



УДК 004.942

Ж.Т. Бельдеубаева, С.Ж. Рахметуллина

Восточно-Казахстанский государственный технический университет им. Д. Серикбаева,
г. Усть-Каменогорск

**РАЗРАБОТКА ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ МОНИТОРИНГА
СОСТОЯНИЯ ПОДЗЕМНЫХ ВОД**

Применение современных информационных технологий при решении задач экологического мониторинга окружающей среды и недр является одним из приоритетных направлений. Для реализации государственной политики по охране окружающей среды в рамках научного проекта группой специалистов ВКГТУ им. Д. Серикбаева выполняются работы по разработке информационно-аналитической системы (ИАС) мониторинга состояния подземных вод (ПВ). Одним из результатов работ является концепция, на основе которой осуществляется разработка ИАС мониторинга состояния подземных вод [1].

ИАС - это современный высокоэффективный инструмент поддержки принятия тактических, стратегических и оперативных управленческих решений на основе оперативного и наглядного предоставления всей необходимой совокупности данных пользователям, ответственным за анализ состояния дел и принятие управленческих решений [2]. Согласно современным стандартам ISO 15288 и ISO 12207, описывающих процессы жизненного цикла создания информационных и программных систем, в основе всех технических решений по созданию функциональной структуры и обеспечивающей части подобных систем находится понятие «архитектура системы» [3, 4].

Этапу разработки ИАС предшествовал этап формирования программы разработки, реализации и внедрения ИАС, опирающийся на ряд принципиальных организационных и технологических аспектов:

- проектирование ИАС должно осуществляться на основе обобщенной комплексной функциональной модели, включающей функционирование системы подземных вод, механизм государственной системы мониторинга состояния подземных вод;
- использование для создания и внедрения ИАС имеющегося потенциала информационных и технических ресурсов;
- использование современных информационных технологий для обеспечения важных для пользователей ИАС возможностей, как открытость и доступность системы;
- обеспечение модульности и расширяемости ИАС;

При разработке архитектуры ИАС были учтены основные задачи, которые система должна решать:

- извлечение, преобразование и загрузка данных;
- построение хранилища данных (ХД);
- построение многомерных (MOLAP) и реляционных (ROLAP) витрин данных;
- анализ данных с использованием аналитического инструментария.

Разработка архитектуры системы является одной из важных задач при проектировании и разработке ИАС, так как обуславливает эффективность ее функционирования. ИАС предназначается для решения следующих основных задач:

- сбор, накопление информации об объектах мониторинга, данных по химическому составу проб воды, нормативным показателям;

- анализ данных с целью оценки и прогнозирования состояния ПВ;
- формирование документов стандартной отчетности;
- обеспечение информацией пользователей ИАС.

Процессы и средства реализации в ИАС. Задачами любой информационно-аналитической системы являются эффективное хранение, обработка и анализ данных. Эффективное хранение информации достигается наличием в составе информационно-аналитической системы целого ряда источников данных. Обработка и объединение информации достигается применением инструментов извлечения, преобразования и загрузки данных. Анализ данных осуществляется при помощи современных инструментов делового анализа данных [5].

Диаграмма, отображающая структуру и состав процессов и средств реализации ИАС показана на рис. 1.



Рисунок 1 - Структура процессов и средств реализации ИАС

ИАС мониторинга состояния подземных вод должна обеспечивать пользователям доступ к аналитической информации, защищенной от несанкционированного использования и открытой как через внутреннюю сеть организации, так и пользователям сети интранет и Интернет.

Таким образом, архитектура ИАС должна обеспечивать решение задач на следующих уровнях:

- сбор и первичная обработка данных;
- извлечение, преобразование и загрузка данных;
- хранение данных;

- представление данных в витринах данных;
- анализ данных;
- Web-портал.

Сбор и первичная обработка данных. Разнообразие источников данных обусловлено тем, что часть данных (по объектам мониторинга, химическому составу проб воды и т. д.) импортируется из республиканской БД «Подземные воды», данные по нормативным показателям, параметры моделирования процессов загрязнения подземных вод и др. вводятся непосредственно в ИАС, либо формируются путем вычислений на основе определенных алгоритмов (например, фоновые показатели). Другим фактором разнообразия источников данных является и необходимость использования данных разного формата (карографические данные, свод правил). В целом, источники данных представляют транзакционные источники данных, ХД, витрины данных. Транзакционные базы данных формируются путем сбора и обработки данных, полученных путем импорта или ручного ввода.

Извлечение, преобразование и загрузка данных. Входные данные представляют собой данные различного формата, поступающие из множества систем оперативной обработки данных (OLTP). Эти данные представляют группу данных, характеризующих:

- свод документов, регламентирующих работу системы мониторинга состояния подземных вод (законодательная и нормативная документация);
- объекты мониторинга (сеть скважин и водозаборов);
- качественную и количественную оценку состояния подземных вод (результаты лабораторного анализа химического состава проб воды).

Часть данных извлекается путем импорта из республиканской БД «Подземные воды», часть - формируется непосредственно в ИАС.

Поступившие данные извлекаются, преобразуются, осуществляется процесс «очистки», проверки на соответствие нормативно-справочной информации, и далее данные загружаются в централизованное хранилище данных.

Хранение данных. Аналитическая база данных. Ядром системы является интегрированная база данных (ИБД), размещаемая на республиканском сервере. На этапе реализации и внедрения ИАС в ИБД консолидируются импортированные данные из республиканской БД «Подземные воды» и данные, формируемые непосредственно через интерфейс ИАС.

Интегрированная база данных включает:

- базу графических данных (карографические данные, графики, диаграммы и др.);
- базу семантических данных (база знаний, формализующих знания экспертов);
- базу математических моделей (математические методы и модели по оценке уровня загрязнения подземных вод и прогнозу их состояния);
- базу библиографических описаний документов (паспортные данные объектов мониторинга, нормативная и отчетная документация).

Данные, полученные из различных источников данных, как правило, не согласованы друг с другом, поэтому на следующем этапе осуществляется их консолидация, в результате чего формируются аналитические базы данных (ХД, витрины данных). Аналитические базы данных являются источником данных для решения задач анализа состояния ПВ.

Данные в ХД должны отвечать следующим требованиям [6]:

1. *Предметная ориентированность.* Все данные о системе подземных вод собираются, очищаются, согласовываются, дополняются, агрегируются и представляются в единой, удобной для их использования форме.

2. Интегрированность. Все данные о разных объектах системы ПВ взаимно согласованы и хранятся в едином общем хранилище.

3. Неизменчивость. Исходные (исторические) данные, после того как они были согласованы, верифицированы и внесены в общее хранилище, остаются неизменными и используются исключительно в режиме чтения.

4. Поддержка хронологии. Данные хронологически структурированы и отражают историю за достаточный для выполнения задач анализа и прогнозирования период времени.

ХД разрабатываемой ИАС формируются на основе гибридной модели ХД (HOLAP модель) (рис. 2), представляющей собой комбинацию реляционной и многомерной моделей и позволяющей сочетать высокую производительность, характерную для многомерной модели, и возможность хранить сколь угодно большие массивы данных, присущую реляционной модели.

Главным принципом построения ГХД является то, что детализированные данные хранятся в реляционной структуре (ROLAP), которая позволяет хранить большие объемы данных, а агрегированные - в многомерной (MOLAP), которая позволяет увеличить скорость выполнения запросов (поскольку при выполнении аналитических запросов уже не требуется вычислять агрегаты).

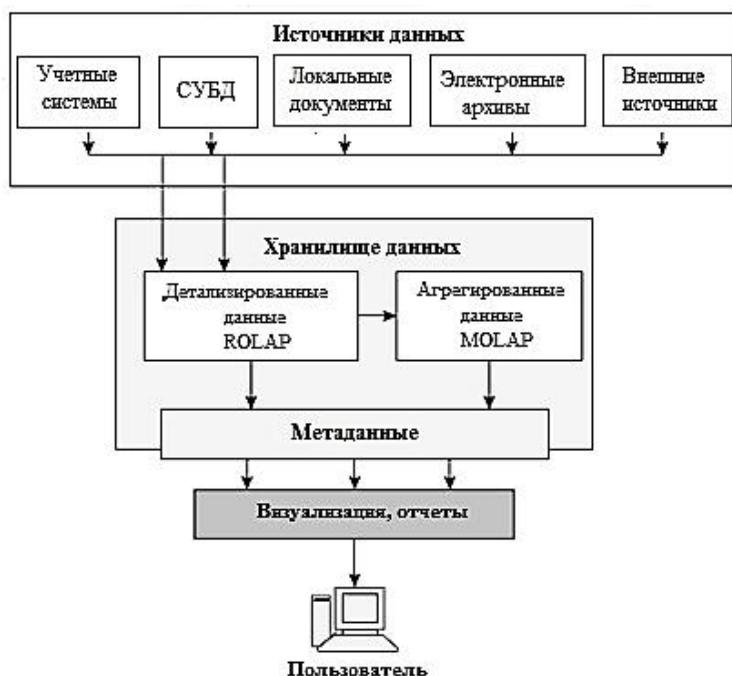


Рисунок 2 - Структура гибридного хранилища данных

Витрины данных. Витрины данных могут быть либо реляционными (ROLAP), либо многомерными (MOLAP). Для максимальной производительности эффективно использовать смешанный способ построения витрин - гибридный (HOLAP).

Использование витрин данных (рис. 3) обусловлено оперативностью и легкостью доступа к данным:

- содержание данных, тематически ориентированных на конкретного пользователя;
- относительно небольшой объем хранимых данных, на организацию и поддержку которых не требуется значительных затрат;

- улучшенные возможности в разграничении прав доступа пользователей, так как каждый из них работает только со своей витриной и имеет доступ только к информации, относящейся к определенному направлению деятельности.

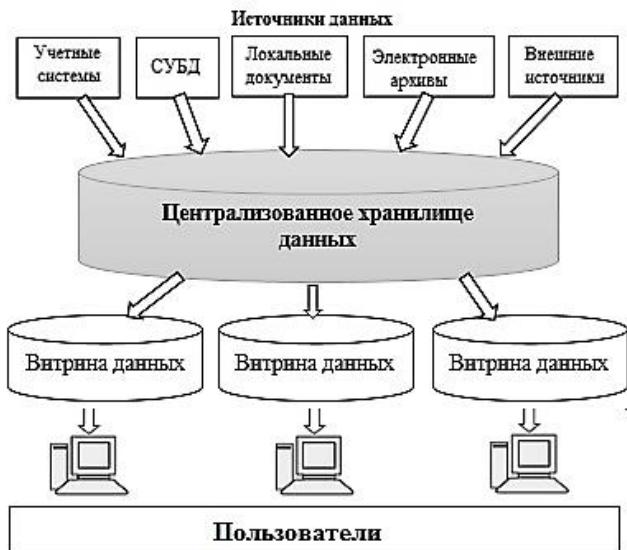


Рисунок 3 - Схема консолидации данных с помощью витрин данных

Анализ данных. Анализ данных с использованием аналитического инструментария. Инструментарий аналитической подсистемы ИАС представляет собой совокупность методов интеллектуального анализа, статистических, математических методов и моделей, позволяющих:

- поиск скрытых закономерностей в данных;
- оценку и прогноз уровня загрязнения подземных вод;
- формирования регламентной отчетности.

Оперативный анализ данных. Оперативный анализ осуществляется при помощи инструментов OLAP-анализа. Для обеспечения работы аналитических приложений исторические и агрегированные данные консолидируются в ИБД и могут быть выгружены в многомерные базы данных, с которыми взаимодействуют OLAP-приложения.

На данное время этот инструментарий OLAP-анализа является наиболее эффективным и популярным. OLAP-анализ позволяет получать доступ к статистическим и организованным данным из источников данных. Технологии OLAP позволяют на основе оперативной базы данных комплекса производить построение многомерных хранилищ данных с целью их экспресс-обработки для принятия управлеченческих решений.

Для анализа информации и формирования отчетности наиболее удобным способом представления является многомерная модель или гиперкуб, ребрами которого являются измерения. Это позволяет анализировать данные сразу по нескольким измерениям, т. е. выполнять многомерный анализ средствами OLAP-систем. OLAP-технология оперативной аналитической обработки данных использует методы и средства для сбора, хранения и анализа многомерных данных в целях поддержки процессов принятия решений [7].

Для обеспечения работы аналитических приложений исторические и агрегированные данные консолидируются в ИБД и могут быть выгружены в многомерные базы данных, с которыми взаимодействуют OLAP-приложения.

Интеллектуальный анализ данных. Хранилище данных ИАС мониторинга ПВ содержит большой объем данных. Это обусловлено сложностью системы ПВ, накоплением в результате многолетних наблюдений огромного количества разнообразного эмпирического материала. Для эффективной работы с огромным объемом данных используются технологии Data Mining.

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей [8].

В общем случае процесс интеллектуального анализа данных (ИАД) для разрабатываемой ИАС состоит из трёх стадий:

- 1) выявление закономерностей в системе подземных вод (свободный поиск);
- 2) использование выявленных закономерностей в системе подземных вод для предсказания неизвестных значений (прогнозное моделирование процессов загрязнения ПВ);
- 3) анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.

Список литературы

1. Бельдеубаев Ж.Т. Концепция разработки информационно-аналитической системы экологического мониторинга подземных вод / Ж.Т. Бельдеубаев, С.Ж. Рахметуллина, Е.М. Турганбаев // Вестник ВКГТУ. - 2015. - № 3. - С. 102-108.
2. Алексеева Т.В. Информационные аналитические системы / Т.В. Алексеева, Ю.В. Амиди, В.В. Дик // Университетская серия, Синергия. - 2013. - С. 384.
3. ГОСТ ИСО/МЭК 15288-2005. Системная инженерия. Процессы жизненного цикла систем. - Введ. 01-01-2007. - М.: Стандартинформ, 2006. - 57 с.
4. ГОСТ ИСО/МЭК 12207-2010. Системная и программная инженерия. Процессы жизненного цикла программных средств. - Введ. 01-03-2012. - М.: Стандартинформ, 2011. - 106 с.
5. Волков И. Архитектура современной информационно-аналитической системы / И. Волков, И. Галахов // Журнал «Директор ИС», 2002. - ООО ИК СИБИНТЕК. - № 3.
6. Алексеева Т.В. Информационно-аналитические системы: Учеб. пособие / Т.В. Алексеева, М.Г. Лужецкий, Е.В. Курганова. - М.: Московская промышленно-финансовая академия, 2005. - 61 с.
7. Лобач Д. Основы OLAP // <http://www.softkey.info>, 2003.
8. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов и др. // Спб.: БВБ-Петербург, 2004. - 336 с.

Получено 13.09.2016

УДК 004.93'1

А.С. Тлебалдинова, М.А. Карменова

Восточно-Казахстанский государственный университет им. С. Аманжолова,
г. Усть-Каменогорск

И.М. Увалиева

Восточно-Казахстанский государственный технический университет им. Д. Серикбаева,
г. Усть-Каменогорск

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ТРУКТУРИРОВАННЫХ СИМВОЛОВ

Термин «распознавание» ментально применяется со словом «образ» и в научных исследованиях представляется в словосочетании как распознавание образов. Сами образы разделяются на следующие категории: абстрактные и конкретные образы [1, с. 360]. Приводя примеры для абстрактных образов, следует представить идеи или аргументы, ну а

примерами для конкретного распознавания служат буквы, символы, рисунки, изображения, сигналы, речь и т. д. В распознавании образов можно разделить общую задачу на три фазы [2]: получение данных, предварительная обработка данных, принятие решения о классификации. Последняя из трех фаз выполняет наиважнейшую, решающую функцию, поскольку на данном этапе происходит представление образов и приближение их к машинному распознаванию.

Далее перейдем к анализу методов, которые применяются в распознавании структурированных символов текстовой информации. Отметим, что технологии распознавания символов реализуются, в основном, тремя традиционными методами [3-5]: структурными, признаковыми, шаблонными. Рассматривая каждый из этих методов, можно сказать, что они ориентированы на свои условия применения, для которых являются эффективными. Так же как и у любых других существующих методов в целом, этим методам присущи свои недостатки. В наибольшей степени это - низкая устойчивость к изменениям масштаба, смещениям, поворотам, смене ракурса и другим искажениям.

Эти недостатки особенно существенно проявились при масштабной эксплуатации программно-технологических систем, использующих в своей основе эти методы, в частности практически у всех систем распознавания структурированных символов точностные характеристики резко падают и становятся ниже технологически приемлемых при искажении аффинными и проекционными преобразованиями. Вместе с тем технологические условия получения информации о маркировке не позволяют полностью устранить эти искажения [6].

Для шаблонных методов характерно то, что происходит сравнение изображения символа со всеми имеющимися в базе системы шаблонами [5, с. 83]. Наиболее подходящим шаблоном считается тот, у которого будет наименьшее количество точек, отличных от исследуемого изображения. Шаблонные методы хорошо распознают дефектные символы (разорванные, склеенные), но основной недостаток шаблонных методов – невозможность распознать шрифт, хоть немного отличающийся от заложенного в систему (размером, наклоном или начертанием).

Признаковые методы [5, с. 85] наиболее распространены. В их основу положено предположение, что можно анализировать не все изображение символа, а только набор признаков, вычисленных по изображению. Подразумевается, что значения признаков несут достаточно информации о символе. Недостатком признакового подхода является то, что распознаванию подвергается не сам символ, а некоторый набор признаков, что может привести к неправильному распознаванию символов.

Структурные методы [4, с. 9-13; 5, с. 83] хранят информацию не о поточечном написании символа, а о его топологии (эталон содержит информацию о взаимном расположении структурных элементов символа). При этом становится неважным размер распознаваемой буквы-образа и шрифт, которым она напечатана. Но главным недостатком в данном случае являются большие ресурсные затраты, требующиеся для реализации данного метода, поскольку при структурном подходе в изображении символов ведется построение скелета, вычисление определенных форм окружностей, угловых и линейных соотношений, пропорций между продольными и поперечными линиями, а также определение пробелов и др.

Из исследования подходов к распознаванию символов описанными выше методами можно сделать выводы о том, что преимуществом первого метода является использование базы шаблонов символов, которые сравниваются с искомым символом изображения, а недостатком данного метода является тот факт, что даже небольшие отличия от заложенных в базе шаблонов символов не позволяют правильно распознать необходимый шрифт. Пре-

имуществом признакового метода является использование вектора признаков изображения, по которым ведется распознавание, недостатком является то, что набор признаков также не дает гарантию точного распознавания изображения. Преимущество структурного метода распознавания заключается в построении скелета символов изображения, который несет информацию о расположении структурных элементов символа, путем вычисления определенных форм, а недостатками являются большие ресурсные затраты.

Рассмотрим более подробно шаблонный метод распознавания структурированных символов, в котором изображения символов представляются в бинарном виде. Точки, принадлежащие данному символу, имеют черный цвет, а точки, имеющие отношения к фону, имеют белый цвет. В обсуждаемом методе шаблон для каждого класса обычно получают, усредняя изображения символов обучающей выборки. Мерой подобия в данном методе выбран коэффициент подобия изображения символа с обобщенным образом S-го класса, выражаемый следующей формулой [5, с. 83]:

$$R_s = \sum_{j=1}^n \left(\frac{\ln P_{js}}{1 - P_{js}} \right) x_j + \sum_{j=1}^n \ln(1 - P_{js}), \quad (1)$$

где R_s - коэффициент подобия эталонного изображения S-го класса символов к опознаваемому символу; P_{js} - вероятность появления черного цвета в j-м элементе эталонного изображения S-го класса; x_j – значение интенсивности, соответствующее j-му элементу распознаваемого символа.

Для черных элементов изображения символа принимает значение $x_j=1$, для белых элементов - $x_j=0$. Изображение символа отождествляется с эталонным классом, давшим максимальный коэффициент подобия R среди всех R_s .

Морфологические методы распознавания структурированных символов. Отметим, что одними из наиболее важных источников информации являются каналы зрительного (визуального) восприятия. Под каналами зрительного восприятия, которые фиксируют необходимые данные и в нужное время, в нужный момент дают возможность к доступу информации, подразумеваются цифровые видеодатчики, видеонаблюдатели или видеокамеры, которые, в свою очередь, под своей функциональностью скрывают сложные, многоуровневые операции и математические методы. Поэтому проблема создания эффективных алгоритмов автоматического анализа изображений, обеспечивающих качественную обработку больших объемов данных, является актуальной. Особенностью разработки таких автоматизированных систем распознавания изображений является создание соответствующего математического описания изображений, которые отражают весь смысл и его содержание. Иначе говоря, данное описание должно показывать только существенные особенности изображения, а не быть зависимым от несуществующих мелочей и деталей.

В рассмотренных традиционных методах распознавания структурированных символов есть свои особенности, а также некоторая ограниченность в условиях их применения. По этой причине также следует рассмотреть методы, которые базируются на применении признаков, инвариантных к аффинным и проективным преобразованиям, которые составляют основу методов морфологического анализа изображений. Морфологический анализ формы изображений, основывающийся на теории множеств, интегральной геометрии, анализе выпуклых функций, стереологии и геометрической теории вероятностей, был разработан J. Serra и Ю.П. Пытьевым [6]. Он позволяет дать количественное описание особенностей геометрической структуры.

Анализируя методы морфологического анализа формы изображений, можно выделить

ключевые моменты, шаги применения и выполнения самих методов: представление сигналов; преобразование сигнала; применение к анализу изображений; улучшение и обнаружение контуров линий; обнаружение пиков, холмов и долин; морфологическое преобразование изображений.

Входными данными для операций морфологии являются бинарное изображение и структурный элемент либо примитив, который относительно друг друга представляется наиболее меньшим элементом. Основными операциями морфологического преобразования являются: параллельный перенос, центральное отражение, дилатация, эрозия, открытие и закрытие. Ниже приведено краткое описание вышеперечисленных операций.

Параллельный перенос предназначен для сдвига всех элементов основного бинарного изображения на заданное расстояние. Центральное отражение основного множества представляет собой поворот основного множества на 180 градусов относительно начала координат. Выполнение дилатации состоит из нескольких шагов:

- осуществление центрального отражения примитива;
- перенос примитива в заданную точку;
- оценка выходного значения.

При эрозии элементу выходного массива присваивается значение 1, если элементы входного изображения и структурного элемента совпадают. Эрозия приводит к сжатию изображения и применяется в основном для удаления и уменьшения некоторых областей, размеры которых меньше структурного элемента.

Открытие (размыкание) представляет собой последовательно выполненные операции эрозии и дилатации. Данная операция к уменьшению размера объекта не приводит, позволяет избавиться от маленьких фрагментов, выступающих во внешнюю границу области.

Операция закрытия (замыкания) представляет собой противоречие операции открытия, т. е. вначале выполняется операция дилатации, затем эрозии. Операция замыкания позволяет замкнуть отверстия области и устраниТЬ заливы области без изменения ее внешних границ.

Таким образом, применяемые морфологические методы предполагают знания конкретных размеров обрабатываемого объекта, кроме того, они являются достаточно чувствительными к зашумленным изображениям.

Отметим, что для каждого рассмотренного метода распознавания структурированных символов присущи свои достоинства и недостатки, и использование этих методов зависит от постановки задачи и оценки их применяемости в той или иной ситуации.

Список литературы

1. Ту Дж. Принципы распознавания образов / Дж. Ту, Р. Гонсалес. - М.: Мир, 1978. - 401 с.
2. Sing-Tze Bow «Pattern Recognition and Image Preprocessing, Second Edition, Revised and Expanded», Dekker, 1992.
3. Ковалевский В.А. О корреляционном методе распознавания // Читающие автоматы. - Киев, 1965. - С. 46-61.
4. Фу К. Структурные методы в распознавании образов / Под ред. М.А. Азейрмана; Пер. с англ. Н.В. Завалишина, С.В. Петрова, Р.Л. Шейнина. - М.: Мир, 1977. - 320 с.
5. Афонасенко А.В. Обзор методов распознавания структурированных символов / А.В. Афонасенко, А.И. Елизаров // Докл. Томского гос. ун-та систем управления и радиоэлектроники. - Вып. 2(18). - 2008. - Ч. 1. - С.83-88.
6. Пытьев Ю.П. Морфологический анализ изображений // Докл. АН СССР. - 1983. - Т. 269. - № 5. - С. 1061-1064.

Получено 7.09.2016