УДК 004.89:004.912

Д.В. Бабин, С.М. Вороной, Е.В. Малащук
Донецкий государственный институт искусственного интеллекта, Украина
aspirantiai@online.com.ua, smv@iai.donetsk.ua

Повышение эффективности извлечения знаний на основе интеллектуального анализа и структурирования информации

В данной работе исследуется проблема взаимодействия «человек – информация». Рассматриваются основные подходы обеспечения этой взаимосвязи: интеллектуальный анализ текстов и предварительное структурирование информации. Проводится анализ этих подходов и описываются тенденции развития информационного пространства и средств его анализа. Предлагается структура системы извлечения знаний из информационных ресурсов с неструктурированной и структурированной информацией.

Введение

Электронная информация играет все большую роль во всех сферах жизни современного общества. Информационные ресурсы Интернета содержат терабайты информации, средние предприятия работают с базами данных и системами документооборота объемом в десятки гигабайт, сотрудники компаний обрабатывают тысячи электронных писем. Человечество столкнулось с новым для себя явлением – информационной перегрузкой. В этой ситуации сложности возникают уже на этапе поиска информации, хотя, казалось бы, алгоритмы поиска и поисковые интернет-системы существуют и развиваются десятки лет. Но они часто оказываются не способными удовлетворить запросы пользователя. Необходимы новые подходы обеспечения взаимосвязи «человек-информация» при приобретении знаний. Можно выделить два основных подхода: интеллектуальный анализ «сырой» информации и подача информации в структурированном виде.

Интеллектуальный анализ информации

Поскольку информация изначально накапливалась в неструктурированном виде, то возникает необходимость её обработки, структурирования и извлечения из неё знаний. По заверениям Дмитрия Ландэ в 2003 году сырые неструктурированные данные составляли не менее 90 % информации, с которой приходилось иметь дело пользователям. Остальные 10 % в основном касались структурированных данных, загружаемых в реляционные СУБД [1].

Технология эффективного анализа текстов получила название Text Mining. Mining в переводе с английского означает «горное дело, добыча (например, полезных ископаемых)». И действительно, технология направлена именно на добычу самых востребованных на сегодняшний день «ископаемых» – знаний (полезной информации).

Технология глубинного анализа текста Text Mining – это инструментарий, который позволяет анализировать большие объемы информации в поисках тенденций, шаблонов и взаимосвязей, способных помочь в принятии стратегических решений. Некоторые сравнивают эту технологию с выдачей читателю в библиотеке только необходимых книг с уже подчеркнутой необходимой информацией.

Технологии глубинного анализа текста исторически предшествовала технология добычи данных (Data Mining), методология и подходы которой широко используются и в методах Text Mining. Поэтому определение, данное изначально для технологии добычи данных, справедливо и для технологии Text Mining. Data (Text) Mining – это процесс обнаружения в сырых данных ранее неизвестных нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Предварительное структурирование информации

Если процесс извлечения информации из неструктурированных данных так сложен, то возникло разумное желание подавать информацию уже в структурированном виде. Незначительные потери ресурсов (материальных, людских, временных) на этапе подачи информации (размещение в Интернете, занесение в БД) позволяют добиться значительной экономии ресурсов на этапе поиска и выявления новых знаний.

И если технология Text Mining оформилась как направление анализа неструктурированных текстов уже к середине 90-х годов прошлого века, то первоначалом подачи информации в структурированном виде можно считать 1999 год, когда специалисты компании Netscape Communications предложили спецификацию RSS. RSS (RDF Site Summary, Rich Site Summary) – это специальный формат, который проектировался на базе XML и изначально был ориентирован на совместное использование заголовков и иного web-содержания [2]. Данная технология позволяет компьютерам автоматически распознавать и отбирать информацию, нужную пользователям, составлять списки тем и предметов, интересующих конкретного пользователя, и следить за изменением соответствующих ресурсов.

Наличие нескольких вариантов расшифровки аббревиатуры RSS объясняется наличием нескольких версий данного формата, имеющих пусть незначительные, но различия. Кроме того, существуют другие форматы структурирования информации, например формат Atom.

Наличие нескольких форматов структурирования информации, на наш взгляд, является серьезным недостатком данного подхода, поскольку усложняет организацию поисковых систем, но всё же эти системы более просты в реализации и дают лучшие результаты, чем системы Text Mining, ввиду того, что работают с уже структурированной информацией.

Введение формата RSS позволило значительно увеличить долю структурированной информации. Годом активного внедрения RSS стал 2004, и на сегодня все

ведущие порталы новостей подают информацию в этом или подобном ему форматах. На очереди внедрение структурирования во все уровни корпоративных информационных систем и системы интерактивного общения.

Обзор существующих систем

На сегодняшний день существует множество систем интеллектуальной обработки информации (глубинного анализа текстов, технологий добычи текстовых данных). Разработкой таких систем занимаются как небольшие частные компании, группы ученых и программистов, так и гиганты компьютерной индустрии.

Так, IBM предлагает свою разработку Intelligent Miner for Text, которая, по сути, является набором отдельных утилит [3].

Language Identification Tool – утилита определения языка – для автоматического определения языка, на котором составлен документ.

Categorisation Tool – утилита классификации – автоматического отнесения текста к некоторой категории (входной информацией на обучающей фазе работы этого инструмента может служить результат работы следующей утилиты – Clusterisation Tool).

Clusterisation Tool – утилита кластеризации – разбиения большого множества документов на группы по близости стиля, формы, различных частотных характеристик выявляемых ключевых слов.

Feature Extraction Tool – утилита определения нового – выявление в документе новых ключевых слов (собственные имена, названия, сокращения) на основе анализа заданного заранее словаря.

Annotation Tool – утилита «выявления смысла» текстов и составления рефератов – аннотаций к исходным текстам.

Компания SAS Institute предлагает систему, способную сравнивать определенные грамматические и словесные ряды в письменной речи человека, с которым вы общаетесь посредством электронной почты, с тем, что было написано им ранее, и выявлять подозрительные несовпадения. Система получила название Text Miner [4].

Среди разработок на постсоветском пространстве стоит выделить систему GALAKTIKA-ZOOM. Этот программный комплекс предназначен для аналитической обработки динамично пополняющихся больших массивов (до десятков миллионов) текстовых документов, находящихся в подключаемых неструктурированных и структурированных электронных базах данных [5].

Естественно, что в стороне не могли остаться разработчики СУБД. Так, средства Text Mining можно увидеть в продуктах Oracle начиная с версии 7.3.3. В версии Oracle9i эти средства развились и получили название Oracle Text – программный комплекс, интегрированный в СУБД, позволяющий эффективно работать с запросами, относящимися к неструктурированным текстам. При этом обработка текста сочетается с возможностями, которые предоставлены пользователю для работы с реляционными базами данных. В частности, при написании приложений для обработки текста стало возможным использовать SQL [6].

Это только некоторые разработки в области интеллектуального анализа информации, иллюстрирующие возможности и широту применения средств Text Mining.

Что же касается RSS (и им подобных) форматов, систем взаимодействия с информацией в данном формате, систем поиска – их множество, и они уже встраиваются непосредственно в браузеры, почтовые программы, операционные системы. Сфера применения также обширна: новости, реклама, информационные рассылки различного содержания (как аналог почтовых рассылок).

Система извлечения знаний

На основе анализа существующих подходов к поиску и анализу текстовой информации разработана структура системы извлечения знаний из разнородных источников информации, приведенная на рис. 1.

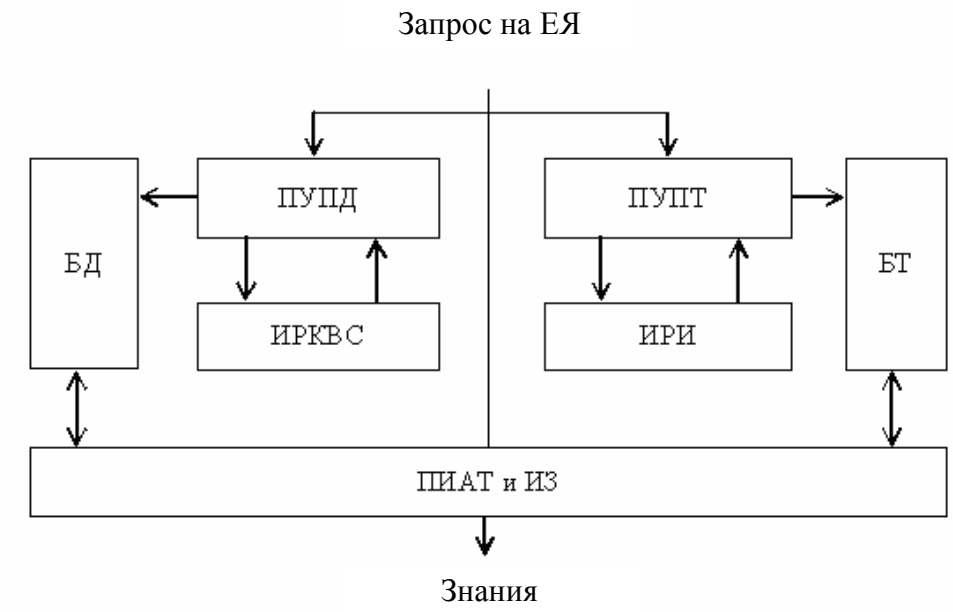

Рисунок 1 – Структура системы извлечения знаний

В состав системы входят:
- подсистема управления поиском неструктурированной информации (ПУПТ);
- подсистема управления поиском структурированной информации (ПУПД);
- информационные ресурсы корпоративной вычислительной сети (ИРКВС);
- информационные ресурсы Интернета (ИРИ);
- база данных структурированной информации (БД);
- база данных неструктурированной информации (БТ);
- подсистема интеллектуального анализа текстов и извлечения знаний.

В состав подсистемы управления поиском текстов [7] входит лингвистический процессор, база знаний синтаксиса запросов поисковых серверов, конструктор поисковых запросов в синтаксисе языков запросов популярных поисковых серверов Интернета. Лингвистический процессор анализирует семантику естественно-языкового запроса и вызывает конструктор поисковых запросов. Результаты поиска загружаются в базу данных неструктурированной информации роботом закачки файлов.

Подсистемы управления поиском структурированной информации содержат лингвистический процессор с конструктором SQL – запросов СУБД, входящих в корпоративную вычислительную сеть. Результаты выполнения запросов сохраняются в БД структурированной информации.

Поиск в интеллектуальной подсистеме анализа текстов и извлечения знаний будет проводиться в два этапа: по структурированной и по неструктурированной информации. При этом первый этап поиска осуществляется всегда. Переход ко второму этапу осуществляется в том случае, если первый этап не даёт необходимых результатов. Обработка документов на втором этапе в ПИАТ и ИЗ осуществляется с учетом конкретной предметной области и возможных вариантов анализа исходных материалов, что приводит к сокращению анализируемых комбинаций слов. Решение о переходе к следующему этапу может приниматься как пользователем, так и автоматически.

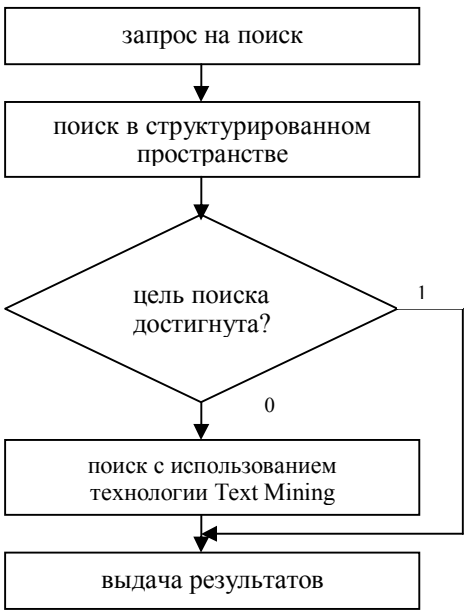

Рисунок 2 – Схема поиска информации

При выполнении поиска с использованием технологии Text Mining проводится структурирование текстов и запоминание в БД, для того чтобы ускорить поиск этой информации в дальнейшем. Таким образом, за счет естественного и искусственного увеличения доли структурированной информации необходимость перехода ко второму этапу поиска будет уменьшаться, увеличивая соответственно скорость поиска информации и извлечения знаний.

Выводы

Оба описанных подхода предназначены прежде всего для преодоления информационной перегрузки. Они разные по своей сути и сложности реализации, и естественно, отличаются результатами.

Предварительное структурирование предназначено для подачи новой информации; это современный подход, дающий хорошие результаты при поиске и извлечении знаний, но, к сожалению, накладывающий ограничения на способ подачи ин-

формации, что делает невозможным использовать ранее размещенную в информационных ресурсах информацию.

Что же касается технологии Text Mining, то она ввиду рассмотрения при анализе любой текстовой информации сложна. А системы на её основе Виктор Шепелев характеризует как «работающие, но не всегда, не везде и не на любых данных; чаще всего, требующие опытного специалиста, который бы трактовал результаты работы, увеличивал объем обучающей выборки, изменял параметры и настройки, оценивал достоверность» [8].

Предложенная в этой статье система призвана объединить оба этих подхода, сделав поиск информации более быстрым и эффективным. Компоненты предложенной системы реализуются в настоящее время в интеллектуальной системе дистанционного обучения по дисциплинам цикла «Компьютерные науки», внедряемой в Донецком государственном институте искусственного интеллекта.

Литература

1. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. – СПб.: Диалектика, 2005. – 272 с.
2. Рассел Кей. Технология RSS – инструменты и методы // Computerworld. – 2004. – № 32. – С. 22-25.
3. Семененко А.В. Варианты корпоративных технологий добычи текстовых данных // Информ. ресурсы России. – 2002. – № 5. – С. 43-47.
4. C-News – издание о высоких технологиях // http://www.cnews.ru.
5. Манченко Е.В., Кишинский И.Ю. GALAKTIKA-ZOOM. Тезисы доклада // Научно-практическая конференция «Проблемы обработки больших массивов неструктурированных текстовых документов». – Москва, Фонд эффективной политики, 21-22 мая 2001 г.
6. Колесов А.Н. Извлекая знания из хаоса информации // PC Week. – 2003. – № 43. – С. 18-20.
7. Вороной С.М., Москалев В.Э. Интегрированные инструментальные средства приобретения, конструирования и обновления знаний интеллектуальных обучающих систем // Проблемы программирования. – 2000. – № 1-2. – С. 484-487.
8. Шепелев В.П. Text Mining как высокая технология словоблудия // Компьютерра. – 2003. – № 43. – С. 40-45.

Д.В. Бабін, С.М. Вороной, Є.В. Малащук
Підвищення ефективності здобуття знань на основі інтелектуального аналізу і структурування інформації
У даній роботі досліджується проблема взаємодії «людина – інформація». Розглядаються основні підходи забезпечення цього взаємозв'язку: інтелектуальний аналіз текстів і попереднє структурування інформації. Здійснюється аналіз цих підходів та описуються тенденції розвитку інформаційного простору і засобів його аналізу. Пропонується структура системи здобуття знань із інформаційних ресурсів з неструктурованою та структурованою інформацією.

D.V. Babin, S.M. Voronoy, E.V. Malaschuk
Promotion of Effectiveness of Knowlege Extraction on the Base of Intellectual Analisys and Information Structuring
In this research the problem of "human-information" interaction is explored. The main approaches of supplying this correlation are taken up such as: intellectual analysis of the texts and preliminary structuring of information. These approaches are analyzed and the tendencies of informational development and means of its analyzing are pointed. The structure of the system of getting knowledge from information sources with structured and non-structured information is proposed.

Статья поступила в редакцию 15.07.2005.