УДК 004.75

ПОБУДОВИ СИСТЕМИ МОНІТОРИНГУ ІНТЕРНЕТ-АКТИВНОСТІ КОРИСТУВАЧІВ КАФЕДРИ УНІВЕРСИТЕТУ НА ОСНОВІ ТЕХНОЛОГІЇ DATA MINING

В. М. Сидоренко, В. О. Морванюк, О. О. Соболєва, О. О. Слабченко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, 39600, м. Кременчук, Україна. E-mail: vnsidorenko@gmail.com

Обґрунтовано концепцію системи та її функціональну структуру. Наведена структура первинних та агрегованих інформативних даних для інтелектуального аналізу: апріорних, що описують користувача, і апостеріорних, що описують його поведінку в мережі. Розроблена структура ETL–процесу та виконано синтез сховища даних. Запропоновано структуру метамоделі математичного забезпечення.

Ключові слова: система моніторингу, інтернет-активність, інтелектуальний аналіз даних, Data Mining, ETL-процес, сховище даних.

MONITORING SYSTEM BUILDING CONCEPTION OF ONLINE USERS' ACTIVITY OF THE UNIVERSITY DEPARTMENT BASED ON DATA MINING TECHNOLOGY

V. N. Sidorenko, V. O. Morvanyuk, O.A. Soboleva, O. O. Slabchenko
Kremenchuk Mikhailo Ostrohradskyi National University
vul. Pershotravneva, 20, 39600, Kremenchuk, Ukraine. E-mail: vnsidorenko@gmail.com

The System concept and its functional structure are based. The structure of primary and aggregated informative data for intelligence analysis is based. A priori, describing a user, and a posteriori, describing user's network behavior. The structure of ETL-process is developed and a Data Warehouse's synthesis is done. The metamodel's structure of the mathematical support is offered.

Key words: monitoring system, Internet activity, intellectual data analysis, Data Mining, ETL-process, Data Warehouse.

КОНЦЕПЦИЯ ПОСТРОЕНИЯ СИСТЕМЫ МОНИТОРИНГА ИНТЕРНЕТ-АКТИВНОСТИ ПОЛЬЗОВАТЕЛЕЙ КАФЕДРЫ УНИВЕРСИТЕТА НА ОСНОВЕ ТЕХНОЛОГИИ DATA MINING

В. Н. Сидоренко, В. А. Морванюк, О. А. Соболева, О. О. Слабченко
Кременчугский национальный университет имени Михаила Остроградского
вул. Первомайская, 20, 39600, г. Кременчуг, Украина. E-mail: vnsidorenko@gmail.com

Обоснована концепция системы и ее функциональная структура. Приведена структура первичных и агрегированных информативных данных для интеллектуального анализа: априорных, которые описывают пользователя, и апостериорных, описывающих его поведение в сети. Разработана структура ETL–процесса и выполнен синтез хранилища данных. Предложена структура метамодели математического обеспечения.

Ключевые слова: система мониторинга, интернет-активность, интеллектуальный анализ данных, Data Mining, ETL–процесс, хранилище данных.

АКТУАЛЬНІСТЬ РОБОТИ. У межах кафедри сучасного університету як основної цілісної його ланки гостро стоїть проблема розподілу інтернет-ресурсів: поряд зі співробітниками основну масу користувачів мережевих потужностей складають студенти, магістри, аспіранти різних форм навчання, різного віку і менталітету, кількісний та якісний склад яких змінюється кожного року. Тому гостро постають питання завантаження мережі і безпечності інформації. Це, в першу чергу, викликане тим, що відвідування розважальних сайтів нерідко пов'язане з отриманням мультимедійної інформації. Останнє потребує швидкісного каналу і, таким чином, обмежує можливості роботи інших користувачів. З іншого боку підвищується ризик зараження вірусами та шкідливими програмами. У зв'язку з цим виникає необхідність застосування засобів моніторингу інтернет-активності користувачів мережі.

Як показав аналіз, застосування існуючого арсеналу програмних засобів моніторингу та обмеження доступу до мережі Інтернет не є достатньо ефективним виходячи з вищезазначених особливостей користувачів мережі кафедри. Виникає необхідність застосування більш гнучкого механізму розподілів ресурсів і пріоритетів, котрий би враховував не тільки статичну і нерідко суб'єктивну апріорну інформацію щодо користувачів, а ґрунтувався б на певних знаннях, отриманих шляхом інтелектуального аналізу апостеріорної інформації щодо поведінки користувачів протягом певного часу. Розв'язання даної задачі потребує побудови системи моніторингу на основі іншої парадигми із застосуванням технології Data Mining і, зокрема, Web Usage Mining [1, 4–16].

Традиційний погляд на процес моніторингу будь-якого об'єкта має на меті оцінку поточного стану і(або) майбутнього стану на основі аналізу накоплених даних, цінність яких у ретроспективі експоненціально знижується. Концепція Data Mining підходить до аналізу «сирих» ретроспективних даних принципово іншим чином.

На основі останніх обчислюється низка агрегатів, які періодично завантажуються у сховище певної структури і де неперервно накопичуються. При цьому точність і достовірність аналітичних висновків тільки збільшується. Як агреговані, так і сирі дані є «клондайком» для застосування методів ІАД для пошуку нетривіальних прихованих знань. Таким чином, процес моніторингу, зокрема інтернет-активності, можна представити схемою, наведеною на рис. 1.

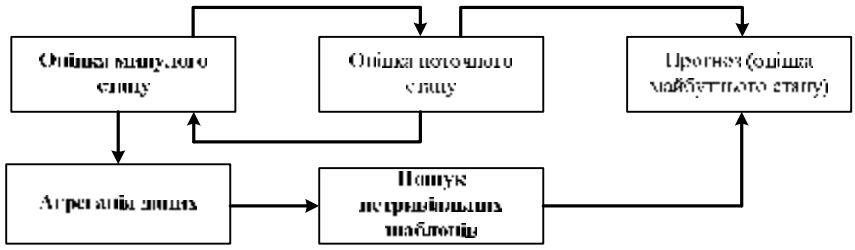

Рисунок 1 – Структура процесу моніторингу інтернет-активності кафедри

Метою роботи є підвищення ефективності розподілу ресурсів комп'ютерної мережі кафедри шляхом застосування методів інтелектуального аналізу даних щодо поведінки користувачів Інтернет.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. Одним із відомих підходів щодо обмеження доступу користувачів до ресурсів мережі є використання вбудованих засобів проксі-сервера SQID, Microsoft ISA Server, або зовнішнього спеціалізованого програмного забезпечення (ПЗ), наприклад, Red Line Software Internet Access Monitor. ПЗ SARG (Squid Analysis Report Generator) містить засоби аналізу лог-файлів проксі-сервера SQUID і формування звітів стосовно адресів сайтів, що відвідувалися, кількості з'єднань, об'єму отриманої інформації тощо. Програмне забезпечення містить певний набір графічних і табличних візуалізаторів [2].

Однак вищезазначені засоби не достатньо ефективні і, головним чином, тому, що вони не містять інструментарію для більш глибокого інтелектуального аналізу даних (ІАД) щодо поведінки користувачів, який би дав можливість виявлення нетривіальних паттернів у поведінці користувачів для побудови шаблонів «благонадійного» і «неблагонадійного» користувача. З цією метою доцільне застосування аналітичних платформ, наприклад, WebAnalyst, Deductor. Такі платформи містять необхідний інструментарій із застосуванням технологій OLAP та Data Mining, надають інтегровану платформу для зберігання та обробки інформації, мають можливість обробки даних з різних джерел, набір вбудованих аналітичних інструментів, здатних реалізувати всі необхідні етапи ETL-процесу: консолідацію, трансформацію, візуалізацію, очищення і завантаження даних до сховища [1, 3, 6].

З урахуванням вищезазначеного і виходячи з організаційної та функціональної структури кафедри, було виконано синтез схеми функціональної структури системи моніторингу інтернет-активності на прикладі кафедри комп'ютерних та інформаційних систем Кременчуцького національного університету імені Михайла Остроградського (рис. 2).

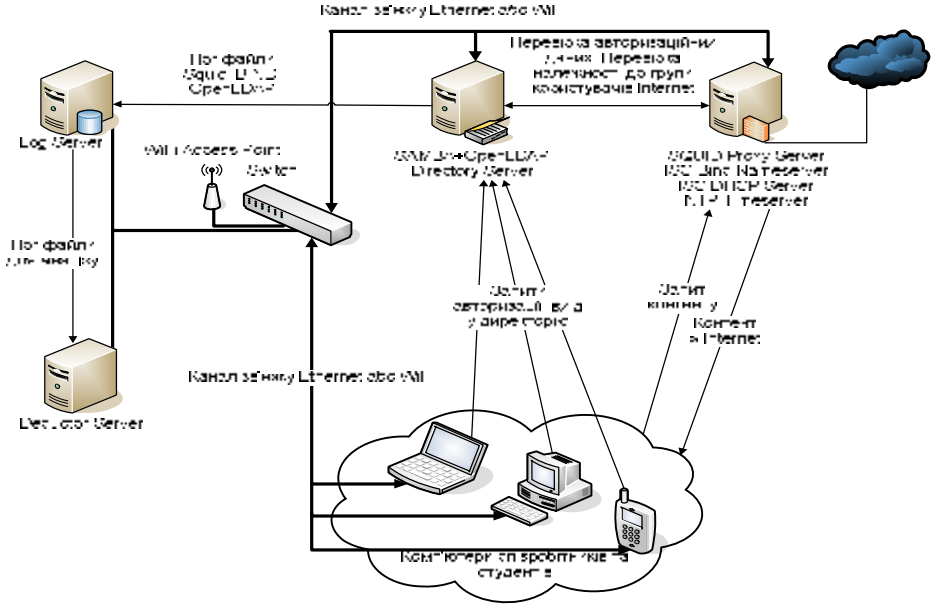

Рисунок 2 – Схема функціональної структури системи моніторингу інтернет-активності кафедри КІС

В основу структури системи моніторингу як програмно-апаратного комплексу було покладено мережу, що використовує для доступу до Інтернет проксі-сервер. В якості альтернативного рішення можливо також використання технології Network Address Translation, але в такому випадку проксі-сервер повинен працювати в режимі transparent і всі запити, що йдуть згідно з протоколом HTTP/HTTPS,

мають направлятися на нього. Для однозначної ідентифікації користувача на різних вузлах мережі використовується загальна база даних (БД) користувачів, що зберігається на сервері-контролері мережі (домену) і, в залежності від операційної системи, може бути реалізована у форматі POSIX користувачів (для *nix-подібних ОС), або в LDAP базі даних (для Windows Server – Active Directory, для *nix-подібних – OpenLDAP+SAMBA). LDAP-база даних використовується частіше, так як підтримує реплікацію і балансування навантаження між серверами. Крім того, необхідною умовою є присутність хоча б одного серверу перетворення імен (може використовуватися вбудований у Windows Server або ISC Bind). Фізично ролі серверів може виконувати один комп'ютер, або кілька – для підвищення надійності. Крім того це може бути віртуальна машина, що забезпечує швидкість відновлення при апаратних збоях. Вибір операційної системи залежить від фінансових факторів та кваліфікації обслуговуючого персоналу. Також можливий варіант змішаного середовища. У випадку вибору Інтернет-шлюзом серверу на *nix подібних ОС в якості проксі-сервера використовують, зазвичай, squid [7] – це високонадійний кешуючий проксі-сервер для протоколів HTTP, FTP, Gopher, HTTPS, котрий розповсюджується за ліцензією GPL на різних платформах, включаючи Windows Server. Відомо, що для Windows Server стандартом являється Microsoft ISA Server, котрий потребує додаткових ліцензійних відрахувань і структура його лог-файлів не досить детальна.

У разі обрання моделі роботи з використанням NAT, слід відзначити, що кількість даних у лог-файлах зменшиться, так як неможливо організувати авторизацію користувача і дані можна буде сортувати тільки за іменем хоста-клієнта.

Побудова системи моніторингу потребує, в першу чергу, визначення кількісного та якісного складу інформативних параметрів, які б дали змогу персоніфікувати користувачів як до користування ресурсами мережі, так і після отримання інформації щодо його поведінки в мережі за певний період.

Головною ідеєю, що покладено в основу моніторингу, є наступна: поряд з існуванням природної сегментації користувачів у просторі апріорних ознак розмірності n, очевидно, існує й інша природна сегментація у просторі розмірності $n+m$, де m – кількість апостеріорних ознак. Останні несуть інформацію щодо поведінки користувачів у мережі. І саме за результатами персоніфікації апостеріорних, а не апріорних сегментів передбачається розробка стратегій розподілу ресурсів і, власне, моніторинг інтернет-активності.

З урахуванням структури БД OpenLDAP і лог-файлу SQUID авторами запропоновано наступну структуру апріорних і апостеріорних даних та факторів, що обчислюються на їх основі, необхідних для розв'язку задач сегментації і класифікації (рис. 3). Останнє визначає як структуру ETL-процесу (рис. 4), так і структуру метаданих сховища даних (СД) (рис. 5). Як видно, інформативні дані мають як числову, так і нечислову природу. Така ситуація потребує акуратного і виваженого підходу до вибору математичного апарату, особливо для побудови моделей зі змішаними параметрами в майбутньому. Відомо [6], що надійність існуючих методів сегментації і класифікації у просторі змішаних ознак досі залишається невисокою. Виходячи із даних обставин, автори запропонували таку структуру інформативних факторів, яку наведено на рис. 3.

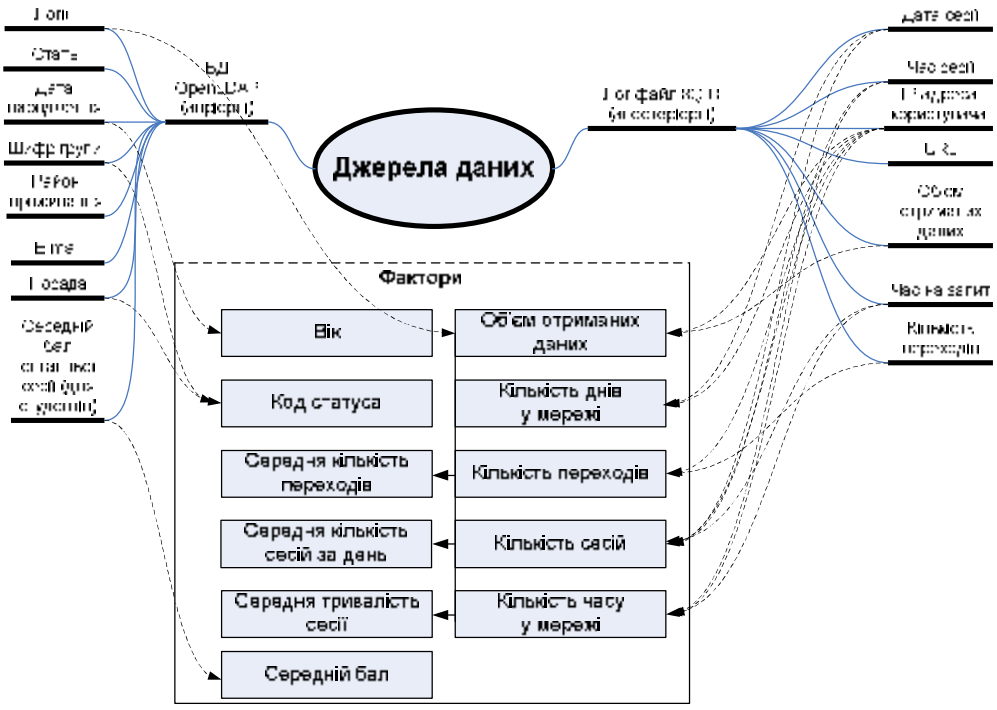

Рисунок 3 – Структурна схема інформативних даних

Поряд із розробкою і реалізацією важливих типових етапів Data Mining – консолідацією даних, реалізацією ETL-процесу, трансформацією, очищенням і завантаженням до сховища даних [7] – актуальним постає питання синтезу структури взаємодії задач інтелектуального аналізу на різних етапах моделювання, що повинно складати основу так званого аналітичного сценарію, чи іншими словами, власне математичного забезпечення системи (рис. 6).

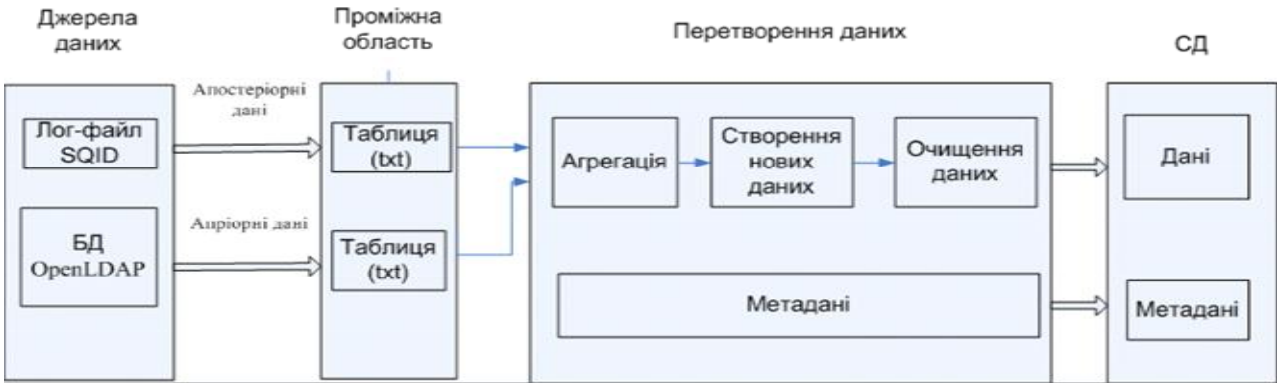

Рисунок 4 – Структура ETL-процесу

На думку авторів на начальній стадії доцільне розв'язання задачі сегментації користувачів мережі – спочатку на основі апріорної інформації, а потім з використанням інформації стосовно поведінки користувачів протягом певного терміну. Це дало б змогу персоніфікації певних груп і їх детального порівняльного дослідження. На послідуючих етапах профіль «благонадійного» і «неблагонадійного» користувача створюється, наприклад, шляхом застосування задач пошуку асоціативних правил і послідовних шаблонів у межах певного сегменту з паралельним аналізом відхилень. Розв'язок задачі класифікації користувачів з метою вироблення певної стратегії стосовно надання прав і пріоритетів можливе через певний термін дослідження поведінки користувачів у межах певного сегменту на основі навчальної вибірки. Отримані знання дають можливість класифікації та прогнозування поведінки нового користувача, що власне і є кінцевою задачею.

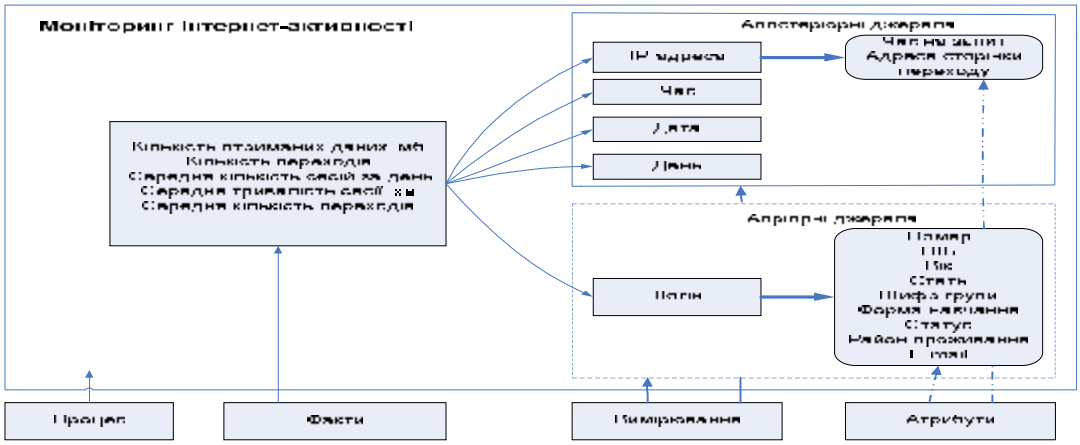

Рисунок 5 – Структура сховища даних

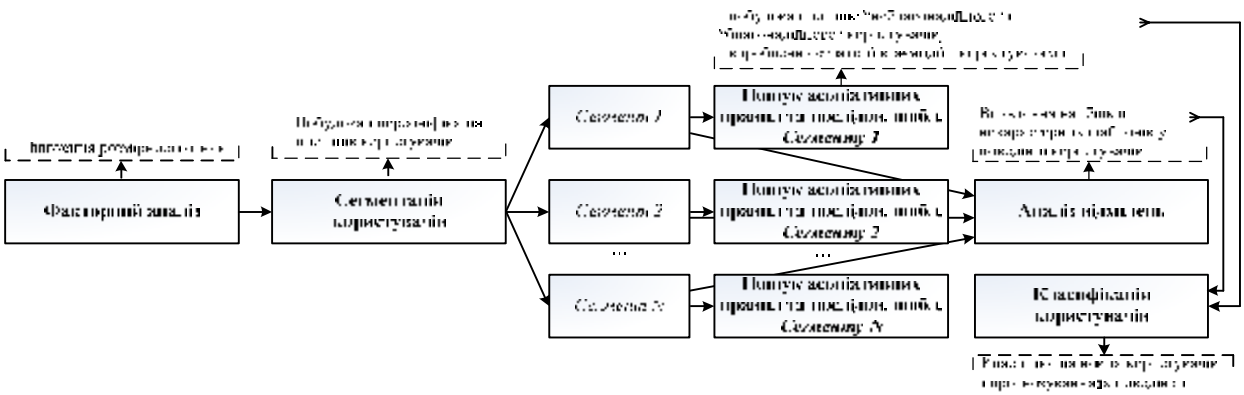

Рисунок 6 – Структура взаємодії задач інтелектуального аналізу на різних етапах моделювання

ВИСНОВКИ. Обґрунтовано і запропоновано концепцію системи моніторингу інтернет-активності на основі технології Data Mining, зокрема структуру інформаційного і математичного забезпечення.

Концепція системи, на відміну від існуючих, базується на іншій парадигмі, сенс якої полягає в наступному. Моніторинг інтернет-активності має на меті не тільки оцінку оперативної ситуації, зокрема по конкретному користувачу, а пошук прихованих нетривіальних закономірностей у поведінці всієї сукупності користувачів у глибокій ретроспективі. Це дає можливість за допомогою методів Data Mining створити профіль користувача, схильного до нецільового використання Інтернету.

Пошук нетривіальних шаблонів і прихованих знань у такій поведінці засобами інтелектуального аналізу дадуть істинний емпіричний фундамент для побудови моделей щодо прогнозування поведінки нових користувачів.

ЛІТЕРАТУРА

1. PolyAnalyst 6.0. Аналитика-это просто. [Электронный ресурс]: DataMining. – Электрон. дан. (1 файл). – Режим доступа: http://www.megaputer.ru. – Загл. с экрана.

2. SARG - анализатор логов SQUID [Электронный ресурс] : SARG–анализатор логов SQUID и генератор отчётов по ним. – Электрон. дан. (1 файл). – 20.09.2005. – Режим доступа: http://www.lissyara.su/articles/freebsd/programms/sarg– Загл. с экрана.

3. BaseGroup Labs технологии анализа данных [Электронный ресурс]: Аналитическая платформа Deduktor. – Электрон. дан. – Режим доступа: http://basegroup.ru/. – Загл. с экрана.

4. From Data Mining to Knowledge Discovery in Databases [Электронный ресурс]: Под ред. Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. – Электрон. дан. (1файл). 1996. – Режим доступа: http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf. – Загл. с экрана.

5. Web Mining Research: A Survey [Электронный ресурс]: Под ред. Raymond Kosala, Hendrik Blockeel. – Электрон. дан. (1файл). 2000. – Режим доступа: www.sigkdd.org/explorations/issue2-1/kosala.pdf. – Загл. с экрана.

6. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.

7. Проксі-сервер SQUID. [Электронный ресурс]: Приклади конфігурації.— Электрон. дан.— Режим доступа: http://www.sqid-cache.org/Doc/config/. – Загл. с экрана.

8. Magdalini P. Eirinaki. New approaches to web personalization. Ph. D. Thesis. Athens university of economics and business, Dept. of Informatics, 2006.

9. Chen M.S., Park J.S., and Yu P.S. Data mining for path traversal patterns in a Web environment // In Proceedings of the 16th International Conference on Distributed Computing Systems. 1996. – P. 385–392.

10. Cooley R., Mobasher B. and Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web, 1997.

11. Elo-Dean S. and Viveros M. Data mining the IBM official 1996 Olympics Web site // Technical report, IBM T.J. Watson Research Center, 1997.

12. Gergely T., Anshakov O., Finn V., Kuznetsov S. Cognitive research: Formal approach. // Series: Artificial Intelligence – Springer-Verlag, 2007.

13. Grigoriev P. QuDA, a Data Miner's Discovery Environment // Technical Report. – Technische Universität Darmstadt, 2003.

14. Mannila H. and Toivonen H. Discovering generalized episodes using minimal ocurrences // In Proc. of the Second Intel Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996. – P. 146–151.

15. Scime A. Web mining: applications and techniques, 2005.

16. Yan T., Jacobsen M., Garcia-Molina H., and Dayal U. From user access patterns to dynamic hypertext linking // In Fifth International World Wide Web Conference, Paris, France, 1996.

17. Хайлова В.В. Анализ Эффективности работы WEB-сайта с применением методов ИАД. //Десятая национальная конференция по искусственному интеллекту с международным участием КИИ–2006 (25–28 сентября 2006 г., Обнинск): Труды конференции. В 3-т. – М: Физматлит, 2006.

18. Информационные процессы и технологии «Информатика — 2010»: Материалы третьей Всеукраинской науч.-практ. конф. молодых ученых и студентов, 26–30 апреля 2010 г. — Севастополь: Изд-во СевНТУ, 2010. – 384 с.

REFERENCES

1. PolyAnalyst 6.0. Analytics – this is easy. [Electronic resource]: DataMining. — The electronic data (1 file). – Mode of access: http://www.megaputer.ru. – Screen with.

2. SARG – log analyzer SQUID [Electronic resource]: SARG – SQUID log analyzer and report generator for them. – The electronic data (1 file). – 20.09.2005. – Mode of access: http://www.lissyara.su/articles/freebsd/programms/sarg – Screen with.

3. BaseGroup Labs technology of data analysis [Electronic resource]: analytical platform Deductor. – The electronic data. – Mode of access: http://www.lissyara.su/articles/freebsd/ programms/sarg – Screen with.

4. From Data Mining to Knowledge Discovery in Databases [Electronic resource]: / Edited Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. — The electronic data (1 file), 1996. – Mode of access: http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf. — Screen with.

5. Web Mining Research: A Survey [Electronic resource]: / Edited Raymond Kosala, Hendrik Blockeel. – The electronic data (1 file). 2000. – Mode of access:

www.sigkdd.org/explorations/issue2-1/kosala.pdf. – Screen with.

6. Paklin N.B., Oreshkov V.I. Business Intelligence: from data to knowledge. — St. Petersburg: Peter, 2009. – 624 p. [In Russian].

7. Proxy server SQUID. [Electronic resource]: Examples of configuration. — The electronic data. — Mode of access: http://www.sqid-cache.org/Doc/config/.— Screen with.

8. Magdalini P. Eirinaki. New approaches to web personalization. Ph. D. Thesis. Athens University of economics and business, Dept. of Informatics, 2006.

9. Chen M.S., Park J.S., and Yu P.S. Data mining for path traversal patterns in a Web environment // In Proceedings of the 16th International Conference on Distributed Computing Systems, 1996. – P. 385–392.

10. Cooley R., Mobasher B. and Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web, 1997.

11. Elo-Dean S. and Viveros M. Data mining the IBM official 1996 Olympics Web site // Technical report, IBM T.J. Watson Research Center, 1997.

12. Gergely T., Anshakov O., Finn V., Kuznetsov S. Cognitive research: Formal approach. // Series: Artificial Intelligence – Springer-Verlag, 2007.

13. Grigoriev P. QuDA, a Data Miner's Discovery Environment // Technical Report. – Technische Universität Darmstadt, 2003.

14. Mannila H. and Toivonen H. Discovering generalized episodes using minimal ocurrences // In Proc. of the Second Intel Conference on Knowledge Discovery and Data Mining. – Portland, Oregon, 1996. – P. 146–151.

15. Scime A. Web mining: applications and techniques, 2005.

16. Yan T., Jacobsen M., Garcia-Molina H., Dayal U.. From user access patterns to dynamic hypertext linking // In Fifth International World Wide Web Conference, Paris, France, 1996.

17. Hailova V.V/ Analysis of the Effectiveness of work-WEB-Site you are using the methods of IAD // Tenth National Conference on artificial intelligence-vennomu with international participation CAI–2006 (25–28 September 2006, Obninsk, Russia): conference proceedings. The 3-m. – M: Fizmatlit, 2006. [In Russian].

18. Information Processes and Technology "Informatics – 2010": Proceedings of the third All-Ukrainian Scientific-Practical Conf. young scientists and students, April 26-30, 2010. – Sevastopol: Publ. SevNTU, 2010. – 384 p. [in Russian].

Стаття надійшла 14.04.2011.
Рекомендована до друку
д.т.н., проф. Гученком М.І.