A Search Engine Based on Valuable Resources and Self-Improvement Network Education Related

Youwei Huang

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

S1296394

6/1/2020

# **HIGHLIGHTS**

1. Search Engine Technology used in learning resources.
2. Definition of valuable learning resources, realized via computer technology.
3. Educational and learning resources collection.
4. Self-improvement network.
5. Verification method of high quality resources.
6. User behavior and big data analysis.

Keywords: Big data, search engine, self-improvement, valuable resources, education, learning resources

# **ABSTRACT**

Before reading this paper, think about a question, how to search for a piece of valuable and useful information that matches what you really want to learn from the internet?

You may say that we can use Google, Bing or Baidu to search the resources that we need by typing some keywords. These tools or websites that we call ‘search engine’, a search engine can use their web crawler technology to sniff the whole visible network and then provide users simple indexes and links as the results to help users find the source of useful resources. At present, the search engine on the internet has good performance and precise keyword search ability. Such a set of tools or websites have already covered most of our daily online searching life. So why does it need us to do the research on new search engines? These search engines have exposed more and more problems in some specific fields and groups. Let’s begin with these problems and then do the research on ‘search engines’ that developed in a special field, education and learning. This paper brings you a new set of search engine technologies and methods on searching valuable learning resources The research starts from the methods to define and normalize the quality and value of online learning resources through different key points. In order to prove such a theory, some experiments will be designed and done in a new resource search engine system.

# **GRAPHICAL ABSTRACT**

General framework of the whole research and the key technologies adopted by the search engine.



Figure 1

Note in figure 1:

1. Normally, search engines are based on Web2.0 technology.
2. DLRV (Degree of learning resource value) is a method to define and improve the definition of resources value that will be mainly explained and discussed in the following parts of the paper.
3. The whole system and research are divided into four parts, collecting data, resources recommendation, self-improvement and data analysis, verification of value data.

# **Background of Web Crawler**

It is a tool for the search engines and alternative information seekers to collect data for indexing and to enable them to keep their databases up to date.The result of crawling is a collection of websites at a central or distributed location.[[1]](#endnote-0)

Generally speaking, if we compare the whole Internet to a huge spider web, a web crawler is like a group of spiders. Each node of the spider web(Internet) contains a lot of information, which is generated in various servers around the world. These spiders’ work is to bring these information back to their home (a database).

# Resource Crawler

Different from the traditional search engine, the learning resource search engine does not use crawlers to obtain web pages, resources and other data. This search engine will be used for special and professional areas, the search engine only obtains and stores the information of learning related resources, and the system will not directly store resources, because it involves copyright issues and system storage performance problems. The so-called resource crawler is a more targeted resource collection container, which is also the central idea of this paper, to create a professional search engine to improve the search quality and user experience for learning resources.

# **Problems**

However, there are many problems in such a technology, which are caused by the quality and value of resources. These 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 can filter some bad information, of course, it can also recommend the content according to the user's interest, but it is still far from the definition of valuable resources.

In the next sections, I will show you how these problems will be magnified to different fields and user groups.

## Problems in specific areas

In the field of education, there are many problems exposed by traditional search engines. Problems of how much values the information contains, is this information relevant to users’ learning, is the information correct and new, how much content can be accepted by the learner. More details of problems to educational fields will be shown in the following list.

1. *Correct or incorrect information?*
2. *How relevant is the resource to the user's learning area or major?*
3. *Is the resource new or old, how is the updating rates?*
4. *Is information safe and healthy?*
5. *Is the information redundant?*
6. *Where are the resources from,how is the reliability and professionalism?*
7. *The cost of learning, how much time and money the users need to learn something?*
8. *How much effective learning percentage?*

## Problems to specific groups

In the field of education, groups of users can be divided by age, level of education or position, family background.

**According to age**, people of different ages have different learning abilities. For example, at present, the Internet learning population is mainly taken by people on average around 32 years old. So most resources will be searched by an adult who may have a family and work for a company. We need to think about whether these resources are practical to him or whether he is interested in them.

Problems list:

1. Whether the resources are healthy, suitable for minors, children?
2. Is it easy for children of different ages to understand?
3. Is it practical for adults?
4. Which age group is more attractive to which resources?

**According to the levels of education**, the algorithms need to take the difficulty level of the resources and users’ education background into consideration.

For the resources, resources need to be divided to different difficulty levels to fit the learners in different learning periods. The resources can be divided into, for example, entry level, junior level and senior level of difficulty.

For the users’ education background. People who receive online education can keep different degrees. The difficulty of the resources are going to fit the levels of the education background.

The difficulty level of these resources will have much impact on the recommendation algorithm of our search engine, because these search results should be close to the ability of different users.

**According to various occupations and their related skills**. The resource search engine requires the search results to fit the users from different posts. So the same keyword may produce different results, because each user’s field is different, that results in these different results.

**To the background of different families**. Internet Education hopes that resources are equal to everyone, so the value of resources should also be reflected in fairness. However, for resources with copyright, we should also follow the corresponding agreements and laws to protect intellectual property, because this is fair to creators. In all, equality will also become a value of the learning resources.

# **Solution**

To solve the above problems can be converted to solve the following two:

1. How to define a valuable learning resource?
2. implement point 1 to create a search engine specially made for searching learning resources.

## Value resource attributes

In particular it is difficult to identify resources within a firm if there is no agreed definition of what ‘valuable’ means.[[2]](#endnote-1)

Valuable resources can generate three types of competitive advantage: cost advantage, the ability to premium price, and volume-based advantage.[[3]](#endnote-2) The above three competitive advantages are used to demonstrate the valuable resources on business and management.

From the analysis of the above problems combined with the advantages that a valuable resource needs, a high-quality learning resource should keep the positive side of all the following characteristics.

1. Cost (both time and money).
2. Professionalism and reliability
3. Practicability
4. Suitability (dynamic, define when search)
5. Popularity
6. Feedback

The 6 characteristics can be used to define a valuable and high-quality learning resource.

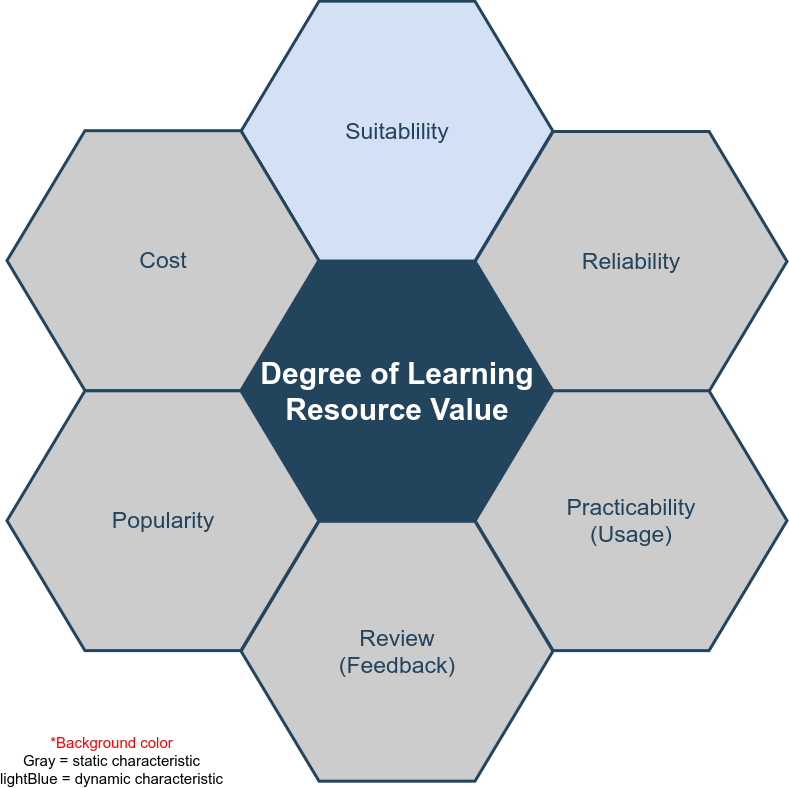


Figure 2

Notes: These six attributes are used to judge the value of resources, of which light blue is the variable attribute and gray attribute is fixed. Fixed attribute does not mean that the property value is constant for a resource.

# Definitions

The following is a description of the common nouns in the full text, for reference.

**DLRV**: Degree of Learning Resource Value. This system measures the value of resources, and value is an index for search engines to use for providing ranking and recommending resources.

**Suitability**: Here, it means resource title, content and tags match search words.

**Cost**: all the cost of finding and using a resource, both time and money.

**Popularity**: Search, click, read and comment quotas

**Review**: User feedback quotas

**Practicability(Usage)**: Amount of a resource usage

**Reliability**: Resource reliability refers to whether the source of resources is reliable and whether it has been certified or audited by authority, this quota is decided by the publisher, recommendation and organization/company.

# Requirements

## Hardware requirements

The whole search engine platform requires several high-performance servers which can undertake billions of requests from users. Users will search useful learning results listed in the browser. We call these web applications or B/S architecture. Generally speaking, the performance of the server and the configuration of hardware devices in the cluster are determined by the number of users and the number of visits. Therefore, in the early stage, there is no detailed requirement report for the whole set of server-side hardware, but we have given some basic requirements.

* 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

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

## 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.

### MVC Design pattern

MVC pattern is a very classic design pattern in software engineering. MVC pattern was first proposed by Trygve Reenskaug in 1978[[4]](#endnote-3), which is widely used in web development. The framework based on this design pattern can be called MVC framework. All the development and implementation described in this paper are based on this design pattern. This set of search engine system adopts MVC framework. MVC is a design pattern that separates model, view and controller. In actual development, model is data level, view is front-end, and controller is part of business logic. MVC design pattern can achieve high cohesion and low coupling, and separate data, view and business. 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.

### Related application software

* MVC framework based on Node and express
* The view layer is based on vue, 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 by the search engine system

## Non-functional requirements

### Performance

Google answers 100 billion searches per month[[5]](#endnote-4), that means the average of a day is at least 3 billion and this is the statistics in 2012. Our learning resource search engine doesn’t need such huge search performance because we are targeting at a special area instead of all the users on the internet. Around 2017, there are more than 30 million children use Google education apps[[6]](#endnote-5), and adults and college students are not included in 30 million, so our system needs at least double of this amount(children) for users’ requests so that it needs to accommodate 60 million users per day. In computer terms, it means DAU(Daily Active User) is at least 60 million. To ensure the smooth requests from the increasing DAU, we raise the performance bottleneck to 100 million DAU.

### Reliability

The operation of all the services are 24 hours, Users get the results within 1 second after starting the search that is to say, the response speed is less than 1 second for each research

### Security

**System layer security:**

1. Firewall between server nodes, access control on blacklist, white-list and iptables technologies.
2. Data backup to prevent the data loss disaster.
3. When main servers crash, use the reserve servers instead.
4. Quick recovery plan for crashed servers.

**Business layer security:**

1. User verification
2. API requests security
3. User behaviour logs
4. Cookie or cache security
5. User privacy
6. Encrypt and decrypt data

# Design

This part is the system design description, including engineering design, system architecture deployment design, unified modeling language.The final implementation, online system and theoretical verification of the system will follow all of the following design principle.

## Software Engineering Process

The design of software process follows the life cycle of software engineering and adopts agile model.



Figure 3

I divide the whole implementation process into 6 parts. The concept of the system is described in the background and problems above parts of the paper. Software requirements will not be described in the paper, there are other requirements documents. Framework design, UML will be described in the next parts of this chapter, development and implementation will not be described, only display through algorithms and methods research, for detail please refer to development documents and code. Testing and verification will be presented in the last part of the paper.

## System deployment structure design

The following is the simulation deployment picture of the search engine system, which follows the normal web deployment mode.



Figure4

Server deployment is distributed, running multiple servers in the same intranet or multiple networks, and 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, which 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 work. 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 in the middle to control access. Part of database server data is stored in high-speed non relational database, such as Redis or MongoDB, to deal with some hight-frequecny search engine requests.

The purpose of CDN is to speed up the existence of static files. The full name of CDN is: a content delivery network, or content distribution network[[7]](#endnote-6), static files can be distributed on multiple nodes of the Internet. When users access, they request the nearest fastest server, which improves the user experience.

The reverse proxy server distributes user requests to upstream servers, which can effectively reduce the possibility of congestion.

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.

There are many servers in the server cluster, most of them are controllers dealing with business logic, and 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 and provide the core business for users to search the accurate resources.

Non-relational database, such as Redis and MongoDB, these 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 is too complex. 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. This will be described and tested in the methodology of the second half of the paper.

## UML

### Database and E-R diagrams.

1. R diagram is a kind of diagrams to display the entity and relations between different data structures, in database, we call them tables. Attention: non-relational database is not included in the diagram.

Any modern web development is based on database driven, and the design of relational database follows E-R diagram. The following is the complete database design diagram of search engine system.



Figure 5

Note: each table(entity) has its own primary key which is named id. ‘is\_effect’ is used to mark a row of data as available or non-available status. It is not safe to completely delete information or to display uncontrollable data, by using ‘is\_effect’ attribute, the system can control the validity of data, hide or display data flexibly.

User table is used to store user’s information, resources table stores resources from the internet, users are the owners the resources, one user to many resources, users use resources to organize a course, one course to many resources.

‘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 evaluation in DLRV system.

### Use Cases



Figure 6

This is the use case diagram of the search engine system which covers 7 main use cases. These are just the most basic just needed functions, and the core use case is the search function. A large number of functional details are not covered here. Users are learners. They can search resources, upload resources manually, set up courses by adding resources as a set, and evaluate other courses or resources.

Note: these use cases: review resources, reference resources, review course and upload will affect the DLRV system define the value of resources. They are called as user behaviour or callback effects.

### Package Diagram & Class Diagram

Package is a [namespace](https://www.uml-diagrams.org/namespace.html) used to group together elements that are semantically related and might change together. It is a general purpose mechanism to organize elements into groups to provide better structure for system model.[[8]](#endnote-7) For the server side of the search engine system, the following figure shows the packages, the system is developed based on object-oriented language. Packages are mapped to different folders or collections. Different packages have dependencies on other packages.

1. Controller, this is a package includes all the controllers, we have talked that the system is designed on MVC mode. Controllers deal with all the business logic, take responsibilities for connecting data and views, accept users’ requests and responses. Controller depends on private libs and public modules.
2. Server, the entrance of the whole system, this package manage 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 too.
3. Config, a package manages the connection configuration of various servers, like mail server, database server, OSS server and all other servers we needed in this 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.
4. Private libs, the package contains all the private modules, plugins which used only in this system. It needs to depend on public modules and Model package.
5. Model, the central component of the pattern. It is the application's dynamic data structure, independent of the user interface.[[9]](#endnote-8) It can manage the data, logic and rules in the system.
6. Modules, public modules are used by the system



Figure 7

Note: in this diagram, Config and Modules two package have the most dependence by other resources, Config contains all the configuration of the system. ‘Modules’ is a class library referenced from public libs and used by the system in many places.

**Controller Package**



Figure 8

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, and the search class solves all the user's search requests. The Config class returns the configuration of the site to the front end. Resource will be responsible for data collection, resource classification, deletion and other functions. Through the Resource class, users can also edit courses and get course categories and lists.

**Private Package**



Figure 9

The core class of private package is the class Public, which means private public library. It is a public library specially developed for only this system, it includes functions such as get random numbers, format date, check user info and etc which are used with high frequency. 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[[10]](#endnote-9). Class DB provides a set of methods to operate database like inset, query and delete.

**Server Package**



Figure 10

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. The figure below explains what is the convention routing mode.



Figure 11

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, which is different from the controller, so it can isolate the access of code and media file, which is a guarantee for security.

### 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

The whole search process of the system is that 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 finally 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 request.

# Algorithms and Methods

This chapter describes the algorithm and some methods in learning resource search engine in order to let users get a good experience. ‘Good User Experience’ means 1. Fast access to search results and 2. Valuable learning resources, which is in line with the theme of the paper. We can simplify these two points to ‘search efficiency’ and ‘resource sorting and filtering’.

Let's first do some research on definitions and methods in traditional search engines(RD). Then it will explain the improved or innovative algorithms used in the learning resource search engine system(DLRV). The algorithm and methodology discussed are also based on the above two ways. The first section reviews the traditional search engine methods to speed up storage/query and rank valuable resources. The second section will put out the weakness of traditional general search engine. Finally, based on the traditional algorithms and methods, aiming at the field of Internet learning resources search engine, DLRV, we improve and enhance the ranking algorithm, storage-query algorithm for this search engine.

## Introduction of Traditional RD Way

RD(Resource Discovery), this is a process of searching valuable information on the Internet.[[11]](#endnote-10) The IETF-RD argues that resource discovery should provide the user consistent, organized view of information.[[12]](#endnote-11) RDS(Resource Discovery Server) can return a set of resources, these resources are links or indexes of web pages from the internet. It is like DLRV server but it provides traditional and general resource discovery service. Various of search engines support RD such as: Google, Bing, Baidu, etc. In particular one keyword or topic can typically contain from thousands to millions resources. When you type ‘English Learning’ keywords on Google, it returns 10,010,000,000 results within only 0.48 seconds. How does search engine do such a fast query? Users are only concern with a few of which are the useful results they need of all the resources results, how does search engine rank them and give a suitable list of valuable results?

### Storage and Query

Google uses its own file system named GFS(Google file system), GFS is a scalable and classical distributed file system for large distributed data-intensive applications[[13]](#endnote-12). GFS has been used in Google since 2003, and GFS is not open source. However, from Google's GFS related paper, we get some basic technical details, which are used in Google's search engine resources storage.

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



Figure 13

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.

### Rank of Results

A general RDS(Resource Discovery Server) can help user quickly rank useful resources through effective filtering and sorting algorithms. The 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.

The above two sections discuss about the traditional storage-query model and competitive searching sorting which are used by RDS. This is also the basic mode adopted by main current search engines.

## DLRV Way

DLRV(Degree of Learning Resource Value), discusses a method helps users to acquire valuable resources. The same with RD is that it definitely contains two parts, storage-query and results-rank. In the above paper, in traditional RD methods, we have discussed distributed file system to store and query the resources and Vector-Space model to compute and rank the value of results. But in the special search engine of learning resources, these methods have some disadvantages.

### Disadvantages of traditional RD way

Note: that the following disadvantages are only proposed in the case of a special learning resource search engine in this paper when comparing with common search engines based on RD.

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.[[14]](#footnote-0)

The biggest difference between learning resource search engine and traditional RDS is the second part, results ranking methods. Go back to chapter **Problems**, we have listed a number of search engine problems in this particular area, learning resources. If we use the basic ideas given in the last section: 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 chapter **Solutions**, 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 in RDS are used commonly.

Therefore, in the learning resource system, the remaining missing attributes become the defects of traditional search engines, and sorting the values of resources based on the above two points does not work well in learning resource search engine system.

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

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.

### Storage and Query

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.

### Relational Database

A relational database is a digital database based on the relational model of data.[[15]](#endnote-13) 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

### 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.[[16]](#footnote-1)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

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

### 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.[[17]](#footnote-2)

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).

### 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

### 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.[[18]](#endnote-14)

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. 23 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, a key value of 512MB can not store all the possible IDs. The value needs at most 512 units.

Absolutely, one hash map index only maps to one value, so we also need to calculate the search frequency of resources, and only store the high frequency resource ID into the value.

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 unit 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)[[19]](#endnote-15), 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

# **Citations**

1. Patil, Yugandhara; Patil, Sonal (2016). ["Review of Web Crawlers with Specification and Working"](http://www.ijarcce.com/upload/2016/january-16/IJARCCE 52.pdf) (PDF). International Journal of Advanced Research in Computer and Communication Engineering. [↑](#endnote-ref-0)
2. Cliff Bowman, Veronique Ambrosini,Identifying Valuable Resources,European Management Journal,Volume 25, Issue 4,2007,Pages 320-329,ISSN 0263-2373 [↑](#endnote-ref-1)
3. Jordan Friedman, U.S. News Data: The Average Online Bachelor's Student, April 4, 2017 [↑](#endnote-ref-2)
4. Reenskaug, Trygve. [THING-MODEL-VIEW-EDITOR: an Example from a planningsystem](http://heim.ifi.uio.no/~trygver/2007/MVC_Originals.pdf) [↑](#endnote-ref-3)
5. 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-4)
6. Natasha Singer. “How Google Took Over the Classroom”. The New York Times, May 13, 2017 [↑](#endnote-ref-5)
7. ["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-6)
8. OMG™ Unified Modeling Language™ (OMG UML®) specifications, Kirill Fakhroutdinov 2007-2016 [↑](#endnote-ref-7)
9. 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-8)
10. 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-9)
11. 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-10)
12. C.M. Rowman, Scalable Internet resource discovery: research problems and approaches, Communications of the ACM 37 (8) (1994) 98–107. [↑](#endnote-ref-11)
13. Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google file system[C] Proc of SOSP 2003.New York:ACM,2003:29-43 [↑](#endnote-ref-12)
14. In the aspect of HTTP, this system uses distributed structure, which divides users' HTTP requests to the upstream server, which is different from the distributed file system here [↑](#footnote-ref-0)
15. [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-13)
16. 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-1)
17. 1TB=1024GB=1024\*1024MB=1024\*1024\*1024KB [↑](#footnote-ref-2)
18. [RAM, virtual memory, pagefile, and memory management in Windows](http://support.microsoft.com/kb/2160852/en-us). Microsoft. [2012-11-26]. [↑](#endnote-ref-14)
19. 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-15)