

УДК 004.89: 004.42

НОВЫЕ ПРОГРАММНЫЕ СРЕДСТВА РАЗРАБОТКИ ДАННЫХ, ПОДДЕРЖИВАЮЩИЕ МЕТОДОЛОГИЮ CORDIET

С.О. Кузнецов (*skuznetsov@hse.ru*)

А.А. Незнанов (*aneznanov@hse.ru*)

Д.И. Игнатов (*dignatov@hse.ru*)

Национальный исследовательский университет

«Высшая школа экономики», Москва

J.Poelmans (*Jonas.Poelmans@econ.kuleuven.be*)

KU Leuven, Leuven

Рассматривается реализация оригинального программного средства разработки данных (*Data Mining*), основанного на методологии итеративного исследования сложно структурированных данных (включая тексты на естественном языке) под управлением онтологий. Описывается архитектура, проектные решения и опыт применения прототипа системы «*Cordiet DMS*».

Введение

Создание программных средств анализа данных является актуальной задачей. К настоящему времени во многих областях человеческой деятельности накоплены огромные массивы оперативной информации. Её анализ затрудняется разнообразием видов, форм и форматов. Информация может быть представлена в виде структурированных баз данных (БД), коллекций структурированных документов, коллекций неструктурированных текстов на естественном языке. Базы данных могут использовать различные языки манипулирования данными и даже используемый в 99% случаев *SQL* имеет, несмотря на стандартизацию, различные диалекты. Документы могут иметь абсолютно разный подход к структуризации и только наиболее современные программные системы используют стандарт *XML*. Но самая тяжёлая ситуация – с неструктурированными текстами на естественных языках. О неполноте и нечёткости данных сказано уже не раз. Актуальной является задача построения систем разработки разнородных данных большого объёма, настраиваемых на предметную область и использующих самые современные методы анализа данных. Используемые сейчас системы

(например, *BM i2 Analyst's Notebook*, *QSR NVivo* и др.) или перекладывают проблемы подготовки исходных данных на пользователя, или слишком тяжело специализируются, при этом либо предлагая ограниченный набор методов анализа, либо огромный набор методов, не связанных серьёзной методологией их применения.

1. Принципы построения и методология *CORDIET*

Cordiet-DMS (Data Mining System) – универсальная программная платформа для создания средств разработки данных и выявления знаний, ориентированных на различные прикладные области. Основой пакета являются новейшие методы и алгоритмы анализа данных, современные технологии работы с крупными коллекциями данных, визуализации и интерактивного анализа данных, построения отчётов. На разработку системы заметное влияние оказала методология *CORDIET (Concept relation discovery and innovation enabling technology)*, изначально разработанная J. Poelmans и P. Elzinga в K.U. Leuven и позволяющая итеративно получать в новые знания из данных под управлением динамической онтологии специального вида (Poelmans et al, 2009, 2011). В качестве методов анализа данных методология *CORDIET* предлагала использовать анализ формальных понятий (*FCA*) для выявления взаимосвязей (Ganter et al, 1999, 2004), скрытые Марковские цепи (*HMM*) для анализа последовательностей (Petrushin, 2007), самоорганизующиеся карты Кохонена (*SOM*) для классификации (Lagus et al, 1996; Poelmans, 2009; Пульманс и др., 2011). К настоящему времени методология переросла частный случай исследования массивов неструктурированных текстов под управлением онтологии и расширена поддержкой других типов входных данных и новыми артефактами анализа данных (на основе мультимодальной кластеризации (Игнатов и др., 2010, 2011; Ignatov, 2011), узорных структур и др.).

Пакет построен на следующих четырёх базовых принципах.

1. Итеративность, обусловленная как методологией, так и динамичностью, присущей исходным данным и онтологией.
2. Разделение процессов получения данных из произвольных источников данных (*data sources*) и анализа неизменяемых снимков данных (*immutable data snapshots*), которые описываются профилями.
3. Выделение четырёх этапов обработки данных:
 - a) доступ к источнику данных и загрузка очередной порции данных;
 - b) доступ к загруженным данным и получение одного или нескольких снимков (по заданному профилю);
 - c) доступ к снимку/снимкам и построение базового артефакта (может использоваться онтология);

- d) доступ к базовому артефакту и анализ различных его представлений и дополнительных артефактов (результатом являются новые знания и, возможно, обновление онтологии).
- 4. Расширяемость с выделением трёх уровней:
 - a) настройки (*custom settings*);
 - b) написания макросов/скриптов (*scripts*);
 - c) создания расширений (*plug-ins*).

2. Архитектурные принципы, проектные решения и технологии

Перечислим базовые архитектурные принципы системы.

1. Многокомпонентная архитектура.
2. Работа на обычных персональных компьютерах.
3. Оптимизация форматов данных и *API* компонентов по скорости.
4. Поддержка как локальной, так и распределённой версий с учётом масштабирования.
5. Использование разнотипных источников данных. Выделение для накопление данных *XML*-хранилища с формализованным *API* и поддержкой различных технологий и систем хранения.
6. Ориентация на работу со снимками (*snapshots*) и неизменяемость артефактов, ставшими входными данными для следующих этапов анализа данных.
7. Расширяемость как на уровне добавления компонентов, так и на уровне «скриптов» внутри компонентов.
8. Выделение трёх главных типов расширений: методов доступа к снимкам (*query parser*), решателей (*solver*), визуализаторов (*visualizer*).
9. Наличие встроенного языка написания макросов (скриптов).

Общая архитектура системы приведена на рис. 1. Можно видеть, что собственно анализ данных вынесен из ядра системы в отдельные компоненты, причём эти компоненты должны существовать в 2 вариантах – локальном и серверном. Распределённый вариант основан на сервисо-ориентированной архитектуре (*SOA*). Это позволяет легко изменять состав конкретного варианта установки и предлагать различные редакции. Основными компонентами локальной версии являются: 1) *XML*-хранилище с поддержкой доступа к внешним источникам данных; 2) индексатор неструктурированных тестовых полей; 3) генератор снимков данных с редактором профилей; 4) редактор запросов к снимкам; 5) редактор онтологии в виде семантической сети; 6) набор решателей и визуализаторов, то есть средств построения и интерактивного исследования артефактов анализа данных; 7) локальная СУБД для хранения снимков, онтологий и артефактов.

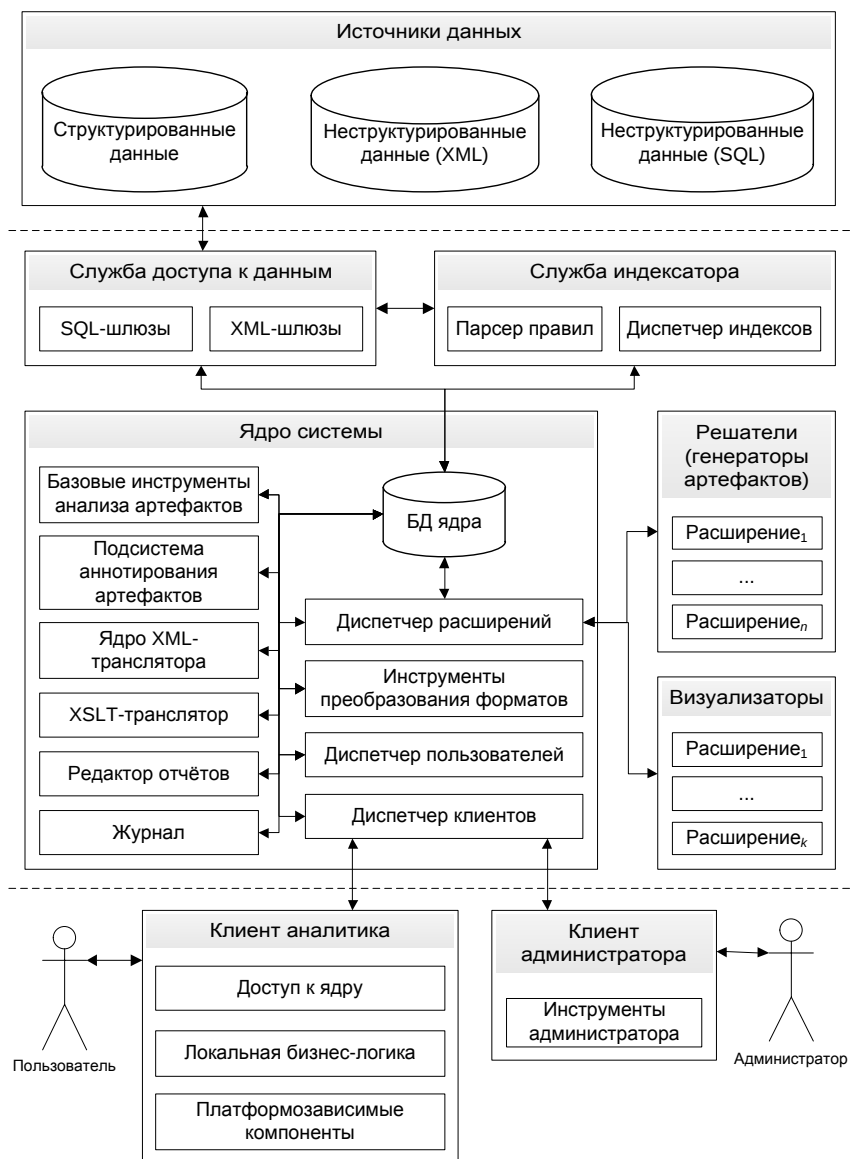


Рис. 1. Общая архитектура платформы Cordiet DMS

Современная многокомпонентная архитектура и использование стандартных форматов межпрограммного обмена (*XML*, *JSON*) позволило использовать большой пул технологий (табл. 1).

Таблица 1. Основные технологии, использованные при создании ПО

№	Назначение	Технологии	Источники
1	Базовые среды разработки	<i>Microsoft Visual Studio, Embarcadero RAD Studio</i>	www.microsoft.com/visualstudio , www.embarcadero.com/products/rad-studio
2	Базовые языки разработки	<i>C++, Delphi, C#</i>	msdn.microsoft.com/ru-ru/visualc/aa336395.aspx , msdn.microsoft.com/en-us/vstudio/hh388566.aspx , www.embarcadero.com/products/delphi
3	Хранилища данных	<i>BaseX, Sedna</i>	www.basex.org , www.sedna.org
4	Визуализация данных в Web-версии	<i>Microsoft SilverLight</i>	www.microsoft.com/silverlight
5	Создание макросов / скриптов	<i>Delphi Web Script, Python</i>	www.delphitools.info/dwscript , www.python.org
	Полнотекстовый поиск	<i>Apache Lucene</i>	lucene.apache.org

3. Решение прикладных задач

Прототип системы использовался при решении задач в областях здравоохранения, противопожарной и правоохранительной деятельности (Пульманс и др., 2011). На рис. 2 представлен пример анализа данных методами *FCA* в специальной версии *Cordiet-FCA*. По сравнению с доступными средствами анализа формальных понятий (*Concept Explorer*, *FCAStone* и др.) система предоставляет возможность интерактивно исследовать контексты большего размера. Построение подрешёток, ассоциативных правил, импликаций и других артефактов с сохранением связей в рамках цепочек исследования позволяет резко повысить эффективность работы аналитика. При разработке базовых алгоритмов использовались результаты последних исследований по сложности решения задач *FCA* (Babin *et al*, 2010).

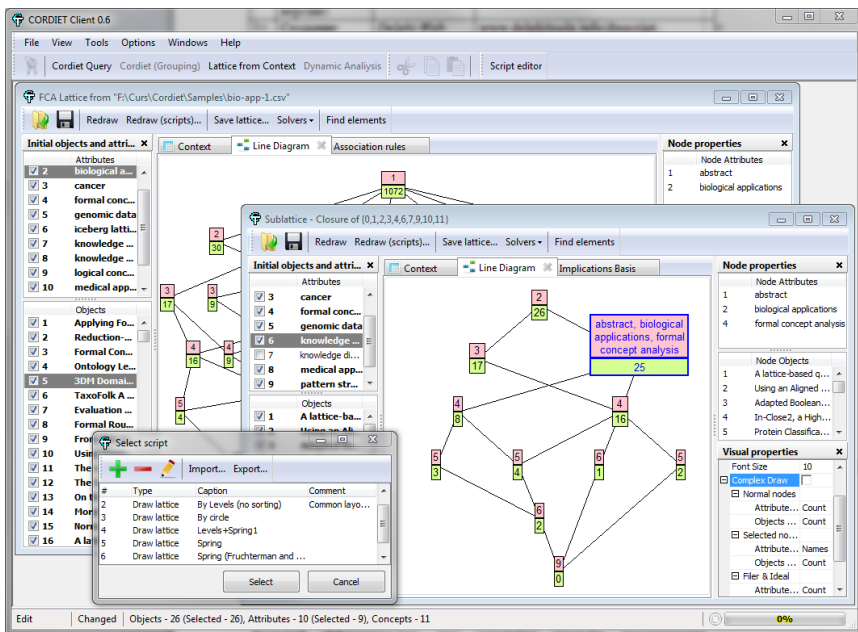


Рис. 2. Пример исследования решётки формальных понятий

Как уже было отмечено, методология предусматривает итеративное изменение метаонтологии с визуализацией основных терминов предметной области. Цикл анализа на примере артефактов *FCA* можно описать следующим образом. Сначала создаются или открываются снимки данных (их может быть несколько, что требуется, например, при динамическом анализе данных). Снимок представляет собой таблицу, где строками являются объекты из *XML*-хранилища, а поля определяются профилем (*profile*). Для работы с профилями существует редактор, который может работать в режиме анализа произвольного *XML/JSON* описания объекта.

Затем на основе снимков по запросу генерируются формальные контексты. Доработка запросов к исходным данным под управлением упомянутой ранее онтологии обеспечивает удобство работы с неструктурированными текстами и способствует выявлению противоречий в данных. Запросы строятся при помощи оригинального языка, делящегося на четыре части: 1) простые правила обработки атомарных полей без шкалирования; 2) правила обработки атомарных полей с номинальными или порядковыми шкалами; 3) правила обработки многозначных (векторных) полей; 4) правила обработки неструктурированных текстов (используется индексатор). Оптимальным является генерация запроса путём выбора некоторых понятий из

онтологии и задания дополнительных параметров. Редактор онтологий предоставляет весь спектр возможностей редактора графов.

Имея набор контекстов, аналитик автоматически получает решётки формальных понятий и возможность построения вторичных артефактов (например, импликаций). Визуализатор решёток предлагает развитые средства поиска объектов и признаков, сравнения понятий, построения подрешёток по заданным множествам объектов и признаков.

Отметим, что с использованием макросов можно существенно изменить поведение существующих решателей и визуализаторов или добавить новые, не выходя из системы. Встроенный редактор макросов/скриптов поддерживает подсветку синтаксиса и предоставляет полноценный контроль за ошибками компиляции и времени выполнения.

Апробированная методология анализа данных с хорошим потенциалом развития, новейшие методы анализа данных, эффективные алгоритмы и современные технологии программирования в рамках общей платформы позволяют решать задачи разработки данных и выявления знаний с большей эффективностью, чем отдельные средства, реализующие конкретные методы анализа, или существующие универсальные системы.

Благодарности В данной работе использованы результаты проекта «Математические модели, алгоритмы и программные средства интеллектуального анализа информации в текстовой и структурной форме», выполняемого в рамках Программы фундаментальных исследований НИУ ВШЭ в 2012 году. Jonas Poelmans is an aspirant of the Research Foundation Flanders.

Список литературы

- [Ganter et al, 2004] Ganter B., Stumme G., Wille R. Formal Concept Analysis, Foundations and Applications. In Lecture Notes in Computer Science, Vol. 3626, Springer, 2005.
- [Ganter et al., 1999] Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
- [Ganter et al., 2001] Ganter B., Kuznetsov S.O. Pattern Structures and Their Projections. In: Proc. G. Stumme and H. Delugach, Eds., 9th International Conference on Conceptual Structures (ICCS 2001), Lecture Notes in Artificial Intelligence (Springer), Vol. 2120, pp. 129-142, 2001.
- [Ignatov et al., 2011] Ignatov D.I., Kuznetsov S.O., Magizov R.A., Zhukov L.E. From Triconcepts to Triclusters // Proceedings of 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining And Granular Computing (RSFDGrC 2011), LNCS/LNAI Volume 6743/2011, Springer-Verlag Berlin Heidelberg, 2011.
- [Lagus et al, 1996] Krista Lagus, Timo Honkela, Samuel Kaski, Teuvo Kohonen: Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration. KDD 1996.

- [Petrushin, 2007] Petrushin V.A. Hidden Markov Models: Fundamentals and Applications, Online Symposium for Electronics Engineers (OSEE), 2007.
- [Poelmans et al, 2009] Poelmans J., Elzinga P., Viaene S., Dedene G. A case of using formal concept analysis in combination with emergent self-organizing maps for detecting domestic violence. In Lecture Notes in Artificial Intelligence, Vol. 5633(XI), Springer, 2009.
- [Poelmans et al, 2011] Poelmans J., Elzinga P., Neznanov A., Viaene S., Kuznetsov S., Ignatov D., Dedene G. Concept Relation Discovery and Innovation Enabling Technology (CORDIET) // CEUR Workshop proceedings Vol-757, CDUD'11 – Concept Discovery in Unstructured Data, 2011.
- [Игнатов и др., 2010] Игнатов Д.И., Кузнецов С.О. Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств// Труды 12-й национальной конференции по искусственному интеллекту, М., Физматлит, Т. 1., 2010.
- [Игнатов и др., 2011] Игнатов Д. И., Кузнецов С.О., Пульманс Й. Разработка данных систем совместного пользования ресурсами: от трипонятий к трикластерам// Математические методы распознавания образов: 15-я Всероссийская конференция. г. Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011.
- [Пульманс и др., 2011] Пулманс Й., Игнатов Д.И., Ван Хуле М.М., Виене С., Дедене Г., Элзинга П. Анализ текстов полицейских отчетов с помощью эмерджентных самоорганизующихся карт и многомерного шкалирования// Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов VI-й Международной научно-технической конференции. Т2. - М.: Физматлит, 2011, стр. 915-925.
- [Babin at al, 2010] Babin M.A., Kuznetsov S.O. On Links between Concept Lattices and Related Complexity Problems // Formal Concept Analysis, Proc. 8th International Conference (ICFCA 2010), Lecture Notes in Artificial Intelligence (Springer), vol. 5986, 2010.