DB-IR Integration and Its Application to a Massively-Parallel Search Engine

Kyu-Young Whang
Compute Science Department
KAIST, Korea
kywhang@cs.kaist.ac.kr

Abstract

Nowadays, as there is an increasing need to integrate the DBMS (for structured data) with Information Retrieval (IR) features (for unstructured data), DB-IR integration is becoming one of major challenges in the database area[1,2]. Extensible architectures provided by commercial object-relational DBMS(ORDBMS) vendors can be used for DB-IR integration. Here, extensions are implemented using a high-level (typically, SQL-level) interface. We call this architecture loose-coupling. The advantage of loose-coupling is ease of implementation. But, loose-coupling is not preferable for implementing new data types and operations in large databases when high performance is required. In this talk, we present a new DBMS architecture applicable to DB-IR integration, which we call tight-coupling. In tight-coupling, new data types and operations are integrated into the core of the DBMS engine in the extensible type layer. Thus, they are incorporated as the "first-class citizens"[1] within the DBMS architecture and are supported in a consistent manner with high performance. This tight-coupling architecture is being used to incorporate IR features and spatial database features into the Odysseus ORDBMS that has been under development at KAIST/AITrc for over 19 years. In this talk, we introduce Odysseus and explain its tightly-coupled IR features (U.S. patented in 2002[2]). Then, we demonstrate excellence in performance of tight-coupling by showing benchmark results. We have built a web search engine that is capable of managing 100 million web pages per node in a non-parallel configuration using Odysseus. This engine has been successfully tested in many commercial environments. This work won the Best Demonstration Award from the IEEE ICDE conference held in Tokyo, Japan, in April 2005[3]. Last, we present a design of a massively-parallel search engine using Odysseus. Recently, parallel search engines have been implemented based on scalable distributed file systems (e.g., GFS). Nevertheless, building a massively-parallel search engine using a DBMS can be an attractive alternative since it supports a higher-level (i.e., SQL-level) interface than that of a distributed file system while providing scalability. The parallel search engine designed is capable of indexing 30 billion web pages with a performance comparable to or better than those of state-of-the-art search engines.

CIKM'09, November 2–6, 2009, Hong Kong, China.
ACM 978-1-60558-512-3/09/11.

Categories & Subject Descriptors: H.2.4 [Database Management]: Systems; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Performance, Design

Keywords: DB-IR integration, tight-coupling, DBMS, search engines

References

[1] Abiteboul, S. et al., "The Lowell Database Research Self-Assessment," *Communications of the ACM*, Vol.48, No.5, pp. 111-118, May 2005.

[2] Whang, K., Park, B., Han, W., and Lee, Y., "An Inverted Index Storage Structure Using Subindexes and Large Objects for Tight Coupling of Information Retrieval with Database Management Systems," U.S. Patent No. 6,349,308, Feb. 19, 2002 (Appl. No. 09/250,487, Feb. 15, 1999).

[3] Whang, K., Lee, M., Lee, J., Kim, M., and Han, W., "Odysseus: a High-Performance ORDBMS Tightly-Coupled with IR Features," In *Proc. IEEE 21st Int'l Conf. on Data Engineering (ICDE)*, Tokyo, Japan, Apr. 5-8, 2005. This paper received the Best Demonstration Award.

Bio

Kyu-Young Whang is a KAIST Distinguished Professor and Professor of Computer Science at KAIST. Previously, he was with IBM T.J.Watson Research Center from 1983 to 1990. Since joining KAIST in 1990, he has been leading the Odysseus DBMS/Search Engine project featuring tight-coupling of DBMS with information retrieval (IR) and spatial functions. An earlier version of this technology played a vital role in starting up NaverCom Co. (currently, NHN Co.) in 1997-2000, which is the number one portal in Korea. Dr. Whang is one of the pioneers of probabilistic counting, which nowadays is being widely used in approximate query answering, sampling, and data streaming. One of the algorithms he co-developed at IBM Almaden (then San Jose) Research Lab in 1981 has been made part of DB2. Dr. Whang is the author of the first main-memory relational query optimization model developed in 1985 and reported in 1990 in ACM TODS in the context of Office-by-Example (OBE). This model influenced subsequent optimization models of commercial main-memory DBMSs. His research has covered a wide range of database issues including physical database design, query optimization, DBMS engine technologies, and more recently, IR, spatial databases, data

mining, and XML. Dr. Whang was the Coordinating Editor-in-Chief of the prestigious VLDB Journal, having served the journal for 19 years from its inception as a founding editorial board member. He is a Trustee Emeritus of the VLDB Endowment and served the international academic community as the General Chair of VLDB2006, DASFAA2004, and PAKDD2003, as a PC Co-Chair of VLDB2000, CoopIS1998, and ICDE2006, and as an editorial board member of journals such as IEEE TKDE, The WWW Journal, and IEEE Data Engineering Bulletin. He served as the Chair of the Steering Committee of the DASFAA International Conference and as a co-founder of the Korea-Japan Database Workshop (KJDB) annually held alternately in Korea and Japan. He is a member of the ACM SIGMOD Dissertation Award Committee and served as a member of many 10-year Best or Influential Paper Award committees of VLDB and IEEE ICDE. He served as an IEEE Distinguished Visitor from 1989 to 1990 and was invited to ACM SIGMOD Distinguished Profile in Databases in 2007. He earned his Ph.D. from Stanford University in 1984. Dr. Whang is an IEEE Fellow, a member of the ACM and IFIP WG 2.6.