УДК 004.912:004.89

Архитектура интеллектуальной системы информационной поддержки инноваций в науке и образовании

Иванов Владимир Константинович, к.т.н., доцент (Тверской государственный технический университет, директор Центра научно-образовательных электронных ресурсов, +79607046888 или +79051272231, mtivk@mail.ru, почтовый адрес: 170024, Тверь, просп. 50-лет Октября, д. 3, корп. 1, кв. 142.)
Палюх Борис Васильевич, д.т.н., профессор (Тверской государственный технический университет, ректор, pboris@tstu.tver.ru).
Сотников Александр Николаевич, д.ф.-м.н., профессор (Межведомственный суперкомпьютерный центр РАН, заместитель директора, г. Москва, заместитель директора,).

Аннотация
Поиск инновационных решений с использованием различных источников данных является важной составляющей многих направлений деловой активности.
Одним из основных трендов развития методологии и решений для поиска инноваций является автоматизированная семантическая обработка больших массивов научно-технической информации, позволяющая осуществлять поиск прорывных технологий и других инновационных идей.
Очевидно, что нужны эффективные методы создания и наполнения электронных коллекций новейших идей и технологий, содержащих не просто их описания, а специальным образом отобранные, классифицированные и ассоциированные данные.
Разработка новых методов поиска готовых решений в базе данных центра обработки данных (data centre) - суть проекта системы информационной поддержки инноваций в науке и образовании.
В настоящей статье описывается один из подходов к поиску информации об инновациях и область его применения. Приведена общая архитектура системы с указанием пилотных компонентов. Описана архитектура соответствующего программного обеспечения, включая его функциональность и поведение системы в течение конкретного сеанса работы. Представлена объектная модель для работы с документами, под которыми понимаются любые текстовые объекты, имеющие отношение к предмету обработки: запросы, результаты поиска, текстовые документы. Даются сведения о текущем состоянии реализуемого проекта.

Ключевые слова: поддержка принятия решений, архитектура программного обеспечения, поисковый алгоритм, инновация, классификация данных, сервис.

UDC 004.912:004.89

Architecture of Intelligent Information Management System for Innovations in Science and Education

Ivanov V.K., Ph.D., Associate Professor (Tver State Technical University, mtivk@mail.ru)
Palyukh B.V., Doctor of Science (Engineering), Professor (Tver State Technical University, pboris@tstu.tver.ru)
Sotnikov A.N., Doctor of Science (Math), Professor (Joint Supercomputer Centre of Russian Academy of Science, asotnikov@jscc.ru)

Abstract
The search for innovative solutions using different data sources is an important part of many lines of business, science and education.
One of the main trends in the development of methodologies and solutions for search innovation is an automated semantic processing large volumes of scientific and technical information that allows you to search for breakthrough technologies and other innovative ideas.
Obviously, we need efficient methods of digital collections creating and filling with the latest ideas and technologies that contain not just their descriptions, but specially selected, classified and associated data.
The development of new methods to search for ready-made solutions in the data center database is the essence of the project Information System to Support Innovation in Science and Education.
This article describes one approach to finding information on innovations and its scope. The common architecture of the software and pilot components are presented. The architecture of the relevant software, including its functionality and behavior of the system during a given session are described. The object model for working with documents is presented. The document means any text object that is relevant to the subject matter of processing: query, search result, text document. Given the information about the current state of the ongoing project.

Keywords: decision support, software architecture, search algorithm, innovation, data classification, service.

Архитектура интеллектуальной системы информационной поддержки инноваций в науке и образовании

Иванов В.К., Палюх Б.В., Сотников А.Н.

Введение

Поиск инновационных решений с использованием различных источников данных является важной составляющей многих направлений деловой активности. Разнообразные исследования в этой области инновационного менеджмента касаются особенностей поиска инноваций в экономике, науке, образовании (см., например, [1], [2], [3] и многие другие). В этой связи несомненную ценность представляют специализированные коллекции научно-технических достижений [4]. Выделим один из основных трендов развития методологии и решений для поиска инноваций - автоматизированная семантическая обработка больших массивов научно-технической информации, позволяющая осуществлять поиск прорывных технологий и других инновационных идей. В качестве примеров приведем несколько известных решений: illumin8 [5], NetBase [6], Orbit [7]. При всех различиях этих и других подобных систем основной паттерн поиска включает в себя отбор материалов по запросу, выделение ключевых понятий в заданной области и соответствующую группировку материалов, фильтрацию результатов, генерацию аналитических отчетов.

Не затрагивая вопросов стратегии внедрения инноваций в конкретных приложениях, отметим ряд принципиальных на наш взгляд особенностей, касающихся реализации непосредственных механизмов автоматизированного поиска инновационных решений:

- Искомые решения часто находятся на стыке смежных областей; отсюда сложности формулировки точного запроса.
- Одновременно с информацией о собственно инновациях желательно получить сведения о примерах применения, рисках, особенностях использования, пользователях, авторах, производителях.
- Наличие альтернатив и необходимость одновременного поиска критериев отбора наиболее эффективных решений.
- Разрозненность и неоднородность сведений об инновациях; преимущественно внутриотраслевой характер.

Очевидно, что полностью задача автоматизированного поиска инновационных решений далеко не решена. Нужны новые эффективные методы создания и наполнения электронных коллекций новейших идей и технологий, содержащих не просто их описания, а специальным образом отобранные, классифицированные и ассоциированные данные.

В настоящей статье описывается один из подходов к поиску информации об инновациях и область его применения, представляется архитектура соответствующего программного обеспечения, даются сведения о текущем состоянии реализуемого проекта программной системы с функциями семантического поиска и интеллектуального анализа данных для предложения инновационных решений.

О цели проекта

Предположим, что пользователю необходимо получить максимально исчерпывающую информацию о возможных инновационных решениях задачи в какой-либо предметной области. Естественные первоочередные действия - выполнение поискового запроса/запросов для поиска научно-технической информации в:

- ресурсах Интернет (издания общероссийских и отраслевых институтов информации, справочники, статьи и обзоры, материалы конференций, ГОСТы, технические регламенты, нормативно-техническая документация, отчеты о НИР/ОКР, рекламные материалы, статистические данные, экспертные оценки).
- специализированных базах данных (патентных, описаний изобретений и полезных моделей, промышленных образцов, реферативной и/или библиографической информации, товарных знаков).

В результате у пользователя в распоряжении будет большое количество данных, в той или иной степени релевантных соответствующим запросам. При этом, как правило, у пользователя нет возможности подробно рассмотреть все имеющиеся результаты. И возникают следующие вопросы:

- Является ли ранжирование результатов, выполненных поисковой системой, корректным с точки зрения ожиданий пользователя?
- Все ли результаты, доступные для непосредственной оценки пользователем, соответствуют ожиданиям пользователя?
- Все ли результаты, соответствующие ожиданиям пользователя, попали в число доступных для непосредственной оценки?
- Все ли искомые решения найдены вообще?
- Могут ли быть обнаружены эффективные решения, которые относятся к другим областям применения, но могут быть успешно использованы как инновации в данной области.

Ответы на эти вопросы может дать выполнение работ по проекту системы информационной поддержки инноваций в науке и образовании. Суть проекта - разработка новых методов поиска готовых решений в базе данных центра обработки данных (data centre) и ее пополнения результатами интеллектуального анализа данных Интернет. Пользователи должны иметь возможность визуально оценить найденные решения в совокупности со связанными объектами. Основной инструмент - приложение для мобильных устройств с переносом большей части ресурсоемких вычислений в облачный сервис.

Общая архитектура

Согласно [8] типовая архитектура программного обеспечения включает слои представления, сервисов, бизнес-логики, доступа к данным, а также сквозную функциональность, которые должны обеспечивать взаимодействие пользователей и внешних систем с источниками данных. На рис. 1 приведена общая архитектура системы с указанием пилотных компонентов (графические элементы-окружности). Обоснование состава пилотных компонентов - реализация полного цикла обработки с ограниченной функциональностью каждого этапа (слоя).

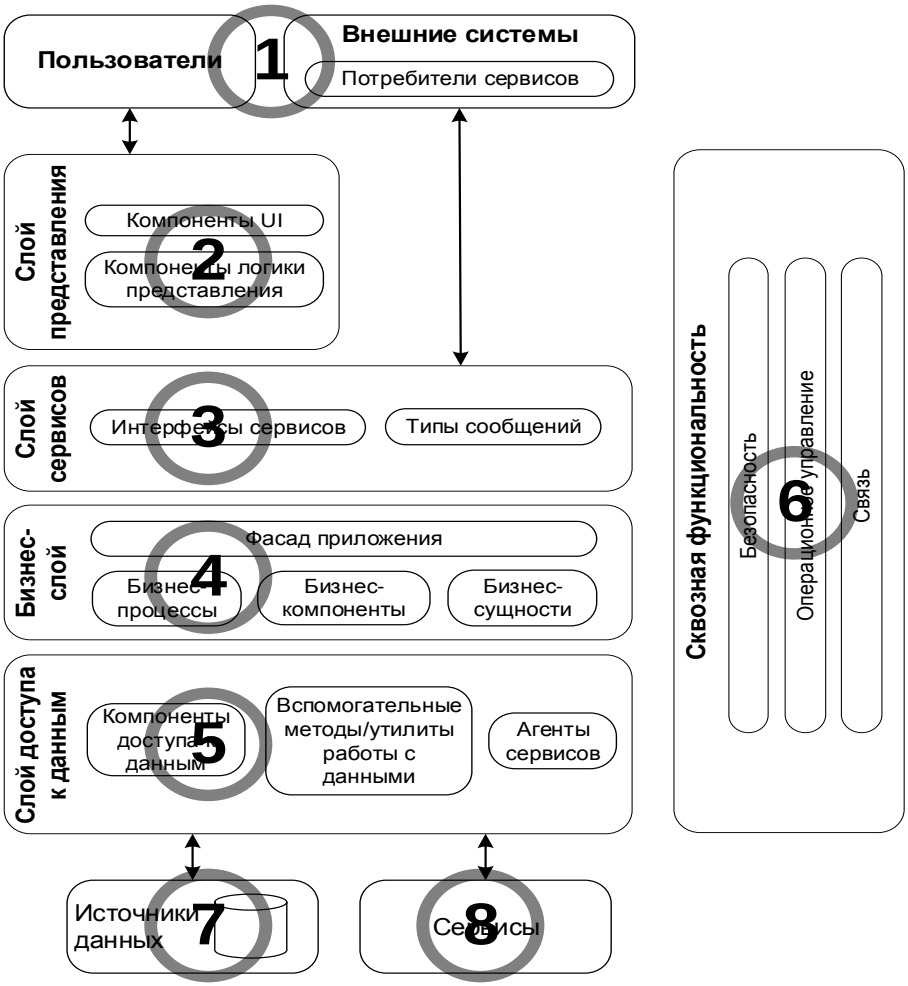

Рис. 1. Общая архитектура

Ниже приведен состав пилотных компонентов с их кратким описанием их функциональности:

1. Пользовательское приложение.
2. Графическая визуализация результатов поиска и работы поисковых алгоритмов, семантику связей между объектами.
3. Сервисы для поиска и резюмирования решений научно-технических и образовательных задач, имеющих инновационный потенциал.
4. Библиотеки классификационных алгоритмов и алгоритмов определения семантически связанных данных.
5. Программная реализация модели векторного пространства документов: объектная модель, библиотеки доступа к документной базе данных, индексатор данных.
6. Подсистема мониторинга - учет и анализ посещаемости и цитируемости ресурсов различными категориями пользователей.
7. Ресурсы Интернет, специализированные базы данных.
8. Реестр инновационных решений научно-технических и образовательных задач.

Архитектура программного обеспечения

Общее представление функциональности программного обеспечения системы показано на рис. 2. Применена нотация диаграммы использования UML с действующими лицами, вариантами использования, ассоциациями между ними, а также с зависимостями между вариантами использования.

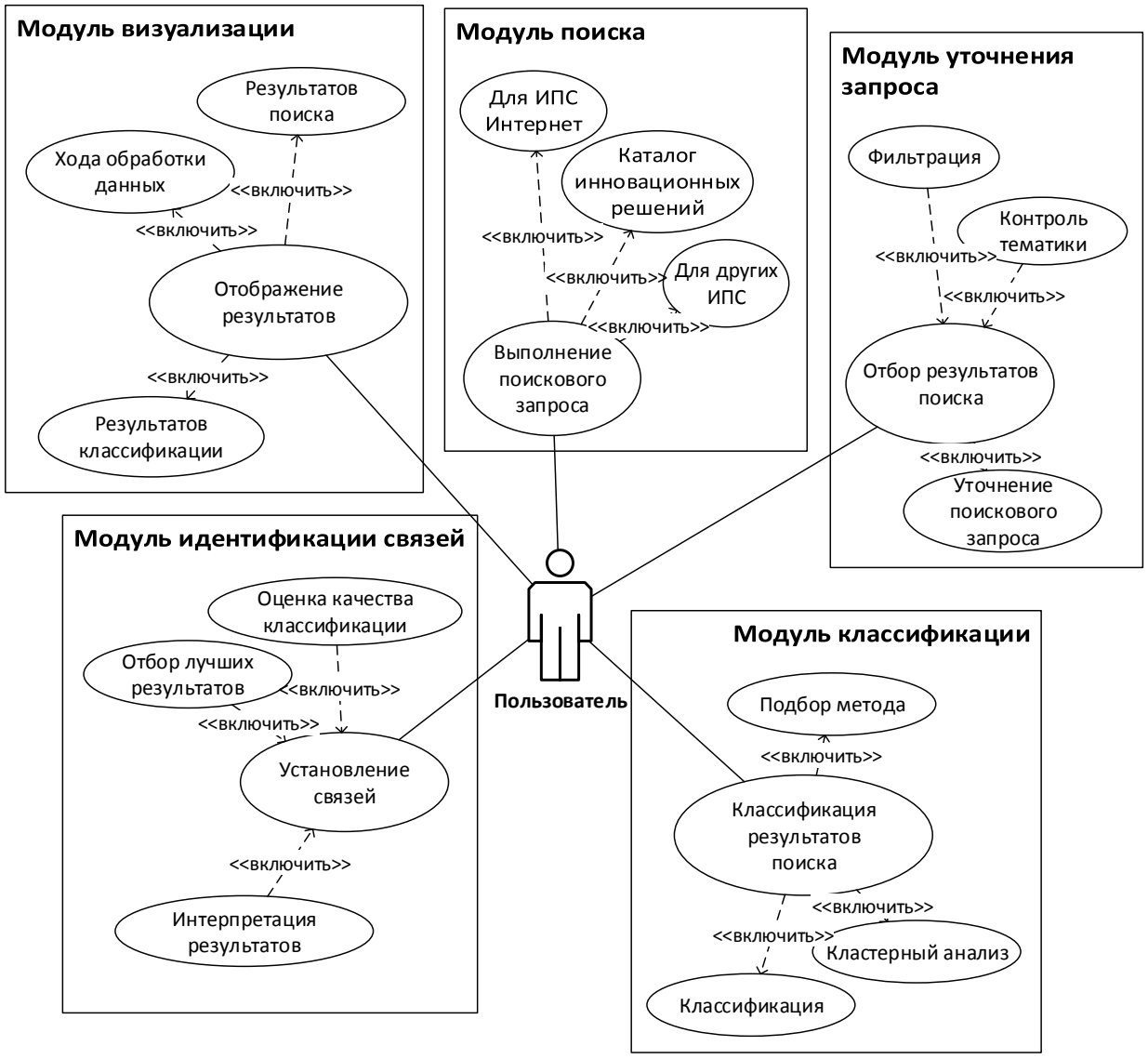

Рис. 2. Функциональность программного обеспечения

Поведение системы в течение конкретного сеанса работы представлено на диаграмме последовательности UML (рис. 3). Изображена последовательность сообщений между взаимодействующими объектами-классификаторами (компонентами и действующими лицами). Отметим два периода активации пользователя: формулировка запроса (начальный шаг) и визуализацию результатов, включая получение вариантов запрошенного инновационного решения и связанных объектов (конечный шаг). Промежуточные шаги отражают алгоритмические аспекты взаимодействия компонентов системы.

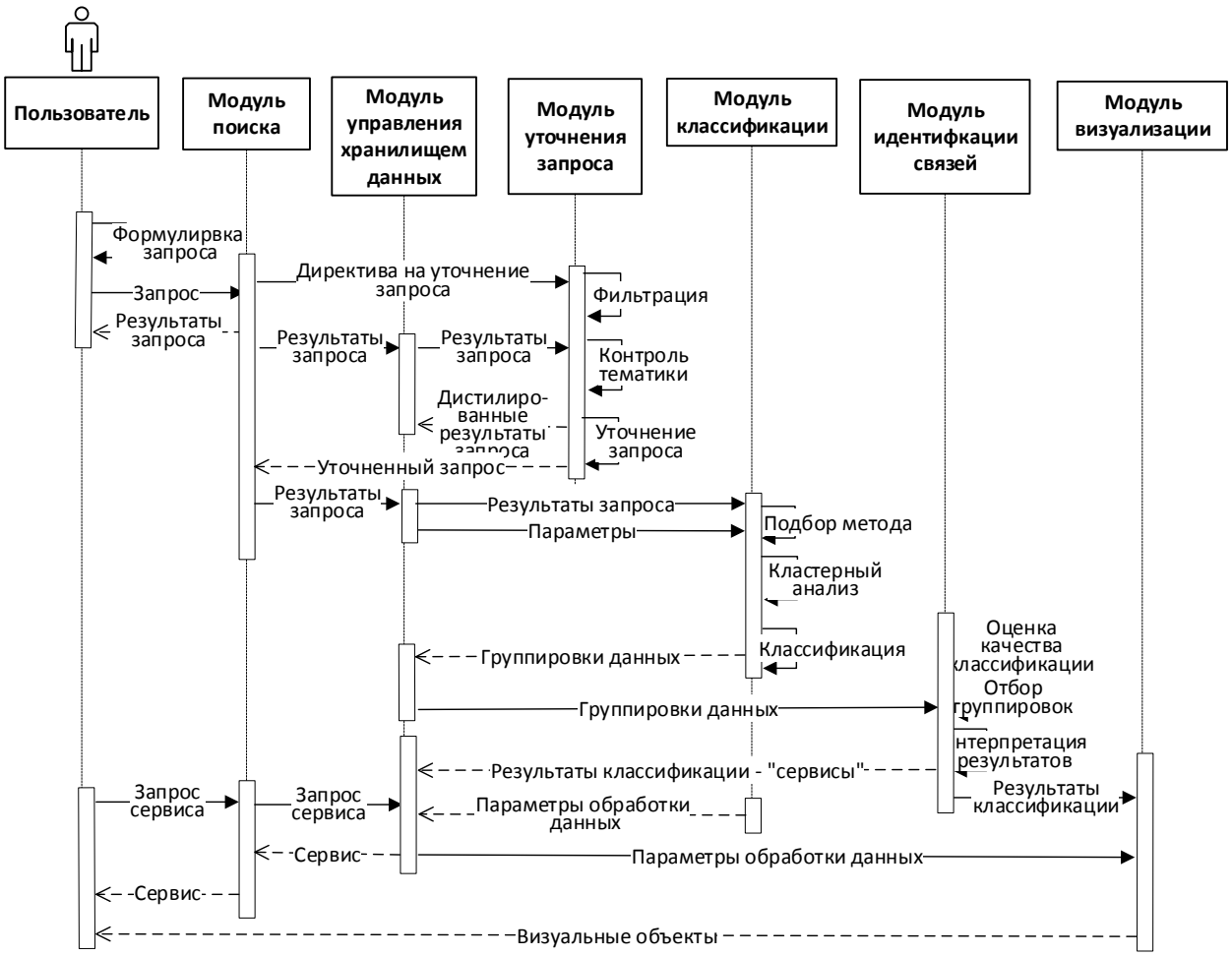

Рис. 3. Поведение компонентов программного обеспечения

Таким образом, основными функциональными компонентами проектируемой системы для интеллектуальной обработки результатов поиска информации будем считать:

1. Модуль поиска. Выполнение поискового запроса для: поисковых систем Интернет, собственного каталога инновационных решений, других поисковых систем. Используется базовый поиск (запросы по атрибутам и по полным текстам), определение местоположения, извлечение данных.
2. Модуль уточнения запроса. Отбор результатов поиска: фильтрация, контроль тематики, уточнение поискового запроса.
3. Модуль классификации. Классификация результатов поиска: подбор метода, кластерный анализ
   текстовых документов и мультимедийных объектов, классификация. Результат - подмножества семантически связанных данных.
4. Модуль идентификации связей. Установление связей: оценка качества классификации, отбор лучших результатов, интерпретация результатов. Модуль завершает генерацию решений, имеющих инновационный потенциал, которая осуществляется в заданном тематическом сегменте или по заданному объекту (промышленному изделию, технологии, продукту).
5. Модуль визуализации. Отображение результатов: результатов поиска, хода обработки данных, результатов классификации.

6. Модуль управления хранилищем данных. Хранение результатов поиска и обработки данных, параметров, промежуточных данных. Само хранилище данных построено на основе модели векторного пространства документов [9] и пополняется в процессе своей актуализации.

На диаграмме последовательности не отображен служебный модуль мониторинга, основными функциями которого являются (а) учет и анализ запрашиваемых ресурсов; (б) агентный мониторинг доступных открытых информационных ресурсов для автономного пополнения хранилища и (в) фоновая индексация документов хранилища.

Объектная модель

На рис. 4 изображена объектная модель программного обеспечения в виде диаграммы классов UML. Данная объектная модель предназначена для работы с документами, под которыми здесь понимаются любые текстовые объекты, имеющие отношение к предмету обработки: запросы, результаты поиска, текстовые документы. Представлены основные (но не все) используемые сущности, а также ассоциации и зависимости между ними. Ниже кратко прокомментированы элементы модели.

Класс *documentGeneral* задает среду обработки документов и порождает классы:

- *documents* - коллекция документов для обработки. Ассоциированный с ним класс *TFIDFmeasure* обеспечивает вычисление мер сходства документов корпуса.
- *finder* - поисковые функции в корпусе документов.
- *reports* - определяет виды отчетов из базы данных документов; выходные формы задаются классом *reportOutput.*

Класс *document* - описывает конкретный документ. Ассоциированные с ним класс *files* - определяет представление документа в файловой системе. Ассоциирован с классом *document.*

Класс *words* - определяет коллекцию слов документа, каждое из которых описывается классом *word*. Лемматизация слов задается классом *stemmer.*

Класс *documentStructure* определяет структуру документа и порождает классы:

- *divisionList* - коллекция описаний составных частей или разделов документа.
- *paragraphList* - коллекция описаний абзацев документа.
- *structure* - описание структуры документа (взаимосвязей между составными частями документа, включая типы связей и их реализацию).

Классы *tableOfContent* и *literatureList* описывают специфические части документов: оглавление и список литературы соответственно.

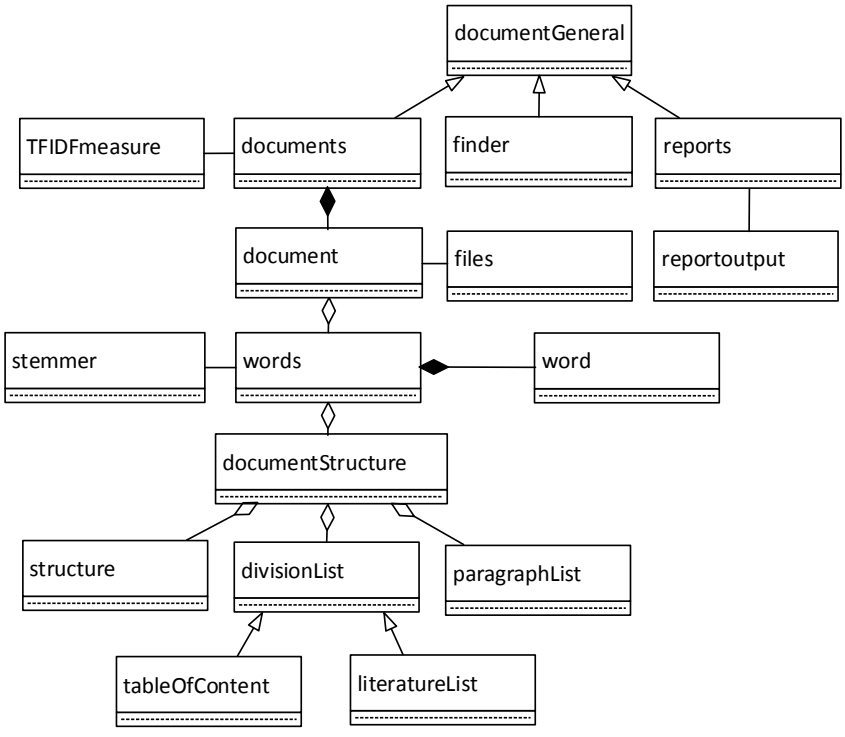

Рис. 4. Объектная модель программного обеспечения (основные классы)

Реализация и приоритеты

Некоторые компоненты обсуждаемой уже реализованы и проходят апробацию в различных приложениях. Так, прототип хранилища данных, построенного с использование модели векторного пространства документов, а также некоторые ключевые элементы модулей классификации документов и идентификации связей прошли успешную апробацию при реализации смежных технологий интегральной оценки качества электронных документов [10] и анализа сходства документов в различных контекстах [11].

Среди основных приоритетных задач разработки представляется важным отметить следующие:

- Разработка унифицированного программного интерфейса доступа к информационным ресурсам.
- Определение методики уточнения запросов и автоматической фильтрации результатов поиска.
- Количественное определение степени пертинентности найденных документов - претендентов на включение в состав искомых инновационных решений.
- Алгоритмизация определения семантического ядра - идентификации связей найденных документов с похожими объектами.

Работы проводились при финансовой поддержке РФФИ (договор № НК13-07-00342\13).

Литература

1. Байгулов, Р.М. Управление промышленным предприятием: специфика поиска инновационных бизнес-идей / Р.М. Байгулов, Е.В. Рожкова // Современные проблемы науки и образования. - 2012. - № 2; Режим доступа: www.science-education.ru/102-5896 (дата обращения: 09.06.2013).
2. Куракова, Н.Г. Создание прорывных инноваций на основе комбинации научных заделов мирового уровня как компетенция инновационного менеджмента / Н.Г. Куракова, В.Г. Зинов // Инновации. - 2012. - № 10. - С. 37-42.
3. Российское образование: тенденции и вызовы: сб. ст. и аналитических докл. - М.: Изд-во «Дело» АНХ, 2009. — 400 с.
4. Антопольский, А., Каленкова, А., Каленов, Н., Серебряков, В., Сотников, А. Принципы разработки интегрированной системы для научных библиотек, архивов и музеев // Информационные ресурсы России. – 2012. - № 1. – С. 2-6.
5. Illumin8. A powerful research tool for innovation & product development. - Режим доступа: http://www.illumin8.com (дата обращения: 19.06.2013) - Загл. с экрана.
6. NetBase Social Media Management System (SMMS). - Режим доступа: http://www.netbase.com (дата обращения: 19.06.2013) - Загл. с экрана.
7. Questel Intellectual Property Portal. - Режим доступа: http://www.orbit.com (дата обращения: 19.06.2013) - Загл. с экрана.
8. Руководство Microsoft по проектированию структуры приложений: 2-е издание, 2009. - 529 с. - Режим доступа: http://www.microsoft.com/architectureguide (дата обращения: 17.06.2013).
9. Salton, G., Wong, A., Yang, C. S. A Vector Space Model for Automatic Indexing. Communications of the ACM (1975), vol. 18, nr. 11, ps 613–620.
10. Иванов, В.К. Критерии интегральной оценки электронных документов в системах подготовки принятия решений // Вестник Тверского государственного технического университета : науч. рецензир. журнал. - Вып.22. - С. 20-26. - Тверь, 2012.
11. Иванов, В.К. Особенности анализа сходства документов в различных контекстах заимствования при подготовке текстовых материалов / Иванов, В.К., Миронов, В.И. // Оценка качества высшего профессионального образования с учетом требований ФГОС и профессиональных стандартов : материалы докладов науч.-практ. конференции. - С. 20-28. - Тверь, 2013.

References

1. Bajgulov, R.M., Sovremennye problemy nauki i obrazovanija, 2012, no. 2, www.science-education.ru/102-5896.
2. Kurakova, N.G., Zinov V.G., Innovacii, 2012, no. 10, pp. 37-42.
3. Rossijskoe obrazovanie: tendencii i vyzovy, Sb. st. i analiticheskih dokl., Izd-vo «Delo» ANH, 2009, 400 p.
4. Antopol'skij, A., Kalenkova, A., Kalenov, N., Serebrjakov, V., Sotnikov, A., Informacionnye resursy Rossii, 2012, no. 1, pp. 2-6.
5. Illumin8. A powerful research tool for innovation & product development, www.illumin8.com.
6. NetBase Social Media Management System (SMMS), www.netbase.com.
7. Questel Intellectual Property Portal, www.orbit.com.
8. Microsoft Application Architecture Guide, 2nd Edition, 2009, 529 p., www.microsoft.com/architectureguide.
9. Salton, G., Wong, A., Yang, C. S., Communications of the ACM (1975), vol. 18, no. 11, pp. 613–620.
10. Ivanov, V.K., Vestnik Tverskogo gosudarstvennogo tehnicheskogo universiteta, 2012, vol. 22, pp. 20-26.
11. Ivanov, V.K. Mironov, V.I., Ocenka kachestva vysshego professional'nogo obrazovanija s uchetom trebovanij FGOS i professional'nyh standartov : materialy dokladov nauch.-prakt. konferencii, Tver, 2013, pp. 20-28.