Scalability of a Distributed Neural Information Retrieval System

Michael Weeks, Victoria J.Hodge and Jim Austin
Computer Science Department,
University of York,
Heslington, York, UK
(mweeks,vicky,austin)@cs.york.ac.uk

AURA (Advanced Uncertain Reasoning Architecture) is a generic family of techniques and implementations intended for high-speed approximate search and match operations on large unstructured datasets [1]. AURA technology is fast, economical, and offers unique advantages for finding near-matches not available with other methods. AURA is based upon a high-performance binary neural network called a Correlation Matrix Memory (CMM). Typically, several CMM elements are used in combination to solve soft or fuzzy pattern-matching problems. AURA takes large volumes of data and constructs a special type of compressed index. AURA finds exact and near-matches between indexed records and a given query, where the query itself may have omissions and errors. The degree of nearness required during matching can be varied through thresholding techniques. The PCI-based PRESENCE (PaRallEl Structured Neural Computing Engine) card is a hardware-accelerator architecture for the core CMM computations needed in AURA-based applications. The card is designed for use in low-cost workstations and incorporates 128MByte of low-cost DRAM for CMM storage.

The E-Science project, Distributed Aircraft Maintenance Environment (DAME), will use AURA technology to process hundreds of gigabytes of aircraft aeroengine diagnostic information. The size of this data means that we are unable to map it to a locally implemented software or hardware CMMs. Therefore, we need to make use of distributed AURA methods that allow a CMM to be striped over multiple PRESENCE cards over a cluster. It is important therefore, that we determine how the performance of the distributed AURA system scales with increasing dataset size.

To investigate the scalability of the distributed AURA system, we implement a word-to-document index of an AURA-based information retrieval system, called MinerTaur[2], over a distributed-PRESENCE CMM. This follows on from a previous paper that compared local and distributed AURA performance [3]. Here we also give updated performance figures for the distributed AURA system, which has since been substantially improved.

MinerTaur comprises three modules, a spell checking pre-processor to identify errors in the user query, a synonym hierarchy to allow paraphrased documents to be matched and a word-document indexing module to identify documents matching particular query words. All modules rely on a fast efficient data structure to under-pin the system. To provide this we implement MinerTaur using binary Correlation Memory Matrices (CMMs) on PCI-based PRESENCE cards in a Beowulf PC cluster named Cortex-1. Cortex-1 consists of seven 500MHz PC nodes connected by 100Mbit Ethernet, six nodes of which contain 28 PCI-PRESENCE cards.

The document corpus used is 571 MBytes in size and contains 476,672 Reuters document abstracts [4]. 62,903 keywords were extracted from the file to index the documents. Mapping this data onto the CMM with a single bit vector representing both word and document, this will fill the whole weights memory available with the Cortex-1 cluster's 28-cards.

References

[1] J.Austin, J.Kennedy, and K.Lees. The advanced uncertain reasoning architecture. In *Weightless Neural Network Workshop*, 1995.

[2] V.Hodge & J.Austin, An Integrated Neural IR System. In, Procs of the 9th European Symposium on Artificial Neural Networks, April 2001.

[3] Michael Weeks, Vicky Hodge and Jim Austin A hardware accelerated novel IR system. In proceedings of 10th Euromicro Workshop on Parallel, Distributed and Network-based Processing, PDP2002. Las Palmas de Gran Canaria, January 9-11th 2002.

[4] Reuters Corpus. Volume 1: English language, 1996-08-20 to 1997-08-19, at http://www.reuters.com/researchandstandards/corpus