УДК 004.42 DOI: 10.17587/it.24.414-420

С. Л. Макаров, канд. техн. наук, доц., доц. Департамента программной инженерии Факультета компьютерных наук, e-mail: smakarov@hse.ru,
Национальный исследовательский университет "Высшая школа экономики"

Информационные технологии поиска аналогов исследования по каталогам диссертаций

Рассматриваются некоторые информационные технологии поиска аналогов исследования, проекта или идеи по каталогам диссертаций и предлагается система автоматического поиска аналогов. Обоснована актуальность разработки системы, описаны методики и методы, лежащие в основе работы системы, и приведены результаты вспомогательных исследований структуры каталогов диссертаций и результаты работы системы. Приводится архитектура системы, структура базы данных, с которой работает система, описание алгоритма работы системы. Рассматривается пользовательский интерфейс и результаты работы системы с точки зрения пользователя и администратора системы. Делаются выводы о преимуществах и недостатках системы и особенностях ее работы и работы с ней, также описывается практическая значимость системы.

Ключевые слова: *информационные технологии, автоматический поиск, автоматизация, программная инженерия, система поиска, анализ проектов, аннотация исследования, каталог диссертаций, пользовательский интерфейс*

Введение

Для поиска аналогов исследования, проекта или идеи в настоящее время используются поисковые системы различных интернет-ресурсов. При использовании обычных информационно-поисковых систем (Яндекс, Google, Mail, Bing и т. д.) существует проблема поискового шума: вероятность найти реальное научное исследование невелика; скорее всего, будут найдены сайты магазинов, форумы, ссылки на бесчисленные блоги и социальные сети и прочие подобные ресурсы. Логично использовать для рассматриваемого типа поиска другие информационные технологии — специализированные ресурсы, исключающие или сводящие к минимуму эффект поискового шума. Такими ресурсами являются каталоги библиотек, в частности, каталоги диссертаций, каталоги научных статей, базы данных, индексирующие научные труды (Scopus, Web of Science, eLibrary и прочие). Однако значимость и вес научного исследования в диссертациях гораздо выше, чем в статьях или трудах конференции; кроме того, многие индексирующие базы данных закрыты извне — воспользоваться ими можно, лишь находясь в институте или университете. Поэтому в данной статье рассматривается поиск именно по каталогам диссертаций.

Система СТАРТLite (Старт Лайт) представляет собой веб-приложение и предоставляет пользователю возможность поиска существующих аналогов предлагаемого проекта или исследования, краткое описание (аннотацию) которого пользователь вводит в систему, по каталогам диссертаций. Среди аналогов могут быть найдены исследования, степень совпадения которых с описанием проекта пользователя является существенной. Таким образом, пользователь может оценить оригинальность своего исследования или проекта, получить первое представление о существующих конкурентах и сделать предварительный вывод о степени новизны своего проекта или идеи для исследования.

Здесь может возникнуть вопрос: почему не воспользоваться существующими каталогами диссертаций напрямую, например — поиском на сайте Российской государственной библиотеки? Во-первых, с помощью подобных каталогов нельзя ввести в поиск текст, можно лишь небольшую фразу или словосочетание, содержащее ключевые термины, интересующие пользователя. Во-вторых, поиск в подобных каталогах является достаточно сложной задачей — нужно разбираться, что именно искать, в каком именно каталоге, по какому полю — по теме, или по заглавию исследова-

ния, или по содержанию, если оно доступно; с учетом или без учета морфологии; изучать информационно-поисковый язык расширенных запросов и т. д. В-третьих, не всегда доступен полный текст или хотя бы оглавление закрытого ресурса, что довольно странно, учитывая, например, постоянную доступность подобной ознакомительной информации, например, для книг на сайте amazon.com. И, наконец, последнее (по порядку, но не по важности): для того чтобы найти что-то хотя бы по двум терминам, например "искусственный интеллект", нужно просмотреть десятки тысяч документов, даже если ограничивать поиск соответствующими фильтрами по техническим наукам, по языку документа и пр. Человеку такая задача не под силу. Предлагаемая система может осуществлять такой поиск автоматически, значительно быстрее человека. Безусловно, у аналогов системы есть и достоинства: возможность искать документы по различным полям: автору, ISBN, ISSN, регистрационному номеру, шифру специальности, системному номеру. Однако и здесь есть проблема — никто заранее не знает значений большинства этих полей.

Методика и методы, реализованные в системе

Для поиска аналогов проекта пользователя система использует набор из методики и методов автоматической обработки текстовых документов, таких как методика автоматического разбора текста с учетом морфологии [1, 2], метод автоматического построения векторной модели документа (простой вектор [3]), метод автоматического построения поисковых запросов с учетом архитектуры каталогов диссертаций, метод вычисления процента совпадения найденных документов с входными данными (косинусная мера [4]). Для поиска аналогов используются два каталога диссертаций: каталог [5] и ресурс Российской государственной библиотеки [6]. Оба ресурса являются открытыми для любого пользователя. Чтобы исключить поисковый шум, поиск ведется только по базе данных диссертаций.

Для морфологического разбора текстовых документов на русском языке в системе используется свободно распространяемый бинарный словарь и модуль на языке php, управляющий этим словарем. Словарь адаптирован к использованию в онлайн-режиме группой разработчиков открытого проекта phpMorphy [7], существующего уже достаточно давно. Именно этот модуль реализует методику автоматического разбора текста с учетом морфологии. С его помощью из текста выбираются наиболее значимые и характерные термины — имена существительные и прилагательные. Исследование разработанной системы показало, что результаты поиска с учетом только имен существительных значительно проигрывают учету как существительных, так и характеризующих их прилагательных. Одной из проблем рассматриваемого модуля является неверное срабатывание на определенные слова — неправильное определение частей речи.

Метод автоматического построения векторной модели документа состоит в учете значимых терминов документа по отношению ко всему документу и базируется на частотном анализе этих терминов. При этом существительные и прилагательные смешиваются в один вектор документа, упорядоченный по убыванию частоты терминов: $V = \{t_1, t_2, ..., t_n\}$, где t_1 — наиболее часто встречаемый термин документа. Проблемой при таком подходе является тот факт, что в заголовках исследований и в их тексте нередко употребляются термины общего назначения, очень мало говорящие о сути исследования, например: система, подход, вид, тип, аннотация и тому подобные. Поэтому необходимо использовать список стоп-слов, куда нужно добавлять подобные малоинформативные термины. На данный момент система использует список стоп-слов, содержащий 82 элемента.

Метод автоматического построения поисковых запросов с учетом архитектуры каталогов диссертаций основывается на исследовании, проведенном для того, чтобы понять, как обращаться с ресурсами [5, 6]. Метод заключается в последовательном формировании поисковых запросов по каждому термину из вектора входных данных в виде адресной строки, которая посылается в систему ресурса [6], обрабатывается, и результат выполнения запроса затем разбирается на очередные 20 источников — названий и ссылок на диссертации, из которых берутся только названия. По этим названиям автоматически формируются ссылки к ресурсу [5] с помощью замены пробелов между словами на знак "+". Адресная строка ресурса [6] содержит очередной термин для поиска и вспомогательные параметры, например: http://sigla.rsl.ru/results.jsp?f=1016&t=3&v0=интеллект&f=1003&t=1&v1=&f=4&t=2&v2=&x=42&y=11&f=21&t=3&v3=&f=1016&t=3&v4=&f=1016&t=3&v5=&bf=4&b=&d=0&ys=&ye=&lng=&ft=&mt=&doi=&dt=&vol=&pt=&iss=&ps=&pe=&tr=Cyr-Common&tro=&cc=a1&i=1&v=tagged&s=2&ss=1003&st=0&i18n=ru&psz=20&bs=20&ce=0w*M&debug=false&c=1&c=c3&c=b3&c=b2&c=b4&c=c4&c=b5&c=b6. Как видно, ресурс [6] исполь-

зует GET-запросы, поэтому в адресной строке браузера располагаются все необходимые параметры, которые используются системой при поиске. В частности, значение параметра v0 — это термин или термины, разделенные знаком +, которые пользователь ввел в строку поиска, а параметр i равен не номеру страницы, которую пользователь в данный момент просматривает на сайте в результатах поиска, а номеру первого источника, который располагается на странице. Если учесть, что ресурс выводит 20 источников на странице результатов поиска по умолчанию, то i будет равен 1 для первой страницы, 21 — для второй и так далее. Благодаря такой структуре ресурса разрабатываемая система может листать страницы поиска и менять поисковые термины, что и используется в рассматриваемом методе. Остальные параметры адресной строки являются служебными и не представляют интереса. Ресурс [5] служит для того, чтобы, скомпоновав название очередной диссертации, полученное с помощью работы с ресурсом [6], автоматически сформировать запрос и ссылку, которая предоставляется пользователю рядом с названием документа и перенаправляет пользователя на краткое содержание документа с оглавлением и возможностью прочитать автореферат, а в некоторых случаях и весь документ.

Метод вычисления процента совпадения найденных документов с входными данными основан на сравнении векторов очередного документа (d_2) и входной информации, переданной пользователем системе (d_1), с помощью определения расстояния между ними с использованием известной формулы косинусов или определения меры схожести двух документов [4]:

$$\mathrm{sim}(d_1, d_2) = \frac{\sum_{j=1}^{N_t} d_{1j} d_{2j}}{\sqrt{\sum_{j=1}^{N_t} d_{1j}^2 \sum_{j=1}^{N_t} d_{2j}^2}},$$

где $d_1 = (d_{11}, d_{12}, \ldots, d_{1N_t})$ и $d_2 = (d_{21}, d_{22}, \ldots, d_{2N_t})$ — векторы документов d_{1j} и d_{2j}, $j = \overline{1, N_t}$ — значения j-го термина в документах d_1 и d_2; N_t — общее число разных терминов в обоих документах. Значение, полученное в результате такого вычисления, умножается на 100, так как выражено в процентах.

Архитектура системы

Система состоит из нескольких модулей (рис. 1, см. третью сторону обложки): модуля для отправки входных данных (index.php), модуля поиска, запускаемого через планировщик и осуществляющего обработку очереди заявок, поиск аналогов, запись результатов поиска и сравнение аналогов с введенными данными в базу данных (novizna.php), модулей отображения результата для пользователя (result.php) и администратора системы (resultadmin.php), модуля подробного отображения результатов для пользователя и администратора (resultkarta.php), модуля интерфейса изменения настроек (settings.php), модуля загрузки настроек и загрузки списка стоп-слов (setsloaded.php) и модулей, связанных со словарем phpMorphy, располагающихся в папке tm2 (бинарный морфологический словарь русского языка и система управления этим словарем). В качестве вспомогательных файлов системой используются: директория js_nova (файлы шаблона, необязательная директория), директория Temp (папка с файлом docounter.txt отображения хода процесса и параметрами перезапуска системы по расписанию), текстовый файл stoplist.txt, содержащий список стоп-слов, и текстовый файл settings.txt (файл настроек системы).

Кроме файловой части система работает с базой данных, состоящей из трех таблиц (рис. 2, см. третью сторону обложки): morph_links (id "заказов", названия, ссылки, процент совпадения входных данных и документа по ссылке и термины документа по ссылке с их частотами), morph_orders (id и входные данные "заказов" на поиск от пользователя системы, статус выполнения "заказов"), morph_results (id "заказов", названия, ссылки, термины аналога с частотами и статус выполнения "заказов"). Термины, хранящиеся в таблицах, доступны для просмотра только администратором системы, причем совпадающие термины входных данных и аналога в интерфейсе выделяются жирным шрифтом. Таблицы базы данных не связаны друг с другом для простоты, однако понятно, что первичный ключ у таблиц для определенного заказа один и тот же, и все операции по выборке, обновлению или добавлению записей в таблицы осуществляются именно по нему.

Для обеспечения работы системы необходимо установить задание для планировщика операционной системы по запуску файла novizna.php через определенный промежуток времени. Рекомендуемый промежуток — 15...20 мин. Для linux-подобных систем это выглядит как создание нового задания для cron.

Интерфейс и алгоритм работы системы

Алгоритм работы системы следующий (см. также [8]). Пользователь заходит на сайт и вво-

дит данные с помощью модуля для отправки входных данных. В качестве входных данных пользователь указывает свой адрес электронной почты для оповещения о готовности результатов, название своего проекта, ключевые слова проекта и аннотацию проекта, которая представляет собой краткое описание предлагаемого решения. Для запуска поиска необходимо заполнить все поля без исключения и затем нажать кнопку "Начать поиск" (рис. 3).

Если все поля заполнены правильно (есть проверка на корректность входных данных), пользователь видит сообщение, показанное на рис. 4, а система принимает введенные данные для обработки, записывает их в базу данных, пополняя таблицу заказов очередной записью, и начинает работу. При этом введенные данные сохраняются под определенным номером, который указывается в этом сообщении, пользователю предоставляется ссылка на результат работы системы, а его данные становятся в очередь. Под работой системы имеется в виду запуск модуля поиска системы с помощью планировщика, который находит необработанный заказ со входными данными в базе данных, строит для него векторную модель с помощью морфологического модуля, выделяет все термины и ищет каждый термин через каталог диссертаций, при этом для каждого найденного документа строит векторную модель, сравнивает ее с моделью входных данных и выдает процент совпадения, который записывает в базу данных. При превышении времени работы системы (600 с) она записывает все параметры ("то место, на котором остановилась") в файл счетчика и "засыпает", пока ее не "разбудит" планировщик. При "пробуждении" модуль поиска считывает все параметры и продолжает работу до тех пор, пока все термины в векторе входных данных не будут перебраны. По завершении работы система высылает письмо со ссылкой, по которой можно посмотреть результат поиска, на адрес, оставленный пользователем. При этом адрес электронной почты, введенный пользователем, не виден ни ему, ни администратору системы и известен только самой системе и ее разработчику, который может зайти в базу данных напрямую через панель управления хостинга.

Форма выходных данных или результатов работы системы, на которую пользователю была дана ссылка (рис. 4), содержит информацию обо всех заказах, содержащихся в системе (рис. 5).

Номер заявки соответствует номеру, который был присвоен введенным данным и сообщен пользователю (см. рис. 4), название проекта — введенному названию, и последнее поле — статус заявки — информирует пользователя о том, была ли его информация обработана (done) или обработка еще не закончена (undone). Пользователю предоставляется возможность нажать на название своего проекта и увидеть предварительные или окончательные результаты работы системы, которые отображаются в карте проекта (рис. 6).

Карта проекта отображает исходные данные: номер заявки, название проекта, ключевые слова проекта, аннотацию проекта и статус заявки, а также упорядоченные по убыванию степени совпадения аналоги проекта, найденные в каталоге диссертаций, и их число (рис. 6). Для каждого аналога выводится его название, автоматически сформированная ссылка на аналог на ресурсе [5], содержание аналога (в данной версии системы оно совпадает с названием, вообще же в идеале имеется в виду полный текст документа) и степень совпадения аналога с исходными данными, выраженная в процентах. В соответствии с настройками системы, которые может редактировать администратор системы, регулируется число выводимых пользователю результатов (в данном случае равное 333) из общего числа найденных аналогов на момент просмотра карты (в данном случае равное 920).

Для администратора системы форма результатов работы системы содержит дополнительную ссылку "Удалить" (рис. 7), при нажатии на которую из всех таблиц базы данных удаляется информация о соответствующем результате работы системы, и ссылку "Настройки системы", с помощью которой можно поменять параметры

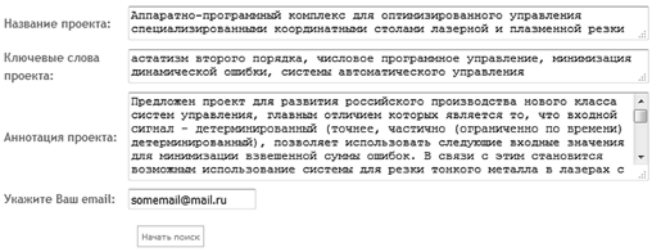

Рис. 3. Интерфейс заполнения входных данных

Спасибо за использование системы! Вашей заявке присвоен номер 1. Через некоторое время можно будет посмотреть результаты анализа по этому адресу, а также Вам придёт письмо на почту с этим адресом по завершении работы системы.

Рис. 4. Сообщение об успешном принятии введенных данных

Список проектов: Назад

Номер заявки	Название проекта	Статус заявки
1	Аппаратно-программный комплекс для оптимизированного управления специализированными координатными столами лазерной и плазменной резки	undone
30	первый проект	undone

Рис. 5. Форма результатов работы системы для пользователя

Данные проекта: Назад

Номер заявки	Название проекта	Ключевые слова проекта	Аннотация проекта	Статус заявки
30	первый проект	методы интеллектуальной обработки информации	автоматизация анализа проектов с применением методов интеллектуальной обработки информации	undone

Список аналогов (всего найдено 920, из них отображено 333):

Название	Ссылка	Содержание	Степень совпадения
"Психическая болезнь" как интеллектуальный проект	ссылка	"Психическая болезнь" как интеллектуальный проект	3.59%
Разработка и исследование системы обработки технологической информации гидрогенерирующих предприятий	ссылка	Разработка и исследование системы обработки технологической информации гидрогенерирующих предприятий	3.59%
Автоматизация лингвистической обработки словарей научно-технической информации	ссылка	Автоматизация лингвистической обработки словарей научно-технической информации	2.94%
Формирование учетно-контрольной информации для природоохранных проектов	ссылка	Формирование учетно-контрольной информации для природоохранных проектов	2.94%
Статистическая обработка данных с использованием априорной информации	ссылка	Статистическая обработка данных с использованием априорной информации	2.94%
Гибкий интеллектуальный интерфейс для систем передачи сложноорганизованной информации	ссылка	Гибкий интеллектуальный интерфейс для систем передачи сложноорганизованной информации	2.55%

Рис. 6. Карта проекта для пользователя

Список проектов: Настройки системы Назад

Номер заявки	Название проекта	Статус заявки	
1	Аппаратно-программный комплекс для оптимизированного управления специализированными координатными столами лазерной и плазменной резки	undone	Удалить
30	первый проект	undone	Удалить

Рис. 7. Форма результатов работы системы для администратора

Данные проекта: Назад

Номер заявки	Название проекта	Ключевые слова проекта	Аннотация проекта	Статус заявки
30	первый проект	методы интеллектуальной обработки информации	автоматизация анализа проектов с применением методов интеллектуальной обработки информации	done

Список аналогов (всего найдено 4033, из них отображено 10):

Название	Ссылка	Содержание	Степень совпадения	
Применение импульсных методов магнитного резонанса в устройствах обработки информации	ссылка	Применение импульсных методов магнитного резонанса в устройствах обработки информации	4.86%	Подробнее
Акустооптический эффект и его применение в системах оптической обработки информации	ссылка	Акустооптический эффект и его применение в системах оптической обработки информации	4.2%	Подробнее
"Психическая болезнь" как интеллектуальный проект	ссылка	"Психическая болезнь" как интеллектуальный проект	3.59%	Подробнее
Разработка и исследование системы обработки технологической информации гидрогенерирующих предприятий	ссылка	Разработка и исследование системы обработки технологической информации гидрогенерирующих предприятий	3.59%	Подробнее
Методы и средства обработки биоэлектрической информации	ссылка	Методы и средства обработки биоэлектрической информации	3.59%	Подробнее
Адаптивные алгоритмы обработки информации в мультиагентных системах	ссылка	Адаптивные алгоритмы обработки информации в мультиагентных системах	3.59%	Подробнее

Рис. 8. Карта проекта для администратора

Данные проекта: Назад

Номер заявки	Название проекта	Ключевые слова проекта	Аннотация проекта	Статус заявки
30	первый проект	методы интеллектуальной обработки информации	автоматизация анализа проектов с применением методов интеллектуальной обработки информации	done

Данные аналога:

Название	Ссылка	Содержание	Степень совпадения
Применение импульсных методов магнитного резонанса в устройствах обработки информации	ссылка	Применение импульсных методов магнитного резонанса в устройствах обработки информации	4.86%

Сравнение частотных характеристик проекта и аналога:

Частотная характеристика проекта		Частотная характеристика аналога	
интеллектуальный 2	проект 2	импульсный 1	**обработка 1**
информация 2	**метод 1**	**информация 1**	**применение 1**
обработка 2	**применение 1**	магнитный 1	резонанс 1
		метод 1	устройство 1

Рис. 9. Результат нажатия на ссылку "Подробнее" для определенного аналога

работы системы, логины и пароли для доступа к таблицам базы данных, имя сервера СУБД (система работает на mysql) и названия самих таблиц. При этом, чтобы получить доступ к режиму администратора, нужно нажать на соответствующую ссылку в верхнем меню системы и пройти аутентификацию.

Карта проекта для администратора также более подробная (рис. 8). Кроме входных данных пользователя и отображения списка аналогов, число которых также можно регулировать в настройках системы (в данном случае оно равно 10), для каждого аналога после названия, ссылки, содержания и степени совпадения выводится ссылка "Подробнее". При нажатии на эту ссылку, кроме уже рассмотренных двух таблиц, отображается третья таблица, в которой приводится сравнение частотных характеристик входных данных (проекта) и выбранного аналога (рис. 9). Способы вывода могут быть настроены администратором системы (в данном случае термины выводятся в два столбца для проекта и аналога), при этом для проекта и для аналога отображаются термины в нормальной форме (для имени существительного, например, это именительный падеж, единственное число) и их частоты; совпадающие термины выделяются жирным шрифтом, и список терминов упорядочивается по убыванию частоты. Если термин находится в списке стоп-слов, он не отображается и не учитывается системой.

Заключение

Система обладает некоторыми недостатками. Помимо упомянутых выше проблем, связанных с не всегда точной работой морфологического модуля, необходимостью обновления списка стоп-слов, главной проблемой системы является зависимость от ресурсов [5] и [6], хоть они и по-прежнему работают и нисколько не изменились за многие годы. Еще одна сложность, которая возникла в результате тестирования системы, — необходимость разгружать сервер хостинга на определенное количество времени. Несмотря на то что сервер достаточно мощный, опытным путем было установлено, что главный скрипт системы может работать только 29 с (которые впоследствии были увеличены до 600 с), после чего должен давать "отдыхать" серверу, сбрасывая промежуточные данные и значения счетчиков циклов в базу данных или в отдельный файл, и возобновлять работу по очередному вызову планировщика cron. При превышении 29 (600) с либо могут исказиться данные, либо сервер автоматически останавливает скрипт, считая его "зависшим", несмотря на то что внутри скрипта есть команды отдыха (sleep) после каждой итерации главного цикла. При этом есть риск наложения работы одного и того же скрипта на самого себя, который необходимо исключать на 100 %. Дело в том, что настройки системы (в коде модуля поиска записано, что он должен работать 600 с с помощью команды set_time_limit(600)) и реальное время работы скрипта могут значительно отличаться — вплоть до 100 %, в связи с чем даже было проведено мини-исследование (основным выводом которого является интервал в заданиях планировщика (cron), равный аргументу функции set_time_limit(), умноженному на 2) со следующими результатами (приведен лишь фрагмент результатов):

выставлено 300 секунд, по факту работает 459.973237991 секунд

выставлено 600 секунд, по факту работает 606.108923197 секунд

выставлено 600 секунд, по факту работает 774.679234982 секунд

выставлено 240 секунд, по факту работает 381.271329165 секунд

выставлено 1 секунд, по факту работает 1.03081297874 секунд

выставлено 600 секунд, по факту работает 680.900939941 секунд

Еще одной проблемой системы является время обработки очередного "заказа" пользователя. Чем больше текст, вводимый пользователем, тем дольше работает система, так как число терминов растет. В ходе тестирования и опытной эксплуатации системы было принято решение искать аналоги, анализируя только названия или темы документов (диссертаций), иначе работа системы может затянуться с учетом вышеописанных сложностей и тонкостей на несколько недель, что будет неприемлемо для пользователя. В зависимости от сложности заказа и серверных настроек системы результаты работы системы приходят на почту пользователя через промежуток времени в среднем от 20 мин до нескольких дней после успешного заполнения полей. Такой временной промежуток обусловлен необходимостью не загружать сервер слишком сильно и не перегружать ресурс [6] поисковыми запросами, а также очередью заявок системы.

Тестирование системы показало, что ресурс [6], несмотря на заявленный морфологический поиск на самом деле предоставляет его не всегда. Достаточно трудно определить, каковы же условия для входных данных, чтобы морфологический поиск учитывался ресурсом полностью; эксперименты выявили, что при вводе в систему ресурса словоформы "интеллектуальный" ресурс выводит одни результаты, а при вводе другой словоформы этого же слова — "интеллектуальной" — ресурс выводит другие результаты, которые пересекаются с первыми, но совсем не на 100 %. Другими словами, то, что есть в первой выборке, отсутствует во второй и наоборот. Это приводит к тому, что разработанная система должна учитывать не только нормальную форму слова ("интеллектуальный"), но и ту форму, которая фактически была использована при вводе входных данных ("интеллектуальной") и, следовательно, работает дольше, чем могла бы. Кроме того, система должна учитывать все словоформы для каждого слова. Вместе с этим тестирование системы выявило различие в кодировках разных почтовых сервисов (hse.ru — UTF-8, mail.ru — KOI8-U и т. д.).

С точки зрения практической полезности система за счет обработки любой информации о предстоящем исследовании позволяет получать полезную и иногда неожиданную информацию, которую нельзя получить с помощью традиционных систем поиска в каталогах диссертаций, так как у пользователя просто не хватает терпения и времени просмотреть и проверить все возможные варианты и комбинации терминов. Кроме того, система обладает максимально простым интерфейсом и выполняет всю работу автоматически, пользователю достаточно ввести входные данные и прочитать через некоторое время письмо со ссылкой

на результаты работы системы. Система доступна по адресу [9].

Список литературы

1. **Использование** phpMorphy — описание и методы. sourceforge / heromantor, последнее изменение от 14.01.2010. URL: http://phpmorphy.sourceforge.net/dokuwiki/manual (дата обращения: 24.12.2017).

2. **Сокирко А. В.** Морфологические модули на сайте www.aot.ru. URL: http://www.aot.ru/docs/sokirko/Dialog2004.htm (дата обращения: 24.12.2017).

3. **Моченов С. В., Бледнов А. М., Луговских Ю. А.** Векторная модель представления текстовой информации: Материалы Международной научной конференции "Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам". URL: http://mns.udsu.ru/conf/report/Mochenov2.pdf (дата обращения: 24.12.2017).

4. **Соколов Е.** Метрические методы классификации (семинары). URL: http://www.machinelearning.ru/wiki/images/9/9a/Sem1_knn.pdf (дата обращения: 24.12.2017).

5. **Научная** электронная библиотека диссертаций и авторефератов disserCat. URL: http://www.dissercat.com (дата обращения: 15.12.2017).

6. **Электронная** библиотека Российской государственной библиотеки. URL: http://sigla.rsl.ru (дата обращения: 14.12.2017).

7. **phpMorphy.** URL: http://sourceforge.net/projects/phpmorphy/ (дата обращения: 14.12.2017).

8. **Макаров С. Л.** Интеллектуальная система автоматизированного поиска диссертаций // Материалы V Всероссийской конференции студентов, аспирантов и молодых ученых "Искусственный интеллект: философия, методология, инновации", г. Москва, МГТУ МИРЭА, 9—11 ноября 2011 г. М.: Радио и Связь, 2011. 272 с.

9. **Система** СТАРТLite. URL: serjmak.com/startlite/ (дата обращения: 15.12.2017).

S. L. Makarov, Ph. D, Associate Professor at the School of Software Engineering
of the Faculty of Computer Science, smakarov@hse.ru
National Research University Higher School of Economics

Information Technologies of a Search for Similar Researches in Dissertations Catalogues

The article represents some information technologies of a search for similar researches, projects or just ideas in catalogues of dissertations and suggests a software web-based system of automated search of the kind. The relevance of the system is justified by impossibility of doing the same work by people and absence of systems which can input a large text for this kind of search. There are several methods described which the system is based upon as well as several mini-researches results which have been conducted in order to understand structures of the dissertations catalogues and to interpret results of the system's work. The system's architecture, database structure and algorithm are presented in the article along with the rules of its running automatically using one of the planning tasks mechanisms of operating systems (cron). There are several user's and the system administrator's interface examples and system outputs provided from the user's and the system administrator's points of view. In the conclusion, key features and difficulties of the system and of the process of using it are outlined together with the outcome of the research with advantages and disadvantages of the system along with mentioning its practical value. Also, as a part of the conclusion there are results of another mini-research about the system's search script lifetime and some outcomes of server and the script settings necessary for getting correct results with the help of the system.

Keywords: *information technologies, automated search, automation, software engineering, search system, project analysis, research summary, dissertations catalogue, user interface*

References

1. ***Ispol'zovaniye*** *phpMorphy — opisaniye i metody* (phpMorphy Manual — Description and Methods), available at: http://phpmorphy.sourceforge.net/dokuwiki/manual (date of access: 12.24.2017) (in Russian).

2. **Sokirko A. V.** *Morfologicheskiye moduli na saite www.aot.ru* (Morphology Modules on www.aot.ru Site), available at: http://www.aot.ru/docs/sokirko/Dialog2004.htm (date of access: 12.24.2017) (in Russian).

3. **Mochenov S. V., Blednov A. M., Lugovskih Yu. A.** *Vektornaya model' predstavleniya tekstovoy informatsiyi* (Vector Model for Representing Textual Information), *Materialy mezhdunarodnoy nauchnoy konferentsii "Sovremenniye informatsionniye tehnologii I pis'mennoye naslediye: ot drevnih rukopisey k elektronnim tekstam"*, available at: http://mns.udsu.ru/conf/report/Mochenov2.pdf (date of access: 12.24.2017) (in Russian).

4. **Sokolov Ye.** *Metricheskiye metody klassifikatsii (seminary)* (Metric methods of classification (seminars)), available at: http://www.machinelearning.ru/wiki/images/9/9a/Sem1_knn.pdf (date of access: 12.24.2017) (in Russian).

5. ***Nauchnaya*** *elektronnaya biblioteka dissertatsiy I avtoreferatov disserCat* (disserCat Scientific Electronic Library of Dissertations and Their Abstracts), available at: http://www.dissercat.com (date of access: 12.15.2017) (in Russian).

6. ***Elektronnaya*** *biblioteka Rossiiskoy gosudarstvennoy biblioteki* (Electronic Library of the Russian State Library, available at: http://sigla.rsl.ru (date of access d: 12.14.2017) (in Russian).

7. **phpMorphy**, available at: http://sourceforge.net/projects/phpmorphy/ (date of access: 12.14.2017).

8. **Makarov S. L.** *Intellektualnaya sistema avtomatizirovannogo poiska dissertatsiy* (Intellectual System for Automated Search for Dissertations), *Materialy V Vserossiiskoy Konferentsii Studentov, Aspirantov I Molodyh Uchyonih "Iskustvenniy Intellekt: Filosofiya, Metodologiya, Innovatsiyi"*, Moscow, Moscow Technological University (MIREA), 9—11 of November, 2011, Moscow, Radio e Svyaz, 2011, 272 p., pp. 202—204 (in Russian).

9. **StartLite** System, available at: serjmak.com/startlite/ (date of access: 12.15.2017) (in Russian).