СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА, КЛАССИФИКАЦИИ И РЕФЕРИРОВАНИЯ ДОКУМЕНТОВ ДЛЯ ИНТЕРНЕТ-ПОРТАЛА

Вячеслав Ланин

***Abstract**: В статье представлено описание предполагаемых подходов к реализации подсистемы обработки информации на Интернет-портале. Основные проблемы связаны с экспоненциальным ростом числа документов, отсутствием семантического индексирования и неструктурированным характером информации. При реализации предлагаемого подхода пользователь получает эффективные интеллектуальные средства поиска электронных документов на основе семантической индексации, автоматической классификации и каталогизации документов с построением семантических связей между ними и автоматического реферирования документов с использованием знаний. Эффективность работы с электронными документами предлагается значительно увеличить за счет их интеллектуального анализа, для которого применяются агентный и онтологический подходы. В соответствии с предлагаемым подходом онтология используется для описания семантики данных документа и его структуры. В процессе анализа документа онтология является центральным понятием – благодаря использованию онтологий из документа можно получить требуемые данные: известно, где искать данные и как они могут быть интерпретированы. Репозитарий онтологий содержит три уровня онтологий: на первом уровне расположены онтологии, описывающие объекты, используемые в конкретной системе и учитывающие ее особенности; на втором в терминах объектов первого уровня описываются объекты, инвариантные к предметной области; объекты третьего уровня описывают наиболее общие понятия и аксиомы, с помощью которых описываются объекты нижележащих уровней. Третий и второй уровни можно разделить на две составляющие: описание структур и описание самих документов.*

***Keywords**: онтология, агент, мультиагентные системы, интеллектуальный поиск, семантическое индексирование, анализ документов, адаптируемые информационные системы, CASE-технология.*

***ACM Classification Keywords**: H.2 Database Management: H.2.3 Languages – Report writers; H.3.3 Information Search and Retrieval – Query formulation.*

***Conference**: The paper is selected from XVth International Conference "Knowledge-Dialogue-Solution" KDS 2009, Varna, Bulgaria, June-July 2009*

Введение

Экспоненциальный рост количества электронных документов, наблюдающийся в настоящее время, наглядно показывает, что традиционные механизмы обработки электронных документов не справляются с потребностями пользователя. Эта тенденция заметна как в сети Интернет, так и в корпоративных сетях. В настоящее время все большую популярность приобретают так называемые информационные порталы (тематические и корпоративные), основная цель которых консолидация информации и знаний.

Одним из таких решений является исследовательский портал – информационно-аналитическая система сбора и аналитической обработки данных об инновационной активности регионов для поддержки принятия эффективных управленческих решений («Исследовательский портал "Инновационное развитие регионов"»). Данные для анализа извлекаются из гетерогенных неструктурированных или слабоструктурированных источников данных, в частности, Интернет-ресурсов, а также оперативных баз

данных. По замыслу система должна обеспечивать интеграцию, согласование, агрегацию и сопровождение ранее разъединенных данных, поддерживать различные формы визуализации данных и результатов анализа, настраиваемые в соответствии с потребностями пользователей. Из этого следует, что поиск и обработка неструктурированных текстовых данных, получаемых из различных источников в разных форматах, становится одной из основных функций разрабатываемой системы.

Таким образом, актуальность задачи вызвана следующими причинами:

- экспоненциальный рост числа документов, делающий невозможной обработку данных традиционными методами без потери качества;
- отсутствие семантического индексирования, что не позволяет приводить интеллектуальную обработку документов в полном объеме;
- неструктурированный характер информации, не позволяющий применить традиционные механизмы ее обработки и анализа.

Рассмотрим перечисленные проблемы более подробно.

Экспоненциальный рост объема информации, содержащейся в Интернете, является причиной все более и более возрастающей трудности поиска необходимых документов (рис. 1) и организации их в виде структурированных по смыслу хранилищ [6]. Пользователю становится все труднее найти необходимую информацию, традиционные механизмы поиска оказываются малоэффективными.

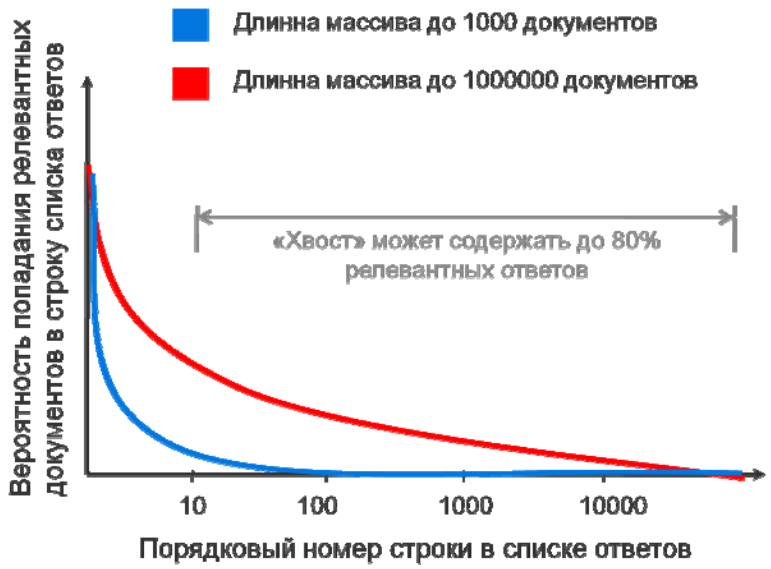

Рис. 1. Проблема поиска информации при росте числа документов

Большинство технологий работы с документами ориентированы на организацию удобной работы с информацией для человека. Но зачастую методы работы с электронной информацией просто копировали методы работы с «бумажной» информацией. В текстовом редакторе присутствуют широкие возможности форматирования текста (представления в удобном для человека виде), но практически отсутствуют возможности для передачи смыслового содержания текста, т.е. *отсутствует семантическое индексирование*. Для эффективного решения задачи поиска необходимо расширить понятие традиционного документа: *с документом необходимо связать знания, позволяющие интерпретировать и обрабатывать хранящиеся в этом документе данные*.

Неструктурированная информация составляет значительную часть современных электронных документов (рис. 2). Системы класса Data Mining работают со структурированными данными, а для

работой с неструктурированным контентом используются системы Text Mining. Фактически они решают одну и ту же задачу для разных типов данных, поэтому предполагается, что эти системы сойдутся в «одной точке».

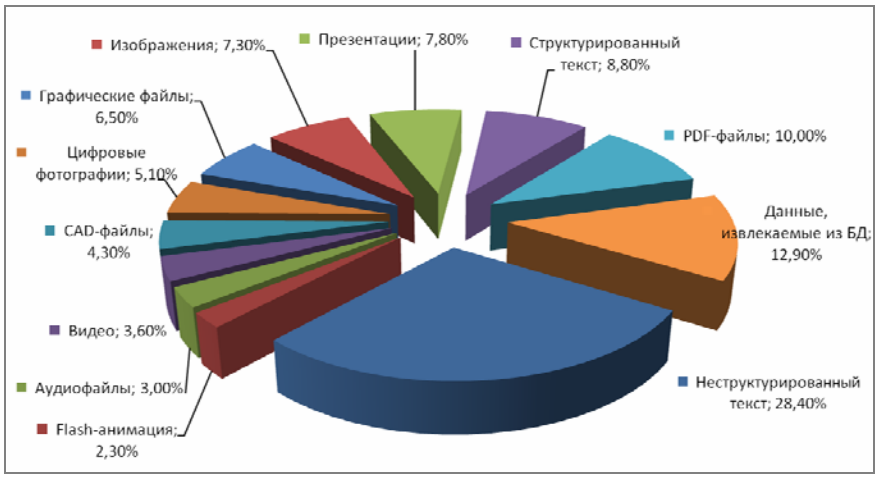

Рис. 2. Распределение категорий документов

Text Mining – это алгоритмическое выявление прежде неизвестных связей и корреляций в уже имеющихся текстовых данных [5]. Важная задача технологии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться как метаданные документа, ключевых слов, аннотаций. Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов. Возможности современных систем Text Mining могут применяться при управлении знаниями для выявления шаблонов в тексте, для автоматического «выталкивания» или размещения информации по интересующим пользователей профилям, создавать обзоры документов.

Реализовать интеллектуальные возможности портала при работе с электронными документами планируется за счет реализации средств и подходов Text Mining.

Подход к семантическому индексированию

На результативность процесса поиска необходимых документов оказывает большое влияние и человеческий фактор: зачастую пользователь не готов к долгому ожиданию результатов поиска, к просмотру и анализу большого объема результирующей выборки. Кроме того, большинство пользователей неэффективно используют поисковое программное обеспечение и, как правило, они игнорируют расширенные поисковые возможности и ограничиваются короткими типовыми запросами.

Для повышения эффективности обработки электронных документов требуется наличия метаданных, описывающих структуру и семантику документов. Одним из возможных подходов к описанию информации, заложенной в документе, является подход на основе онтологий. Под онтологией понимается база знаний специального типа, которая может «читаться» и пониматься, отчуждаться от разработчика и/или физически разделяться ее пользователями [4].

В качестве подхода к семантическому индексированию был выбран онтологический подход [1], в котором онтология может описывать как структуру, так и содержание документа, т.е. *онтология используется для описания семантики данных документа и его структуры*. Учитывая специфику решаемых в данной работе задач, конкретизируем понятие онтологии: будем считать, что *онтология – это спецификация некоторой предметной области*, которая включает в себя словарь терминов (понятий) предметной

области и множество связей между ними, которые описывают, как эти термины соотносятся между собой в конкретной предметной области. Фактически в данном контексте *онтология – это иерархическая понятийная основа рассматриваемой предметной области*.

Онтология документа используется для анализа документов, благодаря ей из документа можно получить требуемые данные: известно, где искать данные и как они могут быть интерпретированы. Если представлять документ с использованием онтологий, то задача сопоставления онтологии и имеющегося документа сводится к задаче поиска понятий онтологии в документе. Как следствие, системе необходимо ответить на вопрос: описывает ли данная онтология документ или нет. На последний вопрос можно ответить утвердительно, если в процессе сопоставления в документе были найдены все понятия, включенные в онтологию. Таким образом, исходная задача сводится к задаче поиска в тексте документа общих понятий на основе формальных описаний.

Репозитарий онтологий содержит *три уровня онтологий*. На первом уровне расположены онтологии описывающие объекты, используемые в конкретной системе и учитывающие ее особенности. На втором уровне описываются объекты, инвариантные к предметной области. Объекты второго уровня описываются в терминах объектов первого уровня. Это выражается в отношениях наследования и меронимии. Объекты третьего уровня описывают наиболее общие понятия и аксиомы, с помощью которых описываются объекты нижележащих уровней. Третий и второй уровни можно разделить на две составляющие: описание структур и описание самих документов, причем документы описываются в терминах структур.

Для решения проблемы выделения общих понятий на основе формальных описаний предлагается агентный подход [2]. Данный подход будет удовлетворять требованиям, предъявляемым к процессу поиска, если при построении системы будут реализованы все преимущества мультиагентных систем.

При использовании данного подхода для каждой вершины онтологии, содержащей общее понятие, создается агент, который проводит поиск данного конкретного понятия. В данном подходе агент рассматривается как система, направленная на достижение определенной цели, способной к взаимодействию со средой и другими агентами. Для признания агента интеллектуальным необходимым условием является наличие у него базы знаний. Таким образом, чтобы определить агентов, действующих в системе, необходимо выбрать способ для описания базы знаний, характера взаимодействия со средой и сотрудничества.

Базу знаний агента для поиска общих понятий онтологии удобно представлять также в виде онтологии. Для предоставления пользователю возможности добавления новых шаблонов необходимо выделить базовые понятия для формирования общих.

Одним из важнейших свойств агентов является *социальность*, или способность к взаимодействию [2]. Как было сказано ранее, для каждой вершины онтологии, содержащей общее понятие, создается агент. Согласно принятой классификации агентов он является *интенциональным*.

Данный агент нацелен на решение двух задач:

1. Весь имеющийся список шаблонов понятия он разбивает на отдельные компоненты и запускает более простых агентов для поиска полученных компонент.

2. Производит сборку результатов из всех списков, полученных агентами более низкого уровня.

Упомянутые выше агенты более низкого уровня являются *рефлекторными*. Они получают шаблон, и их целью становится отыскание в тексте фраз, подпадающих под этот шаблон. Результаты поиска агентов всех уровней заносятся на «доску объявлений».

На данный момент в других системах инструменты онтологического характера применяются в перечисленных ниже направлениях:

- WordNet в сочетании с векторной и булевой моделями информационного поиска;
- традиционные информационно-поисковые тезаурусы в комбинации с разного рода статистическими моделями;
- тезаурус для автоматического индексирования в булевских моделях поиска документов, в задаче автоматической рубрикации, автоматического аннотирования.

Онтологии станут ядром метаданных портала при работе с электронными документами. Четко очерченная предметная область позволяет создать достаточно детализированные онтологии, которые могут быть использованы всеми его подсистемами.

Автоматическое реферирование

На данный момент для автоматического реферирования применяются два подхода. Традиционный подход (*квазиреферирование*), который используют такие системы, как Microsoft Office, IBM Intelligent Text Miner, Oracle Context, основан на выделении и выборе фрагментов текста из исходного документа и соединении их в короткий текст. Подход, *основанный на знаниях*, предполагает подготовку краткого изложения и передачу основной мысли текста, возможно даже другими словами.

Квазиреферирование основано на выделении характерных фрагментов (как правило, предложений). Для этого методом сопоставления фразовых шаблонов, выделяются блоки наибольшей лексической и статистической релевантности. В большинстве реализаций метода применяется модель линейных весовых коэффициентов. Основу аналитического этапа в этой модели составляет процедура назначения весовых коэффициентов для каждого блока текста в соответствии с такими характеристиками, как расположение этого блока в оригинале, частота появления в тексте, частота использования в ключевых предложениях, а также показатели статистической значимости. Таким образом выделяют при основные направления, часто применяемые в совокупности: статистические методы, позиционные методы и индикаторные методы.

Главное преимущество данной модели заключается в простоте ее реализации. Однако выделение предложений (или параграфов), не учитывающее взаимоотношений между ними, приводит к формированию бессвязных рефератов. Некоторые предложения могут оказаться пропущены, либо в них могут встречаться «висящие» слова или словосочетания.

Для реализации второго метода нужны некие онтологические справочники, отражающие соображения здравого смысла и понятия, ориентированные на предметную область, для принятия решений во время анализа и определения наиболее важной информации.

Метод формирования краткого изложения предполагает два основных подхода.

Первый опирается на традиционный лингвистический метод синтаксического разбора предложений. В этом методе применяется также семантическая информация для аннотирования деревьев разбора. Процедуры сравнения манипулируют непосредственно деревьями с целью удаления и перегруппировки частей, например, путем сокращения ветвей на основании некоторых структурных критериев, таких как скобки или встроенные условные или подчиненные предложения. После такой процедуры дерево разбора существенно упрощается, становясь, по существу, структурной «выжимкой» исходного текста.

Второй подход к составлению краткого изложения уходит корнями в системы искусственного интеллекта и опирается на понимание естественного языка. Синтаксический разбор также входит составной частью в такой метод анализа, но деревья разбора в этом случае не порождаются. Напротив, формируются

концептуальные репрезентативные структуры всей исходной информации, которые аккумулируются в текстовой базе знаний. В качестве структур могут быть использованы формулы логики предикатов или такие представления, как семантическая сеть или набор фреймов.

Функция автоматического реферирования является необходимой для разрабатываемого портала. При поиске пользователю необходимо выдать аннотацию документа, по которой пользователь сможет принять решение о полезности данного документа.

Классификация и каталогизация документов

Задача автоматической классификации и каталогизации документов является задачей разбиения поступающего потока текстов на тематические подпотоки в соответствии заранее заданными рубриками. Автоматическая каталогизация электронных документов, а документов размещенных в сети Интернет в особенности, осложнена ввиду следующих причин [8]:

- большой массив документов;
- отсутствие специальных структур, отслеживающих появление новых документов;
- необязательность авторской классификации электронных документов (в отличие от печатных изданий) посредством их аннотирования, приписывания кодов классификатора и т.п.;
- проблема отслеживания изменений документов.

Как и для автоматического реферирования, существует два противоположных подхода к каталогизации. Наиболее эффективными, но сложными в реализации, являются *методы, основанные на знаниях*. При каталогизации текстов на основе знаний используются заранее сформированные базы знаний, в которых описываются языковые выражения, соответствующие той или иной рубрике, и правила выбора рубрик [5]. Другим классом методов для автоматической рубрикации текстов являются *методы машинного обучения*, которые в качестве обучающих примеров могут использовать заранее отрубрицированные вручную тексты.

При реализации системы автоматической каталогизации на портале необходимо решить две задачи:

- *Создание механизма введения и описания рубрик*, как некоторого выражения на основе слов и терминов документов. Задача может быть решена на основе экспертного описания рубрики или методов машинного обучения по уже отрубрицированным коллекциям документов.
- *Анализ языкового материала, контекста употребления* того или иного слова, требующий привлечения обширных знаний о языке и предметной области.

Заключение

Описанные выше подходы применяются при разработке подсистемы управления электронными документами исследовательского портала. Отличительной особенностью является ориентация на явное представление знаний с помощью онтологий. Данный подход позволит реализовать интеллектуальные сервисы для поиска и обработки электронных документов по тематике портала, получаемых из различных источников.

Таким образом, при создании исследовательского портала будут решены следующие задачи:

- семантическое индексирование документов и интеллектуальный поиск данных, соответствующих запросам пользователей и специфике предметной области;
- извлечение информации из неструктурированных документов;
- интеллектуальная классификация и каталогизация и автоматическое реферирование найденных

документов;

– ведение хронологии электронных документов.

Реализация подсистемы существенно снизит трудоемкость поиска необходимой информации, ее анализа и возможности использования в исследованиях.

Благодарности

Работа выполнена при поддержке грантов РФФИ № 08-07-90006-Бел_а и РГНФ № 09-02-00373В/И.

Библиографический список

[1] Ланин В.В. Интеллектуальное управление документами как основа технологии создания адаптируемых информационных систем // Труды международной научно-технической конференций «Интеллектуальные системы» (AIS'07). Т. 2 / М.: Физматлит, 2007. С. 334-339.

[2] Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал, УРСС, 2002.

[4] Хорошевский В.Ф., Гаврилова Т.А. Базы знаний интеллектуальных систем. СПб.: Питер, 2001.

[5] Ландэ Д. Поиск знаний в Internet. Профессиональная работа. М.: Издательский дом «Вильямс», 2005.

[6] Ефремов В. Search 2.0: огонь по «хвостам» // Открытые системы. СУБД №08 (134), 2007.

[7] Черняк Л. Корпоративный поиск 2.0 // СУБД. – 2007. – №07 (133).

[8] Федотов А.М., Барахнин В.Б. Ресурсы интернета как объект научного исследования [Электронный ресурс]. – 2007. – Режим доступа: http://www.rfbr.ru/pics/28320ref/file.pdf.

[9] Weal M.J., Kim S., Lewis P.H., Millard D.E., Sinclair P.A.S., De Roure D.C., Nigel R. Ontologies as facilitators for repurposing web documents / Shadbolt. Southampton, 2007.

Сведения об авторе

Вячеслав Ланин *– Пермский государственный университет, аспирант кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: lanin@psu.ru*