Система управления хранилищем электронных образовательных ресурсов

Егунова Алла Ивановна
к.и.н., доцент кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет имю Н. П. Огарева
ул. Б.Хмельницкого, 39, г. Саранск, 430005, (8342)478691
egunova@fet.mrsu.ru

Аббакумов Андрей Александрович
к.т.н., доцент кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева
ул. Б.Хмельницкого, 39, г. Саранск, 430005, (8342)478691
abbakumov_aa@mail.ru

Воропаева Мария Анатольевна
магистрант кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева
ул. Б.Хмельницкого, 39, г. Саранск, 430005, (8342)478691
vallis_nimbium@mail.ru

Вечканова Юлия Сергеевна
лаборант кафедры автоматизированных систем обработки информации и управления,
Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва
ул. Большевсистска, 68, г. Саранск, 430005, (8342) 478691
yuliya_kolushova@mail.ru

Аннотация

В статье рассматриваются проблемы повышения эффективности образовательной деятельности вуза и использование научной интеллектуальной собственности в преподавательской деятельности. Рассмотрена архитектура автоматизированной системы управления хранилищем электронных образовательных ресурсов. Определены акторы, участвующие в системе и правила доступа к требуемой информации. На основе анализа размещенных ресурсов разработан алгоритм по обработке запросов на естественном языке и реализован поиск релевантных информационным потребностям пользователей документов. Предложена реализация информационного хранилища и поисковой системы в виде J2EE-приложения с использованием фреймворка Spring. Управление ресурсами и организация доступа к источнику данных выполнено при помощи спецификации JPA Hibernate.

The article deals with the problems of improving the effectiveness of the educational activities of the university and the use of scientific intellectual property in teaching. The architecture of the automated storage management system of electronic educational resources is considered. The actors involved in the system and the rules

for access to the required information are defined. Based on the analysis of hosted resources, an algorithm was developed for processing requests in natural language and a search for documents relevant to the information needs of users was implemented. The implementation of the information storage and search system in the form of a J2EE application using the spring framework is proposed. Resource management and organization of access to the data source is performed using the JPA Hibernate specification

Ключевые слова

Мессенджер, слушатель, web-приложение, бот, система информирования
Messenger, student, web application, bot, information system, NoSQL

Введение

Информатизация образования помогает решить главную задачу в сфере государственной образовательной политики в Российской Федерации – повысить качество получаемых знаний. Системный подход предусматривает формирование навыков работы с электронными ресурсами в образовательном процессе, умению находить релевантную потребностям информацию, обеспечивающую компетентность в будущей сфере деятельности студента. Наряду с формированием доступа к электронно-библиотечным системам вузы создают свои элементы цифрового университета: лекции в виде презентаций, видеороликов или оцифрованных документов, виртуальные лаборатории и другие виды электронных ресурсов. Электронные материалы широко используются в образовательном процессе с элементами дистанционных технологий, повышают возможности самостоятельной работы над дополнительным материалом и становятся незаменимыми для обучения студентов с ограниченными возможностями.

Электронные образовательные ресурсы зачастую привязаны к курсам или личным кабинетам преподавателя. Такие курсы как информатика, изучаются на разных направлениях подготовки и однотипный контент присутствует в различных вариациях. В силу различного подхода к методике подачи информации одного и того же курса, и особенностей восприятия каждого обучающегося, студенты усваивают материал распределяясь по уровням компетентности. Факультативные курсы, рекомендованные для изучения по разным направлениям, формируются единые и могут обеспечиваться единой информационной базой, независимо от ведущего преподавателя.

Формируя электронную информационную базу, вуз заинтересован агрегировать научную интеллектуальную собственность для активного применения в образовательном процессе. Научные доклады, статьи конференций полно отражают актуальные тенденции развития наук и требования современных специальностей. Различные формы подачи учебного материала позволяют повысить эффективность учебного процесса на основе индивидуализации и нивелирования уровневого компетентностного разделения субъектного восприятия.

Образование с элементами дистанционного обучения – еще одна форма обучения в современном обществе, предоставляющая гибкие возможности для широкого круга лиц. Электронные формы учебно-методических материалов, таких как лекции, методические указания, рабочие программы, рабочие планы и др., формируют полное представление о структуре курса и способствуют организации самостоятельной работы в удобное для студента время.

Большую роль в учебном процессе играет система, поддерживающая эффективную функциональность при работе с интересующей информацией за счет

удобного и понятного пользователю интерфейса. Такие системы должны не только обеспечивать оперативный поиск ресурсов, но и позволять анализировать хранимые данные, обновлять версии ресурсов, поддерживая их актуальность, предоставлять услуги для управления единой базой и настраивать соответствующие политики безопасности.

Тенденция прогресса информационных технологий приводит к росту объема больших массивов информации, и главная задача такой системы – это получение полной релевантной информации по требуемому запросу. Наибольший интерес представляют программные продукты, содержащие многофункциональный интерфейс по обслуживанию ресурсов учреждений различных профилей и назначения. В частных случаях соответствующий программный продукт реализуется в виде информационно-поисковой системы, связанной с электронным хранилищем специализированных данных.

Разработка информационного хранилища электронных образовательных ресурсов позволяет автоматизировать функции поддержки актуальности и целостности хранимых ресурсов, представления информации в удобном формате для последующей обработки и анализа, создание эффективных средств организации поиска и анализа запросов пользователя на естественном языке.

Теоретическая часть

Большую роль в учебном процессе играет система, поддерживающая эффективную функциональность при работе с интересующей информацией за счет удобного и понятного пользовательского интерфейса. Такие системы должны не только обеспечивать оперативный поиск ресурсов, но и позволять анализировать хранимые данные, обновлять версии ресурсов, поддерживая их актуальность, предоставлять услуги для управления единой базой и настраивать соответствующие политики безопасности.

Формат хранения ресурсов зависит от выбора типа базы данных, предъявляющей требования к определенному устройству информации внутри них. Несмотря на отсутствие строгой структуры, которая характерна для электронных документов, созданных не по шаблону, ставится задача определения хранилища в зависимости от приоритетов операций, наиболее часто выполняемых с данными. В настоящий момент большой популярностью пользуются документно-ориентированные СУБД, которые способны обеспечить гибкую систему хранения документов с иерархической структурой, масштабируемость и улучшение показателей производительности при правильном использовании. Они поддерживают поиск по полям документов, позволяют запрашивать коллекции документов на основании нескольких ограничений на атрибуты, могут осуществлять агрегатные запросы и сортировку результатов. В качестве основного хранилища на практике обычно выступает реляционная СУБД, но допускаются и гибридные подходы с NoSQL СУБД для работы со сложными объектами, интенсивно изменяющими свое состояние.

Не менее важной является задача эффективного поиска ресурсов, основанная на анализе свободных текстовых запросов пользователей. Помимо традиционных методов информационно-поисковых систем, включающих теорию компьютерной лингвистики и статистические модели, находят широкое применение и алгоритмы искусственного интеллекта.

Одним из исследуемых подходов является онтологический инжиниринг, применяемый в целях систематизации информации и ее приведения к структурированной форме [1].

Использование более простых методов для решения поставленной задачи путем установки конкретных фильтров затрудняет процесс поиска при отсутствии

знаний пользователя по принадлежности документа к конкретным критериям. Задание же всех фильтров не обеспечивает наилучший результат, так как сильно сужает область поиска, отбрасывая документы, которые могли соответствовать по содержанию заданному запросу. Таким образом, разработка информационно-поисковой системы, предоставляющей функционал обработки свободных текстовых запросов пользователя является актуальной задачей, а ее успешное решение способствует нахождению документов с большим процентом релевантности в рамках заданного корпуса текстов.

Реализация поиска, основанного на запросах на естественном языке в системе управления хранилищем электронных образовательных ресурсов

В системе управления хранилищем электронных образовательных ресурсов определены следующие группы пользователей: каталогизатор, администратор, привилегированный пользователь (преподаватели и сотрудники вуза, обучающиеся в Университете) и гость.

Каталогизатор выполняет проверку поступающих заявок, полную классификацию и загрузку информационных ресурсов в рабочие каталоги подсистемы хранения. Администратор выполняет проверку загруженных элементов ЭОР на релевантность заполненных параметров, применяет политики безопасности и разграничение доступа к размещаемым данным. В процессе регистрации документов в системе формируются мета-описания, запускается обработка и обновление статистических данных, а также обновление семантического кластера.

Привилегированный пользователь выполняет навигацию каталога по указанному разделу, устанавливает параметры фильтра запроса по интересующему ресурсу, просматривает метаописание ЭОР, получает доступ к просмотру содержимого ресурса.

Гость – сторонний пользователь, незарегистрированный в системе, может выполнять все действия, доступные привилегированному пользователю, но не может получить доступ к просмотру содержимого ЭОР.

Система требует содержать в составе обслуживающего персонала минимум двух работников: каталогизатора и администратора.

Диаграмма вариантов использования хранилища ЭОР представлена на рисунке 1.

Информационная система управления хранением ЭОР состоит из двух подсистем: подсистемы хранения информационных ресурсов и подсистема доступа к ЭОР.

Подсистема хранения информационных ресурсов предусматривает возможность создания, каталогизации, размещения и полного доступа к электронным информационным ресурсам. С данной подсистемой работают каталогизатор и администратор. Кроме того, в данную подсистему включен закрытый автономный модуль обработки информации для организации процессинга индексирования и поддержки запросов поиска [2].

Для корректного функционирования системы необходимо провести ряд операций по обработке образовательных ресурсов, загруженных в хранилище. При отсутствии обработки такие ресурсы будут недоступны в поисковой выдаче пользователю в виду специфики анализа свободных текстовых запросов. Полный цикл обработки загруженных документов включает следующие стадии:

- ааанализ текста документа;
- пподсчёт термов;

– ссортировка термов;
– оопределение ключевых термов;
– обновление количества документов, включающих ключевой терм [3].

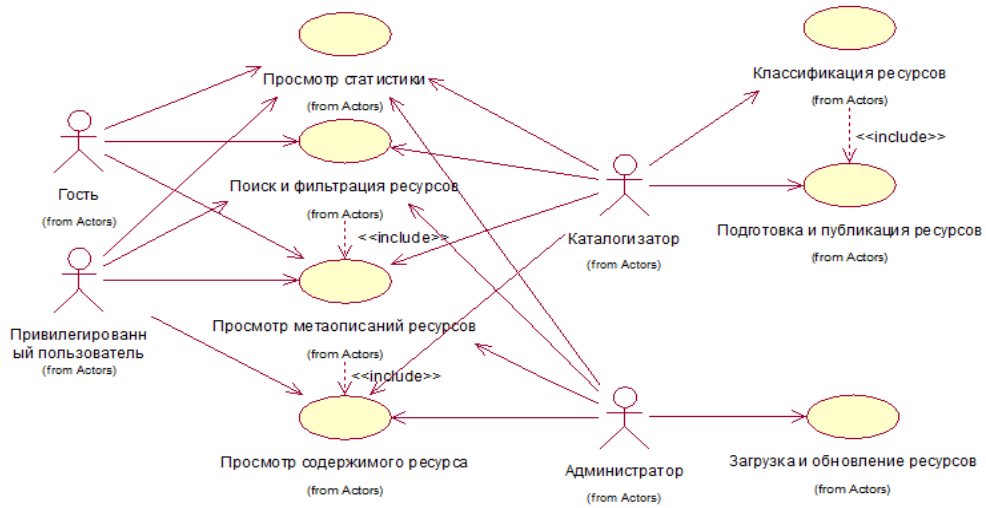

Рис. 1. Диаграмма вариантов использования

При завершении указанных стадий запускается операция «Обновление словаря» для корректного вычисления обратной документной частоты (IDF), необходимой для определения весовых коэффициентов термов поискового запроса

Подсистема доступа включает модуль доступа удаленных пользователей и модуль доступа к фонду хранения ЭОР.

Подсистема хранения информационных ресурсов и поддержки запросов на естественном языке реализована на языке Java с использованием универсальной платформы Spring. Фреймворк Springпредоставляет среду доступа к данным с использованием JDBC или ORM-продуктов, таких как Hibernate, что позволяет эффективно и безопасно управлять сессиями, управлять ресурсами и интегрировано управлять транзакциями.

Слой доступа к данным посредством Spring Data обеспечивает доступ к данным хранилища через набор готовых реализаций для работы с объектами. Для управления сущностями созданы интерфейсы, которые наследуются от интерфейса JpaRepository, предоставляющего набор методов JPA в виде CRUD-операций.

Интерфейсы CourseRepository, DisciplineRepository, FileFormatRepository и JournalRepository реализуют стандартные методы и не имеют расширенной функциональности. Дополнительные методы остальных интерфейсов выполняют запросы на языке HQL (Hibernate Query Language), определенные аннотацией @Query. Поддержка SpEL выражений для использования динамического связывания параметров в аннотациях @Query методов осуществляется посредством аннотации @Param.

Специализированные методы репозитория DictionaryRepository и SemanticClusterRepository реализуют выборку значений ключевых и уникальных значений термов из коллекции документов, подсчет и получение идентификоторов сущности терма, формируют весовые коэффициенты терма для каждого ресурса по идентификатору терма, расширяют терм схожими словами и сортируют элементы семантического кластера. Интерфейс StatisticsRepository реализует подсчет релевантности документов классифицированных дисциплин для расширенного

запроса на основе подсчета суммы коэффициентов термов, входящих в расширенный запрос, для каждого документа соответствующего курса с последующей сортировкой. Список наиболее вероятных курсов определяется классификатором запроса и передается в качестве параметра.

StatisticsRepository использует класс SearchReport для создания пар {Ресурс, релевантность}, на основании которых формируется выдача промежуточных документов, соответствующих информационной потребности пользователя по весовым коэффициентам схемы if-idf. Интерфейс ExtraStatisticsRepository выполняет подсчет числа вхождений терма во все документы заданной дисциплины и формирует запись из обучающих данных классификатора по идентификатору дисциплины и входящему терму. Интерфейс ExtraJournalRepository выполняет выборку записей журнала обработки соответствующих процессу анализа или подсчета статистики, текущего ЭОР.Интерфейс ResourceRepository формирует статус и соответствие ресурсов определенной дисциплине с фиксацией местоположения файлов документов на сервере.

Слой бизнес-логики реализует сервисные интерфейсы для подготовки представления, взаимодействующего с клиентом. Интерфейсы CourseServiceImpl и DisciplineServiceImpl формируют выборку дисциплин и курсов всех определенных направлений. Интерфейс FileFormatServiceImpl отвечает за выборку указанных форматов электронных образовательных ресурсов.

Классы DictionaryService, StatisticsServiceImpl и SemanticClusterImpl работают со словарями термов, рассчитывая и обновляя IDF термов с учетом полученных показателей документной частоты, а также формируя выборку статистических данных для обучения классификатора. Интерфейс ResourceRepository реализует процесс добавления новых ресурсов в систему с извлечением, сортировкой и обработкой термов.

Контроллеры системы управляют запросами пользователя (получаемые в виде запросов HTTP GET или POST, когда пользователь нажимает на элементы интерфейса для выполнения различных действий), вызывая и координируя действие необходимых ресурсов и объектов, нужных для выполнения процессов, задаваемых пользователем. Для обработки запросов для отображения метаданных, характеризующих электронные образовательные ресурсы, используется контроллер DictionaryController. ResourceController выполняет запросы по загрузке ресурсов в хранилище, их постраничное отображение и запуску процесса обработки с ведением журналов действий. Метод processDocuments() выполняет основную нагрузку по анализу образовательного ресурса. Формирование статистики, используемой при поисковых запросах выполняет контроллер StatisticsController.

Все ресурсы хранилища проходят пять стадий, которые записываются в журнал обработки (рис. 2):

- анализ текста документа;
- подсчет термов;
- сортировка термов;
- определение ключевых термов;
- обновление статистики для обучения классификатора;
- подсчет TF;
- подсчет весовых коэффициентов;
- сортировка термов;
- определение ключевых термов;
- обновление поисковой статистики.

После окончания указанных стадий выполняется обновление статуса обработки ресурсов, а в завершении операции в журнал событий добавляется запись "Подсчет статистики" со статусом "Завершено".

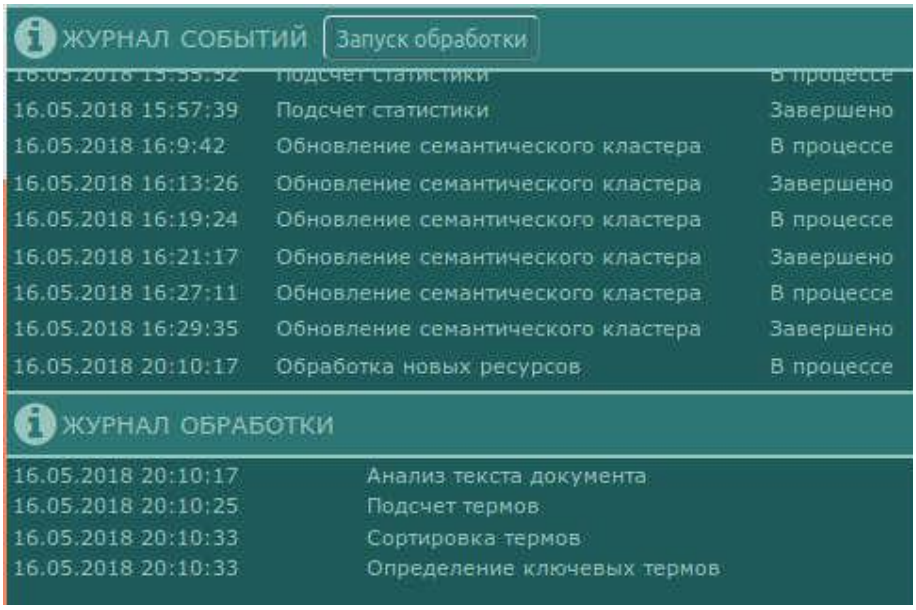

Рис. 2. Процесс обработки ресурса

Заключительный этап регистрации документов – обновление семантического кластера ключевых термов для корректного расширения запроса схожими терминами. Интерфейс по обработке семантического кластера представлен на рисунке 3.

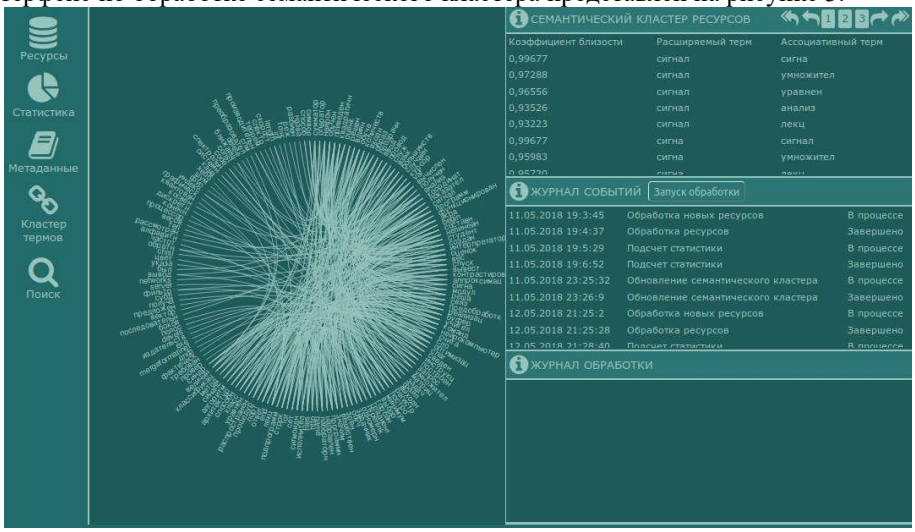

Рис. 3. Процесс обработки ресурса

Левая область представляет собой графическое изображение автоматического тезауруса семантически близких терминов. Для навигации по кластеру необходимо навести курсор мыши на исследуемый терм. В зависимости от таблицы семантических связей возможны три варианта отображения (рисунок 4):

– оранжевый (связь «both») – термы могут расширять другу друга;

– зеленый (связь «source») – терм расширяется понятием на другом конце линии связи;

– красный (связь «target») – терм служит для расширения слова на другом конце линии связи.

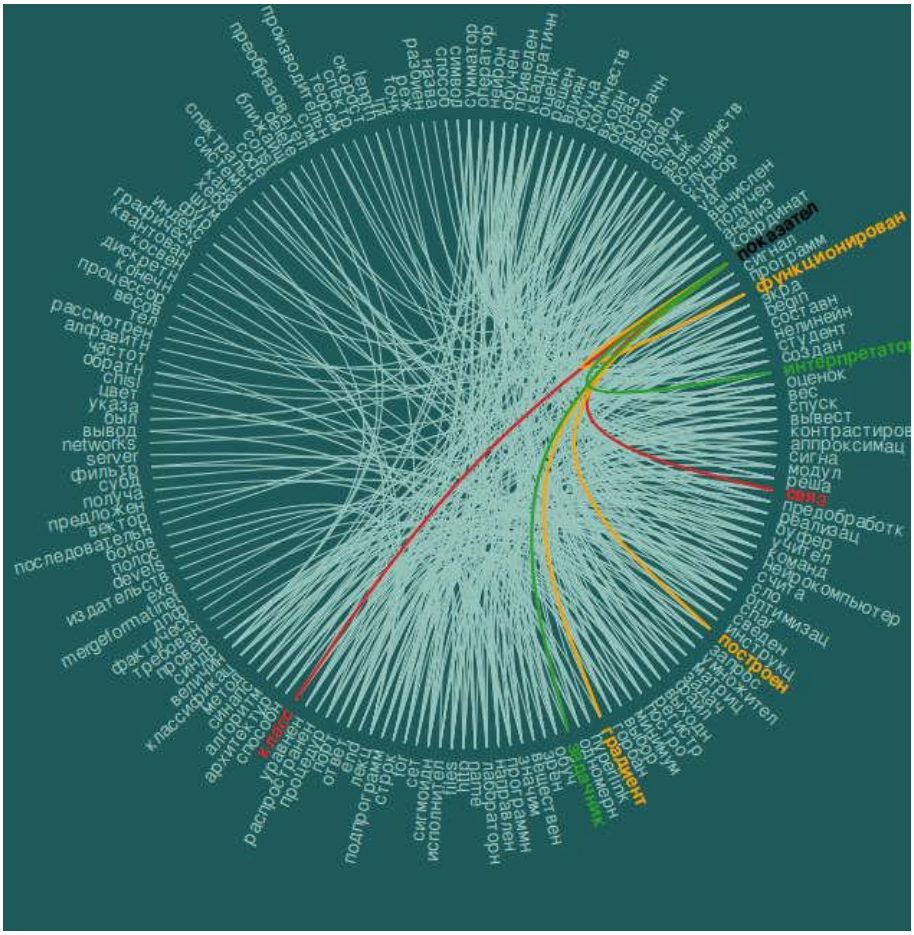

Рис. 4. Отображение связей семантического кластера

Правая область интерфейса по обработке семантического кластера содержит окно «СЕМАНТИЧЕСКИЙ КЛАСТЕР РЕСУРСОВ», отображающее табличный вариант тезауруса с вычисленными значениями косинусной меры между векторами ключевых термов. Окна «ЖУРНАЛ СОБЫТИЙ» и «ЖУРНАЛ ОБРАБОТКИ» служат для управления кластером, обновляя его актуальными сведениями при пополнении коллекции документов новыми ресурсами.

Анализ и оценка разработки

Предлагаемая система информационного хранилища ЭОР реализует основную функцию работы с документами – предоставляет простой и удобный способ доступа к интересующей информации посредством запросов на естественном языке.

Этап предобработки запроса предполагает декодирование последовательности символов для выделения лексем (переменная parse_query) и перевод букв в нижний регистр. Извлеченные лексемы сравниваются со встроенным словарем стоп-слов. При наличии совпадения отсекаются слова, входящие в указанный словарь, а оставшиеся лексемы проходят процедуру обработки стеммером Портера.

На втором этапе осуществляется расширение семантики запроса посредством выбора ассоциативно-связанных терминов из семантического кластера коллекции

документов. Данный кластер содержит пять дополнительных термов с максимальными значениями косинусной меры для ключевых слов, встречающихся в нескольких ресурсах хранилища. Термы, встречающиеся только в одном документе, не нуждаются в поиске скрытых связей, так как достаточно хорошо идентифицируют документ, соответствуя составляющим низкочастотных запросов поисковых систем. Найденные термины добавляются к обработанным лексемам, генерируя расширенное представление запрос

На третьем этапе классифицируется тематика расширенного запроса для сужения области поиска и повышения точности результатов. Подсчитываются весовые коэффициенты наивного байесовского классификатора и выбираются наилучшие категории, наиболее отражающие информационную потребность пользователя.

На заключительном этапе определяется результирующая выборка ресурсов. Документы, соответствующие найденным дисциплинам, ранжируются по найденному на предыдущем шаге коэффициенту классификатора, создавая основные группы, в пределах которых происходит упорядочивание по релевантности.

Заключение

Рассмотрена реализация информационного хранилища электронных образовательных ресурсов с реализацией поиска, основанного на запросах на естественном языке. Описан процесс работы подсистемы хранения информационных ресурсов с разработанной методикой индексирования документов, основанной на модели векторного пространства и статистической меры tf-idf, которая позволяет отсекать «шумовые слова» в пределах коллекции загруженных ресурсов для оптимизации выделения значимых термов [4]. Индексация документов реализует алгоритм автоматизации извлечения ключевых слов из электронных документов после предварительного анализа и очистке текста, расчета весовых коэффициентов для определения релевантности ресурсов для организации процедуры ранжирования результатов поиска.

Алгоритм обработки свободных поисковых запросов пользователя использует гибридный подход с формированием автоматического тезауруса для расширения области поиска по ассоциативно-семантическим связям между ключевыми словами поиска и классификация запроса по дисциплинам документа с выдачей результирующей выборки с достаточной степенью релевантности.

Представленная система позволяет выполнять запросы в свободной форме на естественном языке в различных специализированных информационных хранилищах документов учебных заведениях, для привлечения большего количества пользователей электронных образовательных ресурсов различного формата.

Литература

1. Карпенко А. П., Трудоношин В. А. Оценка релевантности документов корпоративной онтологической базы знаний на основе их иерархической ролевой кластеризации // Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 11. С. 571–584.
2. Егунова А.И., Воропаева М.А. Система управления хранилищем электронных образовательных ресурсов. Синтез науки и общества в решении глобальных проблем современности: Сборник статей Международной научно-практической конференции. Часть 2. – Уфа. МЦИИ ОМЕГА САЙНС, 2016. С. 22 – 25

3. Александров Э.Э., Егунова А.И., Воропаева М.А. Методика обработки текстовых документов для организации информационного поиска с помощью запросов на естественном языке. Научно-технический вестник Поволжья. №1 2018г. – Казань: Научно-технический вестник Поволжья. 2018. – 154 с. С. 102 - 106
4. Аббакумов А.А., Егунова А.И., Воропаева М.А. Информационно-поисковая подсистема хранилища электронных образовательных ресурсов. Научно-технический вестник Поволжья. №3 2017г. – Казань: Научно-технический вестник Поволжья. 2017. С. 100 - 103.