ОСОБЕННОСТИ ОРГАНИЗАЦИИ БАЗЫ ПРЕДМЕТНЫХ И ЛИНГВИСТИЧЕСКИХ ЗНАНИЙ В СИСТЕМЕ АНАЛИТИК

И.П. Кузнецов, А.Г. Мацкевич
Институт проблем информатики РАН
Россия, 117333, Москва, ул. Вавилова, 44 корп. 2
E-mail: igor-kuz@mtu-net.ru

Ключевые слова: лингвистический процессор, представление знаний, семантические сети, естественный язык, базы знаний, интеллектуальные системы

Рассматриваются вопросы организации логико-аналитической системы АНАЛИТИК, осуществляющей автоматическую формализацию текстовой информации с запоминанием результатов в Базе Знаний (БЗ). Для представления информации в БЗ (как предметных, так и лингвистических знаний) используются семантические сети, ориентированные на отображение семантических компонент естественного языка - упоминаемых в текстах объектов (это могут быть ФИО лиц, адреса, телефоны, названия организаций, производимые продукты и т.д.) и их связей. Последующая обработка осуществляется в БЗ на уровне семантических сетей. Решаются прикладные задачи: поиск по признакам и связям, ответ на запросы, выраженные в свободной форме на естественном языке, выявление связанных объектов и др.

Для обработки знаний используются специальные инструментальные средства - язык ДЕКЛ. Он состоит из правил, у которых в левой и правой части - семантические сети. Обработка идет за счет управления вызовом таких правил. Как показывает опыт, использование таких средств заметно упрощает построение прикладных интеллектуальных систем (Спрут, Аналитик, Криминал, ДИЕС и др.)

Однако при работе с большими объемами информации соответствующие сети выходят за рамки возможностей оперативной памяти. Для их хранения используются базы данных. В статье описываются методы организации такого хранения с подкачкой семантических сетей по мере необходимости.

Введение

За последнее время наблюдается лавинообразный рост объемов текстов, получаемых организациями. Требуется их автоматическая обработка. Трудности такой обработки определяются особенностями естественного языка (ЕЯ): наличием большого количества словоформ, синтаксических конструкций, неоднозначностей, умолчаний и др. В связи с этим, уровень формализации текстов в существующих системах (полнотекстовых баз данных, системах на гипертекстовой основе) невысок, что затрудняет, а часто делает невозможной логико-аналитическую обработку информации.

Для содержательной обработки информации предлагается система АНАЛИТИК, основанная на технологии баз знаний (БЗ) и соответствующих методиках обработки текстов для решения прикладных задач. Особенность методик - в переносе сложных этапов лингвистического анализа на уровень обработки знаний, а также в наличии ограничений на выделяемые объекты и глубину семантического анализа [1]. Система базируется на концептуально-лингвистической модели и методиках, развиваемых на протяжении последних десяти лет в ИПИРАН. Уровень полученных результатов сопоставим с передовыми научными исследованиями за рубежом [2].

Система АНАЛИТИК ориентирована на обработку больших потоков текстов с выдачей пользователю (аналитику) необходимой информации в наиболее удобном для него виде. Эта система решает следующие задачи:

- автоматический ввод документов с их делением на части и лексическим анализом;

- автоматическую формализацию текстовой информации с созданием собственной базы знаний (БЗ), имеется в виду направленное извлечение знаний из текстов ЕЯ (русского, английского) с ее использованием на уровне БЗ;

- поиск похожих документов и упоминавшихся в них объектов на основе критерия их семантической близости;

- ответ на запросы на естественном языке;

- поиск объектов по связям с другими объектами;

- использование статистических методов для поиска зависимостей и их выдачей в виде временных и других диаграмм;

- автоматическое заполнение тематических полей баз данных.

Структура системы

Логико-аналитическая система АНАЛИТИК - это аппаратно-программный комплекс, автоматизирующий процесс ввода, формализации и анализа текстовых документов, их использование в задачах поиска и оперативной идентификации. Формализация сводится к выделению значимой информации: интересующих пользователя объектов (например, ФИО лиц, названия организаций, городов и др.), а также их признаков, связей, атрибутов (это могут быть свойства, словесные портреты и др.). Они образуют структуры знаний [3].

Система содержит собственные базы данных и знаний, а также терминологический словарь.

База данных (БД) системы АНАЛИТИК служит для хранения поступающих документов и структур знаний. Документы могут быть:

- в виде текстов естественного (русского) языка;

- в виде информационных карточек.

База знаний (БЗ) системы АНАЛИТИК обеспечивает:

- хранение значимой информации и связей;

- эффективный поиск и анализ информации по связям.

Знания (предметные и лингвистические) в БЗ представляются в виде структур, которые записываются в нотации семантических сетей, дополненных средствами представления событийных компонент и комплексных связей. В результате образуются расширенные семантические сети (РСС).

РСС ориентированы на отображение особенностей семантики ЕЯ - упоминаемых объектов, их связей, а также возможности интеграции множества связанных объектов в один объект, что выражается на ЕЯ в виде форм с отглагольными существительными. Понятие связи рассматривается в широком смысле. Это могут быть не только отношения, но и зависимости. Связанными считаются также объекты, участвующие в одном действии. Группа связанных объектов может быть связана с другой группой, что на ЕЯ выражается в виде глагольных форм с отглагольными существительными.

РСС состоят из однотипных N-арных фрагментов. В каждый из них введена вершина, называемая кодом фрагмента и соответствующая всей представленной в нем информации [4]. Помимо этого, вводится множество "внутренних" вершин, которые порождает сама система по мере необходимости и которые сопоставляются неименованным объектам.

Сети (РСС), представляющие объекты и связи какого-либо документа, образуют так называемые содержательный портрет этого документа. Такие портреты необходимы для обеспечения быстрого и качественного поиска информации по значимым компонентам и связям.

Содержательные портреты (как и документы) шифруются и помещаются в БД, ориентированную на большие потоки информации (сотни мегабайт) и обеспечивающую их быстрый выбор - за счет индексных файлов. Такие портреты подкачиваются в оперативную память по мере необходимости, образуя активную часть БЗ. Итак, система АНАЛИТИК обеспечивает автоматический ввод документов в БД и БЗ.

Лингвистический процессор

Лингвистический процессор обеспечивает автоматическое построение содержательных портретов. Он включает в себя лексикографический, морфологический, синтактико-семантический анализ, а также блок пост лингвистической обработки.

Морфологический анализ необходим, чтобы избавиться от различных форм написания слов, и облегчает поиск. Синтактико-семантический анализ служит для выделения из документа значимых компонент и связей.

Блок лексикографического анализа обеспечивает:

- автоматическое деление текста на самостоятельные части (например, выделение документов из сводок)

- определения начала и конца предложения, а также начала и конца абзаца.

Блок морфологического анализа обеспечивает преобразование слов к единому виду (каноническому) и дает их признаки:

- для каждого слова текста определяется его часть речи, род, число, падеж и другие морфологические характеристики (они зависят от части речи);

- для слов, не содержащихся в морфологическом словаре, дается морфологическая информация - по аналогии с известными словами, имеющими сходную морфологическую структуру;

- для всех слов указываются их формальные признаки, например, слово записано с большой буквы, русскими или английскими буквами, имеет на конце точку и т.д.

Блок морфологического анализа имеет свои каталоги (например, организаций, имен, фамилий и др.), позволяющие добавлять к словам и словосочетаниям дополнительные признаки, например, что слово или группа слов - есть название организации и др.

Блок синтактико-семантического анализа выполняет следующие функции:

- по признакам и контексту выделяет значимые объекты (ФИО людей, организации и др.);

- для каждого выявленного значимого объекта находит в документе связанную информацию (для лиц это их год рождения, пол, адрес и др.).

Данный блок использует терминологический словарь. Результат его работы представляется в виде РСС, образующей содержательный портрет.

К этому портрету добавляются фрагменты, формируемые блоком пост лингвистической обработки. Он осуществляет содержательный анализ информации на уровне РСС с автоматическим выявлением особенностей документа и его значимых объектов, например, автоматическое выявление атрибутов лица, его словесного портрета, формирование по классификатору особенностей события или происшествия. При этом используется терминологический словарь. В результате формируются фрагменты, представляющие значимые характеристики и дополняющие содержательный портрет документа.

Терминологический словарь системы АНАЛИТИК обеспечивает представление типовых классификаторов, используемых в прикладных областях для различения особенностей описываемых объектов и событий. Он содержит следующие виды связей:

- род-вид (класс-подкласс);

- безусловные синонимы;

- условные синонимы (слова совпадают по смыслу при определенном контексте);

- антонимы (противоположные по смыслу);

- взаимоисключающие;

- близкие по смыслу (из одного вытекает другое);

- образующие значимые словосочетания.

Терминологический словарь служит для выявления значимых характеристик документа, расширения пространства поиска и формирования объяснительной компоненты.

Логико-аналитическая обработка

Логико-аналитическая обработка осуществляется на уровне структур знаний (РСС) и ориентирована на логический анализ признаков, связей. Обработка включает в себя различные виды поиска:

- контекстный (по словам с учетом их весов);

- логический (с учетом сходства значимых объектов, их признаков, связей, однотипности описываемых процессов);

- точный (требуется совпадение всех признаков).

Два первых поиска являются нечеткими и могут управляться из запроса, где есть средства указания обязательных и факультативных компонент.

Другие виды аналитической обработки - выявление связей выделенного объекта с выдачей результата в виде графа, построение временных диаграмм и др. Здесь используются как логические, так и статистические методы.

Логический поиск похожих документов и объектов имеет следующие особенности. Задание на поиск допускается в достаточно произвольной форме. Это может быть выделенный объект, запрос на ЕЯ или текстовый документ.

При логическом поиске по запросу или документу система выделяет из задания все значимые объекты, для каждого из них находит признаки (слова, числа), соотносит их к той или иной категории (приметам, адресам и др.), приводит к единому виду, устраняя многозначность. Признакам присваивается степень значимости.

Поиск информации осуществляется с учетом следующих факторов:

- количества и значимости совпавших признаков;

- соотнесенности признаков к той или иной категории (приметы сравниваются с приметами, адреса с адресами и т.д.);

- сильного совпадения по какой-либо категории признаков (например, совпадает большинство примет);

- наличия противоречивых признаков.

По этим факторам подсчитывается степень сходства. Найденная информация выдается пользователю в ранжированном виде с детальным объяснением причин сходства: указанием совпавших и противоречивых признаков, их значимости.

При контекстном поиске задание также допускается в достаточно произвольной форме. Слова приводятся к единому виду - в каноническую форму. Поиск осуществляется по количеству совпавших слов-признаков с учетом их весов. Для каждого документа подсчитывается сумма весов совпавших слов. Результатом является список документов, упорядоченных по этой сумме весов.

Точный поиск требует совпадения всех признаков. Он используется для нахождения похожих объектов, а также для выявления прямых и косвенных связей объекта. Прямые связи - те, которые представлены в содержательном портрете документа, например, принадлежность адреса к тому или иному лицу, участие объектов в одном действии и др. Косвенные связи - это наличие похожих объектов в других документах и их прямые связи. В результате строится граф, иллюстрирующий эти связи.

Организация хранения и доступа к информации

Система АНАЛИТИК обеспечивает ввод документов из различных источников, их автоматическую формализацию и логико-аналитическую обработку, а также удобный интерфейс пользователям для получения результатов. Допускаются удаленные пользователи, их рабочие станции, связанные с сервером приложений в рамках локальной сети. Интерфейс реализован средствами WEB-браузера.

В качестве сетевой модели системы используется трехуровневая архитектура <тонкий клиент/сервер>, которая включает в себя рабочие станции (тонкий клиент), сервер приложений и сервер баз данных (БД). При этом собственно система (приложение) развертывается, управляется и запускается полностью на сервере приложений. В модели используется многопользовательская операционная система и технологии передачи всего пользовательского интерфейса на рабочую станцию клиента.

Такая модель имеет ряд преимуществ по сравнению с традиционной двухуровневой архитектурой <клиент/сервер>, в которой приложение запускается на каждой рабочей станции.

Во-первых, снижается требовательность к ресурсам рабочих станций клиента, упрощается конфигурация программного обеспечения для них.

Во-вторых, использование протокола представления данных отделяет работу приложения от пользовательского интерфейса и посылает по сети только события клавиатуры и мыши, а также обновления изображения на экране.

В третьих, повышается эффективность управления приложением: технология "тонкий клиент/сервер" позволяет системным администраторам разворачивать, управлять и поддерживать приложения с одного рабочего места на сервере приложений за считанные минуты.

В четвертых, повышается производительность. Технология "тонкий клиент/сервер" обеспечивает производительность, сравнимую с производительностью локальной сети, даже на низкоскоростных линиях. Приложения не перекачиваются по сети и не запускаются на пользовательском компьютере. И обмен с сервером данными о нажатии клавиш, движениях мыши и обновлениях дисплея происходит очень эффективно. В результате у администраторов информационных систем есть возможность предоставить пользователям решение, обеспечивающее высокий уровень производительности.

В пятых, повышается защищенность информации, так как не происходит загрузки данных по сети на пользовательский компьютер.

В шестых, использование сервера приложений может существенно сократить число лицензий на сервер БД, используемый системой для хранения знаний, представленных в виде РСС. В двухуровневой модели <клиент/сервер> лицензия требуется на каждую рабочую станцию. Для сервера приложений одно лицензионное подключение требуется на один процесс, который может обрабатывать запросы от нескольких рабочих станций.

Как уже говорилось, вся обработка осуществляется на уровне структур знаний - семантических сетей (РСС). Однако при работе с большими объемами информации необходимые для работы знания выходят за рамки возможностей оперативной памяти. Для их хранения используются базы данных с подкачкой семантических сетей по мере необходимости.

Сервер баз данных (БД) состоит из базы данных и базы знаний системы. База данных служит для хранения поступающих документов в виде текстов. База знаний системы предназначена для хранения значимой информации и связей (содержательных портретов) и обеспечивает эффективный поиск и анализ по связям. Обе этих системы объединены в один банк данных, который ориентирован на работу с большими потоками информации (сотни Мбайт). На основе содержательных портретов строится индексный файл, который тоже хранится в БД. Такой файл состоит из слов (в нормальной форме), выделенных из содержательных портретов, с указанием документов, в которых они присутствуют. На его основе обеспечивается быстрый поиск информации, что является краеугольным камнем при решении любой задачи логико-аналитической обработки.

Банк данных может быть реализован на любой СУБД, поддерживающей стандарт языка SQL-92. Возможно использование и Oracle, и MS SQL Server, и MS Access, и Sybase, и InterBase или другие. Всё зависит от возможностей, требований и традиций конкретных пользователей системы. Обращение к такому банку данных осуществляется с помощью SQL-запросов. Если СУБД поддерживает этот стандарт, то принципиальных возражений в использовании той или иной конкретной СУБД нет. Различия будут в том, как быстро и с каким объемом данных сможет работать БД, а также в различиях условий эксплуатации (лицензирование, сопровождение и поддержка) различных СУБД.

Структура БД следующая: это реляционная база данных, предназначенная для хранения текстов документов, их содержательных портретов и индексов, с помощью которых осуществляется поиск номеров документов по словам содержательных портретов. Пусть, к примеру, на вход системы поступил запрос на ЕЯ. Это может быть любой текст, взятый из какого-либо документа. С помощью лингвистического процессора система строит его содержательный портрет. Далее начинается процесс подготовки данных для решения задачи - поиска ответа. Он состоит в предварительном отборе документов, содержательные портреты которых содержат значимые слова запроса. Для этого используется индексный файл. На его основе выделяется ранжированный список документов, в которых слова из содержательного портрета запроса (с учетом их весов) входят наибольшее число раз. После этого содержательные портреты документов из списка загружаются в оперативную память. Образуется активная часть БЗ, в которой ищется ответ на запрос или решается какая-либо другая аналитическая задача.

Для реализации банка данных используется структура реляционной базы данных, состоящая из трех таблиц: DB_TXT_ZZZ, Net2_sl и Net2_DB.

В таблице DB_TXT_ZZZ хранится информация о документах: номер документа, текст документа, содержательный портрет документа, а также информация о типе документа (уровень доступа, достоверность, дата и т.п.). Для этого в таблице DB_TXT_ZZZ выделено четыре поля: сквозной номер документа (Ndoc), текст документа, его содержательный портрет и тип документа. Эта таблица проиндексирована по полю Ndoc.

В таблице Net2_sl два поля: сквозной номер слова и само слово. Эта таблица проиндексирована по второму полю.

В таблице Net2_DB два поля: номер слова и номер документа (NDoc). Если слово встречалось во многих документах, то в таблице будет много записей с номерами этих документов. У этой таблице два индекса - по номеру слова и по NDoc (на случай, если придется сортировать по документам). Кроме этого, в процессе работы используются списки выделенных объектов. Каждый из них представляет собой таблицу из двух полей, в которых хранятся имена объектов и количество документов, в которых эти объекты встречаются.

Такая организация БД позволит организовать быстрый поиск документов и обеспечит поддержание ссылочной целостности данных при добавлении документов и их удалении из БЗ.

Система АНАЛИТИК может быть настроена на различные приложения в различных областях - где требуется дифференцированная обработка больших потоков текстовых документов. Еще одно приложение - анализ документов, их автоматическая формализация с заполнением полей какой-либо базы данных.

Список литературы

1. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистики и ее приложениям. Том 2. Тарусса 1999.
2. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. AIC, SRI International. Menlo Park. California, 1996.
3. Kuznetsov Igor, Matskevich Andrey. System for Extracting Semantic Information from Natural Language Text. Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том 2. Протвино, Наука, 2002.
4. Кузнецов И.П. Семантические представления. М. Наука. 1986г. 290 с.