МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

# Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

**Писменото наследство и информационните технологии** [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.



ISBN 978–954–9787–25–2

# Повышение релевантности результатов поисковых запросов на основе кратких аннотаций научно-технических текстов

**С. В. Моченов, М. А. Шаронов, Е. С. Волкова**

*Аннотация, информационный поиск, релевантность*

# Relevancy Increase of Search Results on the Basis of Short Summaries of Scientific and Technical Texts

**Stanislav Mochenov, Mikhail Sharonov, Ekaterina Volkova**

This paper examines a relevancy increase of search results when searching specialized scientific and technical literature (articles, reports, author's abstracts, etc.). The relevancy increase is due to giving the user a short text summary generated automatically using algorithms for morphological, syntactic and semantic text analysis.

В настоящее время поиск специализированной научно-технической информации при помощи стандартных информационно-поисковых систем зачастую оказывается малоэффективным. Это выражается в том, что после ввода запроса пользователь получает выборку ссылок на источники, из которых требуется выбрать необходимые ему ресурсы. Для этого понадобится ознакомиться с источником, что требует немалых временных затрат пользователя. Поэтому особую актуальность приобретают исследования, направленные на повышение эффективности поиска научно-технической информации — статей, докладов, авторефератов и др.

Целью данной работы явилась реализация представления результатов поисковых запросов в виде кратких аннотаций текстов, синтезируемых автоматически из исходного текста с помощью описанной далее технологии, с целью сокращения времени на поиск требуемой информации за счет работы не с исходным текстом статьи или доклада, а с кратким вторичным текстом. Это значительно сократит расход времени на знакомство с источниками и увеличит скорость усвоения информации. В результате повысится и релевантность источника в отношении к поисковому запросу.

В данной работе исходными текстовыми данными, среди которых производится информационный поиск, является коллекция научно-технической русскоязычной литературы (статьи, доклады, авторефераты, монографии и др.), собранная на локальном компьютере.

Выделение ключевых слов и получение краткой аннотации базируется на информационной технологии анализа русскоязычного текста [Бледнов 2007: 61–72], протестированной экспериментально с помощью специально разработанного программного комплекса “Текстан”. Она включает в себя следующие этапы:

 

- определение статистических количественных характеристик текста [Бледнов 2007: 37–42];
- предварительный анализ, осуществляющий сокращение объема текстовой информации за счет направленной фильтрации с использованием статистических методов анализа [Бледнов, Моченов, Луговских 2006а: 128–129];
- морфологический анализ, при котором определяются морфологические характеристики каждого слова текста, а также различные интерпретации словоформ каждого слова [Бледнов 2007: 61–62];
- синтаксический анализ, осуществляющий выделение именованных групп — словосочетаний, которым соответствуют синтаксические отношения [Бледнов 2007: 62–64];
- нормализация предложений, где все предложения приводятся к типовой форме [Бледнов 2007: 64–66];
- семантический анализ текста, при котором осуществляется выделение векторов цели предложений и абзацев, установление связей между ними [Бледнов, Моченов, Луговских 2006б: 137–140];
- структуризация текста на основе выделенных векторов целей абзаца.

Для раскрытия важных понятий текста (главных слов из именованных групп) можно использовать цепочку векторов. Таким образом, с помощью механизма развертывания векторов главных слов возможно генерировать новые предложения, из которых впоследствии будет составлена аннотация к анализируемому тексту.

Программный комплекс “Текстан” решает только задачу анализа текста и структурирует его, но не предоставляет возможности поиска документов. Кроме того, он реализован на Delphi6, поэтому возникает необходимость использования более современной платформы. Для того чтобы реализовать поиск документов и параллельный анализ на основе созданной технологии с последующим составлением аннотации, необходимо решить следующие задачи:

- провести анализ имеющихся свободных серверных платформ;
- установить выбранный сервер и настроить его конфигурацию для получения требуемой в работе функциональности;
- адаптировать алгоритмы, используемые при реализации программного комплекса “Текстан” на языке Delphi, под язык PHP и возможности и ограничения выбранного сервера, реализовать алгоритмы на языке PHP;
- протестировать и провести экспериментальное исследование разработанной информационно-поисковой системы в качестве программного обеспечения информационной системы ИжГТУ при поиске информации по научной и учебной тематике.

Сегодня существует множество решений автоматизированного информационного поиска с открытым исходным кодом. Были рассмотрены следующие серверные платформы с открытым исходным кодом: *Lemur Toolkit & Indri Search*

*Engine* [The Lemur Project 2004], *mnoGoSearch* [Lavtech.Com Corp. 2000/2013], *SWISH-E* [Holon.net Web Team 2007], *Zettair* [RMIT University 2009], *Egothor* [Egothor 2013], *Sphinx* [Sphinx Technologies Inc. 2001/2014], *Terrier Search Engine* [University of Glasgow 2011], *Xapian* [Xapian project], *YaCy (P2P)* [YaCy], *Apache Solr* [The Apache Software Foundation 2011/2112], *HSearch* [Bizosys Technologies Pvt Ltd. 2010].

Для реализации веб-приложения взаимодействия с пользователем был выбран язык PHP в связи с его широким применением для разработки веб-приложений. Также важно наличие у PHP библиотек для реализации взаимодействия с поисковыми серверами.

В соответствии с целью работы к серверу предъявляются следующие требования:

- наличие компилятора PHP;
- обработка документов со сложным форматом (DOC, DOCX, PDF);
- токенизация, распознавание языка, настраиваемый анализ текста.

При проведении анализа возможностей и требуемой функциональности серверных платформ был выбран сервер Apache Solr, удовлетворяющий всем перечисленным требованиям и широко применяемый при разработке информационно-поисковых систем. Для выполнения перечисленных выше задач осуществлены следующие работы:

- установлен и настроен сервер Apache Solr 1.2.0 в сервлет-контейнере Apache Tomcat V5.5, сервер Web-приложений Apache V2, а также PHP V5 для разработки Web-приложения;
- для того чтобы посредством системы можно было делать простейшие поисковые запросы, на языке PHP реализована тестовая форма.

Поиск осуществляется среди тестовых данных. Продолжается работа по настройке Apache Tika для импортирования информации из текстовых файлов форматов doc, pdf, txt в базу данных Solr.

Результатом работы станет веб-приложение, позволяющее осуществлять поисковый запрос в стандартной форме. Результаты запроса отображаются в виде списка найденных сервером файлов текстовых форматов doc, pdf и др., содержащихся в коллекции научно-технической литературы на локальном компьютере. Рядом с именем файла отображаются ключевые слова, выделенные из текста данного файла, и составленная краткая аннотация этого текста.

На рисунке представлен интерфейс веб-приложения для работы с тестовыми данными. Уровень детализации аннотации зависит от объема текста во входном файле.

Реализация проекта позволит получать краткие сведения о найденном источнике, а при необходимости обратиться к его полному тексту, нажав соответствующую ссылку. Это должно сократить временные затраты на поиск научно-

технических документов и обращаться только к необходимым пользователю статьям, монографиям, авторефератам и другим текстам.

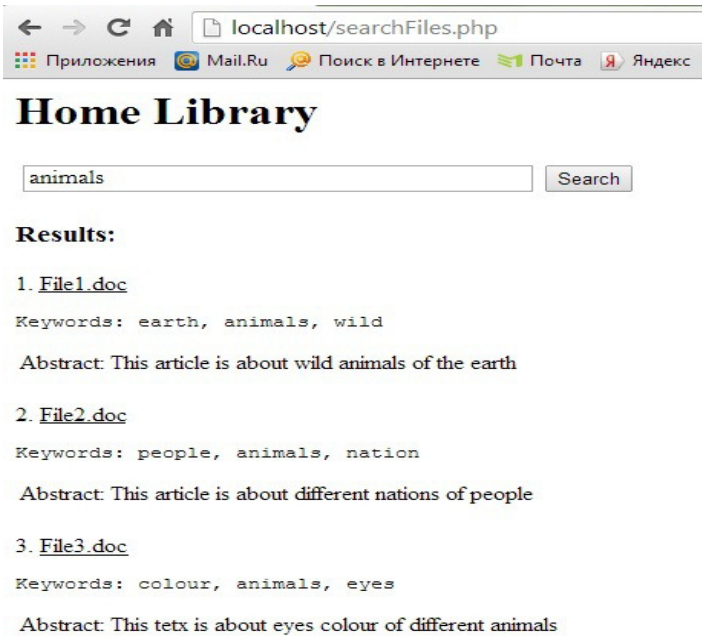


Рис. Интерфейс веб-приложения

Разрабатываемая информационно-поисковая система может быть полезна при поиске файлов наиболее распространенных текстовых форматов в ОС Windows с выводом дополнительной краткой информации об их содержимом. В дальнейшем она может быть модернизирована для эффективного поиска различной информации в сети Интернет (HTML-страниц, текстовых файлов и т. п.).

**Литература**

Бледнов 2007 — *Бледнов А. М.* Разработка и исследование моделей и информационной технологии семантико-синтаксического анализа русскоязычного текста: дис. … канд. техн. наук: 05.13.18, 05.13.01: защищена 30.05.07: утв. 22.07.07. Ижевск, 2007. 120 с.

Бледнов, Моченов, Луговских 2006а — *Бледнов А. М., Моченов С. В., Луговских Ю. А.* Об одном методе статистической фильтрации текстовой информации // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам: материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.). Ижевск: Изд-во ИжГТУ, 2006. С. 126–130.

Бледнов, Моченов, Луговских 2006б — *Бледнов А. М., Моченов С. В., Луговских Ю. А.* Векторная модель представления текстовой информации // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам [Текст]: материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.) / отв. ред. В.А.Баранов. Ижевск: Изд-во ИжГТУ, 2006. С.136–145.

Bizosys Technologies Pvt Ltd. 2010 — *Welcome to HSearch!*. HSearch—The NoSQL Search. October 10. <http://bizosyshsearch.sourceforge.net/index.html>.

Egothor 2013 — *Egothor Search Engine.* Home | www.egothor.org. October, 10. <http://www.egothor.org/cms/>.

Holon.net Web Team 2007 — *About Swish-e. Swish-e*: Home Page. October, 10. <http://swish-e.org/>.

Lavtech.Com Corp. 2000/2013 — *MnoGoSearch for Windows*. Windows search engine software—free trial download. October 10. <http://www.mnogosearch.org/win.html>.

RMIT University 2009 — *Zettair*. Zettair Homepage. October 10. <http://www.seg.rmit.edu.au/zettair/index.html>.

Sphinx Technologies Inc. 2001/2014 — *Open Source Search Server Sphinx*. Sphinx | Open Source Search Server. October 10. <http://sphinxsearch.com/>.

The Apache Software Foundation 2011/2112 — *Apache Solr*. Apache Lucene—Apache Solr. September 16. <http://lucene.apache.org/solr/>.

The Lemur Project 2004 — *Overview of the Lemur Toolkit*. Overview of the Lemur Toolkit. October 10. <http://www.cs.cmu.edu/~lemur/3.1/overview.html>.

University of Glasgow 2011 — *Welcome to the Terrier IR Platform*. University of Glasgow :: School of Computing Science. October 10. <http://www.terrier.org/>.

Xapian project — *The Xapian project*. The Xapian project. October, 10. <http://xapian.org/>.

YaCy — *Web Search by the people, for the people*. YaCy—The Peer to Peer Search Engine: Home. October 10. <http://yacy.net/en/index.html>.