УДК 004.912, 004.738.52

А.В. ПРОХОРОВ, Е.А. БОРВЕНКО

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

МУЛЬТИАГЕНТНАЯ СИСТЕМА ПОИСКА И ТЕМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ В ИНТЕРНЕТ

Рассмотрены вопросы тематического анализа неструктурированной текстовой информации, источниками которой выступают многочисленные ресурсы сети Интернет. Рассмотрены особенности и методы тематического анализа текстовой информации. Представлен краткий анализ последних исследований и публикаций. Предложена структура мультиагентной системы поиска и тематического анализа текстового контента web-ресурсов. Представлены функциональные возможности приложения для тематического анализа текст. Описаны особенности программной реализации и функциональные возможности системы для тематического анализа текста.

Ключевые слова: *текстовая информация, тематический анализ, частотно-контекстная классификация, ключевые слова, мультиагентная система, тематическая принадлежность.*

Введение

В настоящее время значительно увеличилось число прикладных задач, связанных с Интернет, а именно, поисковые системы, порталы, блоги, каналы новостей, википедии, карты и навигация и многое другое (рис.1), что позволяет рассматривать Интернет в качестве источника самых разнообразных знаний. Соответственно появляются различные идеи автоматизированной обработки и извлечения знаний из текстовой неструктурированной информации. Примерами важных и актуальных задач поиска и интеграции информации в Интернет являются следующие: поиск и автоматическая доставка тематических ресурсов; тематическое рубрицирование, реферирование документов, кластерный анализ ресурсов и др.

Одной из основных проблем, возникающих при работе в Интернет в свете постоянно увеличивающихся объемов информации, является поиск документов по их содержанию.

Ставшие традиционными средства контекстного поиска по вхождению слов в документ, представленные привычными поисковыми машинами, зачастую не обеспечивают адекватного выбора информации по запросу пользователя.

Возможность классифицировать найденный материал по тематическим группам у поисковых систем либо отсутствует вообще, либо представлена в крайне примитивном виде.

Рис. 1. Интернет, как источник знаний

Это стимулировало развитие средств, которые получили название тематические навигаторы. Они дают возможность пере движения по связанным тематическим категориям (рубрикам), к каждой из которых может относиться большое число документов, близких по содержанию. Такие системы основываются на различных технологиях автоматического анализа содержания текста.

При этом анализ web-ресурсов связан с решением ряда проблем: зашумленность (сайты нередко содержат значительное число страниц, не относящихся к теме); политематичность. Это приводит к тому, что чаще приходится отбирать не целые тексты, а релевантные теме фрагменты текстов. Содержание многих текстов является переплетением ряда тем. Возникает проблема поиска внутри текстовых документов фрагментов, релевантных заданной теме.

Кроме того, сложность решения этих задач обусловлена необходимостью учета закономерностей совместного употребления слов в документах. Действительно, смысл слова зависит, не от его написания, а от его употребления, т.е. определяется совокупностью всех тех его комбинаций с другими словами, в которых оно встречается в языке.

Целью данной статьи является описание мультиагентной системы для тематического анализа и вычисление степени тематической принадлежности текстового контента web-ресурсов.

Анализ последних исследований и публикаций

Одним из способов, обеспечивающих снижение размерности решаемых задач анализа это переход от основного текста к его представлению в виде множества ключевых слов, приближенно описывающих его содержание. Для повышения точности и адекватности описания содержания документа ключевые слова используются с некоторыми весовыми коэффициентами, которые соотносятся с частотой повторений этих слов в тексте. Это необходимо, прежде всего, для последующей тематической идентификации сравниваемых текстов. Задача классификации в данном случае сводится к задаче отнесения текста к некоторому тематическому классу, описываемому множеством ключевых слов. При этом тематические классы не определены заранее, их формирование, а также идентификация и отнесение текста к той или иной предметной области происходит в процессе анализа текста.

Классификация текстового контента web-ресурсов (страниц и сайтов) может осуществляться различными способами. Всю совокупность представленных на сегодняшний день методов тематического анализа текста можно разделить на две группы: лингвистический (ориентирован на извлечение смысла текста по его семантической структуре и представлен, в свою очередь, лексическим, морфологическим, синтаксическим и семантическим анализом, а, следовательно, зависим от языка и предметной области) и статистический (использование частотного распределения слов в тексте) анализ. Чаще всего встречаются реализации именно статистического анализа в различных вариациях, однако, при этом качество его работы обычно повышается за счет использования методов определения значимости найденных документов, учета морфологии языка или семантических словарей. При этом наиболее часто используется определение TF-IDF, как статистической меры, используемой для оценки важности слова в контексте документа, являющегося частью коллекции документов. Вес некоторого слова в этом случае пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

В работе [1] рассматривается нейросетевой подход к обработке текстовой информации на основе специфического статистического анализа, реализованный в среде TextAnalyst, которая позволяет автоматически сформировать смысловой портрет текста в виде ассоциативной сети основных понятий с их связями, помеченными их числовыми характеристиками.

В работе [2] описан метод автоматического конспектирования естественно-языковых текстов (реализован в системе KONSPEKT), позволяющий формировать сжатые образы исходных текстов и основанный на синтактико-семантическом анализе с выделением полносоставных предложений из исходного текста и последующем тематическом анализе, который проводится на основе онтологии ассоциаций с возможностью генерации дополнительных ключевых слов, детализирующих тематику текста.

Оригинальный метод извлечения ключевых терминов из текстовых документов приводится в [3]. Он основан на определении меры семантической близости терминов с использованием Википедии и алгоритме Гирвана-Ньюмана для обнаружения сообществ в сетях, а результатом его работы является множество ключевых терминов, сгруппированы по темам анализируемого документа.

Способ тематического анализа текстовой информации, основанный на статистическом подходе и учитывающий специфику web-ресурсов, рассматривается в [4], при этом здесь, для каждого ключевого слова ставится в соответствие коэффициент значимости, при вычислении которого учитывается

частота вхождения слова в текст, часть речи, положение слова относительно документа, html-тег, в который входит слово.

Отдельно следует отметить работу [5], где для тематического анализа неструктурированной текстовой информации и эффективного решения задачи поиска документов по образцу используется метод частотно-контекстной классификации тематики текста, который отличается дополнением частотно значимых слов контекстно-связанными с ними словами, что позволяет более точно отобразить тематику текста.

Проведенный анализ показал, что на сегодняшний день остаются актуальными вопросы, как теоретической проработки, так и практики эффективного решения задач тематического анализа неструктурированной текстовой информации произвольного содержания.

На сегодняшний день наибольший интерес и перспективы в решении задач поиска и интеграции информации в Интернет заключаются в эффективном совместном использовании дополняющих друг-друга технологий: мультиагентных систем, онтологических баз знаний и семантического веб. Мультиагентные системы, которые строятся из множества взаимодействующих интеллектуальных агентов, совместно решающих поставленную задачу в распределенных средах эффективно используются при поиске информации как в web-ресурсах, так и в хранилищах полнотекстовых документов.

Однако при этом требуются средства для обеспечения семантической обработки информации, в роли которых выступают онтологии, представляющие собой базы знаний специального вида, содержащие семантическую информацию об описываемой предметной области. В процессе работы интеллектуальные поисковые агенты могут использовать онтологии предметных областей и производить дополнительный анализ текстовой информации web-контента. Основная роль Semantic Web, в основу которого также положен онтологический подход – как раз сделать размещенную в сети Интернет информацию доступной и понимаемой не только людьми, но и агентами, расширить существующую разметку html, предназначенную для визуализации, семантической разметкой – т.е. дать аннотацию каждому ресурсу в сети.

Существует целый ряд исследований, посвященных созданию мультиагентных систем, выполняющих поиск информации как в web-документах, так и в хранилищах полнотекстовых документов. Так в работе [6] предложена мультиагентная система поиска информации, но при этом рассматривается только простой подсчет числа вхождений концептов онтологии в текст web-документа.

В работе [7] рассматриваются только вопросы содержательного поиска и извлечения фактов из документов на основе онтологии предметной области. Эффективному использованию онтологий в решении задач поиска и анализа информации препятствует ряд проблем: трудоемкость разработки онтологий и исключительная сложность в автоматизации этого процесса; аккуратно установленные связи в онтологии с языковыми единицами — терминами предметной области и др.

Таким образом, дальнейшее совершенствование методов и инструментальных средств выделения тематики и вычисления тематической близости документов на основе технологии мультиагентных систем и онтологического подхода представляется весьма перспективным направлением исследований.

Мультиагентная система тематического анализа текстовой информации

Структура предлагаемой мультиагентной системы тематического анализа web-контента представлена на рис. 2. Рассмотрим назначение и особенности взаимодействия агентов.

Агент TopicsManagerAgent, обладая первичным списком ссылок для обхода, осуществляет создание поисковых агентов SearchAgent, назначая им при этом задачу тематической классификации текстового контента web-ресурсов.

Для тематической классификации текстовой информации мы использовали подход, предложенный в работе [5], который предполагает выделение множества ключевых слов, определяющих тематику текста. При этом каждому из них приписывается вес, определяющий значимость данного слова в тематике, т.е. какие-то ключевые слова играют большую роль в определении тематики, какие-то меньшую, но именно такая совокупность слов и определяет тематическую направленность.

Такой подход обеспечивает снижение размерности решаемой задачи за счет перехода от основного текста к его представлению в виде множества ключевых слов (определяется на основе пороговой величины), приближенно описывающих его содержание. Это необходимо, прежде всего, для последующей тематической идентификации сравниваемых текстов.

Отличительной особенностью изложенного в [5] подхода является предположение, что корректное и адекватное машинное представление тематики текста должно включать в себя не только ключевые слова, но и контекст этих слов, т.к. смысл любого слова определяется исключительно в контексте тех слов, которые употреблялись вместе с ним, близко, рядом по тексту.

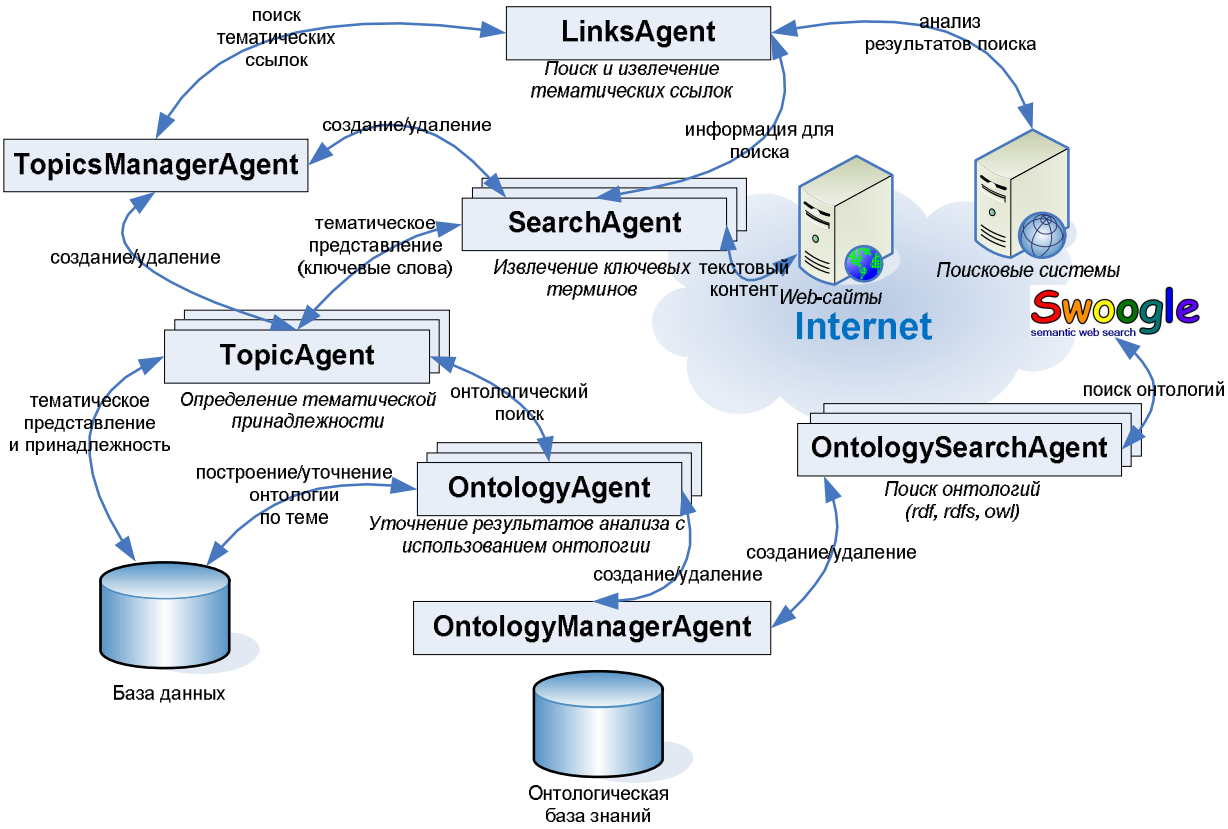

Рис. 2. Мультиагентная система поиска и анализа текстового контента

И сами по себе ключевые слова в отрыве от их контекста не отражают в полной мере тематическую направленность текста. Уточняющие ключевые слова (контекст) определяются путем анализа окрестности каждого ключевого слова с учетом ее размера.

Кроме поиска ключевых слов агент SearchAgent осуществляет передачу информации с анализируемых страниц сайтов агенту LinksAgent, основной задачей которого является извлечение тематических ссылок для расширения дальнейшего списка для поиска и обхода страниц. Критерии важности, используемые этим агентом, собирающим информацию для дальнейшего поиска ресурсов и их тематического анализа, определяются на основе следующих элементов: глубина URL; URL, на которые больше ссылок из других, наиболее часто цитируемых страниц в Internet. На основании полученных списков новых ссылок агент TopicsManagerAgent принимает решение о том создать новый поисковый агент или отдать ссылки для анализа какому-либо из имеющихся агентов SearchAgent.

Сформированное первичное и уточняющее множество ключевых слов (тематическое представление) используется для определения тематической принадлежности (классификации) в агенте TopicAgent. Задача классификации в данном случае сводится к задаче отнесения текста к некоторому тематическому классу, описываемому множеством ключевых слов. При этом тематические классы не определены заранее – их формирование, а также идентификация и отнесение текста к тому или иному классу происходит в процессе анализа текста. На этом этапе агенте TopicAgent взаимодействует с агентом OntologyAgent для уточнения результатов тематического анализа с использованием онтологии (подсчет числа вхождений концептов онтологии в тематическую принадлежность, оценка семантических связей, проверка гипотезы относительно слов из контекста и др.). В случае отсутствия онтологии по данной предметной области в онтологической базе осуществляется запрос к агенту OntologyManagerAgent, который инициирует поиск релевантных онтологий в сети путем создания одного или нескольких агентов OntologySearchAgent. Основной поисковой системой, с которой взаимодействует OntologySearchAgent в данном случае выступает Swoogle (http://swoogle.umbc.edu), в которой на сегодняшний день проиндексировано более 10 тыс. различных онтологий в форматах rdf, rdfs и owl. Возможность повышения качества тематического анализа тесно связана с учетом коррелированности появления слов в тексте, обусловленной наличием между ними семантических связей, которые отражены в онтологиях.

Полученная таким образом информация – тематическое представление (создается новая тема или уточняется уже имеющаяся) и принадлежность заносятся в базу данных. При этом инициируется диалоговое взаимодействие с пользователем, который решает три основных задачи: прямое указание какие ключевые слова будут занесены в данную тему, а какие нет; расширение базы стоп-слов; указание темы, к которой относится данный текст.

Очевидно, что в таком подходе, когда происходит накопление знаний, степень автоматизации (возможность решения этих вопросов системой в автоматическом режиме) и качество тематического анализа будут повышаться с каждым новым обработанным web-ресурсом. При этом соответствующий поисковый агент SearchAgent закрепляется за данной тематикой и в дальнейшем учитывает уже имеющуюся информацию о тематической релевантности уже обнаруженных страниц и выявленных ссылок для определения дальнейшего порядка обхода.

Особенности программной реализации

На данном этапе мультиагентная система поиска и анализа текстовой информации находится в стадии опытной эксплуатации и доработки.

Остановимся на достигнутых результатах. На данном этапе разработано web-приложение системы на основе технологии ASP.NET для тематического анализа и вычисление степени тематической принадлежности текста.

Для работы с онтологиями используется dotNetRDF API.

Рассмотрим структуру системы, указав некоторые основные классы для работы с текстовой информацией:

- DocumentConverter - класс для загрузки документов различных форматов и преобразования их в текст;

- IElement - класс, отвечающий за представление и хранение информационных элементов (слов);

- TextConverter - класс, преобразующий текст в набор связанных информационных элементов (в соответствии с окрестностью каждого информационного элемента);

- HtmlMetaParser – класс для парсинга html-документа и работы с его мета-тэгами;

- Stemming - класс, отвечающий за морфологический анализ (алгоритм Портера);

- WordCounter - класс, отвечающий за базовую обработку текста;

- WordOccurence - класс, отвечающий за частотную классификацию;

- ContextAnalys - класс, отвечающий за контекстный анализ;

- TopicNearness - класс, отвечающий за нахождение тематической близости

и другие классы, в том числе для работы с онтологиями.

Фрагмент полученной в результате проектирования и разработки в СУБД MS SQL Server структуры базы данных представлен на рис. 3.

Таблица «Analysis» представляет собой журнал проведенных тематических анализов (история). Таблица «Categories» предназначена для хранения информации о категориях предметных областей или тем. В «Subjects» хранится информация о темах или предметных областях для той или иной категории. Таблица «Texts» предназначена для хранения информации об анализируемых текстах. Таблица «Stopwords» предназначена для хранения информации о стоп-словах. В «Text_keywords» хранится информация о ключевых словах тега html-страницы keywords. Таблица «Analys_keys» предназначена для хранения информации о первичных ключевых словах и ключевых словах из контекста (уточняющих) найденных при анализе текста. В «Subject_keys» хранится информация о ключевых словах для тематики/предметной области. Именно в эту таблицу мы заносим ключевые слова найденные в результате тематического анализа. Это означает, что данная таблица будет расширяться в ходе анализа текстов с той или иной тематикой – будут добавляться только новые слова найденные в результате анализа. Если слово в таблице уже имеется, то происходит модификация поля вес. В конечном счете, это позволит определить действительное ядро ключевых слов для той или иной тематики – те слова, которые наиболее часто встречаются в релевантных текстах.

1. Subjects (Хранение информации о предметных областях) состоит из 3 полей:

- Subject_id - идентификатор предметной области;
- Category_id - идентификатор категории;
- Subject_name - название предметной области.

2. Categories (Категории предметных областей) состоит из 2 полей:

- Category_id - идентификатор категории;
- Category_name - название категории.

3. Texts (Хранение информации об анализируемых текстах) состоит из 8 полей:

- Text_id – идентификатор текста;
- Text_url – ссылка url на html-страницу;
- Text – текст (если небольшой);
- Text_filename – ссылка на файл;
- Text_caption – заглавие текста;
- Text_description – содержимое тега description html-страницы;

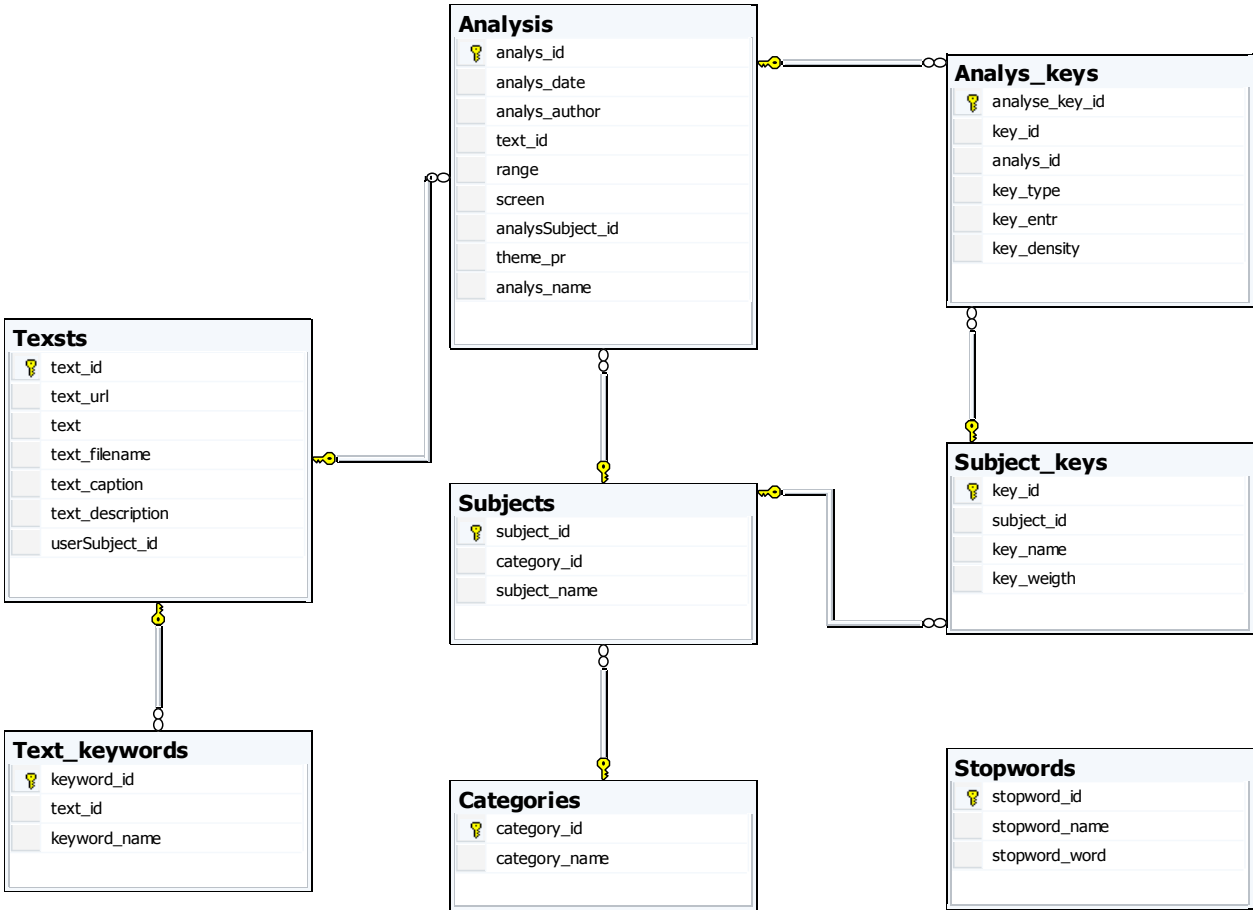

Рис. 3. Фрагмент структуры базы данных

- AnalysSubject_id – тема текста, установленная в результате анализа;

- UserSubject_id - тема текста, заданная пользователем.

4. Text_keywords (Хранение информации о ключевых словах тега html – страницы keywords) состоит из 3 полей:

- Keyword_id - идентификатор ключевого слова;

- Text_id – идентификатор текста, к которому относится ключевое слово;

- Keyword_name - ключевое слово.

5. Analysis (Журнал проведенных анализов) состоит из 8 полей:

- Analys_id - идентификатор анализа;

- Analys_date - дата/время проведения анализа;

- Analys_author - кто проводил анализ;

- Text_id - идентификатор текста;

- Range - процент от максимальной частоты, определяющий порог;

- Screen - окрестность для контекста;

- AnalysSubject_id - тема текста, установленная в результате анализа;

- Theme_pr - тематическая близость.

6. Keys (Хранение информации о первичных словах и ключевых словах из контекста при анализе текста) состоит из 10 полей:

- Key_id – идентификатор ключевого слова;

- Subject_id - идентификатор темы;

- Analys_id – идентификатор анализа;

- Key_name – первичное ключевое слово;

- Key_word – часть речи;

- Key_main – базовая форма;

- Key_entr – частота;

- Key_density – плотность;

- Key_weight – вес;

- Key_type тип ключевых слов (0 – первичные, 1 – из контекста).

7. Key_formsof (Хранение информации о словоформах первичных ключевых слов и ключевых сов из контекста) состоит из 5 полей:

- Keyformsof_id - идентификатор словоформы;

- Key_id - идентификатор ключевого слова;

- Keyformsof_type - тип словоформы (префиксы, окончания);

- Key_formsof – словоформа;

- Key_formsof_name - ключевое слово со словоформой.

8. Stopwords (Хранение информации о стоп-словах) состоит из 3 полей:

- Stopword_id - идентификатор стоп-слова;

- Stopword_name - стоп-слово;

- Stopword_word - часть речи.

Обобщенный алгоритм функционирования web-приложения представлен на рис. 4.

Перед тем, как анализировать тексты в соответствии с используемым в работе подходом необходимо выполнить их предварительную обработку.

Предварительная обработка текстов предполагает лексический и морфологический анализ, а также исключение часто используемых слов – «стоп-слов» (союзов, местоимений и т.д.).

Рассмотрим подробнее этапы предварительной обработки текста:

1) *Лексический анализ.* Заключается в разборе текста на отдельные абзацы, предложения, слова. На этом этапе выделяются отдельные слова из текста.

2) *Исключение часто используемых слов.* В любом тексте существует большое количество слов, используемых в качестве союзов, предлогов, местоимений и т.д., так называемые "стоп-слова", ("stop-words"). Как правило, эти слова не определяют тематику текста, но при этом являются частотно-значимыми, поэтому такие слова необходимо исключать из текста. Данная процедура выполняется на основе предварительно составленной базы наиболее часто встречающихся слов в документах.

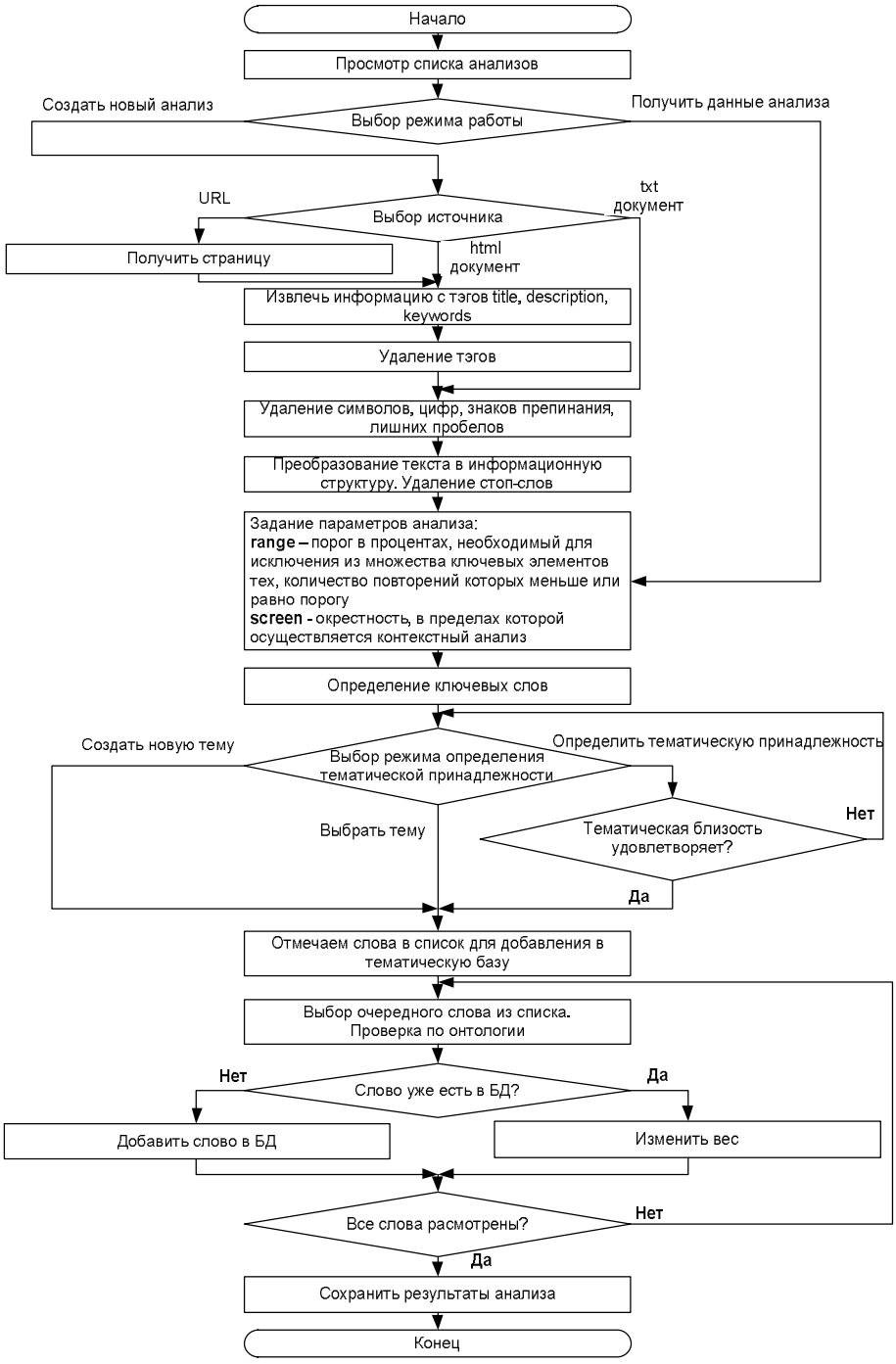

Рис. 4. Обобщенный алгоритм функционирования системы

3) *Морфологический анализ.* Морфологический анализ сводится к автоматическому распознаванию частей речи каждого слова текста. На этом этапе слова приводятся к базовой форме. Для выполнения этой части предварительной обработки текста будем использовать алгоритм Портера, получивший большое распространение и называющийся стемминг.

После предварительной подготовки текста, выполняется его обработка и тематический анализ по используемому методу тематической классификации [5].

На данной стадии развития проекта у пользователя есть возможность выбора нескольких режимов работы с системой: просмотр истории проведенных анализов; создание/редактирование категорий и тем; тематический анализ текста; анализ принадлежности; настройки; регистрация/редактирование пользователей.

На рис. 5 представлена экранная форма создания/обновления предметных областей. Здесь есть возможность просмотра имеющихся предметных областей (тем) в базе по категориям. При этом для выбранной темы в списке отображается ее тематическое представление в виде набора ключевых слов с указанием их веса.

На рис. 6 представлена экранная форма одного из основных режимов работы с системой – «Тематический анализ». Как видно из рисунка нажатие кнопки «Провести анализ» приводит к формированию первичного множества ключевых слов (без учета контекста) и уточненного множества (включая контекст) до и после морфологического анализа. После получения списка ключевых слов (до морфологического анализа) система предоставляет возможность выбрать из них те, которые являются стоп-словами и занести в базу. Таким образом, база стоп-слов системы будет пополняться, что будет повышать качество проведения тематического анализа.

Дальнейшие действия пользователя могут быть связаны с двумя направлениями (рис. 7).

Первое – задать тему – используется в том случае, когда заранее известно к какой теме принадлежит текст или в том случае, когда база знаний системы еще находится в начальной стадии формирования. Если такой темы еще нет в базе, можно ее добавить.

Перед сохранением ключевых слов в базу система предоставляет возможность выбрать, какие слова необходимо занести в базу, а какие нет. По умолчанию все найденные слова будут сохранены.

Второе направление – определение тематической принадлежности – используется тогда, когда в базе знаний системы есть информация о тематических представлениях по предметным областям (минимум одной).

Система выводит вычисленные значения тематической близости на диаграмме (рис. 8). Кроме того, рекомендует предметную область для сохранения в базе.

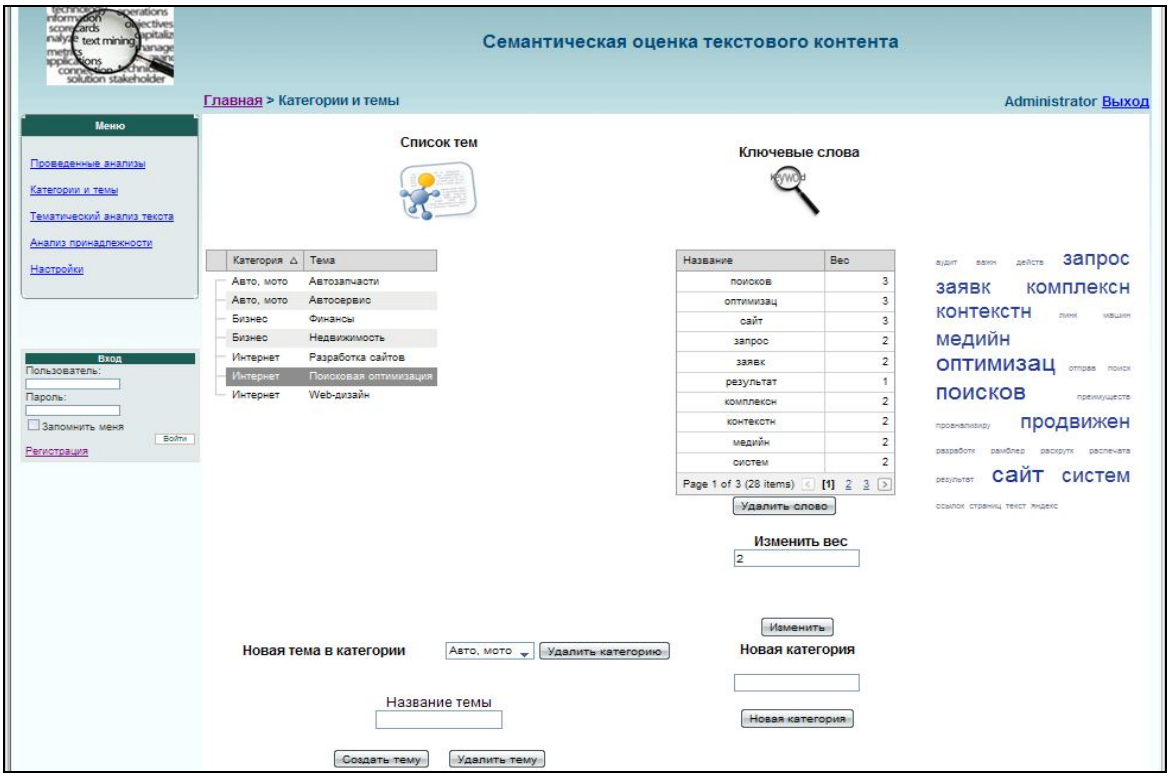

Рис. 5. Создание/обновление предметной области

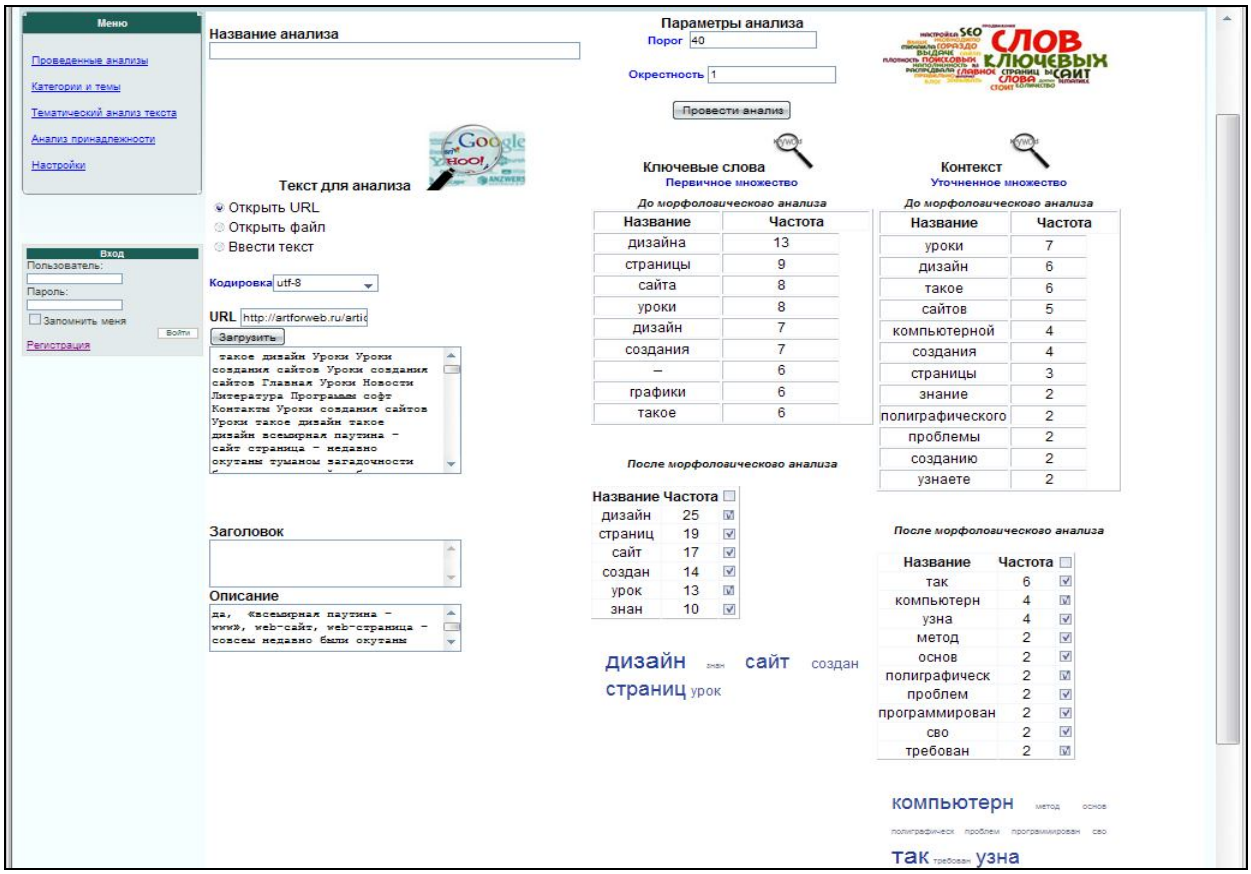

Рис. 6. Результат проведения тематического анализа

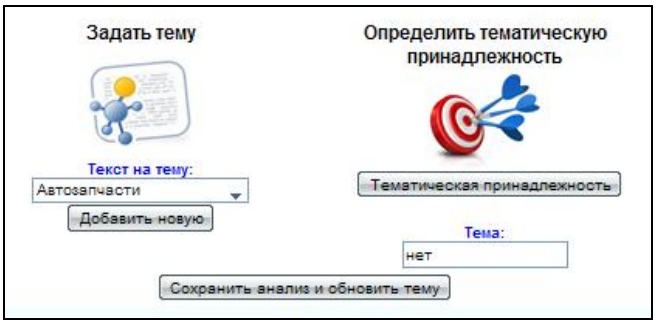

Рис. 7. Режимы работы с результатами тематического анализа

Заключение

В результате работы была спроектирована и программно реализована мультиагентная система для тематического анализа и вычисление степени тематической принадлежности текстового контента web-ресурсов.

Использованный метод частотно-контекстной классификации тематики текста, с возможностью дополнения ключевых слов контекстом, а также уточнение результатов тематического анализа с помощью онтологий позволили получить значительные преимущества. Экспериментальные оценки используемого метода показали, что он извлекает ключевые термины из документов с высокой точностью и полнотой.

В дальнейшем планируется развивать систему и реализовать следующие возможности:

- реализация автоматического (фонового) режима обхода сайтов с целью определения ключевых слов и тематики;

- подключение словарей (содержащих знания о синонимах и степени смысловой близости) для улучшения полноты и точности тематической классификации и для повышения качества проведения морфологического анализа;

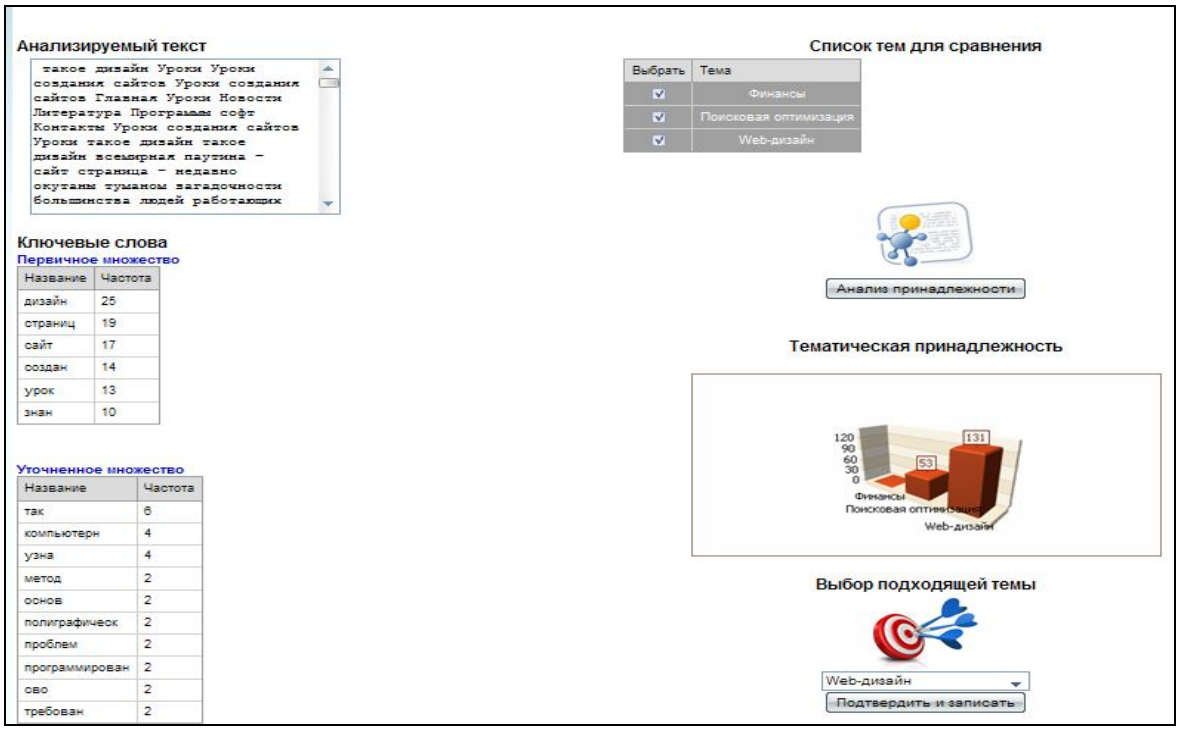

Рис. 8. Результаты анализа принадлежности текста

– разработка адаптивной методики формирования веса ключевого слова для тематики (для слов, часто встречающихся в одном окружении (контексте) следует предусмотреть использование метрики близости);

– взаимодействие через API с поисковыми системами и другими сайтами, которые предоставляют каталогизированную информацию для расширения базы категорий и тем, а также для получения базовых ссылок на сайты по той или иной тематике;

– при анализе web-страниц сейчас учитывается только частота встречаемости слова в тексте, хотя следовало бы предусмотреть возможность использовать и другие признаки, например, такие как «слово встретилось в заголовке», «слово выделено другим цветом, шрифтом» и т.д., поскольку это может повлиять на вес данного слова в рамках документа.

Практическая значимость разработки связана с тем, что в настоящее время имеется огромный комплекс возможных прикладных задач, связанных с автоматизированным извлечением и обработкой знаний из текстовой информации, число таких задач и систем постоянно растет, возникает необходимость решения целого спектра проблем, связанных с повышением эффективности анализа текста. Эти вопросы лежат на стыке компьютерной лингвистики и технологий создания распределенных интеллектуальных систем, которые уже сегодня входят в повседневную жизнь, используя Интернет в качестве источника знаний.

Литература

1. Харламов, А.А. Автоматический структурный анализ текстов [Электронный ресурс] / А.А. Харламов // Открытые системы. – 2002 – № 10. – Режим доступа: http://www.osp.ru/os/2002/10/182010/. – 20.12.2011 г.

2. Гладун, В.П. Тематический анализ естественно языковых текстов [Текст] / В.П. Гладун, В.Ю. Величко, Л.А. Святогор // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». – М.: Изд-во РГГУ.– 2006. – С. 115-118.

3. Гринева, М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов [Текст] / М. Гринева, М. Гринев // Труды Института системного программирования РАН. – 2009. – Т. 16. – С. 155-165.

4. Никитина, Л.А. Об одном способе выполнения тематического поиска информации в Интернет [Текст] / Л.А. Никитина, О.В. Касилов, А.А. Никитин //Вісник міжнародного слов'янського університету. – 2008. – Т. XI, № 1. – С. 25 – 28.

5. Чугреев, В.Л. Анализ текста, применительно к решению задач поиска документов по образцу [Тескт] / В.Л. Чугреев, С.А. Яковлев // Информатизация процессов формирования открытых систем на основе САПР, АСНИ, СУБД и систем искусственного интеллекта (ИНФОС - 2003): Материалы 2-й Межд. науч.-техн. конф. – Вологда: ВоГТУ, 2003. – С. 49 – 52.

6. Рувинская, В.М. Мультиагентные системы для поиска информации в Интернете с использованием онтологий [Текст] / В.М. Рувинская, К.Л. Ма-

нукян // Штучний інтелект. – 2002. – №4. – С. 606 – 613.

7. Загорулько, Ю.А.Семантический подход к анализу документов на основе онтологии предметной области [Текст] / Ю.А. Загорулько, И.С. Кононенко, Е.А. Сидорова // Труды международной конференции Диалог'2006 «Компьютерная лингвистика и интеллектуальные технологии». – М.: Изд. РГГУ, 2006. – С. 468 – 473.

Поступила в редакцию 25.01.2012

Рецензент: д-р техн. наук, проф., зав. каф. информационных технологий проектирования летательных аппаратов Е.А. Дружинин, Национальный аэрокосмический университет им. Н.Е.Жуковского «ХАИ», Харьков.

МУЛЬТИАГЕНТНІ СИСТЕМИ ПОШУКУ ТА ТЕМАТИЧНІ АНАЛІЗУ ТЕКСТОВОЇ ІНФОРМАЦІЇ В ІНТЕРНЕТ

О.В. Прохоров, К.О. Борвенко

Розглянуто питання тематичного аналізу неструктурованої текстової інформації, джерелами якої виступають численні ресурси мережі Інтернет. Розглянуто особливості та методи тематичного аналізу текстової інформації. Представлений короткий аналіз останніх досліджень і публікацій. Запропоновано структуру мультиагентної системи пошуку та тематичного аналізу текстового контенту web-ресурсів. Представлені функціональні можливості програми для тематичного аналізу текст. Описано особливості програмної реалізації та функціональні можливості системи для тематичного аналізу тексту.

Ключові слова: текстова інформація, тематичний аналіз, частотно-контекстна класифікаційція, ключові слова, мультиагентних система, тематична приналежність.

MULTIAGENT SYSTEM OF SEARCH AND ANALYSIS OF TEXTUAL INFORMATION CASE TO THE INTERNET

A.V. Prokhorov, E.A. Borvenko

Questions of a thematic analysis of unstructured textual information sources which are the numerous resources on the Internet. The characteristics and methods of thematic analysis of textual information. A brief review of recent research and publications. The structure of multi-agent system and search for thematic analysis of textual content of web-resources. Submitted the application's functionality for thematic analysis of text. The features of software implementation and system functionality for thematic analysis of text.

Keywords: text information, thematic analysis, frequency-contextual classification function, keywords, mutiagent system, a thematic identity.

Прохоров Александр Валерьевич – канд. техн. наук, доцент, доцент кафедры информационных управляющих систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

Борвенко Екатерина Александровна – магистрант кафедры информационных управляющих систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.