Cooperative Biomedical Knowledge Inference

Chun-Hsi Huang, Sanguthevar Rajasekaran, Longde Yin [1]

Keywords: molecular biology, knowledge inference, semantic network

1 Introduction.

The Human Genome Project (HGP) is extracting information from the DNA strands that constitutes human genetic inheritance. The acquisition of a comprehensive human genome sequence imposes unprecedented impact on basic biology, biomedical research, biotechnology, and medicine. It is crucial the massive genomic data produced are well represented so that useful biological information may be efficiently extracted/inferred. The Unified Medical Language System (UMLS) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). The goal of the UMLS is to facilitate associative retrieval and integration of biological and biomedical information so researchers and health professionals can use such information from different (readable) sources [1].

The UMLS project consists of three core components: (1) the **Metathesaurus**, providing a common structure for more than 95 source biomedical vocabularies. It is organized by concept, which is a cluster of terms, *e.g.*, synonyms, lexical variants, and translations, with the same meaning. (2) the **Semantic Network**, categorizing these concepts by semantic types and relationships, and (3) the **SPECIALIST lexicon** and associated lexical tools, containing over 30,000 English words, including various biomedical terminologies. Information for each entry, including base form, spelling variants, syntactic category, inflectional variation of nouns and conjugation of verbs, is used by the lexical tools [2]. The 2002 version of the Metathesaurus contains 871,584 concepts named by 2.1 million terms. It also includes inter-concept relationships across multiple vocabularies, concept categorization, and information on concept co-occurrence in MEDLINE.

In this research work, we focus on designing a distributed UMLS semantic network to cooperatively infer biological and medical information efficiently from distributed information sources.

2 Software and files.

The system construction bases on a task-based and message-driven model to exploit both task and data parallelism while processing queries. Queries are decomposed into tasks and distributed among processors for execution. Other system support activities are also decomposed into system tasks and distributed as well. When a task is completed, a message is generated to either spawn new tasks or trigger further processing, depending on the property and current status of the task. This process is carried out by two collaborating components: the *host system* and the *slave system*. The host system interacts with the user and processes the information for the slave system, while the slave system performs task execution.

The host system is composed of the following major components. The *language front-end* interacts with the user and decomposes the commands into either knowledge or tasks. All the preprocessing and distributing are carried out in the *command processing module*. The *object-oriented packing module* is the communication channel between processors. When the

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA. E-mail: {huang,rajasek,longde}@cse.uconn.edu

slave module finishes a query, the answer messages are then sent back to the *host answer processing module* of the host system to be merged into a final inference conclusion. Some knowledge is kept in the *host knowledge base* for simple queries. Similarly, the slave system has the following components: the *shared knowledge management module*, the *task execution module*, the *kernel message module*, the *task execution engine*, the *load balancing module*, the *duplicate checking module*, the *slave scheduler* and the *object-oriented packing system*.

3 Figures and tables.

Fig. 1 illustrates the software architecture of the host system. Tests of individual components and the overall performace are being conducted on a local Grid, consisting of three heterogeneous systems: a SUN Cluster, an SGI Origin 3800, and a Dell Pentium Cluster. Preliminary experiments demonstrate promising results.

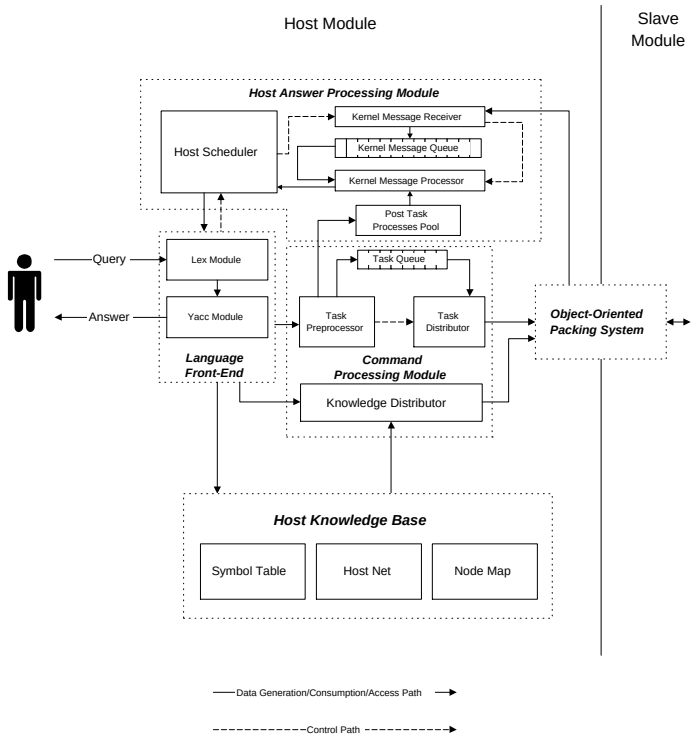

Figure 1: Host System Software Architecture

References

[1] Lindberg, D., Humphreys, B., and McCray, A. The Unified Medical Language System. *Methods Inf. Med. 32*, 4 (1993), 281–291.

[2] McCray, A., Srinivasan, S., and Browne, A. Lexical methods for managing variation in biomedical terminologies. In *Proc. Annual Symposium Compu. Appl. Med. Care* (1994), pp. 235–239.