Методы и средства аналитической обработки информации. Обзор

Д.В. Краюшкин

Аннотация. Рассмотрены функциональные возможности наиболее популярных из представленных на российском рынке программных продуктов информационно-аналитического назначения (информационно-аналитических систем – ИАС), на основе которых определен круг методов обработки информации, наиболее активно применяемых в таких программных продуктах. Полученные результаты являются основой для формирования системы критериев сравнительной оценки ИАС и обоснования их выбора в качестве базовых платформ при решении конкретных информационно-аналитических задач.

Введение

В настоящее время наблюдается устойчивый рост объемов анализируемой и обрабатываемой информации. Очевидно, что основным способом повышения эффективности такой обработки является внедрение специализированных программно-технических комплексов – информационно-аналитических систем (ИАС) [1].

Существует множество программных продуктов, предназначенных для автоматизации информационно-аналитической деятельности, различающихся по количеству и качеству выполняемых функций, производительности, интеграционным возможностям и т.д. Тем не менее, в основе функционирования таких систем лежат преимущественно традиционные алгоритмы и структуры данных, разработанные научно-исследовательскими коллективами еще в 60х – 70х годах XX века (например, [2-4]).

В рамках настоящей работы проведена попытка классификации программных продуктов информационно-аналитического назначения в соответствии с их основными функциональными возможностями. В результате такой классификации определен круг наиболее востребованных на практике научно обоснованных методов обработки информации, применяемых в подобных программных продуктах. Иными словами, определен обобщенный базовый набор функциональных возможностей программных продуктов информационно-аналитического назначения, относительно которого возможно оценить преимущества и недостатки реальных ИАС.

Полученные в настоящей работе результаты позволяют сформировать критерии для сравнительной оценки современных программных продуктов информационно-аналитического назначения с точки зрения их функциональных возможностей, что, в свою очередь, обеспечивает возможность обоснованного выбора программных продуктов при решении практических задач.

Приведенные в рамках настоящей работы сведения являются результатом аналитической обработки и обобщения сведений о программных продуктах, размещенных на официальных сайтах их производителей.

1. Основные классы технологических решений

Основная цель применения информационно-аналитических систем состоит в автоматизации работы специалиста (эксперта) с информацией о предметной области. Такая информация традиционно представляется в двух формах:

неструктурированной (текстовой) и структурированной (фактографической).

В большинстве случаев информация поступает в ИАС в неструктурированной текстовой форме (сообщения информационных агентств, сводки, статьи, иные документы), в различных форматах (плоский текст, HTML, форматы MS Office и т.д.). Работа с такой информацией в большинстве ИАС осуществляется за счет ее преобразования к единому внутрисистемному формату и применения различных методов полнотекстового поиска и выявления статистических взаимосвязей терминов [1]. Результаты такой обработки применяются для аналитического исследования предметной области и формирования выходных документов – отчетов, справок и т.д.

Помимо обработки информации на текстовом уровне, значительная часть ИАС позволяет работать с информацией, представленной в структурированной форме: все сведения в этом случае привязываются к конкретным объектам предметной области и представляют собой либо значения отдельных атрибутов этих объектов, либо связи между объектами. Структурированная информация может быть получена как из соответствующих источников (баз и хранилищ данных, структурированных документов и т.д.), так и выделена непосредственно из текстов документов, поступающих в ИАС, при помощи специализированных формальных моделей и алгоритмов. В целом такой подход обеспечивает более богатые аналитические возможности по сравнению с работой с неструктурированной информацией – идентификацию объектов предметной области, выявление связей, мониторинг, прогнозирование и т.д. [3-7].

Таким образом, принимая за основу классификации современных ИАС группы выполняемых ими функций, можно разделить множество ИАС на следующие классы:

информационно-поисковые системы;

системы анализа текстовой информации;

лингвистические процессоры;

системы визуализации структурированной информации;

системы анализа структурированной информации.

Информационно-поисковые системы (например, Артефакт – www.integrum.ru, Медиалогия – www.medialogia.ru, Айкумена – www.iqmen.ru и др.) применяются для ввода, накопления, систематизации документальной информации, а также предоставления ее пользователям и другим информационным системам. Основным методом работы с информацией в таких системах является поиск документов и статистический анализ встречающихся в них терминов; при этом могут применяться дополнительные механизмы, например, информационно-поисковые тезаурусы, стоп-словари, словари синонимов и т.д.

Системы анализа текстовой информации (например, Аналитический курьер – www.i-teco.ru, Астарта – www.cognitive.ru, RCO KAOT – www.metric.ru, SmartWare – www.smartware.ru и др.), как правило, реализуют большинство функций информационно-поисковых систем, но выполняют также глубокую статистическую обработку текстов документов на уровне терминов и их синтаксической взаимосвязи, что позволяет обеспечить тематический поиск и анализ документов, их тематическую кластеризацию и классификацию, определение близких тем. В целом системы анализа текстовой информации позволяют существенно повысить релевантность поисковых процедур и достоверность результатов аналитической обработки.

Лингвистические процессоры (например, RCO Fact Extractor – www.metric.ru, Арион-Лингво – www.sytech.ru, ClaraBridge Document Analytics – www.epam-group.ru и др.) обеспечивают преобразование массива неструктурированной текстовой информации в массив структурированной информации (как правило – семантическую сеть), отражающую характеристики объектов предметной области и их взаимосвязи. В отличие от информационно-поисковых систем и систем анализа текстовой информации программные продукты этого класса обычно не реализуют прикладные поисковые и аналитические функции.

Системы визуализации структурированной информации (например, Visual Links – www.spi2.ru, ONTOS – www.avicomp.ru, Семантический архив – www.anbr.ru и др.) предназначены для отображения и визуального редактирования структурированной информации, получаемой из различных

источников. Как правило, отображение данных реализуется в виде семантической сети, в некоторых случаях в узлах сети могут отображаться дополнительные сведения (значения атрибутов, изображения, рефераты, фрагменты документов и т.д.); в большинстве подобных программных продуктов реализованы функции атрибутивного поиска объектов и простые аналитические режимы (анализ связей объекта, построение досье и т.д.).

Системы анализа структурированной информации (например, X-Files – www.i-teco.ru, Арион – www.sytech.ru, TAIS/ONTOS – www.avicomp.ru и др.) реализуют большинство функций лингвистических процессоров и систем визуального анализа структурированной информации, а также обеспечивают выполнение в автоматическом режиме интеллектуальных процедур идентификации, расширенного поиска объектов и ситуаций, аналитических процедур, направленных на выявление логических зависимостей между объектами предметной области и ситуациями. Кроме того, системы анализа структурированной информации обеспечивают представление результатов работы в различной форме, удобной для восприятия: в виде схем, таблиц, когнитивных карт, диаграмм.

Следует отметить, что среди современных программных продуктов доля систем, относящихся целиком к одному из выделенных классов, незначительна вследствие того, что решение практических задач требует, как правило, применения различных подходов к обработке информации.

1.1. Информационно-поисковые системы

Информационно-поисковые системы образуют наиболее широкий класс программных продуктов вследствие их универсальности, т.е. применимости к произвольным массивам текстовой информации.

Базовая функция информационно-поисковых систем состоит в сборе документальной информации из различных источников: папок файловой системы, Интернет-ресурсов, реляционных баз данных. Сбор информации может осуществляться в соответствии с некоторым регламентом (режим мониторинга источников). В некоторых системах применяются специализированные сценарии, учитывающие структуру и особенности источников данных. Сбор информации может выполняться в трех основных режимах: загрузка отдельных документов, массовая (пакетная) загрузка данных из источника и выполнение запросов к источнику в режиме реального времени в соответствии с запросами пользователей.

Поступающая в систему документальная информация (исходные полнотекстовые документы, представленные в различных форматах – плоский текст, HTML, RTF, PDF, MS Office и др.) приводится к единым кодировке и текстовому формату, пригодному для дальнейшей автоматизированной обработки. Большинство ИАС обеспечивают работу с текстовыми форматами (плоский текст, HTML, CSV и т.д.), ряд систем допускают ввод документов в двоичных форматах (MS Office, PDF и др.).

Среди накапливаемых в информационно-поисковой системе документов выявляются дубликаты и похожие документы. В большинстве систем обнаруженные дубликаты сохраняются в информационных массивах и выводятся при поиске документов совместно с оригиналами (при этом дубликаты помечаются соответствующим образом); в других программных продуктах дубликаты уничтожаются при выявлении. Как правило, предусматривается несколько методов выявления дубликатов: по совпадению контрольных сумм исходных файлов, по совпадению текстов, по совпадению ключевых слов. После выявления и устранения дубликатов документы индексируются (в некоторых системах – по всем терминам, в остальных – только по основам или по определяемым специальным образом ключевым словам).

Основной прикладной функцией информационно-поисковых систем является полнотекстовый поиск документов – поиск документов по вхождению в них заданных ключевых слов. Практически все программные продукты учитывают при поиске морфологическую изменчивость ключевых слов; в различных программных продуктах применяются специфические языки уточнения запросов, позволяющие использовать в запросах маски * и ?, фиксировать расстояние между ключевыми словами. Кроме того, в большинстве систем предусмотрена

возможность ограничения области поиска при помощи тематического рубрикатора документов, указания даты публикации, автора, источника и др.

В пользовательском интерфейсе для уточнения поискового запроса (сам запрос при этом автоматически не расширяется) зачастую применяются средства помощи (подсказки). Два основных метода, применяемых для уточнения запроса, состоят в использовании словарей синонимов и гипонимов, а также в использовании контекста (наиболее частых терминов, встречающихся в текстах документов в окружении ключевых слов). В обоих случаях пользователю предоставляется перечень дополнительных терминов, которые он по своему выбору может включить в запрос (обычно при помощи мыши).

Обнаруженные в результате выполнения поискового запроса документы отображаются в форме списка ссылок на документы, в некоторых системах снабженных аннотацией. Элементы списка упорядочиваются по дате или по убыванию релевантности, также в списке отмечаются похожие документы. В ряде программных продуктов в пользовательском интерфейсе выводятся количественные оценки релевантности обнаруженных документов, а при обращении к полным текстам документов выполняется подсветка обнаруженных ключевых слов в тексте. К результатам выполнения поисковых запросов допустимо применение теоретико-множественных операций – добавление, удаление документа из подборки документов, редактирование атрибутов документа (тип, информативность, конфликтность и т.д.).

Подборки документов, полученные в результате поиска и последующей обработки, могут быть экспортированы в общепринятых форматах для дальнейшей обработки и принятия решений. В большинстве информационно-аналитических систем применяется экспорт в виде файлов форматов MS Office, в ряде систем применяется экспорт набора исходных документов, снабженных HTML-оглавлением, или набора связанных HTML-страниц. Возможен также экспорт статистической информации (статистических таблиц, графиков и диаграмм), отражающей общие характеристики массива или выборки документов – для этого в большинстве систем используются предопределенные наборы статистических отчетов (распределение документов по темам, источникам, авторам, периодам времени и т.д.).

Некоторые информационно-поисковые системы обеспечивают создание и применение хранимых (сторожевых) запросов (тематический мониторинг) – при изменении состава документов, удовлетворяющих таким запросам, в пользовательском интерфейсе выводится соответствующее сообщение (либо используются альтернативные виды доставки сообщений – электронная почта, файловая система и др.). Кроме того, в ряде систем результаты работы персонализированы – допустимо создание пользователями личных (персональных) папок (подборок) документов с возможностью предоставления доступа к ним другим пользователям для чтения или модификации в целях организации самостоятельной и совместной работы.

1.2. Системы анализа текстовой информации

Функциональные возможности систем анализа текстовой информации включают, как правило, соответствующие возможности информационно-поисковых систем, однако реализуют их на более высоком научно-техническом уровне. Например, индексирование документов осуществляется не только по отдельным терминам (ключевым словам), но и по темам, задаваемым синтаксическими сочетаниями терминов, что существенно повышает релевантность результатов поиска; в качестве поискового запроса может использоваться фрагмент текста на естественном языке, причем при поиске будут отбираться только документы, в которых термины (ключевые слова) находятся в той же синтаксической взаимосвязи, что и в запросе; при поиске учитываются тезауральные отношения не только между отдельными терминами, но и между ключевыми темами.

Тем не менее, существуют и принципиальные функциональные отличия систем анализа текстовой информации от информационно-поисковых систем. Прежде всего, большинство систем анализа текстовой информации благодаря наличию средств синтаксического анализа позволяют устранять омонимию и анафорию в текстах документов. Эта служебная операция

позволяет существенно повысить качество выполнения последующих функций (прежде всего, аннотирования, выявления ключевых тем). Устранение анафории основано на выполнении синтаксического анализа текста с учетом синтаксического окружения (близлежащих терминов и предложений). Устранение омонимии осуществляется в системах, в которых различаются типы терминов, и выполняется за счет синтаксического анализа.

Как уже было отмечено, системы анализа текстовой информации обеспечивают определение ключевых тем документов на основе анализа синтаксических связей между терминами терминов. Темы применяются для выполнения поиска документов на естественном языке, тематического поиска, а также (наравне с отдельными терминами) при рубрикации и кластеризации документов. В некоторых системах реализовано также выделение типовых синтаксических конструкций (время, даты, числа, деньги, географические названия т.д.) для последующего поиска по ним.

Важной аналитической функцией систем анализа текстовой информации является автоматическая классификация поступающих документов в соответствии с предусмотренными заранее рубрикаторами (в отдельных случаях выполняется также классификация произвольной выборки документов, в частности – результатов поискового запроса). В большинстве систем классификация осуществляется по ключевым словам и темам, но встречаются и более сложные методы: определение тем на основе синтаксического анализа, нейросетевые классификаторы (в том числе – с использованием обучающей выборки). Подавляющее большинство программных продуктов выполняют неисключающую классификацию (т.е. один документ может относиться к нескольким разделам рубрикатора).

Второй важной аналитической функцией, присутствующей в большинстве развитых систем анализа текстовой информации, является кластеризация документов – выделение в массиве документов, размещенных в системе, тематически близких групп (кластеров) документов (в некоторых системах реализована возможность кластеризации произвольной выборки документов, например, результатов поискового запроса). Кластеризация осуществляется статистически на основе методов факторного анализа. В отдельных случаях кластеризация выполняется несколькими методами в зависимости от качества результатов (оценка результатов проводится экспертным путем), а результаты кластеризации применяются для отбора похожих и совпадающих документов, а также выявления связанных документов (в большинстве программных продуктов связанными считаются документы, в которых упоминаются одинаковые или похожие термины и группы терминов, однако в ряде систем рассматривается связность документов по степени схожести их текстов, т.е. учитывается возможность заимствования фрагментов текста).

Большинство систем анализа текстовой информации выполняют аннотирование документов – выделение в документе фрагментов текста, в наибольшей степени отражающих содержание документа. Практически все программные продукты, реализующие эту функцию, предусматривают предварительное устранение анафории в текстах документов. Как правило, в качестве аннотации выбирается непрерывный фрагмент текста (N предложений, T слов или M символов), в наибольшей степени соответствующий рубрике, к которой отнесен документ (если документ отнесен к нескольким рубрикам, то для построения аннотации выбирается та из них, к которой он отнесен наиболее уверенно). С задачей аннотирования тесно связано реферирование подборки документов, реализованное в некоторых программных продуктах и выполняемое путем последовательного включения в один выходной документ аннотаций всех исходных документов, входящих в выборку.

Одной из прикладных функций, выполняемых системами анализа текстовой информации, является определение тональности текста по отношению к объекту – реализовано в ряде систем на основе применения встроенных правил и/или статистической классификации по синтаксической структуре. В большинстве случаев применяемые наборы правил в процессе эксплуатации не модифицируются.

Некоторые системы анализа текстов обеспечивают прогнозирование и выявление анома-

лий на основе статистической обработки характеристик потока поступающих в систему документов – активности источников, интенсивности поступления документов по определенной тематике, усредненных показателей потока документов и т.д.

1.3. Лингвистические процессоры

Лингвистические процессоры – программные продукты, предназначенные для преобразования неструктурированной текстовой информации в структурированную форму – представляют наибольший интерес с точки зрения автоматизации обработки больших объемов текстовой информации, т.к. от качества процедур структуризации текстовой информации напрямую зависит, какая информация поступит в структурированное хранилище информации и какими будут исходные данные для прикладных поисковых и аналитических режимов. Ситуация осложняется как отсутствием стандартов, фиксирующих основные требования к исходным данным, процедуре и результатам структуризации, так и недостаточностью объективных критериев оценки качества результатов структуризации. Тем не менее, следует отметить, что в настоящее время класс лингвистических процессоров активно развивается.

Базовыми функциями большинства лингвистических процессоров являются сбор неструктурированной текстовой информации, приведение к единому формату, выявление и устранение дубликатов, устранение омонимии и анафории, аналогичные соответствующим функциям информационно-поисковых систем и систем анализа текстовой информации.

Принципиальной для всех лингвистических процессоров является функция формирования модели данных (знаний) о предметной области. В большинстве программных продуктов для описания системной модели данных применяются специализированные языки, задающие онтологию (набор допустимых типов объектов, их атрибутов и связей между ними) предметной области. В некоторых программных продуктах присутствуют средства визуальной настройки модели данных; в отдельных системах настройка на модель данных выполняется разработчиком и в дальнейшем не может быть модифицирована. Кроме того, отдельные программные продукты позволяют как работать с несколькими моделями данных одновременно, так и ограниченно изменять модель данных без перезагрузки информационных массивов.

На основе системной модели данных и лингвистических механизмов осуществляется выделение в текстах документов описаний объектов. В большинстве систем для выделения описаний объектов применяются наборы синтаксических шаблонов (правил), сформулированных на специализированном языке. В некоторых системах применяются правила, встроенные в код системы. Существуют также программные продукты, в которых правила выделения объектов могут настраиваться при помощи визуального редактора. В отдельных системах возможен анализ посредством одного шаблона нескольких предложений текста. Во всех системах автоматически устанавливаются связи объектов с текстами документов, в которых они обнаружены.

Следующим этапом структуризации текстовой информации является идентификация объектов. Идентификация объектов (определение, относятся ли два и более описаний объекта к одному или нескольким различным объектам) в рамках одного текста является основным отличительным функциональным признаком программных продуктов, ориентированных на работу с фактографической информацией. Для идентификации объектов применяются встроенные или формулируемые на специализированном языке правила, фиксирующие набор ключевых полей, или правила вычисления некоторой числовой меры схожести. Идентификация объектов, выделенных в текстах различных документов, выполняется во всех программных продуктах аналогично идентификации объектов в рамках одного текста.

После обнаружения и идентификации объектов предметной области выполняется выделение в текстах документов связей (отношений) между объектами и построение семантической сети документа. Практически во всех программных продуктах для выделения связей объектов применяются синтаксические шаблоны различной степени детерминированности (фиксирующие синтаксическую структуру описания связи с точностью до каждого слова, до взаимосвязи слов,

до последовательности слов, до последовательности предложений), сформулированные на специализированном языке или встроенные в код системы. В ряде систем используются числовые оценки достоверности связей (в случае, когда допустимо применение к тексту нескольких синтаксических шаблонов).

После завершения обработки отдельных документов выполняется объединение семантических сетей документов в единую базу знаний. В большинстве программных продуктов объединение семантических сетей выполняется за счет применения правил идентификации объектов. Помимо этого, в ряде систем проводится идентификация связей (т.е. в общую сеть не попадают дубликаты связей, выделенные в различных документах). Применяются два подхода к построению общей семантической сети (базы знаний): первый заключается в слиянии одинаковых объектов и связей на уровне данных (т.е. общая семантическая сеть становится неделимой), второй - в раздельном сохранении одинаковых объектов и связей (в этом случае семантические сети отдельных документов сохраняются также отдельно, но объединяются на логическом уровне при выполнении поисковых и аналитических запросов за счет связывания одинаковых объектов).

Объединение семантических сетей отдельных документов в единую базу знаний сопровождается обычно выполнением логического вывода. Некоторые из ИАС обеспечивают построение новых связей между объектами и означивание атрибутов на основании связей и атрибутов объектов, выделенных ранее в текстах документов. Иными словами, такие системы позволяют в автоматическом режиме получать новую (аналитическую) информацию. Эта цель достигается двумя способами: в первом случае на основе встроенных или заданных на специализированном языке правил выполняется автоматическое добавление новой информации в базу данных; во втором - значения атрибутов и связи рассматриваются как потенциальные и отражаются только в пользовательском интерфейсе или вспомогательных разделах базы данных (такие связи и значения атрибутов по усмотрению пользователя могут быть впоследствии включены в базу знаний).

В целях повышения качества структуризации тестовой информации в большинстве лингвистических процессоров применяется ручное редактирование результатов структуризации. Для редактирования результатов в ряде систем предусмотрен специализированный интерфейс (редактора, корректора, оператора); в остальных случаях редактирование результатов структуризации выполняется в рамках пользовательского интерфейса аналитика.

В завершение процесса структуризации текстовой информации лингвистические процессоры, как правило, помимо формирования структурированной информации о предметной области выполняют также и формирование определенной метаинформации – наборов атрибутов объектов и связей, применяемых для статистического анализа структурированной информации (в том числе – с применением OLAP): даты выделения объектов и связей из текстов документов, даты редактирования, количество схожих объектов, вес (частота встречаемости) объекта в документе, информативность (количество связей) объекта в документе, конфликтность (количество связанных объектов) объекта в документе и т.д.

1.4. Системы визуализации структурированной информации

Функции систем визуализации структурированной информации направлены на отображение структурированной информации в виде, удобном для дальнейшего визуального анализа, и выполнение простейших аналитических операций над выбранными наборами объектов.

Базовой функцией систем визуализации структурированной информации является настройка системной модели данных, выполняемая аналогично настройке модели данных лингвистических процессоров. Расширением этой функции является настройка системной модели данных на модели данных источников структурированной информации. В большинстве систем такая настройка осуществляется посредством соотнесения сущностей, атрибутов и отношений системной модели данных с сущностями, атрибутами и отношениями моделей данных источников соответственно. Некоторые программные продукты обладают возможно-

стью ограниченного преобразования (объединения, разделения, выполнения арифметических операций) полей данных.

После завершения настройки моделей данных выполняется сбор структурированной информации из реляционных баз данных и структурированных файлов. Сбор информации может осуществляться в соответствии с некоторым регламентом (режим мониторинга источников). В некоторых системах применяются специализированные сценарии, учитывающие структуру и особенности источников данных. Сбор информации может выполняться в двух основных режимах: массовая (пакетная) загрузка данных из источника и выполнение запросов к источнику в режиме реального времени в соответствии с запросами пользователей. Сбор информации сопровождается идентификацией объектов (записей базы данных или структурированных файлов), полученных из различных источников. Практически всеми программными продуктами идентификация объектов выполняется на уровне совпадения ключевых наборов атрибутов; численные оценки степени схожести объектов не применяются.

К накопленной в системе информации применимы базовые режимы поиска объектов (атрибутивный с использованием морфологии и языка запросов И, ИЛИ, НЕ, масок * и ?; тематический – с ограничением поиска по категории объекта в соответствии с системным рубрикатором; сквозной полнотекстовый – поиск объектов по значениям всех атрибутов). В некоторых программных продуктах выполняется подсветка обнаруженных ключевых слов в атрибутах объектов.

Результаты поиска отображаются в пользовательском интерфейсе в форме таблицы, семантической сети, дерева связей. Пользователю доступно выполнение операций по добавлению, удалению, редактированию, слиянию и разделению объектов и связей, выполнению теоретико-множественных операций над полученными выборками, выделение, копирование, вставка подмножеств, создание новой сети на основе фрагмента существующей; масштабирование сети; фильтрация, отображение и маскировка объектов. Для обнаруженных объектов в большинстве систем реализован поиск всех связей. В некоторых из программных продуктов предусмотрены ограничения на поиск связей объекта по типу связи, по типу связанного объекта, по значениям атрибутов связанного объекта, по глубине связи (если применяется режим поиска опосредованных связей). В отдельных системах визуализации структурированной информации предусмотрено отображение результатов поиска и анализа на карте – реализовано за счет интеграции с геоинформационными системами.

По результатам поиска и аналитической обработки информации возможна генерация форматированных отчетов в соответствии с предусмотренными администраторами шаблонами. В большинстве ИАС применяются средства генерации отчетов, настройка которых выполняется пользователем в графическом интерфейсе на уровне указания последовательности атрибутов объектов, включаемых в отчеты. В некоторых системах применяются специализированные языки описания сценариев формирования отчетов. Частным случаем генерации отчетов является построение досье. Досье объекта представляет собой отформатированный по определенному шаблону перечень фактов, известных об этом объекте (связей объекта). Шаблоны определяются типами досье, задаваемыми администратором (как правило, любой тип досье можно применить к объекту любого типа). Практически во всех программных продуктах выполняется поиск всех связей выбранного объекта, отсеивание излишних, упорядочение оставшихся связей и формулирование фактов в форме предложений на естественном языке со ссылками на источники, из которых получена соответствующая информация.

Большинство систем визуализации структурированной информации обеспечивают сохранение результатов запросов с возможностью последующего продолжения работы с ними; выполнение экспорта и импорта семантических сетей в целях обеспечения совместной работы и обмена данными с другими системами, а также персонализацию работы с системой.

1.5. Системы анализа структурированной информации

К системам анализа структурированной информации относятся наиболее развитые про-

граммные продукты аналитического назначения, ориентированные при обработке информации не на тексты документов и наборы терминов, а на сущности предметной области и их отношения, что позволяет повысить эффективность обработки информации и предоставляет аналитическим работникам более понятный и удобный инструментарий.

Возможности систем анализа структурированной информации существенно превышают возможности систем визуализации структурированной информации как в части работы со структурированной информацией (за счет применения более развитых поисковых и аналитических процедур), так и в части работы с неструктурированной текстовой информацией (в отличие от систем визуализации структурированной информации, не обладающих средствами работы с текстовой информацией, большинство систем анализа структурированной информации реализуют практически полный спектр функций, характерных для систем анализа текстовой информации).

Системы анализа структурированной информации, сохраняя значительную часть функций систем анализа текстовой информации и систем визуализации структурированной информации, обладают рядом дополнительных возможностей.

Основной особенностью систем анализа структурированной информации является возможность применения

- расширенных режимов поиска объектов (например, поиска объектов с заданием запроса на естественном языке – по синтаксису запроса определяется тип искомого объекта и выделяются значения атрибутов; поиска объектов с учетом опечаток – с использованием словарей опечаток либо с указанием в запросе количества символов, которые могут быть набраны ошибочно);
- сценариев запросов – некоторые из современных программных продуктов позволяют автоматически последовательно выполнять поисковые и аналитические запросы в целях выполнения некоторой стандартной аналитической процедуры (например, последовательного применения различных видов поиска для обнаружения объектов, обладающих специфическими свойствами).

Отдельные программные продукты позволяют использовать сценарии запросов в целях мониторинга: при выполнении заданных условий запускается сценарий, выполняющий дополнительные проверки, что позволяет снизить количество ложных тревог при мониторинге объектов и ситуаций.

Некоторые системы анализа структурированной информации выполняют классификацию объектов – распределение объектов по рубрикам системного рубрикатора. Как правило, классификация выполняется после идентификации объектов. Правила классификации формулируются на специализированном языке или фиксируются в коде системы.

Важной функцией систем анализа структурированной информации является поиск фактов (ситуаций). Большинство систем анализа структурированной информации помимо поиска отдельных объектов позволяют выполнять поиск группы определенным образом связанных объектов, образующих факт или ситуацию. Поисковый запрос обычно задается в пользовательском интерфейсе в форме таблицы с добавляемыми разделами, каждый из которых соответствует отдельному объекту; в некоторых системах предусмотрены графические средства формирования запроса или задание запроса в форме описания ситуации на естественном языке.

Ряд функций, характерных для систем анализа текстовой информации, выполняется в системах анализа структурированной информации с применением логико-семантических механизмов более высокого уровня. В частности, это относится к выявлению дубликатов документов на основе анализа совпадения семантических сетей этих документов. Применяются два метода выявления дубликатов: проверка сетей на изоморфизм и вычисление некоторой числовой меры схожести сетей; дубликаты документов сохраняются в базе данных и помечаются соответствующим образом.

Аналогично, процедура определения тональности документа по отношению к объекту реализуется в системах анализа структурированной информации на основе анализа связей этого объекта, выделенных из текста документа. Для определения тональности документа

применяются две методики: формирование вывода на основе сложения тональностей отдельных связей и классификация с использованием нейросетевого классификатора (с предварительным обучением).

Преимущества систем анализа структурированной информации становятся наиболее заметными при рассмотрении их аналитических функций, в частности, поиска сложных зависимостей и опосредованных связей между объектами. В большинстве программных продуктов предусмотрен только поиск связей между объектами, размещенными на одной семантической сети, с единственным ограничением на длину цепочки связей. Более развитые системы обеспечивают поиск цепочек связей между произвольными объектами, содержащимися в базе данных, с ограничениями на длину цепочки, типы входящих в нее связей и типы промежуточных объектов.

Большинство систем обеспечивают мониторинг объектов и ситуаций (работу со сторожевыми запросами). Традиционно мониторинг объектов и ситуаций реализован в форме отслеживания обновлений результатов поисковых запросов, направленных на отбор заданных объектов; таким образом, осуществляется сигнальное информирование заинтересованных пользователей о поступлении новой информации об этих объектах. В отдельных системах отслеживается появление в базе данных описания ситуации, подходящего под заданный шаблон, либо автоматически выполняются уточняющие запросы в соответствии с предусмотренным сценарием (т.е. выполняется мониторинг ситуаций).

Небольшая часть систем анализа структурированной информации выполняет прогнозирование поведения объектов и развития ситуации на уровне статистического выявления типовых предпосылок и следствий, связанных по участникам и времени (наиболее характерные события, которые могут быть связаны по времени). Следует, тем не менее, отметить, что удовлетворительных результатов в прогнозировании поведения объектов и развития ситуаций в процессе подготовки настоящей работы обнаружено не было.

2. Основные методы обработки информации

Приведенный перечень технологических решений, без сомнения, не является исчерпывающим. Как показало исследование современных российских программных продуктов информационно-аналитического назначения, в отдельных информационно-аналитических системах реализованы более сложные механизмы обработки информации, чем рассмотренные в рамках настоящей работы, однако такие решения не являются общепринятыми для таких продуктов в целом.

Таким образом, на основе перечня основных функциональных возможностей, характерных для современных программных продуктов информационно-аналитического назначения и перечисленных в рамках настоящей работы, можно сформулировать перечень наиболее популярных методов обработки информации, применяемых в таких программных продуктах. В различной степени эти методы используются в ИАС различных классов:

- морфологический анализ текстов документов на основе алгоритмов строгого и приближенного морфологического анализа с учетом устранения омонимии и анафории [3-9];
- полнотекстовый поиск на основе индексирования текстов документов с использованием результатов морфологического анализа [1, 3];
- тематический поиск на основе синтаксического анализа текстов документов в целях определения характерных сочетаний терминов – ключевых тем [3, 10];
- статистическая классификация документов по ключевым словам и темам – в качестве признаков классификации применяются встречающиеся в одном тексте сочетания терминов и групп терминов [3-11];
- статистическая кластеризация документов на основе наборов терминов, встречающихся в их текстах, на основе методов кластерного анализа [10, 12];
- аннотирование документов и реферирование подборок документов на основе выбора наиболее информативных фрагментов текста [3, 6, 9];
- определение общей тональности документов и тональности документов по отношению к

конкретным объектам на основании синтаксической структуры текстов [10, 12];

- использование информационно-поискового тезауруса для расширения поискового запроса – как в автоматическом режиме, так и в форме подсказки [6, 13, 14];

- использование статистического анализа для определения соотношения тематик информационных материалов, выявления тенденций, мониторинга ситуации и прогнозирования ее развития [10, 12];

- создание модели знаний о предметной области, учитывающей основные типы объектов, их взаимосвязи и закономерности развития ситуации [5, 15];

- выделение описаний объектов и связей из текстов документов и частично структурированных данных в соответствии с предусмотренной моделью предметной области [8, 15];

- идентификация объектов и фактов, выделенных из текстов документов – как внутри одного документа, так и между различными документами [3, 5, 16];

- выявление и устранение семантических дубликатов – описаний одной и той же ситуации предметной области [8];

- классификация объектов предметной области, выделенных из текстов документов – как по значениям атрибутов и связей, так и на основе статистических процедур [10, 11];

- выполнение логического вывода в массиве данных о предметной области в соответствии с предусмотренными правилами, сокращение пространства перебора за счет алгоритмов теории экспертных систем [5, 15];

- поиск информации (объектов, фактов и ситуаций) с заданием запроса в виде шаблона фрагмента сети или на естественном языке с использованием алгоритмов эффективного поиска на графах [1, 3, 13];

- анализ связей (в том числе – неявных) объектов предметной области и цепочек связей между объектами на основе алгоритмов быстрого поиска на графах и специализированных структур данных [12, 18];

- мониторинг развития ситуации и статистическое прогнозирование поведения объектов предметной области (в том числе – выявление скрытых закономерностей) методами многомерного анализа данных [3, 10, 12, 19];

- отображение пространственных данных, в том числе – средствами геоинформационных систем, за счет применения алгоритмов отображения многомерных фигур на плоскости [2].

Таким образом, в результате выполненного анализа основных функциональных возможностей современных программных продуктов информационно-аналитического назначения, можно сделать вывод о том, что основные научно обоснованные методы обработки информации, применяемые в них, образуют сравнительно небольшую группу. Большинство из этих методов разработаны в 60–70 х годах XX века и являются наиболее значимыми с точки зрения практического применения современных информационных технологий при автоматизации информационно-аналитической деятельности.

Заключение

В результате проведенного исследования определен базовый круг методов, применяемых на различных этапах автоматизированной обработки информации в большинстве современных информационно-аналитических систем, что позволяет проводить сравнительную оценку эффективности реальных ИАС.

Следует также отметить, что в процессе проведения исследования стало очевидным, что применимость методов обработки информации в конкретных информационно-аналитических системах и особенности их реализации в недостаточной степени научно обоснованы. На практике отсутствие научного обоснования применяемых методов затрудняет априорную оценку применимости конкретных ИАС при решении практических задач и приводит к попыткам применения для решения таких задач программных продуктов, которые по своим возможностям не в состоянии обеспечить достижение ожидаемого результата.

Полученные результаты исследования позволяют оценить общий уровень развития современных программных продуктов, применяемых для автоматизации информационно-аналитической деятельности, и определить приоритет-

ные направления дальнейшего развития таких программных средств. Кроме того, на основе обобщенного перечня методов обработки информации, характерных для большинства развитых информационно-аналитических систем, возможно сформировать систему критериев для научного исследования современных программных продуктов и оценки их качества.

Основу такой системы должны составлять наборы критериев, определяющих наличие или отсутствие у оцениваемой системы необходимого объема функций, причем каждому из выделенных в рамках настоящей работы классу ИАС должен соответствовать специфический набор критериев. В дальнейшем, оценив наличие или отсутствие соответствующих функций у конкретной системы, можно, применяя агрегирование оценок по отдельным критериям, не только определить принадлежность этой системы одному или нескольким классам, но и сравнить ее с другими системами этого класса.

В качестве основного направления дальнейших исследований рассматривается разработка системы критериев для сравнительной оценки программных продуктов информационно-аналитического назначения и выбора среди них технологической основы для эффективного решения практических задач.

Литература

1. Сэлтон Г. Автоматическая обработка, хранение и поиск информации, - М.: Советское радио, 1973. - 560 с.
2. Ахо А.В., Хопкрофт Дж., Ульман Дж.Д. Структуры данных и алгоритмы. – М.: Изд. дом «Вильямс», 2001.
3. Белоногов Г.Г, Богатырев В.И. Автоматизированные информационные системы. М.: -1973.
4. Ашманов И., Харин Н. Интеллектуальные технологии обработки текстов. Электронный офис, май-июнь 1997, с. 24-25.
5. Попов Э.В., Фоминых И.Б., Кисель Е.Б., Шапот М.Д. Статические и динамические экспертные системы. – М.: “Финансы и статистика”, 1996.
6. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. // Москва, 1983.
7. Y. Qui, H.P. Frei. Concept based query expansion. ACM SIGIR, 1993.
8. Файн В.С., Рубанов Л.И. Машинное понимание текстов с ошибками. – Москва, 1991.
9. Харин Н.П. Метод ранжирования выдачи, учитывающий автоматически построенные ассоциативные отношения между терминами. НТИ. Сер. 2, 1989, № 9, с. 19-23.
10. Ашманов И., Григорьев С., Гусев В., Харин Н., Шабанов В. Применение статистических методов для интеллектуальной компьютерной обработки текстов. Труды Международного семинара Диалог’97 по компьютерной лингвистике и ее приложениям. Ясная Поляна, 10-15 июня 1997 г. С. 33-37.
11. Jouko Lampinen and Erkki Oja. Clustering Properties of Self-Organizing Maps. Lapperanta University of Technology, Department of Information Technology, PO box 20, SF-53851 Lapperanta, Finland.
12. Дубров А.М., Мхитарян В.С., Трошин Л.И., 1998. Многомерные статистические методы. Москва «Финансы и Статистика»
13. Salton G., Lesk M.E. Computer Evaluation of Indexing and Text Searching. Journal of the ACM. 1968, Vol. 15, № 1, p. 8-36.
14. Susan T. Dumains, George W. Furnas, Thomas K.Landauer. Indexing by Latent Semantic Analysis. Bell Communications Research 435 South St. Morristown, NJ 07960. Richard Rashman: University Of Western Ontario.
15. Искусственный интеллект. – в 3-х кн. Кн. 2. Модели и методы: Справочник // Под ред. Д. А. Поспелова. – М.: Радио и связь, 1990.
16. Eui-Hong (Sam) Han and George Kapyris. Concept Indexing. A Fast Dimensionality Reduction Algorithm with Application to Document Retrieval & Categorization. University of Minnesota, Department of Computer Science / Army HPC Research Center,4-192 EECS Bldg., 200 Union St. SE, Minneapolis, MN 55455 USA.
17. Douglass R.Cutting, David R.Karger, Jan O.Pedersen, John W.Turkey. Scatter/Gather: a Cluster-based Approach to Browsing Large Document Collections.
18. Stiles H.E. The Association Factor in Information Retrieval. Journal of the ACM, 1961, Vol. 8, № 2, p. 271 – 279
19. Christopher D. Manning, Hinrich Schutze. Foundation of Statistical Language Processing. MIT Press.

Краюшкин Денис Вячеславович. Технический директор ООО «САЙТЭК». Окончил Московский государственный университет им. М.В. Ломоносова в 2003 году. Область научных интересов: изучение лингвистических свойств естественно-языковых текстов и исследования в области логико-аналитических алгоритмов их предварительной обработки.