A Web crawler is an Internet bot which helps in Web indexing. They crawl one page at a time through a website until all pages have been indexed. Web crawlers help in collecting information about a website and the links related to them, and also help in validating the HTML code and hyperlinks.

Web crawling is a process that automatically obtains information from the Web pages through the link relationships between them and expands to the entire Web. This process is mainly done by the Web Crawler which usually consists of spider, controller and original page library, as shown in Figure 1. The spider crawls the pages from the Internet, extracts the URLs to URL database, and saves the pages to the original page library

The traditional relational databases store data in the form of a two-dimensional table with strictly row and column format, and emphasize the consistency and integrity of data. While this helps in maintaining relational information, it quickly becomes slow if the columns become too huge and is not an apt choice for applications which do not require large number of columns or relational information. Additionally, purely performance wise, NoSQL databases would perform better.

One such NoSQL DB that can be used is MongoDB due to features such as

(1). Data model convenient to design.

(2). High performance.

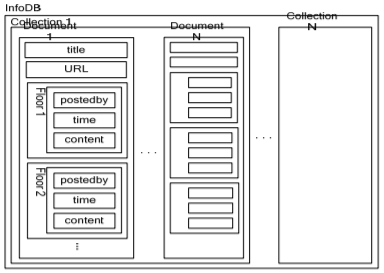
(3). High availability.

(4). Easy scalability.

(5). Rich query language.

According to the official documents, when the amount of data is more than 50GB, the access speed of MongoDB is 10 times faster than MySql.

For Designing our database we can use the following structure



Comparing of NoSQL(MongoDB) and MySQL :

In MongoDB, we put a post with all the floors in one document. Because MongoDB supports schema-free, we don’t have to design the structure beforehand, and it can be modified at run time. The fields in each document do not need to be same, which can be set depending on the actual situation when programming

MongoDB supports embedded document to implement nested, so we can store a post with all the floors in one document that we can efficiently get the whole post by query the id. Relational database stores the post and the floors in two tables, with the post id we query these two tables and get the post and all the floors

Auto-sharding of MongoDB can provide easy horizontal scalability with low cost of hardware, replica set of data replication can provide automatic failover, and order-preserving of sharding makes query fast. Because of strong consistency and integrity of data, the performance of relational database in the distributed data management is not good, resulting in poor horizontal scalability. To store large amount of data, the relational database has to buy advanced server with more storage space which is high-cost.