Шалагинов А.В., Циос С.М., Кадин Е.П. *— рецензент Цурин О.Ф.*
УНК "ИПСА" НТУУ "КПИ", Киев, Украина

Полнотекстовый поиск в базах знаний информационных Веб-порталов

Информационный Веб-портал представляет собой комплексную интеллектуальную систему, которая предоставляет пользователям доступ к определённой области знаний. Многие пользователи не имеют достаточно опыта для быстрого поиска и обработки большого количества информации из интересующей предметной области, которая доступна на Веб-портале. Для упрощения навигации применяют различные классификаторы, как например, фасетный, иерархический или полнотекстовый поиск. Первые два метода классификации удобны при строгом разделении областей знаний, но совершенно не подходят в случае наличия разносторонней информации в хранилище данных. Самым простым и быстрым является полнотекстовый поиск. На ранних этапах развития полнотекстового поиска предполагался просмотр всего содержимого хранилища информации в поиске заданного слова или фразы. В таком виде поиск был практически не применим для Веба. Современные алгоритмы заранее формируют так называемый полнотекстовый индекс – словарь, в котором перечислены все слова и указано, в каких местах они встречаются. В таком случае достаточно задать в поиск нужные слова, и сразу же будет получен список документов, в которых они встречаются.

Основой для хранилища данных практически во всех Веб-порталах используется свободно распространяемая СУБД MySQL. Данная СУБД подходит для полнотекстового поиска в случае небольших объёмов данных и не сложных запросов. Для больших хранилищ данных поиск с помощью MySQL не применим из-за ряда недостатков: отсутствие сортировки, поддержка только VARCHAR и TEXT полей с индексами FULLTEXT, ресурсоемкий процесс, при ключе FULLTEXT добавление данных происходит дольше.

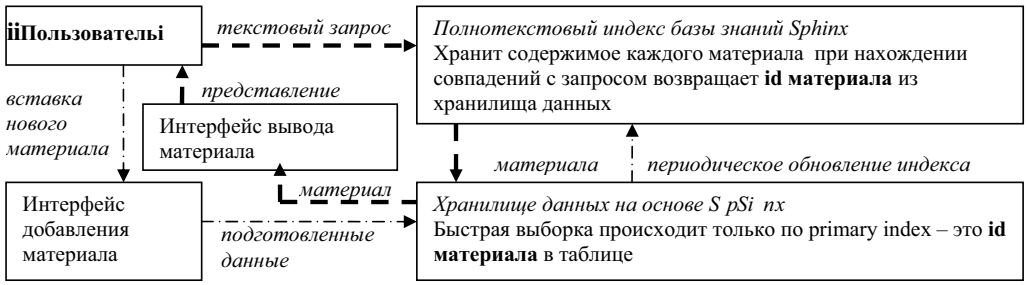

Рис. 1. Построение системы полнотекстового поиска на основе Sphinx

Для ускорения поиска предлагается использоваться систему Sphinx, которая, как видно в таблице 1, во всех тестах значительно обходит по производительности MySQL.

На рисунке 1 приведён вариант взаимодействия MySQL и Sphinx в базах знаний для осуществления быстрого полнотекстового поиска.

Быстрота достигается за счёт внедрения технологии Sphinx, в которой используется монолитный индекс, благодаря которому и достигается высокая скорость поиска (до 250 запросов в секунду на 1 млн. документов).

Таблица 1. Сравнение быстродействия Sphinx и MySQL на базе данных из 60 тыс. док. (100 МБ)

	MySQL(FULLTEXT)	Sphinx
Запрос не существующего в индексе слова	∼11 мс	<1 мс
Запрос 10 одновременно несуществующих слов	∼30 мс	<1 мс
Запрос 30 одновременно несуществующих слов	∼70 мс	<1 мс
Запрос одного слова входящего в индекс	∼22 мс	∼12 мс
Запрос 10 слов из индекса, совпадения найдены	∼50 мс	∼25 мс
Запрос 10 слов из индекса, совпадений не найдено	∼40 мс	∼12 мс

Литература

1. http://sphinxsearch.com/.
2. http://habrahabr.ru/tag/sphinx%20search/.
3. http://www.mysql.ru/docs/man/Fulltext_Search.html.