УДК 004.7

А. В. Смирнов

ОБ ОДНОМ ПРИМЕНЕНИИ СИСТЕМЫ ПОЛНОТЕКСТОВОГО ПОИСКА SPHINX

Статья посвящена применению готовых систем полнотекстового поиска на реальном проекте. Рассмотрена система полнотекстового поиска Sphinx, приведена общая схема работы приложения с интегрированной системой. В качестве демонстрации результатов произведено сравнение скорости работы системы с полнотекстовым поиском и с поиском, написанным на языке SQL.

В настоящее время технологии развиваются стремительно быстро, позволяя увеличивать скорость работы разрабатываемых сервисов. Вместе с тем в глобальной сети постоянно увеличивается количество данных и является важным скорость обработки этой информации. Самостоятельно разрабатываемые системы поиска становятся менее эффективными решениями в виду большого количества данных и роста запросов пользователей. Поэтому актуальным является применение уже готовых систем полнотекстового поиска.

В разрабатываемом проекте (веб-портале для взаимодействия клиентов с CRM-системой [1, с. 5]) для компании BAI, встал вопрос о выборе системы поиска. Компания занимается поставкой авиационных запчастей, деталей и услуг для авиалиний, авиакомпаний, изготовителям авиатехники и ремонтным организациям в различных регионах мира. Портал обеспечивает клиента информацией о большом количестве продаваемых деталей (около 180 тысяч), поэтому важным является высокая скорость предоставления информации. Для решения этого вопроса было решено использовать систему полнотекстового поиска Sphinx [2].

На портале применяется Sphinx версии 2.2.7 от 21 января 2015 года. На данный момент этот поисковик позволяет как индексировать страницы, так и пользоваться поисковым полем в режиме «Predictive Search Box» который позволит пользователям как предлагать варианты поиска, так и предлагать категории источников, в которых данные ключевые слова являются релевантными. Конфигурационный файл системы, настроен на работу под Windows и Linux. В нём указывается сервер и база данных, используемая для работы, описываются источники, по которым будут построены индексы. Здесь наблюдается масштабируемость системы: источники позволяют описывать различные типы баз данных и произвольное количество. Например, сайт сначала может работать на одной базе данных, но в дальнейшем он будет расширяться и потребуется введение дополнительных баз данных либо добавление других источников, поддерживаемого типа (БД, текстовые файлы, HTML файлы, электронная почта), для этого в конфигурационном файле остаётся лишь прописать новые источники для индексации. Общую схему работы веб-портала с интегрированной системой поиска можно увидеть на рисунке 1.

Файл конфигурации разделен на 3 индекса: pages – индексируется таблица pages_content, содержащая контент страниц; parts – индексируется таблица products, в которой хранятся номера и информация о предоставляемых продуктах; parts_delta – дельта-индекс, также, как и parts работает с таблицей products, но индексирует лишь те данные, которые были изменены с момента последней индексации, это сделано для снятия нагрузки с главного индекса и уменьшения ресурсоемкости.

На сервере веб-портала настраиваются дополнительные свойства поиска, такие как поиск с подключённой проверкой морфологии, сортировка по релевантности , ограничение выводов результата, ранжирование.

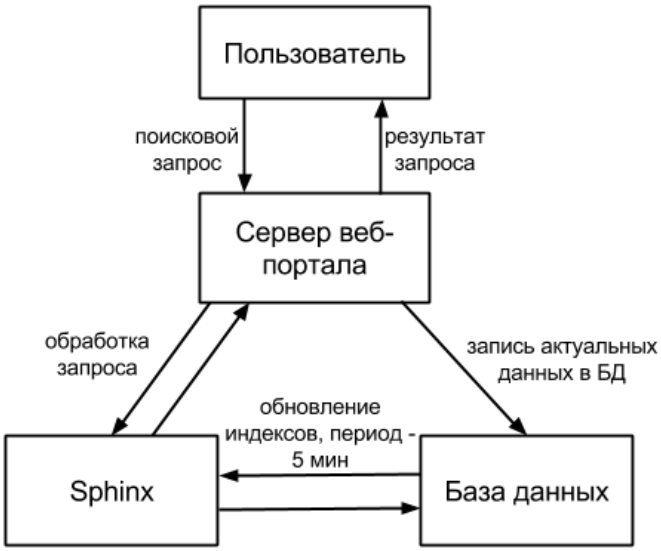

Рисунок 1 – Схема работы системы поиска на веб-портале

К примеру, для того что бы работала морфология, добавляется свойство SPH_MATCH_EXTENDED, при наборе поискового запроса «*23-12*» найдутся все возможные вхождения «23-12» в партийные номера деталей. При включенном свойстве SPH_SORT_RELEVANCE результат сортируется по особому правилу, наиболее релевантные записи отображаются выше. Сама релевантность работает по арифметическим правилам: перемножение и сложение веса индекса, веса поля, количества вхождения искомой строки запроса в документ и количество вхождений в другие документы. Вес полей для поиска можно регулировать, поставив вес поля partnumber выше, чем description. На портале результаты поиска возвращаются в выпадающее меню, поэтому количество отображаемых записей стоит ограничивать, делается это с помощью функции SetLimits, первым параметром указывается номер, с которого начать отображение результата, а вторым количество показываемых записей – SetLimits(0,20).

Таким образом, на веб-портале построена система поиска, полная индексация которой занимает 1.4 секунды, при выполнении поискового запроса – результат приходит менее чем через 0,5 секунды, что в три раза быстрее скорости работы на чистом SQL. Стоит учесть, что на данный момент на портале находится информация о 180 000 деталей, при этом размер индексов на жёстком диске занимает 28 мегабайт, что с текущими размерами жёстких дисков не является большим значением. Полученные результаты говорят о том, что Sphinx является быстрой системой полнотекстового поиска, с оптимальным использованием ресурсов компьютера, в связи с чем её можно применять на проектах с большой нагрузкой.

Литература

1. Гринберг, П. CRM со скоростью света [текст] / П. Гринберг – СПб.: Символ Плюс, 2007 – 528 с.

2. Документация по Sphinx Search [электронный ресурс] // URL: http://sphinxsearch.com/ (дата обращения: 28.04.2015).