

СЕМАНТИКО-ЭНТРОПИЙНОЕ УПРАВЛЕНИЕ OLAP И МОДЕЛИ  
ИНТЕГРАЦИИ xOLAP В SemanticNET (ONTONET)

SEMANTIC-ENTROPIC MANAGEMENT OF OLAP AND xOLAP INTEGRATION  
MODELS IN SemanticNET (ONTONET)

*Миронов Артем Алексеевич,*

*ассистент кафедры ТИССУ МИРЭА, mironov@mirea.ru*

*Мордвинов Владимир Александрович,*

*заведующий кафедрой ТИССУ МИРЭА, vat@mirea.ru*

*Скуратов Алексей Константинович,*

*заместитель директора ФГУ ГНИИ ИТТ «Информика», skuratov@informika.ru*

**Annotation**

Analysis of data repository creation and support experience shows that this very IT industry field is a matter of great difficulties over lack of established semantic information processes and systems theory. The article is aimed at studying xOLAP models, principally directed toward semantic management methods. It touches upon the notions of semantic gaps with reference to xOLAP, their semantic-entropic estimations and regulation.

\*\*\*

Системы поддержки принятия решения представляют собой системы, с помощью которых может производиться выбор решений некоторых неструктурированных и слабоструктурированных задач, в том числе и многокритериальных. Модели этих систем, как правило, являются результатом мультидисциплинарного исследования, включающего теории баз данных, искусственного интеллекта, интерактивных компьютерных систем, методов имитационного моделирования. Как известно, в 1993 г. Е. Коддом (E.F. Codd) для систем поддержки принятия решений специального вида был предложен, и в настоящее время широко применяется, термин **OLAP** (Online Analytical Processing) - оперативный анализ данных, онлайн-аналитическая обработка данных для поддержки принятия решений. Исходные данные для анализа представлены в OLAP в виде многомерного куба, по которому можно получать нужные разрезы — отчеты. Выполнение операций над данными осуществляется OLAP-машиной. По месту размещения

OLAP-машины различаются OLAP-клиенты и OLAP-серверы. OLAP-клиент производит построение многомерного куба и вычисления на клиентском персональном компьютере (ПК), а OLAP-сервер получает запрос, вычисляет и хранит агрегатные данные на сервере, выдавая только результаты. По способу хранения данных различают MOLAP, ROLAP, HOLAP и некоторые другие новейшие модификации, ориентированные на взаимодействие с Web. Опираясь на подходы, изложенные в Интернет публикациях Ю.Кудрявцева [1] и других известных авторов, кратко представим классификационные и другие признаки основных модификаций OLAP моделей (MOLAP, ROLAP и HOLAP).

**MOLAP (Multidimensional OLAP)** — все данные хранятся в многомерной базе данных или в специальном формате, определенном создающим и обрабатывающим OLAP-приложением. Все запросы пользователя транслируются в запросы многомерной выборки (MDX, Express 4GI и другие).

**Достоинства модели:** все данные хранятся в многомерных структурах, что существенно повышает скорость обработки запросов.

**Недостатки:** данные куба «оторваны» от базовой таблицы, необходимы специальные инструменты для формирования кубов и их пересчета в случае изменения базовых значений, слабая и неформализованная зависимость от семантических признаков поиска и извлечения данных. Кроме того, техноло-

гии MOLAP не вполне эффективно решают задачи распараллеливания больших баз данных на множество серверов, задачи работы по технологии тонкого клиента, упорядоченных масштабирования и балансировки серверной архитектуры.

**ROLAP (Relational OLAP)** — Данные, включающие в себя все возможные агрегации, хранятся в реляционных таблицах. Все запросы пользователя транслируются в SQL-операторы выборки из соответствующей таблицы.

**Достоинства модели:** Все данные хранятся внутри одной СУБД в одном формате.

**Недостатки:** Чрезмерное увеличение объема таблицы данных для куба и сложности пересчета агрегированных значений при изменениях начальных данных, слабая и неформализованная зависимость от семантических признаков поиска и извлечения данных. В отношении распараллеливания серверов и вообще обслуживания больших массивов данных недостатки те же, что и в модели MOLAP.

**HOLAP (Hybrid OLAP)** - Базовые данные хранятся в реляционной таблице, агрегированные хранятся в многомерной структуре.

**Достоинства и недостатки:** Модель в значительной мере совмещает достоинства ROLAP и MOLAP, поскольку избавлена от их основных недостатков, кроме слабой связи с семантическими признаками и возможности эффективного распараллеливания серверов, к тому же по скорости модель HOLAP проигрывает модели MOLAP. Также в ней не достигается целостного хранения данных, что еще более усугубляет трудности масштабирования и балансировки серверной архитектуры. Увеличиваются затраты на поддержку и определение типа хранения для подкубов.

Разнообразие версий OLAP достаточно велико и продолжает увеличиваться. Модели OLAP обретают новые классификационные признаки, свойства, изменяющие их особенности, досто-

инства и недостатки, впрочем, оцениваемые в зависимости от специфики решаемых задач. Так, наряду с рассмотренными выше тремя модификациями моделей OLAP в последние годы появились и находят широкое применение SOLAP (Spatial On-Line Analytical Processing) - пространственная аналитическая обработка, предназначенная для изучения пространственных данных. Она объединяет понятия из существенно отличающихся друг от друга сфер знаний, а именно географических информационных систем (ГИС) и OLAP; разработана для интерактивного и быстрого анализа больших объемов данных; R-ROLAP (Real-time ROLAP) - OLAP реального времени, в отличие от ROLAP в R-ROLAP для хранения агрегатов не создаются дополнительные реляционные таблицы, а агрегаты рассчитываются в момент запроса. При этом многомерный запрос к OLAP-системе автоматически преобразуется в SQL-запрос к реляционным данным; JOLAP - Java OLAP, платформенно-независимый стандарт создания, хранения, доступа и обслуживания данных в OLAP-серверах, основанный на технологии Java, является чистым Java API для J2EE TM, который поддерживает создание и поддержание OLAP данных и метаданных.

Необходимость обобщения разновидностей OLAP приводит к рекомендации ввести обозначение всего класса этих систем в виде аббревиатуры xOLAP, подчеркивая тем самым, с одной стороны, их единство на основе извлечения данных из многомерных построений, с другой стороны, расширяющийся спектр их разнообразных функциональных особенностей, обычно определяемый вводимым перед выражением OLAP прилагательным («Multidimensional», «Hybrid» и т.п.), фиксируемым в самом общем виде буквой «х». Таким образом, в целях более обобщенного и экономного изложения, в настоящей публикации объединим всю известную совокупность разнообразных OLAP единым основообразую-

щим термином – аббревиатурой «**xOLAP**», понимая под ним различные системы онлайн-аналитической обработки данных в многомерных кубах для поддержки принятия решений (OLAP), где приставка «x» отображает возможное разнообразие модификаций в части особенностей хранения данных, организации запросов к ним и отчетов и т.д.

Все перечисленные и иные достоинства и недостатки различных xOLAP в значительной мере контрастируют или нивелируются в зависимости от того, какие алгоритмы используются для хранения данных. Так, например, в MOLAP часто задействованы следующие алгоритмы:

- синтаксические алгоритмы, то есть осуществляющие только синтаксические преобразования данных;
- аппроксимирующие алгоритмы, то есть основанные на средних приближенных значениях данных, для которых значения ячейки не всегда совпадают с фактической таблицей;
- алгоритмы вычисления кубов по методам обратных запросов (Iceberg - кубы);
- семантические алгоритмы, работающие на основе семантических признаков, преобразующие структуру куба.

Последнее имеет все более возрастающее значение в условиях эволюционного перерастания хаотичного по своей природе Интернета в Экстранет реализации семантических сетей второго поколения Semantic Web или, что то же самое, ONTONET [2], то есть сетей, где гегемоном управления и поиска является семантика понятий (ключевых слов, метаописаний, тезауруса, морфологии, инфологии, информационных морфизмов). Именно этому аспекту развития и улучшения моделей xOLAP посвящена настоящая статья.

Сохранение и углубление семантики обобщения / специализации (roll-up / drill-down) становится все более актуальным в обеспечении эффективного

моделирования и проектирования xOLAP. Эта парадигма становится не менее важной, чем те, которым уделялось первостепенное внимание в предшествующие годы, а именно, вопросам агрегирования, сортировки, кэширования и группировки данных, сжатия, деконпозиций и реструктуризации кубов и т.д. Даже существенные успехи в решении такого рода задач не снимают остроты вопроса семантики отношений, поскольку по завершению ступени дивергенции проектирования согласно конвергентно/дивергентному методу управления проектами xOLAP восстановление ранее не отрегулированной семантики отношений становится невозможной (даже на ступени трансформации метода, не говоря уже о заключительной ступени конвергенции этого метода), что делает результаты проектирования информационной системы (ИС) неполноценными и тупиковыми.

Более того, возникла и начала удовлетворяться потребность в моделях xOLAP, целевым образом ориентированных на семантические методы управления в статусе главенствующих. Классифицируем такие модели как **SeOLAP (Semantic OLAP)** и будем настоятельно рекомендовать профессиональному сообществу специалистов этой области придерживаться такого классификационного признака, поскольку он вполне передает прозрачную аналогию парадигмы SeOLAP со стремительным становлением семантических сетей второго поколения Semantic Web или ONTONET, где также главенствуют принципы семантического регулирования в части поиска и извлечения данных и знаний. В обеих рассматриваемых ситуациях семантическое управление встает заслоном по отношению к возникновению «взрыва данных» (профессиональный термин теории фактографических информационных систем) и резкому, переходящему в энтропостат, лавинообразному росту энтропии главной функции системы (сети). Вопросы изучения и улучшения моделей семантического управления xOLAP раз-

работаны в специальной литературе недостаточно, хотя в последние годы внимание к ним явно обозначилось. Анализ имеющихся источников в сфере создания и сопровождения хранилищ данных говорит о том, что именно в этой области ИТ индустрии наиболее резко ощущаются трудности, порожденные отсутствием устоявшейся семантической теории информационных процессов и систем. Новым в постановке и развитии такого рода раздела теории информационных процессов и систем является то, что парадигмы этого раздела опираются на учет происходящего на интегративной основе слияния инфологий и морфологий архитектур OLAP и Semantic Web второго поколения, то есть ONTONET [2]. Именно на этой основе преодолевается так называемый «семантический разрыв OLAP», о котором пишут в публикациях последних лет. Речь здесь прежде всего идет о сборе и консолидации данных из разрозненных и несогласованных источников в предметно-ориентированный, интегрированный и независимый от времени (atemporary system, **AtempOLAP**) набор данных. Такое построение уже неизбежно, если возникает опасность коллапсирования фактографической системы по схеме развития «взрыв данных»; оно вообще универсально, если ставится вопрос о перегрузке данных из одного хранилища - донора в другое хранилище - акцептор, созданное независимо от донора. При этом сразу же возникает задача выделения частей и признаков, обладающих абсолютными свойствами **мажоритарности** и **эргодичности**, анализируя и соединяя которые можно **гармонизировать контент** по семантическим признакам. Здесь, согласно отраслевой ВШ РФ «Онтологии ИС» [2], под мажоритарностью функционала ИС понимается одно из важнейших обязательных свойств ИС и сетей, заключающееся в том, что все сигналы, события, команды на входе или в любой части системы или сети согласуются с аналогичными проявлениями на выходе или в других частях сети или системы

(кроме специально обособленных); эргодичность рассматривается как явление, при котором средние значения по времени почти всех возможных реализаций процесса с вероятностью единица сходятся к одной и той же постоянной величине; а гармонизация контента является продуктом систематизации и унификации в результате изменения состава, свойств и признаков составляющих контента, приводящих к росту энтропии и мажоритарности системы. В теории фактографических ИС носителем таких данных, содержащих измеряемые показатели, доступные для хранилища данных, являются OLTP – (Online Transaction Processing) системы обработки данных в реальном времени или, что то же самое, транзакционные системы. *Транзакционная система* - в информатике, система, реализующая транзакции над хранилищем данных. Задача транзакционной системы - обработать как можно больше транзакций в минимальное время с гарантией безошибочных результатов, то есть *перти-нентно* (релевантности для xOLAP абсолютно недостаточно). Решать такого рода задачу можно только с использованием систем под семантическим дирекционным управлением, то есть, используя SeOLAP модельный подход. Этот подход развивает идеологию применения распределенных, многофазных, семантиконесущих и других транзакций в их новых качествах и выстраивая для этого новые модифицированные архитектуры различных версий SeOLAP систем. В настоящую статью предлагаемые авторами альбомы архитектур разновидностей xOLAP и математического описания их информационных морфизмов не внесены ввиду громоздкости изложения, но с ними можно ознакомиться в фондах ОФАП. Традиционный в теории OLAP подход, связанный с исследованиями образующихся множеств **морфем**, **денотатов** (понятий, признаков) вполне универсально пригоден для работы практически со всеми известными авторам разновидностями xOLAP, поэтому здесь подробно не об-

суждается. Однако нетрудно заметить, что в разных OLTP-системах одним и тем же понятиям могут оказаться присвоены разные имена и, наоборот, одни и те же имена могут быть присвоены понятиям с разными концептами, то есть возникают пересекающиеся или еще сложнее соотносящиеся друг к другу денотаты, что и вызывает семантический разрыв.

Вопрос этот осложняется еще и тем, что семантические разрывы бывают разных типов, иногда действующих в совокупности. Чаще всего приходится сталкиваться с пятью следующими их разновидностями:

- семантический разрыв простой модели «двух точек» - модель поиска, классификации и устранения семантического разрыва во взаимодействии функционала «регистратора» и «аналитика»;

- семантический разрыв между параллельными потоками данных кросспотоковой модели «двух точек» - модель, учитывающая и минимизирующая кроссязыковые и кросспотоковые семантические разрывы;

- семантический асинхронный разрыв – модель «двух точек» с охватом обеих выше обозначенных версий, но осложненная несвоевременной передачей как самих данных, так и информации от регистратора к аналитику;

- семантический кроссязыковый разрыв морфизма хранилища данных и витрины данных – все возможные варианты и их комбинации из числа рассмотренных выше;

- семантический кроссязыковый и/или кросспотоковый разрыв в процессе формирования отчетов для анализа пользователем – тот же набор возможных вариантов, причем все, без исключения, составляющие этих явлений и процессов могут быть успешно подвергнуты системной декомпозиции, улучшению и последующему интегративному объединению в единую комплексную модель семантического регулирования функционала xOLAP (SeOLAP).

В кибернетическом аспекте последняя задача может эффективно решаться введением в архитектуру xOLAP дирекционных подсистем семантического управления метаданными (что учтено авторами статьи при формировании альбома архитектур xOLAP – см. в ОФАП). Дирекционная семантическая подсистема, видимо, должна проектироваться с учетом того, что вся система SeOLAP в целом должна иметь общий репозиторий, в котором хранятся как семантические слои, так и права пользователей (логинов) на объекты семантических слоев, при этом типы семантических слоев зависят от типов используемых источников данных. Для создания и поддержки семантических слоев используются различные приложения, число которых на рынке IT быстро увеличивается (в статье обзор опущен).

На основе преодоления семантических разрывов всех возможных разновидностей и их комбинаций, в сущности, осуществляется интеграция xOLAP и Semantic Web с образованием SeOLAP. По мнению некоторых авторитетных авторов, единение Semantic Web – скорее парадигма, чем исчерпывающее решение вопроса о снятии проблемы семантических разрывов xOLAP. Действительно, сегодняшнее развитие Semantic Web сконцентрировано главным образом в направлении дополнительной метаинформации к документу, чем на семантической интеграции самих данных. Появившийся недавно новый вариант развития Fusionsoft Semantic Net преодолевает это затруднение, поскольку он ориентирован на семантическую интеграцию данных. К тому же обеспечивается автоматическая навигация по признакам дополнительно вводимой семантически связанной информации, причем по гетерогенным и распределенным признакам без необходимости писать программный код. Следует заметить при этом, что платформы Fusionsoft Semantic Net и Web дополняют друг друга. Развивая линию на создание и достаточно широкое использо-

вание отраслевой унифицированной классификации ИС, авторы настоящей публикации предлагают обозначать описанную выше модификацию как **FSeOLAP (Fusionsoft Semantic OLAP)**, относя к этому названию результат инфологического строительства с использованием Fusionsoft Semantic Net.

Еще один пример интегративно-го строительства xOLAP с применением семантических принципов управления – **TransSeOLAP** или **Semantic OLAP Transformer** на основе использования технологии Panorama.

В классификации ИС по синергетическим признакам [2] авторы склонны относить разновидность ИС к трансформерам, поскольку в этой модели используется парадигма map/reduce, согласно которой каждая задача трансформируется в map-фазу (где к каждому входному значению применяется некоторое преобразование) и reduce-фазу (в которой множество входных значений агрегируется по некоторой функции). Это позволяет эффективно распараллеливать задачи на множестве серверов. Map-процессы запускаются над подмножествами исходных данных и выполняются абсолютно независимо друг от друга. Reduce-процессы обрабатывают результаты map-фазы, разбивая их по значениям ключей на непересекающиеся блоки, что также позволяет выполнять их независимо. Таким образом, каждая из фаз может обрабатываться на любом количестве серверов параллельно. Технология позволяет строить всевозможные графики и вращать «кубы». Реализуются архитектуры тонкого клиента. Все множество кубов, создаваемых пользователями Google Docs, вычисляется на серверах (облаке) Google с использованием наиболее выгодным образом масштабирования и балансировки серверной архитектуры. Технология, позволяющая тысячам пользователей одновременно создавать OLAP-кубы на сотнях серверов, называется Panorama PowerApps. Для достижения этого результата Panorama использует платформу MapReduce Google.

Можно привести и дополнительно классифицировать немало других интересных и перспективных вариантов развития инфологий и архитектур xOLAP, так или иначе связанных с реализацией семантических принципов управления фактографическими информационными системами.

С позиций дальнейшего углубления теории информационных процессов и систем в моделировании функционала всех этих систем присутствует единый принцип изучения и упорядочения *информационных морфизмов* как системы с пользователем, так и на подсистемных уровнях и уровнях слияния и взаимодействия на единых семантических принципах различных баз данных с различных серверов, да еще, возможно, в интенсивном многопоточном запараллеленном режиме. Соответственно, в математические модели информационных морфизмов всех этих уровней и комбинаций взаимодействий должны органически вписываться в качестве основоопределяющих составляющие семантико-энтропийного регулирования.

Информационный морфизм представляет собой класс эквивалентности. Информационный морфизм интерпретируется здесь как гомоморфизм свободного моноида в информационном поле, генерируемого из сообщества морфологических, иногда и синтаксических, схожеств и признаков, способных к кластеризации, что принципиально важно в условиях главенствования семантических признаков и принципов управления.

Возникающие носители, в частности, упомянутые выше семантические слои xOLAP, могут обладать или не обладать устойчивостью по отношению к информационной среде. При появлении устойчивого носителя может происходить фиксация возникшего типа носителя в случае возможного его использования по отношению к информационной структуре более высокого порядка. Отсюда вероятностная модель информационного морфизма  $V$  между двумя подсистемами, слоями или отображе-

ниями взаимодействий (например, в формировании отчетов)  $A$  и  $B$  в информационной среде определяется следующим образом:

$$V_i = C_i/E_a + k * E_b, \quad (1)$$

где  $C_i$  - относительное количество информации вида  $I$  в дуплексном (самый общий случай информационного обмена между объектами  $A$  и  $B$ ) информационном пространстве;  $E_a$  и  $E_b$  - относительные (долевые) распределения информации в потоках в направлениях от  $A$  к  $B$  и от  $B$  к  $A$ ;  $k$  - сложный коэффициент, в первом приближении равный натуральному числу  $e$  в степени произведения:  $-L(G_{ai} - G_{bi})$ , где  $L$  - коэффициент Лагранжа,  $G_{ai}$  и  $G_{bi}$  - характеристические коэффициенты информационных потоков в направлениях от  $A$  к  $B$  и от  $B$  к  $A$ .

Модель позволяет отследить основные закономерности *информационного морфизма*. Показателем упорядоченности в модели является *информационная энтропия* взаимодействующих объектов, что является классикой семантико-энтропийных оценок и регулирования.

В самом общем виде семантико-энтропийные оценки и регуляторы используют понятие *обобщенной энтропии* [3], которое в первоначальном ее виде (опуская промежуточные выкладки и рассуждения) найдено авторами настоящей статьи слишком общим и неточно отображающим все внутрисистемные взаимодействия SeOLAP и модификации. Возникновение семантических разрывов провоцируется не только сугубо морфологическими причинами, но и проблемами многопоточности, кроссязыковыми и другими, описанными выше.

Найдено, что в отношении основной версии модельного представления SeOLAP достаточно универсален и продуктивен семантико-энтропийный анализ с использованием комплексной реализации так называемых условной энтропии, взаимной энтропии и энтропии объединения (впрочем, не лишено интереса применение других разновид-

ностей, например, энтропии потока, кросс-энтропии и других).

Остановимся на этом вопросе подробнее. Выше упоминалось, что в моделях всех xOLAP отклик всегда должен быть пертинентным (релевантным он уж точно является, но не наоборот). В этой ситуации лучше воспользоваться не энтропией информации, а так называемой условной энтропией, в названии которой озвучена некая назначенная пользователем условная зависимость вероятностей различных событий друг от друга. Вид этой зависимости определяет форму математического описания такой энтропии, которая может оказаться весьма сложной. Разнообразия здесь много. Поэтому здесь приводится наиболее очевидный случай такой зависимости, аналогичный Марковской модели первого порядка. Последовательность дискретных случайных величин  $\{X_n\}_{n \geq 0}$  называется цепью Маркова (с дискретным временем), если:

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n) \quad (2)$$

Здесь в простейшем случае условное распределение последующего состояния цепи Маркова зависит только от текущего состояния и не зависит от всех предыдущих состояний (в отличие от цепей Маркова высших порядков). По аналогии условной энтропией первого порядка (филогенетической информативностью) является энтропия для алфавита, где известны вероятности появления одной буквы после другой (то есть вероятности двухбуквенных сочетаний) или определена вероятность и даже неизбежность возникновения одного события, свойства, явления за другим (например, неизбежность принятия решения о релевантности отклика, если он пертинентен). Для двухбуквенной (двухуровневой) зависимости первого порядка условная энтропия в самом простейшем случае может иметь следующее математическое описание:

$$H_1(S) = -\sum_i p_i \sum_j p_i(j) \log_2 p_i(j) \quad (3)$$

где  $i$  — это состояние, зависящее от предшествующего символа, и  $p_i(j)$  — это вероятность  $j$ , при условии, что  $i$  был предыдущим символом.

Википедия, кстати, приводит удачный пример использования условной энтропии для описания информационных потерь при передаче данных в канале связи с помехами. При этом вводится временное понятие неких частных энтропий, что несколько упрощает воспроизводимые далее рассуждения. Эти рассуждения по аналогии весьма просто переносятся на модели xOLAP, где понятие помехи в канале заменяется сбоем семантического свойства, приводящем к семантическому разрыву или к риску разрыва.

Через частную и обобщенную условные энтропии полностью описываются информационные семантические сбои и риски (назовем их для краткости

	$b_1$	$b_2$	...	$b_j$	...	$b_m$
$a_1$	$p(b_1 a_1)$	$p(b_2 a_1)$	...	$p(b_j a_1)$	...	$p(b_m a_1)$
$a_2$	$p(b_1 a_2)$	$p(b_2 a_2)$	...	$p(b_j a_2)$	...	$p(b_m a_2)$
...	...	...	...	...	...	...
$a_i$	$p(b_1 a_i)$	$p(b_2 a_i)$	...	$p(b_j a_i)$	...	$p(b_m a_i)$
...	...	...	...	...	...	...
$a_m$	$p(b_1 a_m)$	$p(b_2 a_m)$	...	$p(b_j a_m)$	...	$p(b_m a_m)$

Вероятности, расположенные по диагонали описывают вероятность истинного исхода транзакций, а сумма всех элементов столбца даст вероятность появления соответствующего символа на принимающей стороне -  $p(b_j)$ . Тогда семантические риски, приходящиеся на  $a_i$ , описываются через частную условную энтропию вида:

$$H(B | a_i) = - \sum_{j=1}^m p(b_j | a_i) \log_2 p(b_j | a_i) \quad (4)$$

Для вычисления совокупного риска от всех выявленных или возможных aSA может использоваться обобщенная условная энтропия:

$$H(B | A) = \sum_i p(a_i) H(B | a_i) \quad (5)$$

$H(B|A)$  означает энтропию со стороны источника, аналогично рассматривается  $H(A|B)$  - энтропия с при-

последующего изложения **aStvantic Agent - aSA**) при передаче данных и любой транзакции, если условно считать ее канальной. Для этого в предметной области «телематика» используют понятие так называемых канальных матриц. Воспользуемся им, опять же прибегая к методу аналогий. Так, в телематике для описания потерь со стороны источника (то сеть известен посланный сигнал), рассматривается условная вероятность  $p(b_j|a_i)$  получения приемником символа  $b_j$  при условии, что был отправлен символ  $a_i$ . В рассматриваемом случае замена (риск замены) истинного значения и понимания  $a_i$  на  $b_j$  и есть проявление семантического разрыва в xOLAP. При этом матрица совокупности значений aSA по аналогии с теорией телематики обретает следующий вид:

нимающей стороны: вместо  $p(b_j|a_i)$  всюду указывается  $p(a_i|b_j)$  (суммируя элементы строки можно получить  $p(a_i)$ , а элементы диагонали означают вероятность того, что был отправлен именно тот семантический посыл, который получен, то есть вероятность отсутствия риска семантического разрыва).

Опираясь на правила аддитивности и объявленные в начале статьи обязательные эргодичность и мажоритарность функционала SeOLAP различные внутрисистемные и межслойные (в отношении семантических слоев) энтропийные оценки можно свести в некую общую «суммирующую» энтропию. Для рассматриваемого здесь примера оценки возникновения возможных семантических разрывов, вплоть до коллапса системы, также может быть определена та-

кая энтропия объединения всех задействованных в единую систему кубов и их витрин, а также единого для них репозитория, если таковой создан. Взаимная энтропия, или *энтропия объединения*, предназначена для расчета энтропии взаимосвязанных систем (энтропии совместного появления статистически зависимых сообщений) и обозначается  $H(AB)$ , где  $A$ , характеризует передаю-

$p(a_1b_1)$	$p(a_1b_2)$	...	$p(a_1b_j)$	...	$p(a_1b_m)$
$p(a_2b_1)$	$p(a_2b_2)$	...	$p(a_2b_j)$	...	$p(a_2b_m)$
...	...	...	...	...	...
$p(a_ib_1)$	$p(a_ib_2)$	...	$p(a_ib_j)$	...	$p(a_ib_m)$
...	...	...	...	...	...
$p(a_mb_1)$	$p(a_mb_2)$	...	$p(a_mb_j)$	...	$p(a_mb_m)$

Для более конкретного случая, когда исследуется информационный морфизм взаимодействия двух подсистем исключительно на семантическом уровне, матрица необязательно должна быть квадратной. Очевидно, сумма всех элементов столбца с номером  $j$  даст  $p(b_j)$ , сумма строки с номером  $i$  есть  $p(a_i)$ , а сумма всех элементов матрицы равна 1. Совместная вероятность  $p(a_ib_j)$  событий  $a_i$  и  $b_j$  вычисляется как произведение исходной и условной вероятности

$$p(a_i, b_j) = p(a_i)p(b_j | a_i) = p(b_j)p(a_i | b_j). \quad (6)$$

Как показал первоначальный опыт работы с развернутым выше математическим описанием, вполне универсально условные вероятности оцениваются по формуле Байеса. Теорема Байеса - одна из основных теорем элементарной теории вероятностей, которая определяет вероятность наступления события в условиях, когда на основе наблюдений известна лишь некоторая частичная информация о событиях. По формуле Байеса можно более точно пересчитывать вероятность, беря в учет как ранее известную информацию, так и данные новых наблюдений.

Формула Байеса имеет следующий вид:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}, \quad (7)$$

щее звено архитектуры или инфологии системы, а  $B$ - приемное.

Взаимосвязь переданных и полученных данных в самом общем виде описывается вероятностями совместных событий  $p(a_ib_j)$ , и для полного описания характеристик этого обобщенного процесса требуется только одна матрица вида:

$p(a_1b_1)$	...	$p(a_1b_m)$
$p(a_2b_1)$	...	$p(a_2b_m)$
...	...	...
$p(a_ib_1)$	...	$p(a_ib_m)$
...	...	...
$p(a_mb_1)$	...	$p(a_mb_m)$

где  $P(A)$  - априорная вероятность гипотезы  $A$  (смысл такой терминологии см. ниже);  $P(A | B)$  - вероятность гипотезы  $A$  при наступлении события  $B$  (апостериорная вероятность);  $P(B | A)$  - вероятность наступления события  $B$  при истинности гипотезы  $A$ ;  $P(B)$  - вероятность наступления события  $B$ .

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Это свойство весьма привлекательно при исследовании морфизмов xOLAP, для которых реализуются алгоритмы вычисления кубов по методам обратных запросов (Iceberg - кубы). События, отражающие действие «причин», в этом случае называют гипотезами, так как они — предполагаемые события, повлекшие это событие. Безусловную вероятность справедливости гипотезы называют априорной (насколько вероятна причина вообще), а условную - с учетом факта произошедшего события — апостериорной (насколько вероятна причина оказалась с учетом данных о событии). Важным следствием формулы Байеса является формула полной вероятности события, зависящего от нескольких несовместных гипотез, которые иногда могут иметь место для модельных описаний, объединяемых в единую систему

различных платформ и технологий (как это показано выше целым рядом примеров).

$P(B) = \sum_{i=1}^N P(A_i)P(B | A_i)$  - вероятность

наступления события  $B$ , зависящего от ряда гипотез  $A_i$ , если известны степени достоверности этих гипотез (например, измерены экспериментально).

Таким образом, определяются все данные для вычисления энтропий передающей и принимающей стороны:

$$H(A) = -\sum_i \left( \sum_j p(a_i b_j) \log \sum_j p(a_i b_j) \right) \quad (8)$$

$$H(B) = -\sum_j \left( \sum_i p(a_i b_j) \log \sum_i p(a_i b_j) \right) \quad (9)$$

Взаимная энтропия вычисляется последовательным суммированием по строкам (или по столбцам) всех вероят-

ностей матрицы, умноженных на их логарифм:

$$H(AB) = -\sum_i \sum_j p(a_i b_j) \log p(a_i b_j) \quad (10)$$

Путем несложных преобразований также получаем

$$H(AB) = H(A) + H(B | A) = H(B) + H(A | B) \quad (11)$$

Взаимная энтропия обладает свойством информационной полноты - из нее можно получить все рассматриваемые величины.

Достижимые в результате моделирования полнота вероятностей событий и информационная полнота дают основания применять подходы и результаты показанного здесь моделирования к достаточно широкому спектру разновидностей xOLAP, в том числе FSeOLAP, TransSeOLAP и другим, представленным в альбоме классификаций OLAP.

#### Литература

1. Кудрявцев Ю. «Обзор алгоритмов Molap», режим доступа: [http://www.citforum.idknet.com/consulting/BI/molap\\_overview/](http://www.citforum.idknet.com/consulting/BI/molap_overview/) по состоянию на 19.01.2009.
2. Мордвинов В.А. Онтология моделирования и проектирования семантических информационных систем и порталов: Справочное пособие. - М.: МИРЭА, 2005. - 237 с.
3. Получение знаний для формирования информационных образовательных ресурсов // Иванников А.Д., Кулагин В.П., Мордвинов В.А., Найханова Л.В., Овезов Б.Б., Тихонов А.Н., Цветков В.Я. М.: ФГУ ГНИИ ИТТ «Информика», 2008. - 440с.