TIPSTER II ACTIVITIES AT HNC

Marc R. Ilgen, David A. Rushall
HNC Software, Inc., 5930 Cornerstone Court West, San Diego, CA 92121 USA
email: mri@hnc.com, dar@hnc.com

The arrival of the information age has brought with it new challenges for handling the vast quantities of electronically available information. The ARPA and Intelligence Community sponsored TIPSTER program has risen to this challenge. New technologies have been developed for attacking problems in information retrieval, information extraction, and multilingual information processing. As a premier developer of neural network based technologies, HNC has played an important role in successfully bringing innovative approaches to the problem of information retrieval, including multilingual information retrieval. The purpose of this brief paper is to summarize HNC's research accomplishments and to make recommendations for further study.

The HNC approach to information retrieval is based on context vectors, which are unit vectors in a high dimensional vector space. The relative directions of these vectors encode the meaning and context of information that is to be retrieved. Neural network based training laws are used to adjust the components of these vectors in an iterative fashion. The basic context vector methodology has been implemented in a system called MatchPlus, which serves as a stand-alone information retrieval application as well as the technological core upon which additional context vector applications have been built.

One problem with the MatchPlus system has been the large computation requirements of the learning law. In order to reduce the severity of this problem, HNC has developed a one step learning law that approximates the behavior of the original learning law at a fraction of the computational cost. Preliminary performance results with this learning law have been quite encouraging. More complete performance results can be generated when additional funding becomes available.

HNC has also developed a preliminary prototype of an English-Spanish MatchPlus system. This system uses redundant hash table addressing to store context vectors for a multilingual vocabulary. This system can easily be extended to additional languages, such as Japanese, Russian, Chinese, etc. The multilingual MatchPlus approach is able to perform information retrieval in multiple languages without the necessity of specifying any grammatical information about the language, thereby greatly reducing system development time.

HNC's context vector technology has also been extended to the problem of information routing and filtering, resulting in a COTS product known as CONVECTIS. CONVECTIS is currently being used by DataTimes (a large commercial information provider) as the core technology for routing information to customers based on customer specified "interest profiles." The use of context vectors results in a natural method for specifying degree of relationship between incoming news feeds and customer interests. It is also a natural mechanism for detecting novel information themes, a fact that could be advantageously used by the Intelligence Community.

HNC has also responded to the explosion of information available on the Internet by developing a system of autonomous retrieval agents based upon the MatchPlus technology. These retrieval agents can be scheduled by the user to access information at specified, possibly repeated times, thereby removing the requirement that the user actually be present during the retrieval operation. This system is currently being extended to handle multiple Internet and for-fee database protocols, resulting in a useful system for information retrieval from heterogeneous information sources.

Since the context vector methodology requires only that a finite vocabulary of entities be defined for the domain of information, this technology has been extended to other media, images in particular. The ICARS system is being developed to solve the difficult problem of image retrieval and image content characterization. This system attaches context vectors to cluster centroids of feature vectors composed of Gabor features and color information.

Preliminary tests have demonstrated that this system is able to retrieve relevant images without the need for complex image understanding algorithms.

All these applications have demonstrated the tremendous versatility of the context vector technique. Context vectors allow information to be represented in a universal meaning space, opening up the possibility that text, images, and information in other media can be represented in a unified fashion. This technology has also been transitioned to the commercial sector, thereby satisfying a major objective of the funding agencies. In summary, HNC's context vector methodology has spawned a wide variety of solutions for government and commercial customers and holds considerable promise for serving as the technological foundation for future information retrieval and information management projects.