УДК 681.3.01

Постніков О. С.

ВИКОРИСТАННЯ ПОСЕРЕДНИКА ПРИ ПОШУКУ В РОЗПОДІЛЕНИХ РЕПОЗИТОРІЯХ НАВЧАЛЬНИХ ОБ'ЄКТІВ

У статті розглянуто проблеми ефективного пошуку в розподілених репозиторіях навчальних об'єктів. Зосереджено увагу на тих системах, що використовують посередника при пошуку. Досліджено можливі шляхи покращення ефективності й релевантності пошуку в гетерогенних розподілених репозиторіях. У праці окреслено декілька підходів до покращення пошуку, а саме агрегаційний репозиторій, метапошукові інструкції, статистика запитів та векторна модель пошуку.

Вступ

Останнім часом галузь інформаційних технологій усе активніше використовує бази знань. Одним з пріоритетних напрямів є системи підтримки навчання, в яких все актуальнішими стають міжвузівські розподілені децентралізовані системи зберігання навчальних об'єктів [6]. При реалізації таких систем особливу увагу звертають на ефективний пошук. Проте поточний стан справ у галузі має тенденцію до розвитку функціональності й підтримки стандартів, втрачаючи масштабованість та ефективність. Мета цієї праці — описати деякі з можливих підходів до підвищення ефективності пошуку в розподілених репозиторіях навчальних об'єктів.

З-поміж різних систем керування базами даних ми взяли до розгляду лише ті системи, які виконують такі вимоги: націленість на збереження навчальних об'єктів [10], відкритість інтерфейсів, можливість відокремити репозиторій від системи керування навчанням (LMS[1])-

Найбільш розвинутими й вартими уваги виявилися такі системи, як Blackboard Content System, TopClass LCMS[2], CanCORE, EduSource, та реалізації специфікацій ADL SCORM: eCore та Joint ADL Co-Lab.

Під час аналізу кожного рішення окремо та ситуації на ринку електронної освіти в цілому, були вибрані такі критерії оцінки: можливість пошуку за метаданими, підтримування розподіленого сховища, можливість ранжування результатів пошуку.

Перший критерій означає, що мови запитів на проведення пошуку аналізувалися як у даних (можливо, повнотекстовий пошук, fulltext search), так і метаданих. Термін «розподілений» (другий критерій) означає фізичну розподіленість даних по серверах і можливість баз даних зберігати дані в різних форматах. Третій критерій є важливим тому, що пошукові системи, які працюють з великим обсягом даних, повинні вміти не тільки обирати документи, що формально задовольняють пошуковий припис, а й видавати результати пошуку відповідно до певного евристичного рангу.

Незадовільний стан рішення задачі пошуку в галузі розподілених репозиторіїв навчальних об'єктів змусив шукати аналогій в інших суміжних галузях.

Посередник при пошуку

Цікавим і багатонадійним методом пошуку в розподілених базах знань, на наш погляд, є використання посередника при пошуку (search intermediary). У багатьох джерелах такого посередника називають «мобільним агентом» [4]. Це зумовлено тим, що посередник при пошуку «мандрує» різними серверами й збирає інформацію відповідно до запиту, який був посланий користувачем системи. Код, що виконується, може змінюватися на різних серверах відповідно до його внутрішньої структури та інтерфейсів, але результати пошуку отримуються і зберігаються в єдиному форматі. Загальну схему використання посередника при пошуку зображено на рис. 1. Вищий шар — сервіси доступу (access services) та бізнес-об'єкти, які безпосередньо доступні користувачам системи. Нижній шар — сервіси забезпечення (provision services), репозиторії навчальних матеріалів.

Фактично, створюється ще один рівень аб-

[1] LMS — Learning Management System.
[2] LCMS — Learning Content Management System.

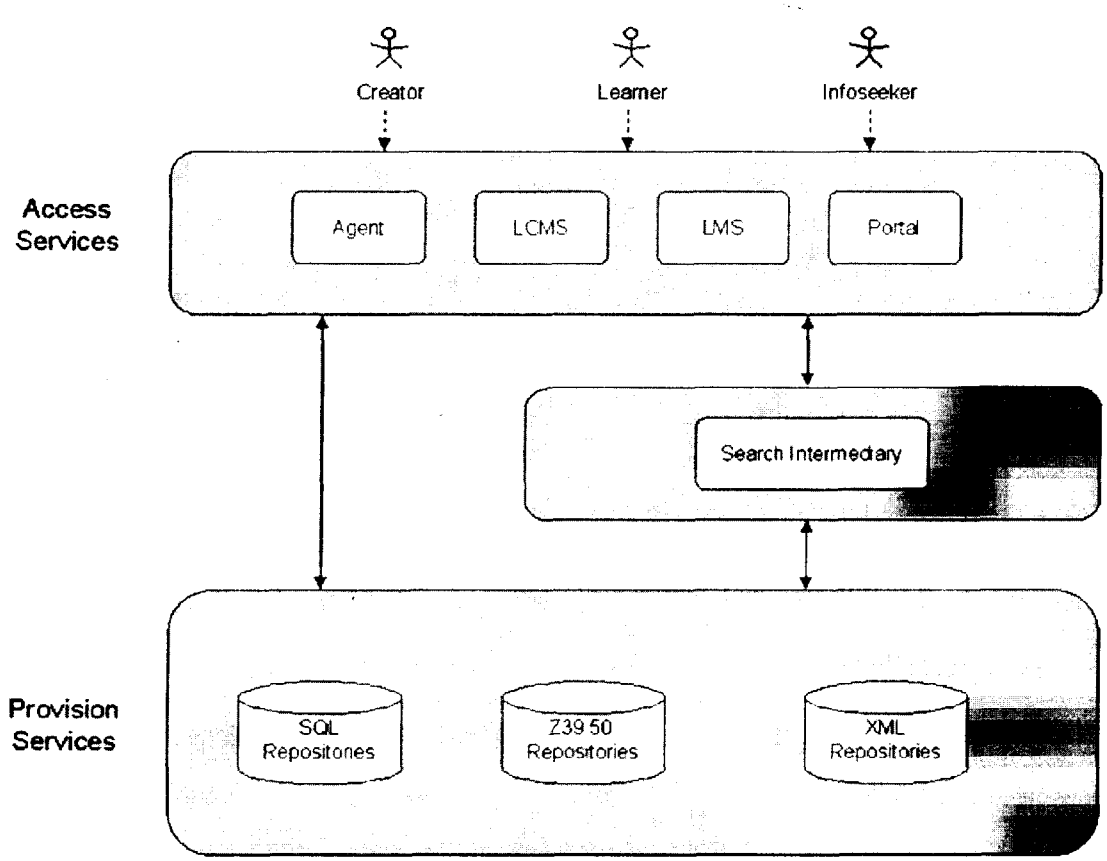

Рис. I. Використання посередника при пошуку в системах підтримки навчання

стракціїміж користувачами і даними. Посередник при пошуку дає змогу відокремити процес пошуку від акторів системи й шару бізнес-об'єктів. Такий підхід надає великі можливості для створення нових інтелектуальних моделей збереження і пошуку даних у навчальних середовищах. Більше того, посередник стає необхідним, коли користувач повинен мати можливість працювати з системою в режимі off-line.

Під час пошуку в різнорідних розподілених репозиторіях із використанням посередника постають такі запитання:

- Який рівень деталізації результатів пошуку потрібен користувачеві?
- Яким чином може система визначати релевантність пошуку і час його зупинки?
- Чи можна звузити область пошуку для кожного конкретного випадку?

Підходи до покращення процесу пошуку

Для вирішення цих питань можна, на наш погляд, запропонувати ряд підходів.

Агрегаційний репозиторій. Додатковий репозиторій, що збирає метадані (metadata) з усіх репозиторіїв системи, обирає з метаданих важливі й зберігає їх локально. «Важливість» даних— уявна категорія, яка відповідає на запитання: «Наскільки часто за певним критерієм/метаданим ведеться пошук?» Важливість можна визначати при розробці метаданих шляхом приписування вагових коефіцієнтів або шляхом збору статистики пошукових приписів під час роботи системи. У першому випадку відповідальність покладається на експертів, у другому — на алгоритм відбору. Для збору інформації варто використати стандарт «протоколу для збору метаданих OAI'» [2]. У разі створення такого репозиторію необхідно забезпечувати підтримку «живої» (up-to-date) інформації про підлеглі сховища, оновлення і видалення метаданих. При дублюванні даних обов'язковою є чітка система унікальних ідентифікаторів (наприклад, «Постійний ідентифікатор ресурсів IMS»[2]). Ще однією проблемою стає непрозорість і стохастичний характер пошуку. Один і той самий запит необов'язково повертатиме дані за один і той самий час, і не завжди від-

The Open Archives Initiative Protocol for Metadata Harvesting.
[1] IMS Persistent Location-Independent Resource Identifier.

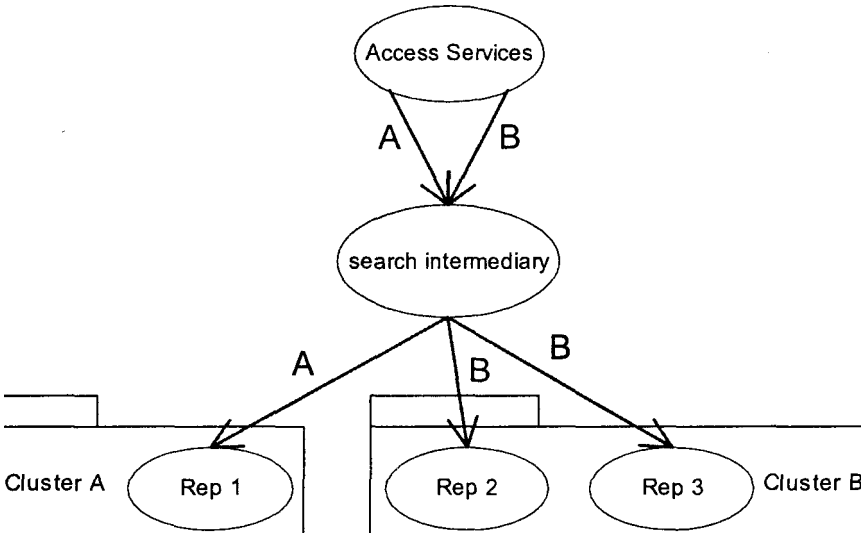

Рис. 2. Використання посередника для розподілу репозиторію на кластери

повідь буде повною. Алгоритми, які дають змогу ефективно впоратися з такими питаннями, на цей час реалізовані в системах, де використовується кешування (див., наприклад, [1]). Такий агрегаційний репозиторій реалізований у багатьох бібліотечних системах і мало поширений у репозиторіях навчальних об'єктів.

Метапошукові інструкції. Метапошукові інструкції - додаткова інформація, яка надходить разом із запитом на оригінальній мові пошуку та дає посередникові можливість ефективніше розподілити задачі. Такими інструкціями є, наприклад, вказівка репозиторію, де шукати, або максимально можлива кількість знайдених документів. Для імплементації таких інструкцій мова пошуку має підтримувати розширення. Приміром, мова запитів до XML-баз даних XPath не може бути розширена, а XML Query — може. Слід зазначити, що мова запитів від об'єктів бізнес-рівня не тотожна мові запитів від посередника до репозиторіїв (див. рис. 1). Тому підхід з використанням метапошукових інструкцій обмежений реалізацією репозиторіїв нижчого рівня.

Інший варіант використання можливий за умови функціонального розподілу репозиторіїв. Якщо всі репозиторії працюють спільно з клієнтами (акторами або посередниками) одного типу, тоді можлива кластеризація навчальних об'єктів по репозиторіях. Припустимо, що є об'єкти двох взаємовиключних типів — A і B. На рис. 2 зображено варіант, коли за домовленістю репозиторій 1 зберігає лише об'єкти типу A, а репозиторії 2 і 3 — об'єкти типу B. Посередник сортує запити і звужує коло пошуку до відповідного кластеру (якщо, звичайно, в запиті вказано потрібний тип). Недоліки такого методу полягають в обов'язковому дотримуванні меж кластерів. Тобто доступ до репозиторіїв обмежується лише для визначених клієнтів, які визнають домовленості про розбиття на кластери. Найближчий аналогтакого підходу — кластерний пошук Vivismo.

Статистика звернень. Посередник може зберігати статистику звернень (профіль користувача) до різних репозиторіїв, тобто ранжувати різні сховища за частотою, з якою користувач працював із певним сховищем. Ймовірність того, що наступний пошук буде схожим на попередні, досить значна, і варто починати пошук саме з найбільш використовуваного репозиторію. Цей підхід добре працюватиме при врахуванні обмеження на кількість знайдених документів. Аналогом цього підходу є Google Page **Rank.**

Векторна модель пошуку. Така модель може бути використана в таксономічному пошуку, якщо звичайної булівської моделі «так/ні» недостатньо. Припустимо, всі елементи метаданих визначають векторний простір (кожне ключове слово, шо належить певній таксономії,— один вимір у векторному просторі). До кожного елемента метаданих в документі і до кожної умови у пошуковому приписі ставляться у відповідність вагові коефіцієнти, які складають два вектори. Для кожного документа релевантність тоді визначається не за принципом «підходить/не підходить», а як векторна відстань між векторами запиту і документа [3]. Такий підхід дає змогу природним шляхом сортувати результати пошуку за релевантністю. Але для великих документів і складних метаданих побудова вагових коефіцієнтів стає проблематичною, оскільки має бути здійсненалюдиною-експертом.

Реалізація

Наша реалізація є репозиторієм навчальних об'єктів, який відповідає стандартам IMS Digital Repositories Interoperability, IMS Metadata та IMS Content Packaging і частково — ADL SCORM. У системі використаний посередник при пошуку, який керує одним сховищем даних у форматі ХМЬта одним—у форматі MS SQL. Система побудована на платформі Microsoft.NET та написана мовою C#. Загальну схему потоків даних у системі зображено на рис. 3. «Рідною» мовою інтерфейсів є мова запитів до XM L-документів XQuery.

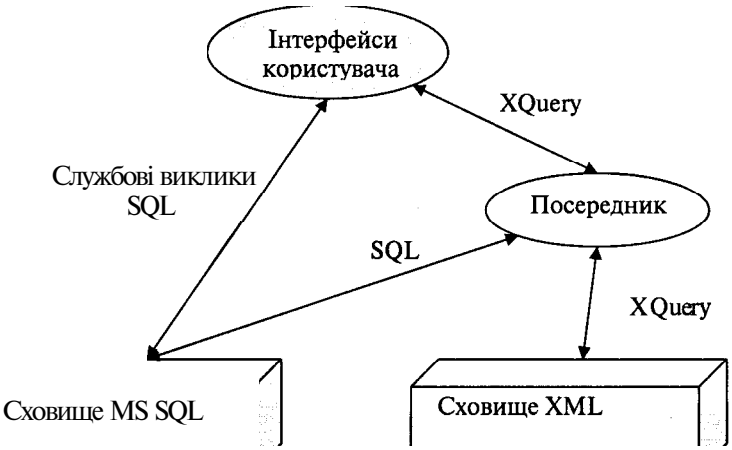

Рис. 3. Потоки даних в реалізованій системі

Проведені практичні експерименти з використанням посередника дали непогані результати й засвідчили можливості подальшого підвищення інтероперабельності, масштабованості й ефективності системи.

Висновок

Останнім часом увага спільноти розробників систем підтримки навчання поступово переходить від обговорення форматів і загальних схем функціонування до обговорення практичного досвіду впровадження [8]. Став зрозумілішим розподіл задач у цій галузі, частково завдяки зусиллям консорціуму I MS та робочої групи IEEE LTSC [9]. Пропонована праця є спробою окреслити підходи до підвищення ефективності пошуку в розподілених репозиторіях навчальних об'єктів. Сфера використання обмежується навчальними системами, де впроваджено посередника при пошуку, який і зумовлює значні можливості для розвитку ефективних алгоритмів пошуку в розподілених системах. Розглянуті підходи (агрегаційнийрепозиторій, метапошукові інструкції, збір статистики звернень та векторна модель пошуку) окреслюють можливі покращення в побудові розподілених репозиторіїв та розширення пошукових мов для потреб гетерогенних розподілених репозиторіїв навчальних об'єктів. Ми не доводили розгляддо конкретних конструктів, а лише намітили сутність кожного з підходів. Загалом, питання пошуку в розподілених репозиторіях у системах підтримки навчання перебуває на межі галузей і, безперечно, має ще великий простір для досліджень.

1. *James Pilkow & Margaret Recker.* A Simple Yet Robust Caching Algorithm Based on Dynamic Access Patterns.- http://citeseer.ist.psu.edu/pitkow94simple.html
2. The Open Archives Initiative Protocol for Metadata Harvesting,— http:// www.openarchives.org/OAI/openarchives-protocol.html
3. *Juha Puustjarvi & Paivi Poyry.* Searching Learning Objects from Virtual Universities.— http://lear.inrialpes.fr/people/triggs/ events/iccv03/cdrom/womtec03/WOMTECM28.pdr
4. *Papastavrou S., Samaras G., Pitoura E.* Mobile Agents for Distributed WWW Access,- http://www.cs.uoi.gr/~pitoura/distribution/Mobile/de99.ps
5. *Гаврилова Т. А., Хорошевский В.* Ф. Базы знаний интеллектуальных систем,— СПб.: Питер, 2000.— 340 с.
6. *M. Tamer Ozsu and Patrick Valduriet* Principles of Distributed Database Systems,- Second Edition, 320—380,- Prentice Hall, 1999.
7. *Bryan Chapman.* Learning Content Management Systems,— http://www.internettime.com/Learning/lcms/
8. *Wolfgang Nejdl, Boris Wolf.* EDUTELLA: Searching and Annotating Resources within, an RDF-based P2P Network,— http:// semanticweb2002.aifb.uni-karlsruhe.de/ proceedings/ Research/Nejdl.pdf
9. *Colin Smythe.* IMS Abstract Framework: A review,- http://www.imsglobal.org/af/IAFReviewvl.pdf
10. IEEE LTSC WG12: Learning Object Metadata,- http://ltsc.ieee.org/wgl2/

O. S. Postnikov

USING SEARCH INTERMEDIARY IN DISTRIBUTED REPOSITORIES OF LEARNING OBJECTS

In the paper the problems of efficient search in distributed repositories of learning objects are considered. The emphasis is set on systems which use search intermediary. Possible ways to improve search performance and relevance in heterogeneous distributed repositories are investigated. Paper outlines several approaches to search improvement, namely aggregational repository, metasearch instructions, request statistics and vector search model.

IEEE Leaning Technology Standards Committee.