УДК 004

ИНТЕЛЛЕКТУАЛЬНЫЕ МЕТОДЫ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ.

Кулаков В.В., Мельник Е.В., Коломойцева И.А.
Донецкий Национальный Технический Университет
кафедра прикладной математики и информатики
E-mail: BBCoders@mail.ru

Аннотация

Кулаков В.В., Мельник Е.В., Коломойцева И.А. Интеллектуальные методы извлечения информации из текстов на естественном языке. *Рассмотрены методы извлечения знаний с помощью анализа текстов. Разработана система для автоматического извлечения данных на основе данных методов.*

Общая постановка проблемы

Задача анализа текстовых документов ориентирована на извлечение знаний и является в настоящее время актуальной проблемой, затрагивающей различные сферы человеческой деятельности, поскольку ее решение позволит полностью автоматизировать процесс обработки, классификации и систематизации информационного ресурса.

Бурный рост объема текстов, в которых ведется поиск, привел к тому, что те статистические методы, которые сделали возможным быстрый и эффективный поиск по большим массивам неструктурированных данных, стали “мешать” эффективности этого поиска. Преимущества этих методов - отсутствие необходимости подробного семантического описания предметной области и содержательного анализа текстов, породило и его ограниченность.

В последнее время на первый план выходит задача предварительной “когнитивной обработки” текстов. С одной стороны, борьба с “переизбытком” выдаваемой пользователю информации вылилась в то, что современные системы анализа текстов от задачи информационного поиска: найти документ по заданной тематике, переходят к задаче извлечения информации из текстов (informationextraction) и более глубокого анализа извлеченной информации - извлечения знаний (datamining). В результате пользователь по своему запросу получает не “мешок” текстов, а некоторые “суммирующие” данные, определенным образом структурированные. Если приписать тексту некоторые “семантические” метки, это позволит частично уменьшить количество “шума” - текстов не из той области знаний, которая интересует пользователя, и поможет решить проблему омонимии. Таким образом, на первый план выдвигается задача разработки специальных языков и систем, описывающих понятийную структуру той или иной области знаний - тезаурусов и онтологий.

Назначение данной работы: извлечение фактов (в данном случае – характеристик объектов поиска); извлечение мнений (opinionmining/extraction, анализ комментариев и отзывов пользователей об исследуемых продуктах с целью выявления их основных достоинств и недостатков).

Целью исследовательской работы является исследование методов автоматического анализа текстовых документов на естественном языке и разработка системы для извлечения основных характеристик и формирования приближенного мнения об анализируемых объектах.

Анализ существующих решений

На данный момент проблема извлечения фактов из естественных текстов остается открытой, поэтому существует малое количество готовых продуктов, решающих данную

проблему. Многие из них являются коммерческими проектами, но есть и свободно распространяемые продукты. Пожалуй, самым известным решением является продукт фирмы Айтеко «Аналитический курьер».

«Аналитический Курьер» является инструментом аналитической разведки, который позволяет быстро погружаться в новые предметные области. Уникальной особенностью системы является совместное применение различных методов извлечения знаний в одном сценарии, например, сначала производится кластерный анализ подборки сообщений, затем строится семантическая сеть тем для выбранного кластера, после чего делается частотный анализ временного ряда сообщений по взаимосвязанным проблемам и др.

В системе реализованы уникальные по качеству методы анализа мнений и определения тональности публикаций.

К достоинствам системы относятся высокая степень автоматизации и адаптивности методов извлечения знаний, а также минимальная стоимость ее эксплуатации по сравнению с аналогами.

Данная система имеет широкие возможности, но является коммерческим проектом и довольно таки сложна для работы с конкретными предметными областями.

Также существует коммерческий продукт RCO Fact Extractor Desktop – это персональное приложение для Windows, которое предназначено для аналитической обработки текста на русском языке и выявления фактов различного типа, связанных с заданными объектами – персонами и организациями. Основная сфера применения программы – это задачи из области компьютерной разведки, требующие высокоточного поиска информации, например, автоматический подбор материала к досье на целевой объект или же мониторинг определенных сторон его активности, освещаемых в СМИ.[3]

Помимо приложения Fact Extractor Desktop компания также предлагает библиотеку Fact Extractor SDK, которая должна предоставлять возможность встраивать функционал по анализу текстов в другие приложения.

Естественно, библиотека также распространяется на коммерческой основе.

Следующей известной системой для анализа текста является проект AOT (www.aot.ru). Данный проект является бесплатным (распространяется под лицензией LGPL). Каждый желающий может бесплатно использовать библиотеки в своих программах, в том числе и в коммерческих приложениях.

Их технологии базируются на многоуровневом представлении естественного языка, которое, в свою очередь, было заимствовано у системы ФРАП.

Компоненты, составляющие языковую модель, - лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Выделяются следующие компоненты:

- Графематический анализ. Выделение слов, цифровых комплексов, формул и т.д.
- Морфологический анализ. Построение морфологической интерпретации слов входного текста.
- Синтаксический анализ. Построение дерева зависимостей всего предложения.
- Семантический анализ. Построение семантического графа текста.

Изначально, рассматривалась возможность использования библиотеки AOT (словари, а также управление словарями) в разрабатываемой системе, однако по некоторым объективным причинам (главная из которых - излишняя стратификация, изоляция текстов от данных, которые ими передаются) это решение было сочтено нецелесообразным решаемой задаче.

Описание системы

Была спроектирована система автоматизированного извлечения фактов из естественных текстов на основе интеллектуальных методов анализа. Разработанная, в ходе данной работы, система состоит из следующих основных частей (рис. 1):

- менеджер потока;
- html-парсер;
- интеллектуальный анализатор;
- база объектов;
- база словарей;
- модуль управления словарями;
- модуль управления базой объектов;
- модуль обучения.
- клиентский модуль

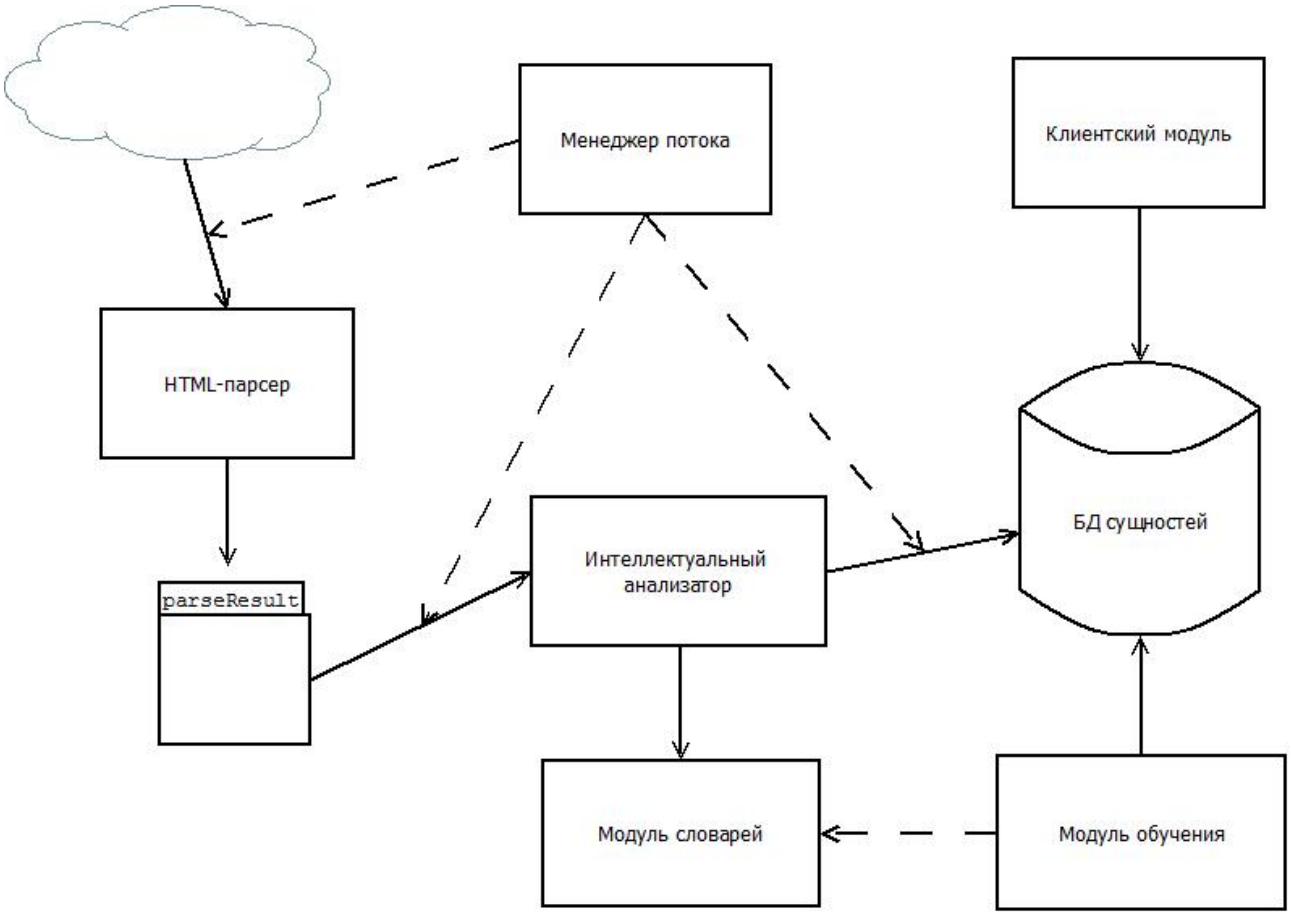

Рисунок 1 –Схема системы автоматического извлечения информации

Менеджер потока занимается выбором web-страниц из сети интернет, перенаправлением извлеченных данных html-парсеру, а также передачей результатов работы html-парсера интеллектуальному анализатору.

Html-парсер приводит текст к необходимому для анализа виду, а также извлекает базовые характеристики объекта поиска.

Интеллектуальный анализатор обрабатывает отзывы об объекте поиска, а также добавляет данные в базу объектов.

База объектов хранит информацию об объектах и связях между ними, также в базе содержаться шаблоны представления объектов.

Модуль словарей (рис. 2) предполагает работу с базой словарей.

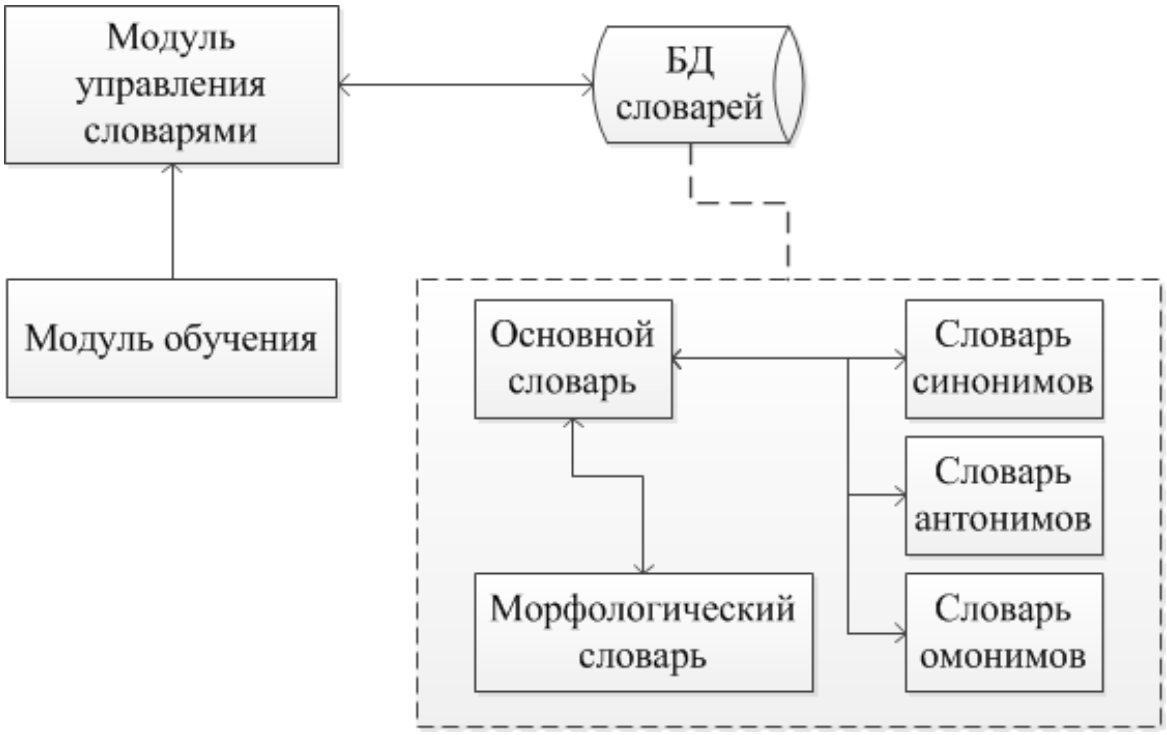

Рисунок 2 – Модуль словарей

База словарей представляет собой базу слов, объеденных тезаурусами (морфологические, синонимические, антонимические связи и омонимия).

Модуль управления словарями представляет интерфейс управления базой данных словарей, позволяющий взаимодействовать с интеллектуальным анализатором и базой объектов.

Модуль обучения имеет интерфейс для ручного редактирования информации, а также предполагает возможность автоматического обучения. Промежуточным вариантом работы модуля обучения является такой процесс, при котором по данным, которые не смог обработать анализатор, составляется "список вопросов" для администратора системы.

Клиентский модуль предназначен для того, чтобы обеспечить возможность пользователю максимально эффективно использовать информацию, накопленную в результате работы системы (рис. 3), содержит элементы управления, с помощью которых пользователь системы может указать необходимые параметры поиска.

Также предусмотрены модули GUI для работы администратора системы, которые предоставляют интерфейсы для управления процессом сбора данных, управления словарями системы, управления процессом обучения, как в ручном так и в автоматическом режимах.

Предметной областью нашего исследования является компьютерное оборудование. Пользователь данной системы вводит название интересующего его объекта и система предоставляет ему основные характеристики данного продукта, а также информацию, основанную на анализе отзывов о данном продукте. Исходные данные берутся со страниц сети Internet, содержащих как информацию о продукте, предоставленную производителем, так и отзывы пользователей. Анализ отзывов может дать довольно точную оценку для

продукта, а возможность использовать большое количество отзывов для каждого объекта может помочь компенсировать зашумленность результатов.

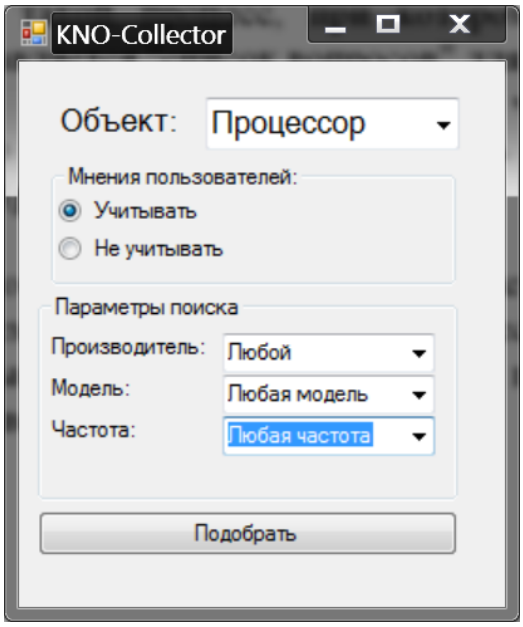

Рисунок 3 – Интерфейс клиентского модуля

Выводы

Проблема автоматического анализа текстов в данный момент является очень перспективной. Разрабатываемая нами система призвана помочь конечному пользователю получить наиболее достоверные знания об объекте поиска.

На данный момент реализовано:

- базовый функционал для модуля словарей и модуля обучения;
- разработана масштабируемая система первичного разбора данных с html-разметкой;
- создана базовая версия интеллектуального анализатора.

В дальнейшем планируется:

- усовершенствовать алгоритмы работы анализатора с целью повысить универсальность разрабатываемой системы;
- расширить возможности модуля словарей;
- усовершенствовать модуль обучения для улучшения его работы в автоматическом режиме.
- расширять набор html-парсеров для взаимодействия с различными ресурсами.

Список литературы

1. Автоматическая Обработка Текста [Electronicresourse] / Интернет-ресурс. - Режимдоступа :www/ URL: http://www.i-teco.ru/ac.html. - Загл. с экрана.

2. Айтекотехнологиибезпробелов [Electronicresourse] / Интернет-ресурс. - Режимдоступа :www/ URL: http://aot.ru/technology.html. - Загл. с экрана.

3. RCO Fact Extractor Desktop [Electronicresourse] / Интернет-ресурс. - Режимдоступа :www/ URL: http://www.rco.ru/product.asp?ob_no=1131. - Загл. с экрана.