Distr.
GENERAL

ECE/CES/SEM.54/14 (Summary)
6 April 2006

RUSSIAN
Original: ENGLISH AND RUSSIAN ONLY

СТАТИСТИЧЕСКАЯ КОМИССИЯ и ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ ОРГАНИЗАЦИИ ОБЪЕДИНЕННЫХ НАЦИЙ
КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

ЕВРОПЕЙСКАЯ КОМИССИЯ
СТАТИСТИЧЕСКОЕ УПРАВЛЕНИЕ ЕВРОПЕЙСКИХ СООБЩЕСТВ (ЕВРОСТАТ)

ОРГАНИЗАЦИЯ ЭКОНОМИЧЕСКОГО СОТРУДНИЧЕСТВА И РАЗВИТИЯ (ОЭСР)
СТАТИСТИЧЕСКИЙ ДИРЕКТОРАТ

Совместный семинар ЕЭК ООН/Евростата/ОЭСР
по вопросам управления статистическими
информационными системами (УСИС)

София, Болгария, 21-23 июня 2006 года

Тема ii): Распространение и отношения с клиентами

РАСПРОСТРАНЕНИЕ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ ЧЕРЕЗ ИНТЕРНЕТ

Вспомогательный документ, подготовленный Стефано де Франчиси, Джузеппе Синдони, ИСТАТ, Италия, и Леонардо Тининини, IASI-CNR, Италия

Резюме

1. Вебхранилище статистических данных представляет собой систему, конкретно предназначенную для распространения статистических данных через Интернет. Данные накапливаются в хранилище данных (Kimball, 1996) и доступны через Интернет благодаря функциям гипермедийной навигации, позволяющим пользователю выбирать и динамически визуализировать данные в различных форматах.

2. Существует тесное соответствие между статистическими базами данных и хранилищами данных (Shoshani, 1997): a) микроданные соответствуют таблицам фактических данных; b) макроданные соответствуют кубам данных, а общие признаки макроданных соответствуют осям куба; c) признаки категорий и их классификационные

иерархии соответствуют координатам и уровням координат; d) функции свертывания данных обеспечивают суммирование, т.е. позволяют перейти с более подробного на менее подробный уровень агрегирования; e) напротив, операции углубления позволяют перейти с менее детального на более детальный уровень агрегирования. К сожалению, некоторые особенности статистических данных, касающихся бизнеса, требуют расширения технических средств хранилищ данных за счет включения специфических моделей и структур. Одна из особенностей касается обследований, основанных на выборках. Как правило, в отношении агрегатов, полученных на основе выборочных микроданных, требуется проводить проверки значимости, которые не нужны в традиционных хранилищах данных. Другая особенность касается соблюдения конфиденциальности и риска вторичной идентификации. В связи с этими требованиями традиционные системы хранения данных не могут использоваться для распространения статистических данных через Интернет без соответствующей доработки, поскольку они позволили бы пользователям беспрепятственно знакомиться со всеми измерениями конкретного факта, не предоставляя функций для проверки на соответствие упомянутым выше принципам.

3. В настоящем документе описывается обобщенная методика распространения статистической информации через Интернет, основополагающая концептуальная модель пространственно-временных многомерных данных и обеспечиваемые функции хранения данных. Эта методика использовалась для разработки и реализации обобщенной системы, называемой "DaWinci/MD" (Sindoni G., Tininini L., 2006), используемой в настоящее время в Институте статистики Италии (ИСТАТ) для распространения данных.

4. Система DaWinci/MD основана на модели для статистических таблиц, разработанной для обеспечения доступа к данным через Интернет. Эта модель основана на принципе декомпозиции информационного пространства на базовые многомерные таблицы, характеризуемые парой компонентов: **объектом интереса**, а именно аргумента таблицы (например, сводного признака) и набором **классификационных признаков**, т.е. измерений, используемых для классификации аргумента таблицы. Каждая базовая таблица может иметь несколько пространственно-временны́х вариантов (пространственно-временны́х многомерных таблиц) за счет прикрепления **временно́й метки даты** определения **пространственного контекста**. Соответственно при выборе пространственно-временно́й многомерной таблицы для визуализации через Интернет каждый шаг отбора последовательно определяет четыре компонента таблицы. Затем типичный запрос в систему определяет все или часть компонентов четырехзвенной группы "*t, s, o, c*", где *t* - временна́я метка, *s* - пространственный контекст, представляющий собой комбинацию характеристики конкретного территориального уровня *d* и географического района *a*; а *o* - предмет и *c* - набор, возможно, пустой, классификационных признаков.

5. Изложенный выше метод декомпозиции таблиц служит также концептуальной основой модели хранения информации, которая отражает наличие пространственно-временно́й таблицы в базе метаданных системы. В этом случае характеризующее хранилище четырехзвенное множество "*t, d, o, c*" означает, что у базовой таблицы, определяемой как "*o, c*", существует парная таблица для временно́й отметки *t* и для всех уровней детализации территории до уровня *d*.

6. Статистические объекты и классификации организованы в иерархии, в которых действуют отношения специализации между основными и производными объектами (или классификациями). Для облегчения навигации в базе данных в иерархии можно включать "виртуальные" объекты (или классификации). Они концептуально объединяют более конкретные объекты и не соответствуют ни одному из атрибутов хранилища данных. В целом чем более родовой характер носит данный объект, тем большим будет число базовых многомерных таблиц, в которых содержится ссылка на этот объект.

7. Объект, набор классификационных признаков и пространственно-временно́й контекст определяют набор имеющихся многомерных таблиц, т.е. набор, содержащий все таблицы со специфицированным или более конкретным объектом, с избранными или более конкретными классификациями, и в виде, позволяющем обработку только до требуемого территориального уровня.

8. Каждая таблица совместима с одним или несколькими пространственными контекстами и зависит от максимального уровня территориальной детализации, определенного в метаданных системы. Географический район и уровень территориальной детализации выбираются одновременно. Они не являются независимыми, и варианты имеющегося выбора зависят от иерархии конкретной территории. Выбор большего или меньшего уровня территориальной детализации соответственно уменьшает или увеличивает число таблиц, соответствующих избранному объекту и набору классификаций. С каждым последующим выбором одного из значений конкретного параметра происходит уменьшение числа остающихся возможных значений соответствующего параметра и соответствующих многомерных таблиц. Это не позволяет пользователям запрашивать таблицы, содержащие чувствительные данные или лишенные смысла комбинации признаков или данные, не подлежащие распространению. В свою очередь удаление одного из выбранных параметров обычно уменьшает число ограничений и потому увеличивает число возможных совместимых наборов.

9. Таблица, выбранная из числа всех совместимых таблиц по данному критерию отбора, визуализируется на одной или более нескольких Интернет-страницах, в зависимости от числа связанных с ней классификационных признаков. Имея перед собой

визуальное изображение таблицы, можно удалить или добавить какой-либо классификационный признак, увеличить или уменьшить уровень территориальной детализации или классификаций, а также изменить географический район. Удаление какого-либо классификационного признака приводит к визуализации менее подробной статистической таблицы. В контексте модели хранилища данных, эта процедура соответствует операции свертывания визуально представленного куба информации (таблицы), т.е. представлению (*n-1*)-размерного подкуба на основе n-размерного куба. Напротив, добавление классификационных признаков соответствует повышению уровня детализации визуализируемого куба, а именно переход к (*n+1*)-размерному суперкубу от *n*-размерного куба, что позволяет получить подробную статистическую таблицу.

Справочные материалы

Sindoni G., Tininini L. (2006) Statistical warehousing on the Web: navigating troubled waters. *Proceedings of the International Conference on Internet and Web Applications and Services. IEEE Computer Society Press*.

Shoshani A. (1997). OLAP and Statistical Databases: Similarities and Differences. *Proceedings of the PODS 1997 Conference*.

Kimball R. (1996). *The data warehouse toolkit*. John Wiley & Sons.
