

Др
14129

Одеський національний університет імені І. І. Мечникова

(повне найменування вищого навчального закладу)

Навчально науковий інститут інформаційних та соціальних технологій

(повне найменування інституту/факультету)

Економічної кібернетики та інформаційних технологій

(повна назва кафедри)

Дипломна робота

бакалавра

на тему: «Розробка технологій управління потоками запитів у
інформаційно-пошукових системах»

« Development technologies for flow requests management in information retrieval systems »

Виконав: студент денної форми навчання

Напряму підготовки 6.030502 «Економічна кібернетика»

Шутенко Вікторія Олександрівна

(ПП)

Керівник ст.викл.Максимов О.С.

Рецензент к.економ.наук, доц. Шумлянська О.І.

Рекомендовано до захисту:

Протокол засідання кафедри

№ 11 від 30.05.2017 р.

Захищено на засіданні ЕК № ____

протокол № __ від _____ 2017 р.

Оцінка _____ / _____ / _____
(за національною шкалою, шкалою ECTS, бали)

Завідувач кафедри

Тюрин О.В.

(підпис)

Голова ЕК

Гохман О.Р.

(підпис)

Одеса – 2017

790247

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ I УПРАВЛІННЯ ТА ОБРОБКА ВЕЛИКИХ ОБ'ЄМІВ ДАНИХ ТА СУЧАСНІ ТЕХНОЛОГІЇ ПОБУДОВИ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМ НА БАЗІ ТЕХНОЛОГІЇ ПОТОКОВОЇ ОБРОБКИ	5
1.1. Інформаційно-пошукові системи як інструмент доступу до великих об'ємів даних	5
1.2. Аналіз структури, властивостей та видів запитів до ІПС	13
1.3. Управління потоками запитів в інформаційно-пошукових системах.....	18
ВИСНОВКИ ДО ПЕРШОГО РОЗДІЛУ	21
РОЗДІЛ II. АНАЛІЗ МЕТОДІВ ОБРОБКИ ПОТОКІВ В ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ	23
2.1. Повнотекстовий пошук. Його опис та застосування.....	23
2.2. Аналіз застосування методу графів гіпер-дерева у сучасному інформаційному середовищі	27
2.3. Різноманітні технології обробки потоків запитів у різних програмних продуктах	31
ВИСНОВКИ ДО ДРУГОГО РОЗДІЛУ	55
РОЗДІЛ III РОЗРОБКА ТЕХНОЛОГІЙ УПРАВЛІННЯ ПОТОКАМИ ЗАПИТІВ В ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ НА ПРИКЛАДІ ТОВ «ТЕКОМ ДЕВЕЛОПМЕНТ»	56
3.1. Аналіз діяльності та інформаційних ресурсів підприємства ТОВ «Теком Девелопмент».....	56
3.2. Опис удосконаленої технології управління потоками запитів до інформаційно-пошукових систем	64
3.3 Побудова прототипу ядра пошукової системи на базі розробленої технології.....	67
3.4 Порівняльний аналіз результатів роботи інформаційної системи по розробленій технології з існуючими системами	69
ВИСНОВКИ ДО ТРЕТЬОГО РОЗДІЛУ	80
ВИСНОВКИ	81
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:	83
ДОДАТКИ.....	87

ВСТУП

Сучасний етап розвитку економіки України характеризується необхідністю розвитку перспективних напрямків науки і техніки і підвищенням ефективності виробництва з метою доведення якості продукції до світових стандартів. Для вирішення цієї задачі керівництву підприємств потрібна оперативна достовірна інформація про фактичний стан виробництва, потребах у ресурсах, ситуації на ринку і т. і. Одним з ефективних напрямків удосконалення управління підприємством є розробка та впровадження сучасних інформаційно-управлінських систем і технологій. Нові інформаційні технології управління підприємством є важливим і необхідним засобом, який дозволяє: швидко, якісно і надійно виконувати отримання, облік, зберігання і обробку інформації; значно скоротити управлінський персонал підприємства, який займається роботою по збору, обліку, зберіганню і обробці інформації; забезпечити у потрібні терміни керівництву і управлінсько-технічний персонал підприємства якісною інформацією; своєчасно і якісно вести аналіз і прогнозування господарської діяльності підприємства; швидко і якісно приймати рішення по усіх питаннях управління підприємством.

Актуальність теми дипломної роботи полягає в тому, що існують великі об'єми інформації, багато запитів до цих об'ємів. Виникає потреба обробки великої кількості запитів у потоках. Сучасні технології обробки цих даних потребують вдосконалення.

Основною метою даної дипломної роботи є розробка технологій управління потоками запитів в інформаційно-пошукових системах.

Мета дослідження передбачає постановку та рішення наступних задач:

1. Виявлення факторів, що впливають на побудову ефективних пошукових систем;
2. Аналіз закономірностей та перспектив використання та розвитку сучасних технологій обробки потоків запитів в інформаційно-пошукових системах;

3. Дослідження підходів, методів та технологій у сучасних умовах розвитку інформаційно-пошукових системах;
4. Розробка прототипу програмного забезпечення аналітичного ядра обробки запитів інформаційно-пошукових систем.

Об'єктом дослідження є сукупність інформаційно-пошукових систем, присутніх на ринку України.

Предметом дослідження виступають процеси побудови технологій обробки потоків в інформаційно-пошукових системах.

Інформаційну базу дослідження склали дані аналізу роботи підприємств України, публікації в загальноекономічних і спеціалізованих журналах, матеріали з офіційних сайтів, а також результати досліджень, проведених за офіційними даними звітності підприємств України.

ВИСНОВКИ

Масштаби інформаційних ресурсів і їх кількість постійно розширюється. Інформаційний пошук в таких об'ємах інформації не може бути ідеальним.

Вважається, що ідеальна пошукова машина повинна відповідати наступним вимогам:

1. простота у використанні
2. чітко організований і оновлюваний індекс.
3. швидкий пошук в базі даних і швидке реагування.
4. надійність і точність результатів пошуку.

Швидкість обробки потоків є, на нашу думку, важливою проблемою в розвитку сучасного виробництва, бізнесу та інформаційних технологій взагалі, тому дана тематика і розробка технологій, які б швидко обробляли потоки запитів в інформаційних системах є актуальними на сьогодні.

Технології пошуку та обробки інформації розвиваються кожного дня, з'являються все нове і нове програмне забезпечення, що забезпечує обробку інформаційних потоків. Розглянуті нами програмні засоби мають свої доліки та недоліки, але ніщо не є перевершеним і кожній системі необхідні свої функції чи особливості роботи програми. З розвитком інформаційних технологій з'являються все нові підходи до обробки. Але так, як ніщо не досконале - це є моментом, що слід розвивати і розробляти свої технології обробки потоків запитів, що і було наведено у Розділі 3.

Нами була розроблена технологія по обробці потоків запитів в інформаційно-пошукових системах. Не існує ідеального підходу, тому вивчивши ті, що були наведені у Розділі 2, ми розробили власну технологію. Найбільшою її перевагою є час, за який система оброблює потоки, що є дуже важливим показником для кожного підприємства. Також Байесовські підходи для ймовірнісної оцінки попадання запитів до певних категорій є досить точними і дають гарні статистичні дані.

Апробація даної системи була проведена на прикладі торгового центру «Сіті Центр». Були проведені експерименти з клієнтами центру, що показали кращі часові результати, ніж аналогічна система. Суть полягає у тому, що клієнти даного торгового центру, приходячи до нього створюють запити по цілі знаходження певного магазину, чи магазинів певної категорії. Згідно статистики, кожного дня там проходять більш ніж 10 000 потенційних покупців. Якщо кожен з них буде відправляти до системи запит, то система буде оброблювати потоки запитів довгий час. З розробленою нами технологією ці потоки будуть оброблюватися у процесі надходження та давати відповідь дуже швидко, тим самим задовольняючи потреби всіх клієнтів та торговельних точок.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Mahanti, D. Eager, and C. Williamson. Temporal locality and its impact on web proxy cache performance. *Performance Evaluation Journal: Special Issue on Internet Performance Modelling*, 42(2/3):187–203, September 2000.
2. Bollacker K., Lawrence S., Giles C., CiteSeer: An Autonomous {Web} Agent for Automatic Retrieval and Identification of Interesting Publications, *Proceedings of the Second International Conference on Autonomous Agents*, ACM Press, 1998, pp. 116--123
3. Carey Williamson. On filter effects in web caching hierarchies. *ACM Transactions on Internet Technology*, 2(1):47–77, February 2002.
4. Chen L., Sycara K., WebMate: A Personal Agent for Browsing and Searching, in *Proceedings of Conference on Autonomous Agents*, ACM, 1998, Pages 132-139
5. D. Florescu, A. Y. Levy, D. Suciu, and K. Yagoub. Optimization of run-time management of data intensive web-sites. In *VLDB*, pages 627–638, 1999.
6. Etzioni O., Weld D., *Intelligent Agents on the Internet: Fact, Fiction, and Forecast*, 1995
7. Franklin S., Graesser A., *Is It an Agent, or just a Program?: A Taxonomy for Autonomous Agents*, *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, Springer-Verlag, 1996
8. Gilbert, Aparicio, *The Role of Intelligent Agents in the Information Infrastructure*, IBM 1995
9. Glover E., Lawrence S., Birmingham W., Giles C., Architecture of a Metasearch Engine that Supports User Information needs, *Proceedings of CIKM-99 Conference*, pp. 210-216, ACM, 1999 Информационно-поисковые системы в Internet. Редакция 14 17.03.2006
10. Gopal Pandurangan and Eli Upfal. Can entropy characterize performance of online algorithms? In *Symposium on Discrete Algorithms*, pages 727–734, 2001.

11. International Standards Organization. Database language SQL—Part 2: Foundation (SQL / Foundation). ISO/IEC 9075-2:1999, Sept. 1999.
12. J. Shanmugasundaram, E. J. Shekita, R. Barr, M. J. Carey, B. G. Lindsay, H. Pirahesh, and B. Reinwald. Efficiently publishing relational data as XML documents. *VLDB Journal*, 10(2-3), 2001.
13. Jennings N., Sycara K., Woodridge M., A Roadmap of Agent Research and Development, *Autonomous Agents and Multy-Agent Systems*, 1, pp275-306, 1998
14. K. Forbus. Qualitative physics: Past, present, and future. In *Exploring Artificial Intelligence*. MorganKaufmann, 1988.
15. L. Fegaras and D. Maier. Optimizing object queries using an effective calculus. *TODS*, 25(4), 2000.
16. Lawrence S., Bollacker K., Giles C., Digital Libraries and Autonomous Citation Indexing, *IEEE Computer*, pp. 67-71, June 1999
17. Lawrence S., Bollacker K., Giles C., Indexing and Retrieval of Scientific Literature, *Proceedings of CIKM 1999 Conference*, pp. 139-146
18. Lawrence S., Giles C., Accessibility of Information on the Web, *Nature*, vol.400, pp. 107-109, 1999
19. Lawrence S., Giles C., Context and page analysis for improved web search, *IEEE Internet Computing*, July 1998, pp.38-46
20. Lawrence S., Giles C., Searching the Web: General and Scientific Information Access, *IEEE Communications Magazine*, January 1999, pp 116-122
21. M. F. Fernandez, A. Morishima, and D. Suciu. Efficient evaluation of XML middle-ware queries. In *SIGMOD*, 2001.
22. Oates T., Prasad N., Lesser V., *Networked Information Retrieval as Distributed Problem Solving*. University of Massachusetts
23. P. A. Bernstein, S. Pal, and D. Shutt. Context-based prefetch for implementing objects on relations. In *VLDB*, 1999.
24. P. Seshadri, H. Pirahesh, and T. Y. C. Leung. Complex query decorrelation. In *ICDE*, Feb. 1996.

25. Peter J. Denning. Working sets past and present. *IEEE Transactions on Software Engineering*, SE- 6(1):64–84, January 1980. 7
26. S. Floyd and V. Paxson. Difficulties in simulating the internet. *IEEE/ACM Transactions on Networking*, 9(4):392–403, August 2001.
27. Salton G., Buckley C., *Term Weighting Approaches in Automatic Text Retrieval*, Cornell University Technical Report 87-881
28. Salton G., *Historical Note: The Past Thirty Years in Information Retrieval*, Cornell University Technical Report 87-827
29. Salton G., *Mathematics and information retrieval* Cornell University Technical Report 78-332
30. Shudong Jin and Azer Bestavros. Temporal locality in web request streams. Technical Report BUCSTR-1999-014, Boston University Computer Science Department, July 2002.
31. Stolt H., *Agents, filter and search engines an evaluating survey on technologies for effective search for information from internet resources*, Graduation thesis, Umea University, 1997
32. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.
33. T. Mayr and P. Seshadri. Client-site query extensions. In *SIGMOD*, 1999.
34. Terrence Kelly and Jeffrey Mogul. Aliasing on the world wide web: Prevalence and performance implications. In *Performance Track, 11th Intl. World Wide Web Conference*, 2002.
35. Балабанов И.Т. *Электронная коммерция*. СПб.: Питер, 2001.
36. Зажигалкин А.В. *Международно-правовое регулирование электронной коммерции: Автореф. дис. ... канд. юрид. наук.* – СПб., 2005.
37. Мальковский М. Г., Грацианова Т. Ю., Полякова И. Н., *Прикладное программное обеспечение: системы автоматической обработки текстов, Учебное пособие для студентов факультета ВМиК МГУ, Москва, МГУ, 2000*
38. Черный А. И., *Введение в теорию информационного поиска*, М., Наука,

1975

39. Aglets Workbench, <http://www.trl.ibm.co.jp/aglets/information.html>
40. Bergman K., The Deep Web: Surfacing Hidden Value, BrightPlanet.com LLC, <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>
41. Brin S., Page L., The Anatomy of a Large-Scale Hypertextual Web Search Engine, Stanford University, <http://www.db.stanford.edu/pub/papers/google.pdf>
42. Hermans B., Intelligent Software Agents on the Internet: an inventory of currently offered functionality in the information society and a prediction of future
43. developments, <http://www.hermans.org/agents>, 1996
44. Inktomi Corp., Web Surpasses One Billion Documents, press release issued January 18, 2000, <http://www.inktomi.com/new/press/billion.html>
45. JKQML, <http://alphaworks.ibm.com/tech>
46. Metadata Activity Statement, <http://www.w3c.org/Metadata/Activity.html>
47. Mladenic D., Personal WebWatcher: Implementation and Design Technical Report IJS-DP-7472, Department of Intelligent Systems, J.Stefan Institute, Slovenia, 1996, http://www.cs.cmu.edu/afs/cs/project/theo4/textlearning/www/pww/papers/PW_W/pwwTR.ps.Z.
48. Open Archives Initiative <http://www.openarchives.org>
49. Research Index Computer Science Directory, <http://citeseer.nj.nec.com/directory.html>
50. Resource Center Software AG <https://www.softwareag.com/corporate/rc/default.asp>
51. Seberg E., Etzioni O., The MetaCrawler Architecture for Resource Aggregation on the Web, <http://www.cs.washington.edu/research/metacrawler>, 1998
52. W3C Metadata Activity <http://www.w3c.org/RDF/>