ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 648.8;796:378

В.И. Аверченков, Е.А. Леонов, В.А. Шкаберин, Н.Н. Ивкина,
Ю.В. Крышнев, В.О. Старостенко

МЕТОДЫ ПРОГНОЗИРОВАНИЯ ПАВОДКОВЫХ НАВОДНЕНИЙ НА ОСНОВЕ МОНИТОРИНГА БАССЕЙНОВ ОТКРЫТЫХ ВОДОЕМОВ[1]

Рассмотрены механизмы комплексного анализа данных на основе информации о состоянии уровня воды в притоках крупных бассейнов рек для своевременного предупреждения масштабных паводков и предотвращения их последствий.

Ключевые слова: противопаводковые системы, мониторинг бассейнов водоемов, системный анализ, обработка данных, уровень воды, сложные распределенные системы.

Основной задачей аналитического центра для мониторинга открытых водоемов и противопаводкового предупреждения (рис. 1) является оперативный анализ данных и прогнозирование ситуации в наблюдаемом регионе. Результатом работы аналитического центра являются визуальное представление аналитических данных, специализированные отчеты об актуальной ситуации и разработанных прогнозах. Вся данная информация предоставляется специалисту по чрезвычайным ситуациям посредством специально разработанного веб-интерфейса, в котором также имеется возможность настройки подсистемы экстренного оповещения об ухудшении ситуации и неблагоприятных прогнозах [1 - 3].

Рис. 1. Аналитический центр для мониторинга открытых водоемов и противопаводкового предупреждения

[1] Статья подготовлена в рамках гранта РФФИ «Исследование и разработка метода и автоматической системы противопаводкового мониторинга уровня воды открытых водоемов» (проект № 13-01-90351).

Основой для анализа являются данные о состоянии наблюдаемых объектов, формируемые в центре хранения измерений. Для получения этих данных осуществляется постоянная синхронизация через сеть Интернет с заданными параметрами точности получения измерений.

При синхронизации данных между аналитическим центром и центром хранения измерений выполняется потоковая обработка данных измерений с целью уменьшения их объема. Для этого данные выбираются через неравные интервалы времени Δt_n, при которых значения показаний датчика превысили установленный порог чувствительности Δl_n (рис. 2). Размер порога чувствительности зависит от топологии русла и поймы водного бассейна в месте установки контрольно-измерительного комплекса (КИК). Эта информация хранится в базе знаний о бассейнах водоемов. Количество интервалов и размер чувствительности при сборе данных устанавливаются администратором в зависимости от полноты знаний о месте наблюдения.

$$S_t : M_d, Y_\Delta,$$

где S_t – функция дискретизации сигнала измерений $m_{di} =$, причем для каждого измерения $_\Delta t$ – *const*; Y_Δ – множество пороговых значений для интервалов измерений, $Y_\Delta \ni y_\Delta = (y_n;$ – преобразованное множество измерений $m^*_{di} =$; $_\Delta y$ – *const.*

Установка различных порогов чувствительности для разных уровней обусловлена не только возрастающей опасностью при повышении уровня воды, но и тем, что пойма и тем более терраса водоема имеют более пологий рельеф, следовательно, незначительное повышение уровня воды может приводить к масштабным по площадям затоплениям. Поэтому анализ и прогнозирование для данных порогов уровня воды должны быть значительно точнее.

Таким образом, из потока выбираются только наиболее значимые измерения. При этом чем быстрее изменяется состояние наблюдаемой местности, тем плотнее по времени выбираются контрольные значения.

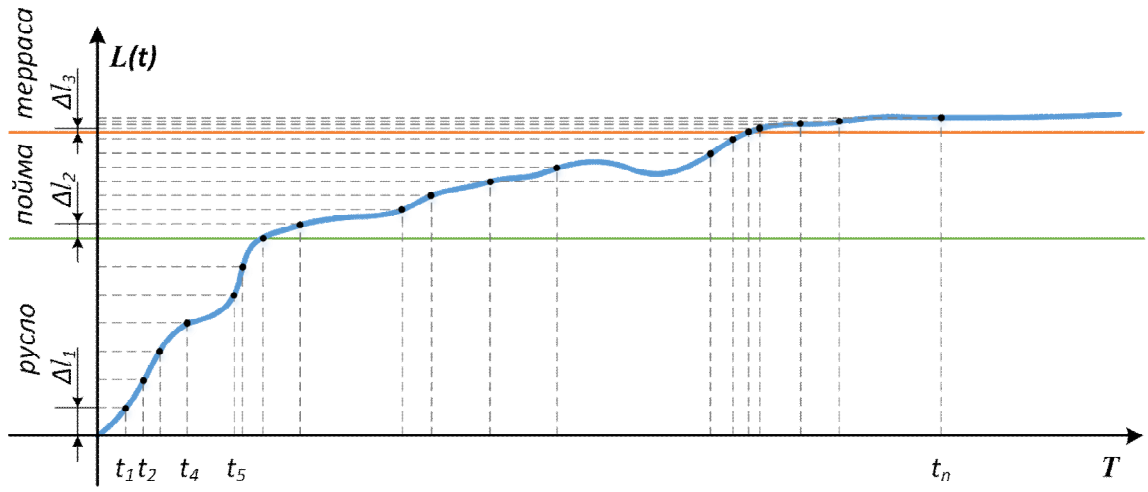

Рис. 2. Пример неравномерной дискретизации данных

Применение неравномерной дискретизации потока измеряемых параметров позволяет в значительной степени уменьшить объем синхронизируемых данных, так как большую часть времени между паводковыми периодами уровень водоемов находится в пределах русел и изменяется незначительно. Такие преобразования необходимы для уменьшения объема исходных данных для последующего анализа и обработки, что значительно ускоряет применяемые методы.

Преобразованные данные хранятся в базе статистических и аналитических данных. При этом неактуальные измерения, имеющие давность более года, используются для создания прогноза в виде статистических данных. Такие устаревшие данные необходимы для построения объективных моделей возможного развития ситуации, а также в связи с циклическим повторением состояний объектов наблюдения. Поэтому даже после дискретизации данных их объем является избыточным для проведения статистического анализа, в связи с чем после потери актуальности все данные измерений подвергаются дополнительной обработке, в ходе которой оставляются лишь округленные среднесуточные измерения. Однако участки измерений, приведшие к развитию чрезвычайной ситуации, оставляются в полном объеме, так как для точного прогнозирования подобных ситуаций необходима более полная и детальная картина их развития.

Также в аналитическом центре имеется база знаний о бассейнах водоемов, используемая не только для корректного сжатия данных при синхронизации с центром хранения измерений, но и для построения корректных моделей при прогнозировании ситуации. Для этого в ней хранится информация обо всех открытых водоемах наблюдаемого региона в виде специальных онтологий.

Описание каждого водоема можно представить в виде кортежа:

$$r = \{t, l, P,$$

где t – тип водоема (река, водохранилище, озеро и т.п.); l – общая протяженность; P – множество контрольных точек измерения состояния, $P' \subseteq P$; R_a – множество водоемов, являющихся притоками; O_r – множество населенных пунктов, расположенных на побережье водоема.

Для прогнозирования возможного ущерба от наводнения в базе хранятся знания о населенных пунктах и технических объектах, расположенных вблизи побережья водоема.

База знаний пополняется различными способами. Любые хранимые знания могут пополняться и корректироваться администратором аналитического центра через специально разработанный веб-интерфейс. Однако данный процесс является чрезвычайно трудоемким, в связи с чем в аналитическом центре имеется специальная подсистема мониторинга Интернета и обнаружения новых знаний. Данная система самостоятельно формирует запросы к поисковым сервисам в сети с целью обнаружения документов, содержащих необходимую информацию. После разбора и анализа найденных документов модуль обнаружения знаний выделяет из документов факты, которые формализуются в виде узлов и связей онтологии базы знаний [2].

Полное покрытие бассейна автоматизированными измерительными комплексами является масштабной и ресурсоемкой задачей. Поэтому в системе имеется возможность занесения среднесуточных измерений контрольных гидропостов в ручном режиме администратором аналитического центра.

В сети существует значительное количество сервисов, которые предоставляют неполные данные об уровне воды в различных участках рек. Множество сайтов предоставляют данные метеоизмерений, такие как изменение температуры воздуха в течение дня, количество и характер атмосферных осадков, а также дают прогнозы по этим данным. Такие данные не могут быть использованы для официальных прогнозов и не дают полной статистической картины по состоянию водных бассейнов. Однако постоянный мониторинг этих ресурсов и комплексное объединение всех источников данных может значительно расширить представление аналитика о ситуации в интересующих его водных бассейнах.

Регулярный сбор и накопление информации из официальных открытых источников позволяет не только провести оценку текущей ситуации, но и повысить точность прогнозирования на основе статистических методов анализа. Алгоритм получения таких данных представлен на рис. 3 [3].

Исходными данными являются координаты долготы и широты объектов, для которых необходимо получить дополнительные измерения. Эта информация выбирается из базы знаний аналитического центра. Так как большинство сервисов получения метеоданных ориентированы на конечных пользователей и информация о данных измерений предоставляется по населенным пунктам, то предварительно осуществляется запрос на нахождение ближайшего населенного пункта по координатам к геоинформационным ресурсам, таким как Google Maps или Яндекс.Карты. Далее из базы данных доверенных источников выбираются адреса сервисов необходимых измерений, к которым осуществляются запросы на получение данных. Данные получаются в виде веб-документа, который разбирается на основании составленного ранее шаблона. Шаблон определяет структуру возвращаемого документа и инструкции по его разбору для получения отдельных данных измерений без остального контента документа.

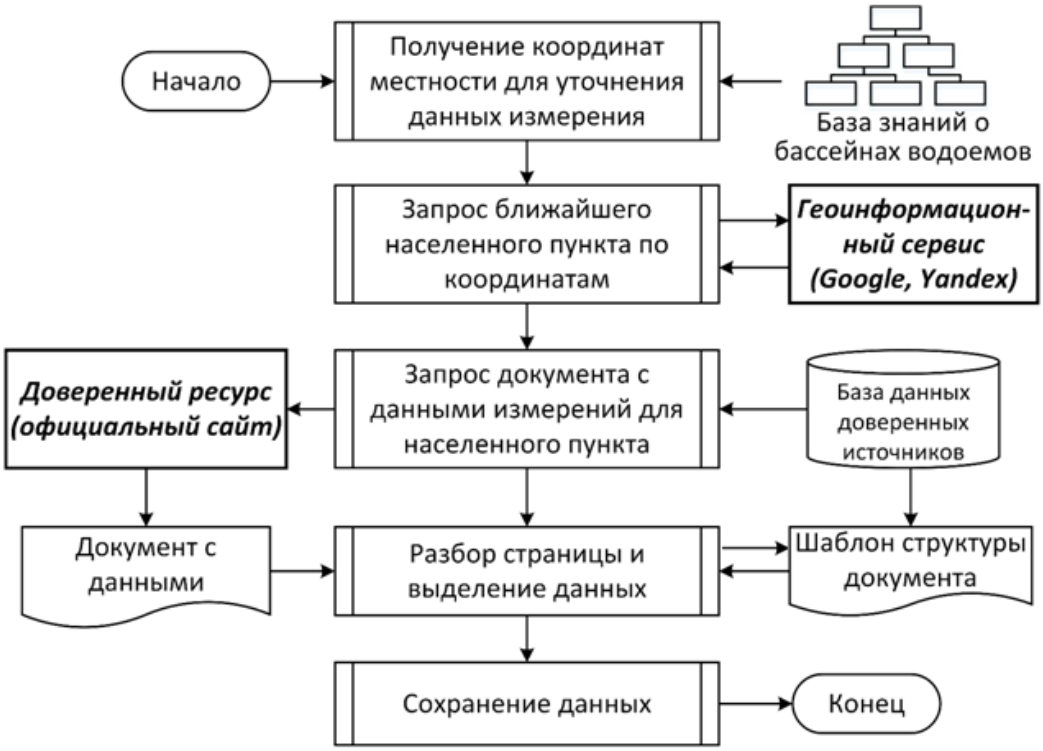

Рис. 3. Алгоритм получения данных измерений из сторонних источников

После этого все полученные данные измерений сохраняются в базе статистики. Вся статистика и результаты измерений, хранимые в аналитическом центре, маркируются по источнику их получения, так как от этого зависит степень доверия к данным, что учитывается в методах прогнозирования и анализа ситуации.

Одним из центральных узлов системы является модуль прогнозирования. Одной из основных его задач является расчет уровня воды в наблюдаемых точках на основании знаний о водных бассейнах, накопленных статистических данных и текущей ситуации. Простые однофакторные методы экстраполяции данных для этой задачи не подходят, так как уровень воды в водоеме зависит не столько от текущей динамики, сколько от сопутствующих факторов. Текущая динамика роста уровня воды является проявлением многих факторов, таких как скорость роста температуры (при наличии снежного и ледяного покрова), объем атмосферных осадков, география бассейна рек и водоемов (притоки), топология местности (скорость и направление стока вод), геологические данные бассейна реки (определяют скорость просачивания и отвод вод в почве) и многие другие [7].

Полный комплексный анализ всего бассейна является чрезвычайно сложной многофакторной задачей, для решения которой должны быть разработаны эффективные математические модели среды, учитывающие прямое взаимовлияние различных факторов.

Однако в первом приближении задача прогнозирования может быть решена путем применения статистических методов анализа временных рядов, в частности многофакторной экстраполяции с учетом динамики воздействия факторов [8].

Каждый прогнозируемый показатель y_t (t=1,2…,n) можно рассматривать как функцию от нескольких факторов-аргументов в виде нелинейной многофакторной модели (степенного типа):

$$y_t = a_0 x_{1t}^{a_1} x_{2t}^{a_2},$$

где a_0, a_j – коэффициенты модели при j=1, 2, … , m.

Данная модель преобразуется в линейную путем логарифмирования.

Коэффициенты a_0, a_j определяются с помощью метода наименьших квадратов:

$$\sum_t (y_t - \hat{y}_t)^2, \qquad (1)$$

где – вид исследуемой функции.

Тогда из системы нормальных уравнений, представляющих собой частные производные по a_0, a_j, равные нулю, можно сформировать следующую систему:

$$\frac{dy}{da_0} = 2\sum_t (y_t - a_0 - a_1 x_1 - \cdots - a_j x_j - \cdots - a_k x_k)(-1) = 0;$$

$$\frac{dy}{da_j} = 2\sum_t (y_t - a_0 - a_1 x_1 - \cdots - a_j x_j - \cdots - a_k x_k)(-x_j) = 0.$$

В результате решения данной системы уравнений находятся такие a_0 и a_j, при которых выражение (1) стремится к нулю.

Все используемые факторы-аргументы являются хранимыми измерениями, имеют перспективные оценки значений на прогнозируемый период и являются линейно независимыми.

Факторы считаются независимыми (мультиколлинеарными), так как линейный (парный) коэффициент корреляции любых двух факторов менее 0,8. В модели были оставлены только те, которые имели больший коэффициент корреляции с функцией *y*.

Коэффициент парной корреляции определяется по формуле

$$r = \text{-}, \qquad (2)$$

где S_{11}, S_{22}, S_{12} – соответственно остаточные дисперсии для функции, фактора-аргумента и их произведения:

$$S_{11} = \sum y^2;$$
$$S_{22} = \sum t^2;$$
$$S_{12} = \sum yt.$$

В нелинейной модели для каждого *j*-го фактора-аргумента определяются оценки Стьюдента по формуле

$$t_a.$$

Для нахождения коррелирующих факторов выбираются наименьшие значения оценки *min* t_{a1} и сравниваются с табличным значением t_p при *n-k-1* степенях свободы и выбранном уровне значимости p=0,05. Если минимальная из рассчитанных оценок $t_a \geq t_p$, то модель оставляется в полученном виде. Если же $t_a < t_p$, то фактор a_1 исключается из модели как незначимый.

С оставшимися факторами строится новая модель, определяются новые значения оценок Стьюдента, после чего находится минимальная из них и т.д. до тех пор, пока в модели не останутся все значимые факторы.

Тесноту связи между функцией и факторами-аргументами можно установить с помощью квадрата коэффициента множественной корреляции:

$$R^2 = \frac{a_1 S_{yk_1} + a_2 S_{yk_2} + \dots + a_m S_{yk_m}}{S_{yy}};$$

$$S_{yk_m} = \sum_t y_t \dots$$

Квадрат коэффициента множественной корреляции показывает, какая часть общего рассеяния зависимой переменной может быть объяснена функцией линейного или нелинейного вида.

Статистическая надежность многофакторной регрессионной модели (или коэффициента детерминации) устанавливается с помощью известного критерия Фишера:

$$F = \frac{(n \cdot}{(m-}$$

где n – число данных; m – число факторов-аргументов в модели; R^2 – квадрат коэффициента множественной корреляции.

Если расчетное значение критерия Фишера превышает табличное при (n-m и m-1 степенях свободы и уровне значимости p=0,05, то модель признается статистически надежной и значимой.

Многофакторная регрессионная модель может быть использована для прогнозирования не более трехлетнего периода упреждения. Ошибки прогноза определяются по следующим формулам:

$$\delta_a = \frac{\Sigma}{};$$

$$\delta_0 = \left[\sum_t \frac{|y_t - y_t^*|}{y_t} 100\%\right] / n.$$

При нахождении коэффициента парной корреляции по зависимости (2) необходимо учитывать временную задержку влияния фактора-аргумента на прогнозируемое значение, так как некоторые из факторов могут иметь инерцию. Для этого при поиске корреляции выполняется временное смещение между фактором-аргументом и прогнозируемым показателем и выбирается постоянная задержка с максимальным значением корреляции.

Предлагаемая регрессионная модель многофакторной экстраполяции не является исчерпывающей и требует уточнения путем создания экспертами предметной области математических моделей, определяющих прямые зависимости между факторами и прогнозируемыми показателями. Также точность прогнозирования может быть повышена посредством использования в качестве аппроксиматоров искусственных нейронных сетей и генетических алгоритмов.

Аналитический центр имеет веб-интерфейс для специалиста по чрезвычайным ситуациям, который предназначен для отслеживания текущей ситуации в наблюдаемых районах, а также имеет возможность создания оповещений (рис. 4). Основной частью интерфейса являются интерактивные карты, на которых показаны пункты наблюдений. Каждый из них отмечается цветом, зависящим от оценки текущей ситуации системой в месте наблюдения. Просматриваемые карты являются полностью интерактивными, позволяя изменять масштаб, область просмотра, вид схемы и спутниковых снимков. При выборе отдельного поста измерений уровня водного бассейна можно просмотреть статистику данных за последний период в виде графиков показаний датчиков. Графики также являются интерактивными и позволяют гибко менять масштаб и область просмотра данных, накладывать различные параметры измерений для экспертного анализа возможности развития ситуации. Интерфейс построен как многооконное веб-приложение, что позволяет обеспечить доступ к нему с любого компьютера, подключенного к Интернету.

Многооконный интерфейс позволяет одновременно просматривать данные интересующих объектов наблюдения, выполнять запрос отчетов и обработку данных, параллельно осуществляя работу с системой. Веб-приложение построено на базе технологий и пара-

дигмы web 2.0, все интерфейсы взаимодействуют с сервером посредством асинхронных запросов. Для длительных процедур используются вызовы SSE. Репликация изменений, вносимых администратором аналитического центра, инженером по обслуживанию сети КИК и специалистом по чрезвычайным ситуациям, осуществляется посредством организации запросов, построенных на технологии web-socket. Таким образом, все интерфейсы являются веб-приложениями реального времени.

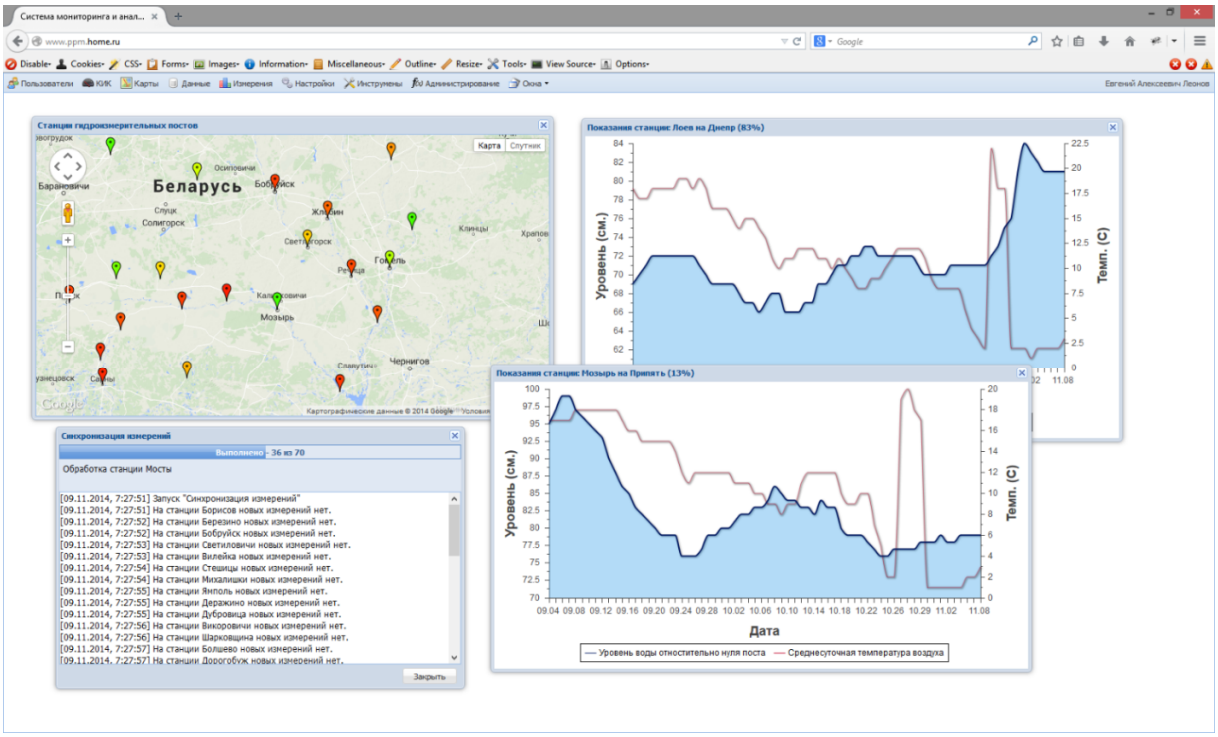

Рис. 4. Пример веб-интерфейса аналитической системы мониторинга состояния бассейнов водоемов

Аналитический центр мониторинга бассейнов открытых водоемов и противопаводкового предупреждения имеет возможность пополнять знания о наблюдаемых водоемах, а также собирать дополнительные данные посредством подсистемы мониторинга Интернета, что обеспечивает частичную автоматизацию деятельности администратора и возможность пополнения данных из открытых источников информации (геоинформационные системы, метеорологические службы и др.) для проведения комплексного анализа и прогнозирования.

Для прогнозирования контрольных показателей предложена многофакторная регрессионная модель экстраполяции на основе анализа временных рядов. Однако качественная диагностика модели и расчет погрешности прогнозирования требуют значительного времени наблюдений, так как наблюдаемые явления имеют четко выраженную цикличность сезонных изменений в течение года.

Для полноценной реализации предложенных подходов требуются значительные капитальные вложения для построения масштабной сети контрольно-измерительных комплексов, а также проведение длительного мониторинга и сбора статистических данных для проверки моделей прогнозирования, что требует объединения данных измерений различных служб на государственном и межгосударственном уровнях, так как бассейны водоемов не разделены по политическим границам государств.

СПИСОК ЛИТЕРАТУРЫ

1. Аверченков, В.И. Разработка принципов создания автоматизированной системы для мониторинга уровня воды в открытых водоемах / В.И. Аверченков, В.А. Шкаберин // Вестн. Брян. гос. техн. ун-та. – 2013. – №4. – С. 143-148.

2. Аверченков, В.И. Автоматизация процедур противопаводкового мониторинга уровня воды открытых водоемов / В.И. Аверченков, В.А. Шкаберин, Я.И. Лепих, В.И. Сантоний, Ю.В. Крышнев // Изв. ВолГТУ. – 2014. – Вып. 20. – №6 (133). – С. 92-97.
3. Аверченков, В.И. Исследование методов автоматизации противопаводкового мониторинга уровня воды открытых водоемов / В.И. Аверченков, В.А. Шкаберин, Н.Н. Ивкина, Я.И. Лепих, Ю.В. Крышнев // Михайло-Архангельские чтения: материалы VIII Междунар. науч.-практ. конф. (15 нояб. 2013 г., г.Рыбница). – Рыбница, 2013. – С. 310-311.
4. Игнатович, Н. IBM MQSeries: архитектура системы очередей сообщений / Н. Игнатович // Открытые системы. – 1999. – № 9-10.
5. Cios, Krzysztof J. Data Mining: A Knowledge Discovery Approach / Krzysztof J. Cios. – Springer, 2007. – P. 123.
6. Gray, J. Transaction Processing: Concepts and Techniques / J. Gray, A. Reuter // The Morgan Kaufmann series in data management systems. – 1993.
7. Leonov, E.A. Architecture and Self-learning Concept of Knowledge-Based Systems by Use Monitoring of Internet Network / E.A. Leonov, V.I. Averchenkov, A.V. Averchenkov, Y.M. Kazakov, Y.A. Leonov // Knowledge-Based Software Engineering: Communications in Computer and Information Science. – Springer International Publishing, 2014. - Vol. 466. - P. 15-26.
8. Тюрин, Ю.Н. Статистический анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров; под ред. В.Э. Фигурнова. – М.: ИНФРА – М, 1998. – 528с.

Материал поступил в редколлегию 19.11.14.