Руслан Мамедов

РАЗРАБОТКА ЭЛЕКТРОННОЙ БИБЛИОТЕКИ НАУЧНЫХ ТЕЗИСОВ

Abstract

The main problems of databases with big quantity of information are information management and convenience of finding the desired information. The system administrator should be able to easily add and manage materials. The client part of the site uses an advanced system of filters to search for materials on multiple criteria. To improve search capabilities and optimize the speed of search queries using a separate search engine - Sphinx, which allow searching for data based on morphology. There is ability to index PDF documents when you add them.

Keywords: HTML, PHP, SPHINX, MySQL.

В наше время существует множество готовых веб-порталов (хостингов), доя размещения на них своих материалов. Такие как depositfiles.com, failiem.lv, rapidshare.ru и множество других. Но такие веб-порталы не годятся нам для размещения на них наших материалов, так как у них нет толкового поисковика, поиск происходит лишь по названию файла.

Существуют также веб-порталы как atlants.lv, на которых можно размешать свои публикации. На данных типов веб-порталах существует хороший поиск по множеству критериев, таких как Автор, ключевые слова, содержание и другие. Но на данных типах веб-порталов существует другая проблема. За скачивание файла человек должен заплатить деньги, что в свою очередь нам не подходит, так как к базе с нашими файлами должны иметь доступ студенты в любое время и бесплатно.

В связи с этими критериями и уже имеющегося у нас в наличии веб сервера, мы будем разрабатывать свой веб-портал. На который сможем без проблем заливать нужные нам публикации и скачивать их в любое для нас время, а также организуем хороший поиск по множеству критериев.

Для написания данного веб-портала нам потребуется такие средства как HTML, PHP, SPHINX, MySQL, Adobe. Как и любой дизайн, наш веб-портал будет отвёрстан и отображен на HTML языке, для дальнейшего программирования.

Само ядро нашего веб-портала будет программироваться на языке PHP, так как это 1 из самых распространённых языков программирования, а также является одним из бесплатных языков.

Для хранения информации о публикациях, ссылки на публикации и различных других учебных материалов, а также информацию о студентах и преподавателей имеющих

те или иные права доступа к файлам, мы будем использовать бесплатную базу данных MySLQ. (www.mysql.com)

Все публикации и учебные файлы будут сохраняться в PDF формате, поэтому мы в дальнейшем углубимся в изучении Adobe разработки – PDF формата.

И для самого главного ингредиента нашего веб-портала – поисковика, мы будем использовать поисковой сервис Sphinx.

И так обсудим, почему же именно Sphinx поисковой сервис мы будем использовать, почему же нельзя воспользоваться простым, встроенным поиском MySQL?

Поисковой сервис Sphinx – один из самых крупных и быстро действующих поисковых сервисов на данный момент, в тоже время он является бесплатным, что позволяет нам его использовать для своего веб-портала. (www.sphinxsearch.com)

Сравнив скорости индексирования и поиска между встроенным функциями MySQL – FullText и Lite, и поисковой сервис Sphinx мы сразу определили, что он действительно лучше. Для испытания мы взяли 500мб документов, 3 миллиона записей в базе данных. Поиск происходил по любому ключевому слову, в результате мы получили по нашему запросу 134 записи. FullText с использованьем морфологии справился с заданием за 5 мин, в свою очередь Lite за 30 сек, но он не распознаёт морфологию, и поисковой сервис Sphinx показал нам удивительный результат, справившись с задание менее чем за секунду, используя морфологию.

Как и во всём вещь не может быть идеально, у любого предмета есть свою плюсы и минусы.

Основными плюсами поискового сервиса Sphinx является:

- Возможность активации морфологического поиска.
- Возможность использования расширенного синтаксиса запросов.
- Высокая скорость поиска.
- Минимальная нагрузка на сервер.
- Размер индекса примерно в два раза меньше, чем у MySQL FullText (со стандартными настройками).
- Простая поддержка произвольных разделителей в качестве границы слова.

Основными минусами поискового сервиса Sphinx является:

- Отложенная индексация. Содержимое только что добавленного материала становится доступным в поиске не сразу, а только после запуска дельта-индексирования. Обычно, его запускают по планировщику каждые 10 минут.
- Необходимость полной реиндексации. Примерно раз в сутки индекс нужно перестраивать полностью, поскольку дельта-индексы не обладают требуемой производительностью для постоянного использования.
- Операции индексации все равно немного нагружают MySQL для извлечения информации, по которой строится индекс.

И так обсудим, почему же мы выбрали именно PDF формат во всех наших публикациях, а не DOC или другие имеющие форматы.

Формат PDF - Portable Document Format это стандартный формат для электронных документов.

Чем же тогда так хорош PDF формат?

Главное преимущество PDF формата – это машинно и платформа-независимость. (www.wikipedia.org) То есть написанной статье в формате PDF всё равно, на какой оперативной системе открываться, будь это Windows, Linux, MacOS или другая оперативная система. Формату PDF также всё равно, какая системная плата у вас, он всё равно будет прочтён.

Вторым преимуществом формата PDF является бесплатность программ для просмотра файлов, сохраненных в нем, а также компактность и размеры PDF-файлов.

Также PDF – позволяет практически без ограничений сочетать в документе текст, векторные и растровые рисунки и даже различные интерактивные элементы.

Для автоматизации нашего веб-портала и улучшения качества поиска информации, при загрузке PDF документа система читает его, превращает в обычный текст (без картинок, оформления и т.п.) и записывает в базу данных, затем либо по планировщику, либо по событию (добавление pdf файла) запускается дельта-индексирование для обновления поискового индекса.

Углубившись в разбор PDF формата, и идентификацию текса мы столкнулись с проблемой, не все кодировщики одинаково кодируют текст. Некоторые примитивные кодировщики делают из текста картинку, и тогда текст не прочитывается как текст. А игнорируется как картинка. Тем самым мы выяснили, что надо все файлы переконвертировать все публикации одним конвектором и в дальнейшем им пользоваться.

Выводы:

1. PDF является стандартным формам для электронных документов.
2. В данный момент автоматизировать идентификацию текста не получится из-за различных кодировок PDF формата.
3. На данный момент наилучшим поисковым сервисом является Sphinx, который также может проводить идентификацию текста PDF формата.

Библиография

1. http://ru.wikipedia.org/wiki/Portable_Document_Format
2. http://www.mysql.com/why-mysql/
3. http://sphinxsearch.com/about/sphinx/