Рабинович Борис Ильич

Система обработки потоков данных

В современной модели развития любого бизнеса своевременный доступ к необходимой информации является одним из определяющих факторов его развития и успеха. Именно поэтому все большую популярность завоевывают информационные системы, функциональным назначением которых является предоставление пользователям доступа к хранящейся в них информации [Integrum, Garant и т.п.]. Основной особенностью этих систем является либо узкая специализация и привязка к определенной области, либо жесткие требования к входным потокам данных.

В данной работе рассматриваются методы, позволяющие создать систему, обрабатывающую различные потоки информации, существующей в электронном виде, методы загрузки этой информации в хранилище знаний, некоторые способы поиска и анализа накопленной информации. Система представляет собой среду обработки потоков данных (СОПД). Данные, обрабатываемые СОПД, могут быть следующих типов: текст на естественном языке, структурированная информация (таблицы и биллинги), графические образы рукописных текстов, статические и динамические изображения, звуковые ряды. Для таких типов данных, как графические образы рукописных текстов, таблицы, изображения, звуковые ряды описываются существующие методы автоматической обработки, существующие на сегодняшний день, но не применяющиеся в СОПД. Для загрузки этих типов данных в СОПД предлагается с помощью лингвистического процессора (ЛП) [1] обрабатывать естественно-языковое описание (ЕЯО) экземпляров этих типов данных, создаваемых человеком-оператором на этапе загрузки. ЕЯО, пройдя через ЛП, преобразовывается в расширенные семантические сети (РСС) и поступает в БЗ системы. Поиск связей и анализ осуществляется по РСС этих документов.

Для содержательной обработки информации предлагаются соответствующие методики обработки текстов. Особенность методик - в переносе сложных этапов лингвистического анализа на уровень обработки знаний и глубине семантического анализа [2].

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 [3].

В качестве исходных данных так же описаны Интернет и внешние БД. Одним из ограничений системы является локализация ЛП под определенную предметную область. Для настройки ЛП под другую предметную область потребуется участие в этой работе специалиста по работе со знаниями.

Данная Система будет полезна во многих областях, где требуется осуществить формализацию информации, обработку текстов на ЕЯ, обработку биллингов. Созданные режимы анализа накопленных знаний позволяют решать ряд аналитических задач. Визуализация во всех режимах наглядно отображает полученные в результате аналитической обработки информации данные [4]. Кроме этого, при необходимости, возможно создание новых аналитических режимов, что существенно расширяет область применения Системы.

Литература

1. 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. – М.: Издательство «Наука», 2002.

2. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Протвино. – М.: Издательство «Наука», 1999.
3. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. AIC, SRI International. Menlo Park. California, 1996.
4. Рабинович Б.И., Гнидо Е.И., Кузнецов И.П., Мацкевич А.Г. Частотный анализ биллингов телефонных переговоров в Логико-Аналитической системе «Аналитик». Научно-техническая конференция профессорско-преподавательского, научного и инженерно-технического состава. МТУСИ. Москва, 29-31 января 2002 г. Тезисы докладов. – М.: Издательство ООО «Инсвязьиздат», 2002. Стр. 409-410.