Чертов О. Інтелектуальний аналіз демографічних даних / Чертов О. // Вісник ТНТУ. — 2010. — Том 15. — № 3. — С. 132-140. — (приладобудування та інформаційно-вимірювальні технології).

УДК 004.931: 517.518.3

О. Чертов, кандидат технічних наук

Національний технічний університет України «Київський політехнічний інститут»

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДЕМОГРАФІЧНИХ ДАНИХ

Резюме. *Розглянуто теоретичні передумови та практичні аспекти застосування інтелектуального аналізу даних до опрацювання демографічних даних. Проаналізовано відмінності між інформаційними системами різного типу (облікового, аналітичного, з можливостями інтелектуального аналізу даних) у галузі статистики населення.*

Ключові слова: *інтелектуальний аналіз даних, виявлення знань у базах даних, статистика населення, демографічні дані, мікрофайл.*

O. Chertov

INTELLECTUAL ANALYSIS OF DEMOGRAPHIC DATA

The summary. *In the paper, we discuss theoretical background and practical aspects of applying data mining to demographic data processing. The differences between information systems of various kind (OLTP, OLAP, and ones with data mining capabilities) in vital statistics area are analyzed.*

Key words: *data mining, knowledge discovery in databases, vital statistics, demographic data, microfile.*

Вступ. Обсяг цифрової інформації у світі зростає за експонентним законом, причому якщо в 2003 р. вважалося, що цей обсяг подвоюється через кожні три роки [1], то в 2009 р. маємо ще вищу оцінку – через кожні 18 місяців [2]. Для адекватного аналізу інформації, що накопичується, потрібні відповідні методи та системи, в описі яких прийнято використовувати прикметник «інтелектуальний».

Інтелектуальний аналіз даних (ІАД) – це найуживаніший і, на думку автора, найбільш адекватний переклад англомовного терміну «data mining». Цей термін отримав свою алегоричну назву від сполучення двох слів: *data*, тобто дані чи, в ширшій інтерпретації, – інформація, знання, та *mining*, тобто гірнича справа, видобуток корисних копалин. Наразі під *інтелектуальним аналізом даних* розуміють нетривіальний процес виявлення коректних, раніше невідомих, потенційно корисних і придатних для інтерпретації закономірностей в даних [3, с. 58]. Системи ІАД знайшли широке застосування в наукових дослідженнях (пошук маркерів у молекулярній генетиці та розшифровка генома людини, аналіз складних хімічних сполук), інженерній практиці (діагностичний контроль технічного стану високовольтного обладнання), комерційній сфері (аналіз кошика покупця в роздрібній торгівлі, виявлення лояльності користувачів мобільним зв’язком, попередження шахрайства з кредитними картками чи взяттям кредитів) тощо [4, 5].

Але в галузі статистики, зокрема статистики населення і демографії, на даний момент методи ІАД використовуються спорадично, відповідні системи не інтегровані з системами прикладної статистики та системами, призначеними для оперативного аналізу даних.

Тому *актуальними* є роботи, направлені на застосування методів ІАД саме до демографічних даних і взаємно доповнююче використання можливостей систем інтелектуального та класичного статистичного аналізу даних.

Аналіз останніх досліджень. Як правило, до ІАД відносять такі задачі (tasks), чи, в іншій термінології, закономірності (regularities) або техніки (techniques): класифікація (classification), кластеризація (clustering), прогнозування (forecasting), асоціація (associations), візуалізація (visualization), виявлення відхилень (deviation detection),

оцінювання (estimation), аналіз зв'язків (link analysis), підбиття підсумків (summarization) [6].

Необхідно відзначити принципову відмінність методів ІАД від такого близького на перший погляд (стосовно аналізу історичних статистичних даних) напряму досліджень, як аналіз часових рядів. Справа в тому, що перед застосуванням методів ІАД ми можемо навіть не здогадуватися, які саме змінні в даних, що аналізуються, корелюють одне з одним чи мають певний причинно-наслідковий зв'язок. Більше того, точні параметри статистичної моделі (наприклад, коефіцієнти моделі ARMA чи ваги нейронної мережі) нас, як правило, не цікавлять, проте цікавлять неявні закономірності, що існують у даних. Наприклад, які комбінації товарів одночасно купують клієнти в супермаркетах [7, с. 174].

Наразі використовується ціла низка стандартів в ІАД [6, 8, 9], основні з яких ми записали у вигляді табл. 1.

Таблиця 1. Стандарти ІАД

Область застосування	*Назва*	*Розробник: короткий опис*	*Поточний стан*
Процеси пошуку закономірностей	*CRISP-DM (CRoss Industry Standard Process for Data Mining) [10, 11]*	*SPSS, DaimlerChrysler, NCR: методологія процесів ІАД та розроблення відповідних проектів*	*Найпоширеніший, але розроблення версії 2.0 призупинена на початку 2007 р.*
	SEMMA – sample, explore, modify, model, assess [12]	*SAS: організація набору функціональних засобів для реалізації ключових задач ІАД*	*Покладено в основу SAS Enterprise Miner*
XML-стандарти (зберігання та пересилання моделей)	*Predictive Model Markup Language (PMML) [13, 14]*	*DMG: опис моделей ІАД та статистичних моделей*	*Версія 4.0 вийшла в червні 2009 р.*
	Common Warehouse Model for Data Mining (CWM-DM) [15]	*OMG: опис метаданих для обміну між сховищами даних і системами ІАД тощо*	*Версія 1.1 вийшла в березні 2003 р.*
	XML for analysis (XMLA) [16]	*Microsoft, Oracle (Hyperion), SAS: використовуючи SOAP, визначає доступ до даних при взаємодії (через Інтернет) клієнта з постачальником послуг OLAP чи ІАД*	*Підтримується, зокрема, такими продуктами, як Oracle Essbase, Microsoft Analysis Services, SAP NetWeaver*
Уніфікація інтерфейсів	*SQL/MM Part 6: Data mining [17]*	*ISO: визначає SQL-процедури обчислень із використанням ІАД*	*Діючий стандарт ISO/IEC 13249-6:2006*
	Microsoft OLE DB for Data Mining [18]	*Microsoft: розширення OLE DB для застосування методів ІАД у структурі реляційних баз даних*	*Підтримується, починаючи з версії Microsoft SQL Server 2000*
	Java Data Mining [19]	*Java Community Process: визначає об'єктну модель та Java API для об'єктів і процесів ІАД*	*У 2006 р. була розроблена версія 2.0 (JSR 247)*
Web	*Semantic Web [20]*	*W3: інтегроване середовище для представлення інформації в машинозчитувальній формі, придатній для ІАД*	*Спирається на XML, RDF, OWL та активно розвивається*

Світова статистична спільнота починає усвідомлювати переваги застосування методів ІАД у своїй повсякденній діяльності. Про це свідчить, наприклад, розробка Бюро перепису США системи DataFerrett (Federated Electronic Research, Review, Extraction, and Tabulation Tool) для ІАД та здобування знань. На момент підготовки даної статті була доступна бета-версія цієї системи [21].

Наведемо постановку конкретної задачі ІАД з класифікації демографічних даних і розглянемо типові шляхи її розв'язання за допомогою сучасних інформаційних технологій.

Постановка завдання. Маємо багатовимірний масив демографічних даних, отриманих у результаті визначеного спостереження в галузі статистики населення. Причому, нехай первинні статистичні дані зведено до мікрофайлів (див. табл. 2), де кожному респонденту r_k, $k=\overline{1,\mu}$ співставлено його значення z_{kn} атрибутів u_n, $n=\overline{1,\eta}$.

Таблиця 2. Дані мікрофайлу, оформлені у вигляді таблиці

		Атрибути			
		u_1	u_2	…	u_η
Респонденти	r_1	z_{11}	z_{12}	…	$z_{1\eta}$
	r_2	z_{21}	z_{22}	…	$z_{2\eta}$
	…	…	…	…	…
	r_μ	$z_{\mu 1}$	$z_{\mu 2}$	…	$z_{\mu\eta}$

Таким чином, передбачається, що будь-яка первинна неструктурована інформація впорядкована та занесена до мікрофайлу наведеної в табл. 2 структури.

Метою роботи є демонстрація на конкретному прикладі потужних можливостей взаємно доповнюючого аналізу зазначених даних за допомогою методів аналітичного та інтелектуального опрацювання інформації.

Вихідні дані. Держкомстат України ще не готує мікрофайли зі статистичними даними, тому в якості вихідних даних візьмемо 5-відсоткові мікродані, що стосуються перепису населення, проведеного Бюро перепису США [22] у 2000 р. При цьому ми вирішили обмежитися даними по штату Каліфорнія як найчисельнішому за населенням. Якби Каліфорнія була окремою державою, то вона б займала 35-е місце у світі за кількістю населення.

Відомо [23], що за останні роки Каліфорнія має від'ємне сальдо внутрішньої міграції, тобто кількість тих, хто виїхав до інших штатів перевищує кількість прибулих більше, ніж на 1,3 млн. осіб (з 2000 до 2008 р.). Припустимо, що губернатор Каліфорнії вирішив компенсувати зазначені втрати населення шляхом підвищення народжуваності й поставив задачу знайти основні чинники, які б могли допомогти це зробити.

Із загальних міркувань, без проведення додаткових соціологічних досліджень, можна припустити, що на рішення завести дитину впливають: матеріальні статки родини, кількість уже наявних дітей в сім'ї та етнічне походження подружжя.

Із вказаного мікрофайлу даних по Каліфорнії (з урахуванням вагів елементів початкової вибірки) були відібрані сім'ї з працездатним подружжям, обоє віком від 18 до 50 років, з відомою етнічною належністю. Загальний розмір цієї вибірки склав

112347 осіб, причому виявилося, що за етнічними показниками в ній присутні тільки «чисті сім'ї», тобто чоловік і дружина належать до однакової етнічної групи. Розподіл виділених родин за етнічною ознакою наведено в табл. 3.

Таблиця 3. Загальна кількість сімей-представників різних етнічних груп у вибірці даних по Каліфорнії (5-відсоткові мікродані перепису населення 2000 р.)

Етнічна група	Кількість сімей-представників
Західна Європа	42247
Східна Європа та СРСР	5065
Латинська Америка	33749
Західна Індія	189
Центральна та Південна Америка	74
Північна Африка та Південно-Західна Азія	2629
Райони Сахари	491
Південна Азія	2414
Інші райони Азії	13701
Тихоокеанське узбережжя	454
Північна Америка	11334

Для спрощення розв'язуваної задачі в якості параметра, що визначає матеріальний стан сім'ї, візьмемо лише один атрибут, а саме: «Family Total Income in 1999» («сукупний дохід родини за 1999 р.»).

Три архітектурні рівні інформаційних систем у галузі статистики населення. З архітектурної точки зору сучасні системи опрацювання даних статистики населення повинні мати три рівні [24, с. 228]:

– інформаційно-обліковий (OLTP) рівень [25], що забезпечує базову функціональність – уведення даних і матеріалів відповідних статистичних спостережень і опитувань, їх структуроване (як правило, за допомогою СКБД – системи керування базами даних) зберігання та облік, контроль у первинному та зведеному вигляді, розповсюдження результатів через різноманітні регламентні вихідні таблиці;

– інформаційно-аналітичний (OLAP) рівень [26], за допомогою якого користувачі можуть швидко будувати нерегламентні таблиці й проводити інші аналітичні дослідження статистичних даних, шукаючи передбачувані закономірності їх розподілу;

– рівень ІАД, котрий бере на себе найбільш громіздку та рутинну аналітичну операцію з пошуку прихованих закономірностей, що, можливо, існують у даних, які аналізуються.

Виділені рівні суттєво відрізняються один від одного (див. табл. 4).

Таблиця 4. Порівняльні характеристики різних рівнів інформаційних систем статистики населення

Характеристика	*Інформаційно-обліковий рівень*	*Інформаційно-аналітичний рівень*	*Рівень ІАД*
Характер (рівень) даних	*В основному первинні*	*В основному консолідовані*	*Первинні та консолідовані*
Мінливість даних	*Висока (з кожною транзакцією)*	*Низька*	*Низька*
Типова операція	*Зміна інформації*	*Аналіз даних*	*Пошук закономірностей*
Звіти	*Регламентні*	*Нерегламентні, але за визначеним переліком атрибутів*	*Нерегламентні, перелік атрибутів може визначатися динамічно*
Дані, що оброблюються	*Тільки поточні, історичні, як правило, в архіві*	*Історичні та поточні*	*Історичні та поточні*
Базова структура	*Таблиця / первинний ключ*	*Куб / вимір*	*Кластер, клас, асоціативне правило тощо*
Пріоритет	*Продуктивність*	*Гнучкість*	*Інтелектуальність, різноплановість*

Розглянемо послідовно, яким чином на зазначених рівнях може бути розв’язана поставлена задача.

Можливості інформаційно-облікових систем (на прикладі). Інформаційно-облікові системи дозволяють здійснювати підрахунок розрахункових показників (типу розмір житлової площі на одного члена домогосподарства), будувати регламентні вихідні таблиці та звіти й виконувати нерегламентовані запити, тобто репрезентативні можливості таких систем досить обмежені. Тому якщо ще на стадії проектування системи не була передбачена побудова необхідної вихідної таблиці, то отримати на практиці відповідні дані дуже проблематично.

Наприклад, можливостей інформаційно-облікової системи, яка була створена для опрацювання даних Всеукраїнського перепису населення 2001 р. [27], було б замало для розв’язання поставленої раніше задачі. Серед усіх вихідних таблиць, що будувалися цією системою, можна виділити такі, які хоча б якимось чином близькі до поставленої задачі: таблиця 2.2 «Розподіл населення за шлюбним станом, статтю і віком», таблиця 2.3 «Розподіл жінок за віком, шлюбним станом та кількістю народжених дітей», таблиця 5.7 «Розподіл населення найбільш багаточисельних національностей за шлюбним станом, статтю та віком», таблиця 7.3 «Розподіл членів домогосподарств за віком та статтю», таблиця 7.5 «Розподіл індивідуальних домогосподарств, які складаються з двох і більше осіб, за розміром та кількістю дітей віком до 18 років», таблиця 7.6 «Розподіл індивідуальних домогосподарств за їх розміром та національністю членів домогосподарств», таблиця 7.18 «Розподіл сімейних осередків у різних типах індивідуальних домогосподарств за розміром та кількістю дітей віком до 18 років».

Але в усіх наведених таблицях відсутня інформація про дохід родини, а також ще одна чи кілька інших характеристик, принципових для розв’язуваної задачі.

Можливості інформаційно-аналітичних систем (на прикладі). Ввівши такі виміри, як: «кількість дітей в сім’ї», «сукупний дохід родини за 1999 р.», «етнічне походження» та, взявши в якості міри «кількість респондентів» чи безпосередньо – «кількість сімейних осередків», можна побудувати багатовимірний куб за сімейними осередками,

як, наприклад, це було зроблено в автоматизованій системі «Перепис-2001 Аналітик» [28]. Користуючись побудованим кубом, можна б було за лічені секунди отримати різноманітну інформацію про співвідношення сімей за виділеними атрибутами. Як приклад, див. табл. 5.

Таблиця 5. Співвідношення між розмірами доходів та кількістю дітей для кожної етнічної групи за вибіркою даних Каліфорнії (5-відсоткові мікродані перепису населення, 2000 р.)

Етнічна група	Дохід *10^5, долари США								
	є борги	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	>7
Західна Європа	1,261	1,343	1,224	1,319	1,590	1,331	1,518	1,174	1,084
Східна Європа та СРСР	0,442	1,280	1,192	1,311	1,632	1,422	1,679	1,392	1,788
Латинська Америка	1,867	1,989	1,478	1,687	1,869	1,903	1,215	2,237	1,800
Західна Індія	-	1,341	1,347	2,000	1,828	0,464	-	-	-
Центральна та Південна Америка	-	1,050	1,246	2,640	1,000	-	-	-	-
Північна Африка та Південно-Західна Азія	0,826	1,506	1,469	1,272	1,724	1,644	2,045	1,076	1,909
Райони Сахари	-	1,543	1,575	1,854	1,910	1,588	0,000	0,000	1,459
Південна Азія	2,355	1,309	1,103	1,400	1,587	1,368	1,651	1,157	1,841
Інші райони Азії	2,221	1,483	1,356	1,478	1,508	1,565	1,120	2,084	1,070
Тихоокеанське узбережжя	-	1,754	1,753	1,646	2,281	-	2,000	-	-
Північна Америка	1,945	1,474	1,259	1,168	1,591	1,035	1,267	1,191	-

На основі даних табл. 5 можна зробити два попередніх висновки. Перший – матеріальне стимулювання народжуваності має певні перспективи лише для вихідців зі Східної Європи та СРСР. Другий – наше припущення, що матеріальні статки та етнічне походження родини мають вирішальний характер при прийнятті рішення подружжям про народження нової дитини, є дещо перебільшеним. Хоча, без сумніву, зазначені фактори не є останніми.

Ситуація, в якій ми опинилися, є достатньо типовою. Справа в тому, що інформаційно-аналітичні системи дозволяють легко досягти результату за рахунок швидкого отримання різноманітних аналітичних звітів, але лише у випадку, коли нам відомий правильний шлях до отримання необхідних даних.

Можливості систем інтелектуального аналізу даних (на прикладі). Для виявлення прихованих закономірностей, що зв’язують дані, які ми досліджуємо, спробуємо розділити їх на класи.

Параметрами класифікації візьмемо «сукупний дохід родини за 1999 р.» та «кількість дітей у сім'ї». Початкове співвідношення між інтервалами даних (кількість дітей змінюється від 0 до 10, сукупний дохід міг досягати мільйонів доларів США) негативно впливає на процес класифікації при використанні Евклідової метрики. Тому обидва параметри було пронормовано, що забезпечило однаковий вплив кожного з них на процес формування класів.

Вихідна множина даних була розбита на 20 класів (рис. 1). Саме така кількість забезпечила формування «ефективних» класів. Використання ж меншої кількості груп призводить до утворення дуже великих сукупностей, в яких важко визначити закономірності.

Навіть візуальний аналіз рис. 1 дозволяє виділити два класи (з номерами 6 та 20), в яких збільшення матеріального добробуту, як правило, викликає прийняття родиною рішення про заведення ще однієї дитини. Аналіз етнічного походження родин, наприклад, класу №6 (див. табл. 6) показує, що цей атрибут не є розділяючим. Тому потрібен подальший аналіз респондентів виділених класів з урахуванням, наприклад, освіти, вікового діапазону подружжя тощо.

Аналогічна ситуація має місце і в загальному випадку: методи ІАД не замінюють дослідника, а лише вказують перспективні напрями подальшого аналізу даних.

Для попереднього оцінювання корисності атрибутів для розв'язання поставленої задачі можна скористатися можливостями інформаційно-облікових та аналітичних систем, а потім знову повернутися до рівня ІАД.

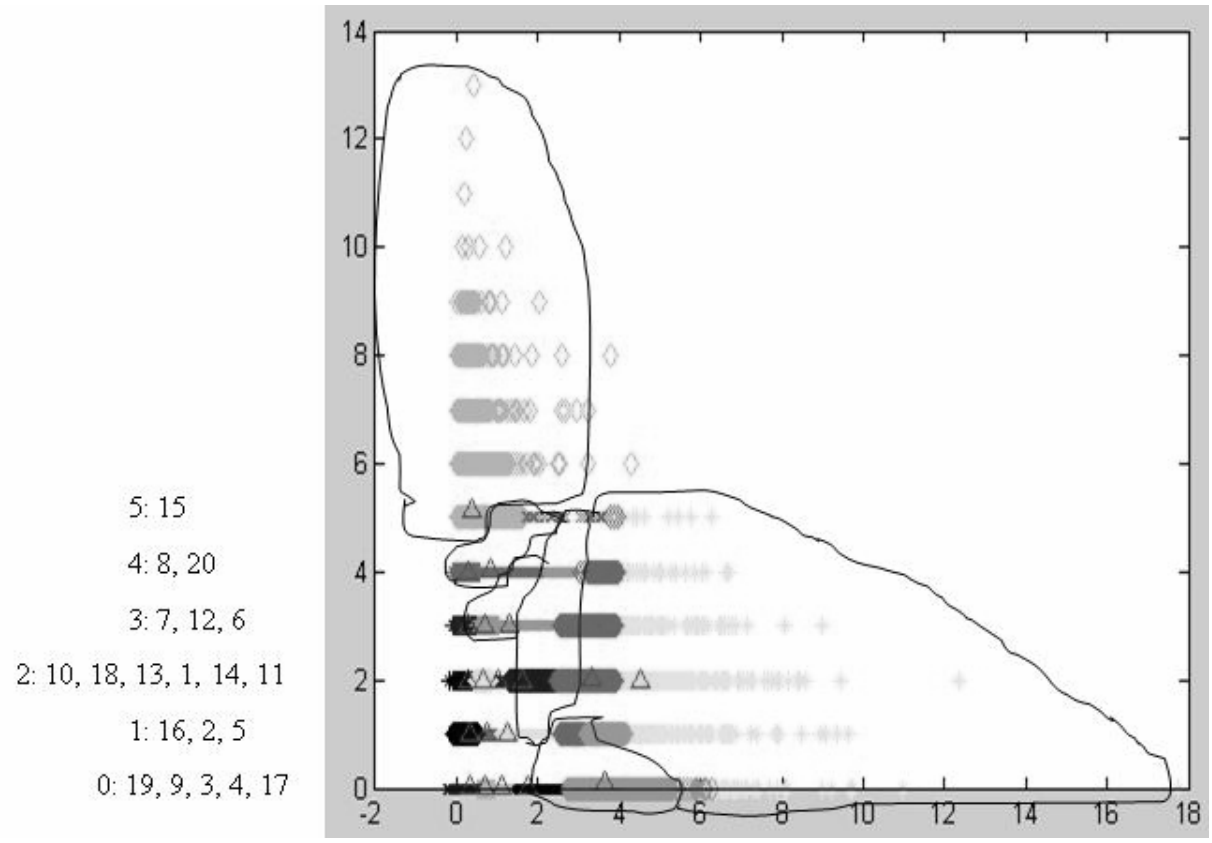

Рис. 1. Класифікація множини даних (вісь абсцис – нормований дохід, вісь ординат – кількість дітей у сім'ї, трикутники позначають центри класів, до двокрапки вказана середня кількість дітей у родинах відповідних класів)

Таблиця 6. Співвідношення між розмірами доходів та кількістю дітей для кожної етнічної групи за вибірці даних Каліфорнії
(5-відсоткові мікродані перепису населення 2000 р.)

Етнічна група	Відсоток від загальної кількості сімей з таким же етнічним походженням	Абсолютна кількість сімей
Західна Європа	2,43	1028
Східна Європа та СРСР	2,67	135
Латинська Америка	1,00	337
Західна Індія	2,12	4
Центральна та Південна Америка	2,70	2
Північна Африка та Південно-Західна Азія	2,85	75
Райони Сахари	2,04	10
Південна Азія	1,45	35
Інші райони Азії	2,17	297
Тихоокеанське узбережжя	3,08	14
Північна Америка	1,91	217

Висновки та перспективи розвитку досліджень. Наведений у статті пошук розв'язання конкретної задачі з опрацювання даних мікрофайлу перепису населення висвітлив основні недоліки та переваги кожного з трьох виділених архітектурних рівнів інформаційних систем у галузі демографії та статистики населення.

На інформаційно-обліковому рівні за допомогою регламентних звітів просто знаходяться відповіді на найбільш масові, типові запитання користувачів (типу статево-вікового складу населення певної території). Можливості інформаційно-аналітичного рівня нам знадобляться тоді, коли потрібно швидко співставити значення атрибутів респондентів, які слід динамічно визначити, особливо, якщо мова йде про історичні дані чи різні адміністративно-територіальні розподіли. ІАД є незамінним під час пошуку прихованих закономірностей в даних, хоча, звичайно, він лише автоматизує рутинні операції і не може замінити роботу дослідника.

Викладений у статті підхід зі взаємно доповнюючого застосування різних рівнів опрацювання статистичних даних планується впровадити в автоматизовану систему опрацювання даних Всеукраїнського перепису населення, проведення якого призначено на 2011 р., та в Інтегровану систему опрацювання статистичних даних України, розроблення котрої розпочато в рамках проекту розвитку системи державної статистики України для моніторингу соціально-економічних перетворень при підтримці Міжнародного банку реконструкції та розвитку [29].

Література

1. Lyman P., Varian H.R. How Much Information, 2003, [Електронний ресурс]. – Режим доступу: http://www.sims.berkeley.edu/how-much-info-2003
2. Gantz J.F., Reinse, D. As the Economy Contracts, the Digital Universe Expands. An IDC Multimedia White Paper, 2009. – Режим доступу: http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm
3. Frawley W.J., Piatetsky-Shapiro G., Matheus C.J. Knowledge discovery in databases: an overview, AI Magazine, 1992. – Vol. 13, N. 3. – P. 57–70.
4. Hand D., Mannila H., Smyth P. Principles of Data Mining. – MA, Cambridge: MIT Press, 2001. – 546 p.
5. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям / Н.Б. Паклин, В.И. Орешков. – СПб: Изд-во Питер, 2009. – 624 с.

6. Чубукова И.А. Data Mining: учебное пособие. – М.: Интернет-университет информационных технологий - ИНТУИТ.ру, БИНОМ: Лаборатория знаний, 2008. – 384 с.
7. Laxman S., Sastry P.S. A survey of temporal data mining // SADHANA, Academy Proceedings in Engineering Sciences. – The Indian Academy of Sciences, 2006. – Vol. 31. – P. 173–198.
8. Grossman R.L., Hornick M.F., Meyer G. Data mining standards initiatives // Communications of the ACM, 2002. – Vol. 45, N 8. – P. 59–61.
9. Kadav A., Kawale J., Mitra P. Data Mining Standards, 2008, 33 p. [Електронний ресурс]. – Режим доступу: http://www.datamininggrid.org/wdat/works/att/standard01.content.08439.pdf
10. Shearer C. The CRISP-DM model: the new blueprint for data mining // Journal of Data Warehousing. – Seattle, Washington: The Data Warehousing Institute, 2000. – Vol. 5 (4). – P. 13–22.
11. CRoss Industry Standard Process for Data Mining [Електронний ресурс]. – Режим доступу: http://www.crisp-dm.org/
12. Matignon R. Data Mining Using SAS Enterprise Miner. – Wiley, 2007. – 581 p.
13. Guazzelli A., Zeller M., Lin W.-C., Williams G. PMML: An Open Standard for Sharing Models // The R Journal, 2009. – Vol. 1 (1). – P. 60–65.
14. PMML Version 4.0 [Електронний ресурс]. – Режим доступу: http://www.dmg.org/pmml-v4-0.html
15. Common Warehouse Metamodel, v1.1 [Електронний ресурс]. – Режим доступу: http://www.omg.org/spec/CWM/1.1/PDF/
16. XML for Analysis [Електронний ресурс]. – Режим доступу: http://www.xmla.org/
17. ISO/IEC 13249-6:2006 [Електронний ресурс]. – Режим доступу: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38648
18. Integrating Data Mining with SQL Databases: OLE DB for Data Mining / A. Netz, S. Chaudhuri, U. Fayyad, J. Bernhardt // Proceedings of the 17th International Conference on Data Engineering (ICDE'01). – Heidelberg: IEEE Computer Society, 2001. – P. 379–387.
19. JSR 247: Data Mining 2.0 [Електронний ресурс]. – Режим доступу: http://www.jcp.org/en/jsr/detail?id=247
20. W3C Semantic Web Activity [Електронний ресурс]. – Режим доступу: http://www.w3.org/2001/sw/
21. Welcome to the DataFerrett for TheDataWeb [Електронний ресурс]. – Режим доступу: http://dataferrett.census.gov/
22. U.S. Census 2000. 5-Percent Public Use Microdata Sample Files [Електронний ресурс]. – Режим доступу: http://www.census.gov/Press-Release/www/2003/PUMS5.html
23. Table 4: Cumulative Estimates of the Components of Resident Population Change for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2008 (NST-EST2008-04). U.S. Census Bureau (22.12.2008) [Електронний ресурс]. – Режим доступу: http://www.census.gov/popest/states/tables/NST-EST2008-04.xls
24. Чертов О.Р. Учетные и аналитически информационные системы в области статистики населения / О.Р. Чертов // Наукові праці. – Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2010. – Т. 134. Вип. 121. Комп'ютерні технології. – С. 225–229.
25. Weikum G., Vossen G. Transactional information systems: theory, algorithms, and the practice of concurrency control and recovery. – San Diego: Morgan Kaufmann, 2001. – 852 p.
26. Thomsen E. OLAP solutions: building multidimensional information systems, 2nd ed. – N.Y.: John Wiley & Sons, 2002. – 661 p.
27. Перший Всеукраїнський перепис населення: історичні, методологічні, соціальні, економічні, етнічні аспекти / Н.С. Власенко, Е.М. Лібанова, О.Г. Осауленко та ін. – К.: ІВЦ Держкомстату України, 2004. – 558 с.
28. Чертов О.Р. Система многомерного анализа данных Всеукраинской переписи населения 2001 года / О.Р. Чертов // Россияне в зеркале статистики: Всероссийская перепись населения 2002 года: Международный симпозиум, 30-31 марта 2004 г.: труды симп. – М.: Изд-во Федеральной службы государственной статистики, 2004. – С. 234–238.
29. Development of State Statistics System for Monitoring Social & Economic Transformation Project [Електронний ресурс] / The World Bank. – Режим доступу: http://web.worldbank.org/external/projects/main?Projectid=P076338&theSitePK=40941&piPK=64290415&pagePK=64283627&menuPK=64282134&Type=Overview

Отримано 04.09.2010 р.