ВИДЫ ПОЛНОТЕКСТОВОГО ПОИСКА ДЛЯ ИНТЕРНЕТ-ПОРТАЛОВ ЗНАНИЙ

Бубнов Н.С.
Научный руководитель: д-р техн. наук, проф. Глоба Л.С.
Институт телекоммуникационных систем НТУУ «КПИ», Украина
E-mail: lgloba@its.kpi.ua

Аннотация — Рассмотрены три вида полнотекстового поиска, применимые для организации эффективного поиска на интернет-порталах знаний.

1. Введение

Поисковые системы используют в качестве основного тематический поиск, то есть выдают ссылки, основанные на входящих в запрос словах. Данный вариант, при всем его удобстве, не обладает «интеллектом», то есть поисковая система не понимает, что же конкретно ищет пользователь, поиск осуществляется механически по совпадению слов. Изменить ситуацию могут некоторые виды полнотекстового поиска, речь о которых пойдет ниже.

2. Основная часть

Полнотекстовый поиск в базах данных является одним из востребованных механизмов доступа к содержимому любой современной информационной системы, которые хранят метаинформацию, а зачастую, и сами документы, в базе данных. Современные веб-сайты, по сути, являются интерфейсом, способом организации доступа к базам данных. По мере накопления документов в системе неминуемо возникает проблема организации эффективной навигации по системе, чтобы посетитель сайта смог за минимальное количество кликов найти нужный документ. Помимо стандартной, зачастую ручной, навигации с использованием рубрикации (тематической, по типу материалов, категории пользователей и т.д.), полнотекстовый поиск является одним из самых эффективных методов навигации, особенно для новых пользователей, незнакомых с устройством сайта.

Теперь рассмотрим поисковые решения, на которые следует обратить внимание при выборе поисковой системы для интернет-портала знаний.

Sphinx search engine — один из самых мощных и быстрых из всех открытых рассматриваемых движков. Особенно удобен тем, что имеет прямую интеграцию с популярными базами данных и поддерживает развитые возможности поиска, включая ранжирование и стемминг для русского и английского языка.

Lucene.Net — это высокопроизводительная, масштабируемая библиотека для информационного поиска (ИП). ИП относится к процессу поиска документов, информации в документах или метаданных о документах. *Lucene* позволяет добавлять возможности поиска в различные приложения. Этот открытый проект реализован *Apache Software Foundation* и может быть использован на условиях лицензии *Apache Software*. Таким образом, *Lucene* в настоящее время и на протяжении уже нескольких лет является самой популярной свободной библиотекой для ИП [1].

Следующая диаграмма наглядно показывает процесс поиска в приложении на основе *Lucene.Net*.

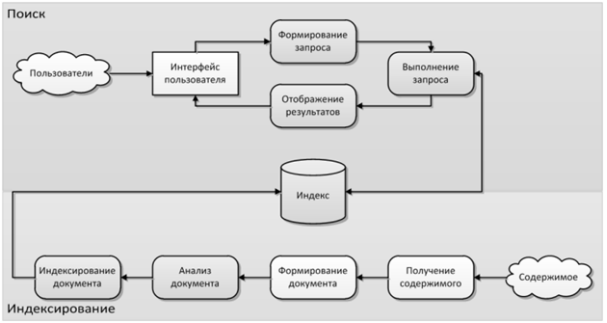

Рис. 2

Xapian — это пока единственный претендент на конкуренцию *Lucene* и *Sphinx*, выгодно отличающийся от них наличием «живого» индекса, не требующего перестройки при добавлении документов, очень мощным языком запросов, включая встроенный стемминг, проверку орфографии, и даже поддержку синонимов.

3. Заключение

Окончательно принять решение, применим ли конкретный поисковик в нашем проекте, можно будет лишь после детального исследования и тестов, однако некоторые выводы можно сделать уже сейчас.

Sphinx позволяет индексировать большие объёмы данных в базе *MySQL* с большой скоростью индексации и поиска, но требует выделения дополнительного сервера или даже кластера [2].

Если необходимо встроить поисковый модуль в приложение, то лучше всего подойдет *Lucene*. Однако необходимо учитывать достаточно медленную индексацию и необходимость частой оптимизации индекса (то есть, требовательность к CPU и скорости диска).

Xapian — достаточно хороший и качественный продукт, однако менее распространённый и гибкий, чем остальные. Требует ручной доводки и модификаций для встраивания в собственный код или использования как отдельного поискового сервера.

4. Список литературы

[1] McCandless M. Lucene in Action / M. McCandless, E. Hatcher, O. Gospodnetić. — Stamford: Mannig Publication Co., 2010. — 532 p.
[2] Aksyonoff A. Introduction to Search with Sphinx / A. Aksyonoff. — Beijing: O'Reilly Media, 2011. — 146 p.

TYPES OF A FULL-TEXT SEARCH FOR ONLINE KNOWLEDGE PORTALS

Bubnov N.S.
Scientific adviser: Globa L.S.
Institute of Telecommunication Systems
National Technical University of Ukraine "KPI", Ukraine

Abstract — Three types of a full-text search, that are applicable for the efficient search on the web portals of knowledge, are considered.