Query on DLRV

There is no need to use distributed storage in the learning resource system, especially the distributed file system. We use a combination of relational database and non-relational database to save data. The database saves the information and attributes of the original resources, including the title, introduction, link, file type, price, and thumbnail of the resources. The following is a diagram of the storage system in DLRV way.



Figure 14 …

In the DLRV resources system, resources are stored in the traditional web storage mode, ‘database center driver’. It means all the basic information and relationships of resources are stored in the databases. Databases are divided to three parts in this system, one relational database, one non-relational database in disk, one non-relational database in memory cache.

### 5.2.1 Relational Database

A relational database is a digital database based on the relational model of data.[[23]](#endnote-20) This means that the things stored in such database become more complex. In addition to entities, it also stores the relationships between entities. The basic unit of a data structure in a database is a table. Each instance of the data is called a row, which is stored in a table. In DLRV, relational database is used as our main storage container because the information of the learning resource itself is a specific entity and it has a one-to-one, one to many relationship with publishers or reference objects. The figure 15 shows some relationships and entities in the database, the resources stored in data are absolutely not independent individuals. These complex relationships are reflected in the business logic function points of this search engine website, such as reference resources, publish resources, comment resources and other more functions.

However, relational database is not very effective in searching, sorting and analyzing, especially in the case of large amount of data, so most of the role of relational database here is to store resource information. The following will introduce how relational database cooperates with non-relational database to carry out data quickly, how to improve the speed of user search and reduce the time cost. Time cost is one of the evaluation attributes of resource value.



Figure 15

### 5.2.2 Non-Relational Database On Disk

The characteristic of non-relational database is that the storage structure in it is very simple, without dependent relationship between data and data. So when retrieving a piece of data, its reading speed is relatively faster, especially in the case of large amount of data. From figure 14, it is a set of storage solutions. The figure shows three databases, two of which are permanent storage database which store data on hard disks and one is cache database which stores data in memory. Between the three, the reading speed of data is:

Cache Database(Non-relational database) > Hard Disk Non-relational database > Hard Disk Relational database

Please refer to query performance of different modern databases in the following data table.[[24]](#footnote-2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Operation | Oracle | MySql | MsSql | Mongo | Redis | GraphQL |
| Insert | 0.091 | 0.038 | 0.093 | 0.005 | 0.010 | 0.008 |
| Update | 0.092 | 0.068 | 0.075 | 0.009 | 0.013 | 0.012 |
| Delete | 0.119 | 0.047 | 0.171 | 0.015 | 0.021 | 0.018 |
| Select | 0.062 | 0.067 | 0.060 | 0.009 | 0.015 | 0.011 |

Table …

Among these databases, Mongo, Redis and GraphQL are all non-relational databases, our search engine system uses Mongo to store tags and short words as the keys. Therefore, after getting the user's keywords, the speed of searching these short words in the non-relational database is very fast, and these short words will be further indexed, which will also be described in this paper later. And a unique index can further speed up queries.



Figure 16

As figure 16, this is the ‘key-value’ storage data structure In Mongo. The ‘key-value’ data structure is very normal in non-relational database which means one key to one value, we can find value by key. The ‘key-value’ data structure actually is stored as hash map on the computer so its time complexity is O(1).

Because the same keyword may map to multiple resources, which means a keyword can query multiple resources. In this system, the key is a keyword(tag, title and short words etc) and the value is a set of resource ids, ids are stored like [id1, id2, id3, id4].

Note, although there are multiple resource ids here, they are stored in one space as a string split by a special table like ‘,’ or ‘|’ which is easily divided to array by any computer language.

This is very important here, which is a key to accelerate keywords query. If in a relational database, these resource ids are related to keywords, and they will be foreign keys of resource data. This requires the database to search multiple rows, read the entire library or use search optimization algorithms, such as binary tree search. But if we only use one ‘key-value’, the database will only query once, the time complexity is O(1). There is only a hash map searching in the database.

In addition, MD5 is used to encode the key in ‘key-value’. MD5 is an algorithm for inputting variable length information and outputting 128 bits of fixed length. The purpose of storing keywords in this way is to ensure that the key stored is always 128 bit space taken. This makes the storage for key is space complexity O(1) too. But MD5 is not decode-able, does this affect the search? Let’s see the following table first, it is the storage example for ‘key-value’ in our search engine system. The left side are the keys they are MD5 encoded. They can be tags, keywords, titles and any other short words. The right side are the corresponding ids for the resources.

|  |  |
| --- | --- |
| **Key(keywords, tags, title)** | **Value(resource id)** |
| ba0a6ddd94c73698a3658f92ac222f8a | 1,2,3 |
| c31b32364ce19ca8fcd150a417ecce58 | 4,5,6,7,8,9,11,12 |
| 4dbe9ff7f2742c912b53b9feab9f343e | 6 |

The searching process is as the following diagram, figure 17. In this flowchart, we can see the steps as follows: first, enter the search keyword. Then we encode the keyword by using MD5 to get the key, use the key map to value of the IDs for resources, and then get the resources’ related information according to the IDs list. Here, ID is a primary key in the MySQL resource table, so the search speed is fast.



Figure 17

### 5.2.3 Non-Relational Database Cache

The cache database of the system adopts Redis, which is a database based on memory, which can also store the data as key-value in database persistently. Cache database can be used to speed up the search.

1. Redis can store a large amount of data. It supports the 2^32 keys in hash map, and the maximum size of each key or value is 512M.
2. Redis can be scalable and distributed.

The reason for using the cache non-relational database is that although MongoDB mentioned in the previous section can speed up the reading query, when the amount of data reaches one-day 100 million rows storage, the storage space will be insufficient first, and then the query speed will be significantly reduced.

The effect of using cache database like Redis is that:

1. Using distributed storage can divide data to several servers, to balance the storage pressure.
2. The speed of reading and writing memory is far higher than that of hard disk, the frequently read and write data should be stored in memory.

Now here we get a problem. Although the memory speed is much faster than the hard disk, on the other hand the storage space is far less than the disk space. Take a personal computer with 16GB memory and 1TB hard disk as an example, the memory is only 1% of the hard disk.[[25]](#footnote-3)

So we only store part of the content in memory, so what kind of data needs to be stored in the cache non-relational database and how to store these data effectively(storage methods).

### 5.2.4 Three Layered Search

Question: What kind of data needs to be stored in memory? It is explained in figure 18.

To sum up, the data stored in memory are the keywords with high search frequency, we put these keywords in cache non-relational database as keys, and the corresponding values are the IDs of the resources. We have already discussed ‘key-value’ mode in the previous section. As shown in Figure 19, when a user searches for resources, the input keywords first enter the cache non-relational database, query the IDs of the resources, and then enter the relational database to get relevant resources information. The whole process is direct, no longer passing through the disk non-relational database. When the keywords can’t be found in the cache non-relational database, then it searches in disk non-relational database. Note that the disk non-relational database saves all possible keywords and tags, but that doesn’t mean cover all the resources. If the controller still can’t find the keywords in disk non-relational database, it will search in the relational database directly finally. Normally, the process will not go to the last strategy, searching directly in relational database.

For the whole storage structure of the layered search, logically, this search engine system storage structure can be divided into three layers by query speed.

The third layer is the super high speed layer, the layer where the query occurs first, the second layer is the high speed layer, and the first (basic) layer is the normal layer. The speed of query in layers is from low to high. Please refer to figure 20 to understand the 3-layers’ search storage structure.



Figure 18



Figure 19



Figure 20

### 5.2.5 Cache Switch Method

Memory is limited, even if we use distributed storage. We need to make sure that the memory is used within a certain range and that the original operating system runs smoothly. In this way, we need a set of algorithms or methods to ensure that the storage size of the non relational database memory area is kept within a reasonable range. Therefore, it is impossible to save all the keywords as indexes in the cache non-relational database. Here we refer to the idea of page switching in operating system area.

About paging, it is a technology of memory management in an operating system that enables the main memory of computer to use data stored in auxiliary memory. The OS will divide the data in auxiliary storage (usually disk) into fixed size blocks called “pages”. When it is not needed, the page is moved from main memory, when necessary, the data is retrieved and loaded into main memory.[[26]](#endnote-21)

Page is a storage unit and exchange unit in OS paging technology. In our search engine system, a ‘page’ is a key-value unit in redis(a cache non-relational database). As the keyword is encrypted by MD5, as described above, an MD5 encrypted hash value is 128bit, so it can be understood that a page needs at least 128bit. However, please note that the resources’ IDs are also stored in a key-value unit(a page), so the resources’ IDs need extra storage space. In redis a value of string can be 512M. Google allows about 25 billion results for each search. We can take this number as the max search results. Therefore, the value in a ‘key-value’ unit can store 25 billion IDs at most. Let's calculate the storage space of a key-value in the maximum case.

Conditions:

1. 25 billion IDs in a value
2. The max length of id is 11
3. 1 English character needs 1 byte with UTF-8 encoding.

Result:

From the calculation results, the max value 262260MB is more than 512MB which means a value needs at most 512 times (262260/512=512).

Absolutely, one hash map index only maps to a list of resources id, so we also need to calculate the search frequency of resources, and only store the high frequency resource ID into the value so that this value will not be more than 512MB.

So far, we need to pick up high-frequency keywords, high-frequency resource IDs. The former is for the lack of memory for keywords storage, and the latter is to solve the problem of limited storage in redis.

In order to achieve effective page switch(key-value unit switch) work, the commonly used algorithms are as follows.

LFU(Least Frequently Used), LRU(Least Recently Used)[[27]](#endnote-22), FIFO(First In First Out) and Clock. Among the above, LFU is suitable for this search engine system.

By using LFU rule, in cache non-relational database we remove the lowest search frequency key-value unit. For example, we store the frequency with IDs in value for each key. A table shows keywords, IDs and frequency in cache non-relational database.

|  |  |  |
| --- | --- | --- |
| **Keywords** | **ID** | **Frequency** |
| Key1 | 1,2,3,4,5 | 10000 |
| Key2 | 6,7,8,9,10 | 20000 |
| Key3 | 1,3,9,11,12 | 30000 |
| Key4 | 12,23,45,222,657,12321 | 50000 |

The switching process is shown in figure 21.



Figure 21

If, these are the units already in the cache database and the database is full. At this time, a new keyword key5 that is not in the cache is searched more than the lowest keyword key1, then key1 will be replaced by key5

## 5.3 Rank based on DLRV

The biggest difference with normal search engines is that a learning resources search engine needs to be evaluated more strictly and on more features.

In the section ‘Value resource attributes’. We give a basic model to evaluate learning resources, which contains six characteristics (Figure 2). We use these six directions to evaluate the value of a learning resource. This basic model is explained below. We call it the DLRV model. We get the value of resources by quantifying the six characteristics, and then return them to users according to the value from high to low.

### 5.3.1 Static and Dynamic

Characteristics are divided into static and dynamic. Dynamic means that the value of this characteristic is generated dynamically, and different values are generated each time because of different conditions. It has no direct relationship with the resource itself, and different conditions can produce different results even on the same resource. A dynamic characteristic is never stored in any database and a static characteristic is stored in the database.

Among the six characteristics, only one is dynamic, which is the ‘suitability. It is related to the user's search keywords. The other characteristics, cost, reliability, practically, popularity, feedback are all static. Which means they are determined by the status of a resource itself. And the values of static characteristics are stored in the database statically. Static is not constant. It is just that static characteristics will not be changed in real time and will not change due to user search conditions. Static value will also be changed due to the change of resource quality and nature, such as views, evaluation, publisher, etc.

The methods to calculate the final value of learning resources in a **DLRV** system are as follows:

The total value of a learning resource, *V(j)*, is the dynamic value of characteristics, *Dynamic(j)*, multiply by static value of characteristics, *Static(j)*.

### 5.3.2 Resource Evaluation

1. **Suitability**

Suitability is a dynamic characteristic of **DLRV**, which means, the value of this characteristic is not determined by the resource itself, it is not a static value, and will change as well each search behaviour.

Suitability means: the results can match the user's search keywords and purposes. If it is different from what users want, or even the opposite result, we call it an unsuitable resource. How to fix what users want? We can set all the keywords into an array, as *[ k(1), k(2), k(3), k(4) ... k(i) ]*, and set the resources as *[ r(1), r(2), r(3), r(4) ... r(j) ]* this step is the same with traditional **RD** search engine in section ‘Introduction of Traditional **RD** Way, Rank of Results’. We still use keywords array to match the information and content of resources. Because these keywords represent the general purposes of the users.

The concept and method are shown above. Add the frequency of each keyword in the resource title, tag and description. The higher the frequency of occurrence, the higher value of resource suitability. Finally, accumulate the number of times keywords appear in the title, description and tags to get a value of suitability. The value of the suitability should more than *0*.

*S(j)* is the value of suitability for *r(j)*, *r(j)* is a resource.

1. **Reliability**

Reliability is determined by the source of resources, which we call resource publishers. Resource publishers can be divided into personal accounts or unit accounts, and personal account can be divided into certified or uncertified account. The unit must be certified, unit is generally an organization or an enterprise.

The reliability value for a resource is calculated as follows, a user can be generated as user(i):

In this function, *R(j)* is the reliability value of a resource *r(j)*.

This is a piece wise function:

1. For the first piece, if a user is not certified (variable *cer(i) = 0*) then the *R(j)* is 0.
2. The second piece is under the condition that a user is certified but not a unit account (not an organization or a company), then the *R(j)* equals *k1* multiply the total usage of all this user’s resources so far. If a user *u(i)* publishes resources *[ r(1), r(2), r(3) ... r(k) ]*, the reference or usage of each resource is *[ u(1), u(2), u(3) ... u(k) ]*

*k1* is a weight number. It can be adjusted according to the experiment to fix the value of reliability.

1. The same way to understand the third piece. It is under the condition that a user is a unit account as an organization or a company. *k2* is another weight number set for unit account. The same, we need to sum all the usage of the resources the user has published.
2. The count of the usage in this system is how many times a resource itself has been referenced in courses or other resources.
3. **Practicability**

Practicability is also called usage. The more times a resource is used, the more practical it is. The value of practicability of *r(j)* is *U(j)*. *U(j)* is the total number of times a resource is referenced.

1. **Feedback**

User feedback is a direct reflection of a resource quality. Users' evaluation of resources can be divided into good or bad. Good feedback can add resource value and the bad feedback reduce the value. We use the proportion of positive comments to reflect the value of resources. If a resource has no feedback, we set this value 0.

In the function, *F(j)* is the value of feedback. *p(j)* is the positive feedback which is called good feedback. *n(j)* is the negative feedback which we call bad feedback. *t(j)* is the total feedback, it equals *n(j)* plus *p(j)*. So the value *F(j)* actually is the rate of good feedback for a resource.

1. **Popularity**

Popularity is a trend that describes the resources’ increasing views. The value of popularity is *P(j)*.

*V* is a number of accumulative views of a resource. *Vd* is the accumulative views of today, *Vd-1* is the accumulative views of last day. The number of views added today is the value of popularity.

1. **Cost**

Cost includes time and money. *C(j)* is the value of cost, *T(j)* is the value of time cost, *M(j)* is the value of money cost.

1. **Static Value**

The characteristics: reliability, practicability, feedback, popularity and cost are the static characteristics. They don't change because of the query keywords, so not dynamically changed when searching the resources. They are the static values stored with resources. We store the following value of all the static characteristics.

Static value of a learning resource adds reliability, usage, positive feedback multiply popularity and then the value is divided by cost of the resource.

Note, in the static function, we multiply Popularity and Feedback values, which means, if *P(j)* keep increasing but there is no Feedback, *P(j)* has no effect to the static value, according to *F(j)* formula.

1. **Dynamic Value**

There is only one dynamic characteristic among the six characteristics in the DLRV model. Suitability is changed dynamically when users search the resources. Different query keywords can course different suitability value.

1. **General formula**

Refer to section ‘Static and Dynamic’:

In the section, we have described the general formula is multiple the static value and dynamic value. So *V(j)* means the final value of a resource, it equals to suitability multiply the total static value.

# 6 EXPERIMENTAL STUDY

In the field of software engineering, software testing and quality assurance is an important part of software development. The method of testing is to use virtual data to test the functions of software. All the data obtained in the test can be used to evaluate the software performance, functional logic, algorithms and security.

## 6.1 Purposes

The main purpose of testing is to meet all the requirements and qualify the quality of the software.

We report the test in this section for the following purposes:

* To test the performance of 3-layered storage structure.
* To test the accuracy of DLRV ranking results.

The first purpose is to verify the query-storage part of the methods. The second part is to verify the DLRV ranking algorithm. These two parts are the core of this learning resource search engine, so we test them in this paper. Other requirements and functions can be verified or tested through similar methods, we will not repeat here.

The expected results of the test are

* Less search response time.
* More high-value resources are ranked topper.

## 6.2 Test Methods

Software testing is based on requirements and specifications of design. There are some common and mature testing methods in software engineering and we briefly introduce and apply some methods to learning resource search engine.

### 6.2.1 Black-box Testing

Black-box treats the software as a "black box", examining functionality without any knowledge of internal implementation, without seeing the source code. The testers are only aware of what the software is supposed to do, not how it does it.[[28]](#endnote-23)

Black-box testing will be used in this search engine system. It is used for testing the ranking part of the search engine. Black-box testing is very suitable for the whole user searching process, from keywords to the list of all the most valuable resources. The search function is a black box, testers don’t need to understand the principle of the searching algorithms and ranking methods. The testers only verify the value of the resources and rank results. Resources with high value are ranked at the top in the list, while those with low value are arranged at the back. Value itself is a more subjective concept. Later in this section, we will discuss how to measure the user's recognition of resource value.

### 6.2.2 White-box and Control Group

White-box testing (also known as clear box testing, glass box testing, transparent box testing, and structural testing) verifies the internal structures or workings of a program, as opposed to the functionality exposed to the end-user. In white-box testing, an internal perspective of the system (the source code), as well as programming skills, are used to design test cases. The tester chooses inputs to exercise paths through the code and determine the appropriate outputs.[[29]](#endnote-24) [[30]](#endnote-25)

We use white-box method to test the storage-query part, especially the 3-layered storage structure. This is not strictly a white box test. According to the previous section, white-box testing asks users to choose different test paths according to the logic of the code. However, here we are not changing the input paths to get the expected results for the code logic of a function. In this part, we query the same amount of data for different storage structures, and design different experiments through the same keyword and the same network. Anyway, the logic is very similar with the white-box, we need to change the structure by changing the code which means the code is transparent to testers. Testers need to test searching business by following different logic codes for different storage structure.

The purpose is to test the query performance and storage performance of the search engine system through different experimental control groups and verify a best storage solution.

## 6.3 Test Conditions

In the test, there are variables and constants in the same experimental group. We define the invariant conditions in the experimental environment.

### 6.3.1 Storage and Query Condition

In testing the storage and search process, we should ensure that some environments are fixed:

* Same local network
* A server with the same configuration
* A PC with the same configuration, the same browser

Test controls the network environment to prevent network problems from affecting the query speed. The server configuration is the same because different server configurations will course different test results. All test data must be based on the same set of server configuration, including hardware and software. The client used in the test must also be the same, less strict than the previous two items.

Local Network: a home router, the server and PC are connected to the same router through wired LAN port, 100M fiber.

Server Software

* Linux operating system Ubuntu distributed, version 20.04.
* NGINX Tengine 2.2.3 proxy HTTP server.
* Relational database MySQL5.7
* Non-relational database MongoDB
* Non-relational database in memory Redis
* Node.js latest version

Server Hardware

* 16G DDR4 PC Memory Cards
* Intel i7 7700
* NVME SSD 512GB Samsung pm961
* ASUS motherboard
* Intel 1000M network card

Client Side

* Chrome Browser
* JS script, automated testing page

## 6.3 Results and Analysis

# 7 COCLUSION

7.1 Summary

7.2 Contribution

1. Research search engine technologies used on learning resources.
2. Define the value of learning resources via quantitative methods.
3. Propose resources storage and rank in a search engine system.
4. Build a web application that supports searching of education and learning resources.
5. Design a verification method to locate high quality resources.
6. Apply user behavior and big data analysis to discover resource value.

7.2 Drawback and Future Work

# CITATIONS

1. Patil, Yugandhara; Patil, Sonal (2016). ["Review of Web Crawlers with Specification and Working"](http://www.ijarcce.com/upload/2016/january-16/IJARCCE%2052.pdf) (PDF). International Journal of Advanced Research in Computer and Communication Engineering. [↑](#endnote-ref-0)
2. ["Googlebot"](https://support.google.com/webmasters/answer/182072?hl=en). Google. 2019-03-11. Retrieved 2019-03-11. [↑](#endnote-ref-1)
3. ["BingBot Crawl Activity Surging?"](https://www.seroundtable.com/bingbot-crawling-much-16273.html). Retrieved 2016-07-16. [↑](#endnote-ref-2)
4. Dell Zhang \*, Yisheng Dong, An efficient algorithm to rank Web resources, 2000 Published by Elsevier Science B.V

    2000 Published by Elsevier Science B.V [↑](#endnote-ref-3)
5. C.M. Rowman, Scalable Internet resource discovery: research problems and approaches, Communications of the ACM 37 (8) (1994) 98–107. [↑](#endnote-ref-4)
6. Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google file system[C] Proc of SOSP 2003.New York:ACM,2003:29-43 [↑](#endnote-ref-5)
7. Dell Zhang \*, Yisheng Dong, An efficient algorithm to rank Web resources, 2000 Published by Elsevier Science B.V

    2000 Published by Elsevier Science B.V [↑](#endnote-ref-6)
8. Cliff Bowman, Veronique Ambrosini,Identifying Valuable Resources,European Management Journal,Volume 25, Issue 4,2007,Pages 320-329,ISSN 0263-2373 [↑](#endnote-ref-7)
9. J.B. Barney Journal of Management, 17 (1991), pp. 99-120 [↑](#endnote-ref-8)
10. Cliff Bowman, Veronique Ambrosini, Identifying Valuable Resources, European Management Journal, Volume 25, Issue 4, 2007, Pages 320-329, ISSN 0263-2373. [↑](#endnote-ref-9)
11. Reenskaug, Trygve. [THING-MODEL-VIEW-EDITOR: an Example from a planningsystem](http://heim.ifi.uio.no/~trygver/2007/MVC_Originals.pdf) [↑](#endnote-ref-10)
12. A modern front-end programming framework, which combines and compiles JS, CSS, and HTML to display the views on the browser. [↑](#footnote-ref-0)
13. Sullivan, Danny. ["Google: 100 Billion Searches Per Month, Search To Integrate Gmail, Launching Enhanced Search App For iOS."](http://searchengineland.com/google-search-press-129925) Search Engine Land. August 8, 2012 [↑](#endnote-ref-11)
14. Natasha Singer. “How Google Took Over the Classroom”. The New York Times, May 13, 2017 [↑](#endnote-ref-12)
15. Collier, Ken W. (2011). Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing. Pearson Education. pp. 121 ff. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)" \o "ISBN (identifier)) [9780321669544](https://en.wikipedia.org/wiki/Special:BookSources/9780321669544" \o "Special:BookSources/9780321669544). What is a self-organizing team? [↑](#endnote-ref-13)
16. ["What is Agile Software Development?"](http://www.agilealliance.org/the-alliance/what-is-agile/). Agile Alliance. 8 June 2013. Retrieved 4 April 2015. [↑](#endnote-ref-14)
17. ["Globally Distributed Content Delivery, by J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman and B. Weihl, IEEE Internet Computing, Volume 6, Issue 5, November 2002"](https://people.cs.umass.edu/~ramesh/Site/PUBLICATIONS_files/DMPPSW02.pdf)(PDF). [Archived](https://web.archive.org/web/20170809231307/http://people.cs.umass.edu/~ramesh/Site/PUBLICATIONS_files/DMPPSW02.pdf) (PDF) from the original on 2017-08-09. Retrieved 2019-10-25. [↑](#endnote-ref-15)
18. Cluster, or server cluster. It is a kind of deployment method that combines multiple servers into a server group to provide different services. The servers in the cluster can exchange data through API. [↑](#footnote-ref-1)
19. OMG™ Unified Modeling Language™ (OMG UML®) specifications, Kirill Fakhroutdinov 2007-2016 [↑](#endnote-ref-16)
20. Burbeck, Steve (1992) [Applications Programming in Smalltalk-80:How to use Model–View–Controller (MVC)](https://web.archive.org/web/20120729161926/http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html) [↑](#endnote-ref-17)
21. Loosely Coupled: The Missing Pieces of Web Services by [Doug Kaye](https://en.wikipedia.org/w/index.php?title=Doug_Kaye&action=edit&redlink=1" \o "Doug Kaye (page does not exist)) [↑](#endnote-ref-18)
22. Dell Zhang \*, Yisheng Dong, An efficient algorithm to rank Web resources, 2000 Published by Elsevier Science B.V

     2000 Published by Elsevier Science B.V [↑](#endnote-ref-19)
23. [Codd, E. F.](https://en.wikipedia.org/wiki/Edgar_F._Codd" \o "Edgar F. Codd) (1970). "A Relational Model of Data for Large Shared Data Banks". [Communications of the ACM](https://en.wikipedia.org/wiki/Communications_of_the_ACM" \o "Communications of the ACM). 13 (6): 377–387. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)" \o "Doi (identifier)):[10.1145/362384.362685](https://doi.org/10.1145/362384.362685). [↑](#endnote-ref-20)
24. The database performance table is from Roman Čerešňák, Michal Kvet, Comparison of query performance in relational a non-relation databases, Transportation Research Procedia, Volume 40, 2019, Pages 170-177, ISSN 2352-1465, [↑](#footnote-ref-2)
25. 1TB=1024GB=1024\*1024MB=1024\*1024\*1024KB [↑](#footnote-ref-3)
26. [RAM, virtual memory, pagefile, and memory management in Windows](http://support.microsoft.com/kb/2160852/en-us). Microsoft. [2012-11-26]. [↑](#endnote-ref-21)
27. Donghee Lee; Jongmoo Choi; Jong-Hun Kim; Noh, S.H.; Sang Lyul Min; Yookun Cho; Chong Sang Kim. [LRFU: a spectrum of policies that subsumes the least recently used and least frequently used policies](http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=970573&isnumber=20937). IEEE Transactions on Computers [↑](#endnote-ref-22)
28. Patton, Ron (2005). [Software Testing](https://archive.org/details/softwaretesting0000patt) (2nd ed.). Indianapolis: Sams Publishing. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)" \o "ISBN (identifier)) [978-0672327988](https://en.wikipedia.org/wiki/Special:BookSources/978-0672327988" \o "Special:BookSources/978-0672327988). [↑](#endnote-ref-23)
29. Limaye, M.G. (2009). [Software Testing](https://books.google.com/books?id=zUm8My7SiakC&pg=PA108). Tata McGraw-Hill Education. pp. 108–11. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)" \o "ISBN (identifier)) [9780070139909](https://en.wikipedia.org/wiki/Special:BookSources/9780070139909" \o "Special:BookSources/9780070139909). [↑](#endnote-ref-24)
30. Saleh, K.A. (2009). [Software Engineering](https://books.google.com/books?id=N69KPjBEWygC&pg=PA224). J. Ross Publishing. pp. 224–41. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)" \o "ISBN (identifier)) [9781932159943](https://en.wikipedia.org/wiki/Special:BookSources/9781932159943" \o "Special:BookSources/9781932159943). [↑](#endnote-ref-25)