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.



Note: 1. Normally, search engines are based on Web2.0 technology. 2. DLRV is a method to define and improve the definition of resources value that will be explained 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.[[1]](https://docs.google.com/document/d/1q7Ahy-41LIjzPQOm6ajPzNK1zNBpsJaUb_YRFQVc-zc/edit" \l "heading=h.q56unx7o5zgl) The result of crawling is a collection of websites at a central or distributed location. [[1]](https://docs.google.com/document/d/1q7Ahy-41LIjzPQOm6ajPzNK1zNBpsJaUb_YRFQVc-zc/edit" \l "heading=h.q56unx7o5zgl)

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).

# **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, 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. [[2]](https://docs.google.com/document/d/1q7Ahy-41LIjzPQOm6ajPzNK1zNBpsJaUb_YRFQVc-zc/edit" \l "heading=h.q56unx7o5zgl) 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

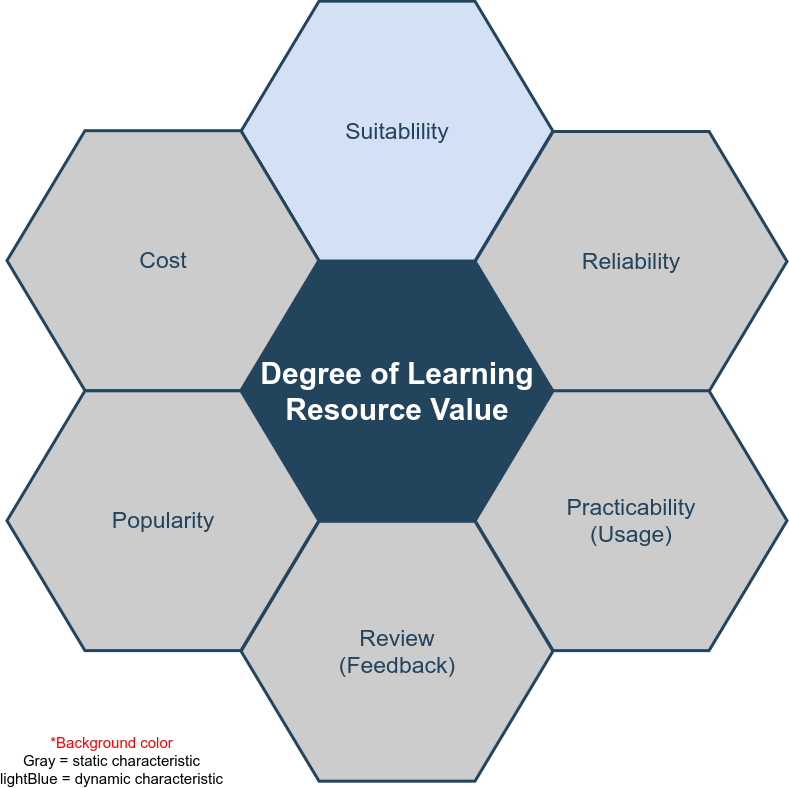
In particular it is difficult to identify resources within a firm if there is no agreed definition of what ‘valuable’ means.[[3]](https://docs.google.com/document/d/1q7Ahy-41LIjzPQOm6ajPzNK1zNBpsJaUb_YRFQVc-zc/edit" \l "heading=h.q56unx7o5zgl)

Valuable resources can generate three types of competitive advantage: cost advantage, the ability to premium price, and volume-based advantage.[[2]](https://docs.google.com/document/d/1q7Ahy-41LIjzPQOm6ajPzNK1zNBpsJaUb_YRFQVc-zc/edit" \l "heading=h.q56unx7o5zgl) 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.



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.

# **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 requirement report for the whole set of server-side hardware, but we have given some basic requirements.

* 基于X86体系的双核处理器E3
* 大于16GB的RAM
* 80GB的disk storage
* 独立的数据库服务器
* 基于Linux的分布式部署服务器，有容器与虚拟化技术的准备，但实验阶段不会使用
* 模块化的数据分析等微服务

User-end or test-end, we require PC and mobile devices to test all the webpages and functions in various browsers

## Software requirements

软件需求的复杂度远高于硬件方面，所有的算法，技术细节，功能需求都是通过软件编程得到实现和验证的。我们可以采用常规的WEB开发环境，工具，语言和相关的SDK。

### MVC 模式

MVC模式是软件工程中非常经典的，被广泛运用于web开发的一种设计模式，基于这种设计模式实现的框架可以称之为MVC框架，论文中所描述的所有的开发与实现均是基于这套设计模式的，这套搜索引擎系统采用了MVC框架。MVC是一种将Model，View，Controller分离的设计模式，在实际的开发中，Model是数据层面的，View是前端，Controller是处理业务逻辑的部分。MVC的设计模式可以做到高内聚，低耦合，把数据，视图，业务分离。MVC提高了开发效率，代码整洁度，具有更高的可扩展性。使用这种模式的目的就是可以让这套搜索引擎在试验中容易优化和扩展功能。

相关应用软件

* 基于Node.js和Express的MVC框架
* 基于Vue的视图层，也可以叫做前端，用户使用
* Nginx的反向代理服务器，HTTP服务器
* MySQL数据库，关系型数据库，存储大量搜索引擎将会产生的数据

## Non-functional requirements

### Performance

Google answers 100 billion searches per month[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 [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 1 second for each research

### Robustness

### Security

# **Citation**

*[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.*

*[2]. Jordan Friedman, U.S. News Data: The Average Online Bachelor's Student, April 4, 2017*

*[3]. Cliff Bowman, Veronique Ambrosini,Identifying Valuable Resources,European Management Journal,Volume 25, Issue 4,2007,Pages 320-329,ISSN 0263-2373*

*[4] 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.*

*[5] Natasha Singer. “How Google Took Over the Classroom”. The New York Times, May 13, 2017*