

Report on KBCS-2000*

A Sen

School of Tech & Computer Sc.
Room No. A-234
TIFR, Mumbai
asen@tifr.res.in

PS Dhekne

Computer Centre
D Block Room No. G-7
BARC, Mumbai
dhekne@magnum.barc.ernet.in

S Sanyal

School of Tech, & Computer Sc.
Room No. A-221
TIFR, Mumbai
sanyal@tifr.res.in

This report reviews KBCS-2000, a conference organized by the National Centre for Software Technology, India in December 2000. The theme of the conference was artificial intelligence. After an overview, some promising areas, such as Machine Translation, Intelligent Search Techniques and alternate modes of interaction with machines through speech and script, are described in some detail. The report also summarizes a few invited talks and panel discussions. Finally, it discusses trends and prospects for artificial intelligence technology in the context of KBCS-2000.

1 Introduction

KBCS-2000, an International Conference on knowledge based computer systems, was held in Mumbai, India, December 17 - 19, 2000. It was organized by the National Centre for Software Technology (NCST), India. This is a bi-annual event, with the present one being the third in the current series. About 170 delegates from 11 countries attended the conference. The delegates included scientists from academic and research organizations, students and representatives from the information technology (IT) industries and the Indian government. The aim of KBCS-2000 is to periodically review the progress in the field of artificial intelligence (AI), which despite known limitations and uncertainties, is emerging as a thrust area for providing solutions to IT problems in the coming years.

As is apt for an emerging field, the conference was fairly broad-based, with the aim of acting as a forum for exchange of a wide range of complementary and conflicting ideas within the vast spectrum of AI. The conference also acted as an opportunity for government and industry representatives to meet with the scientists, so that the former can have concrete ideas about the progress and can decide upon future investments appropriately in this 'venture' area. About 50 papers on 12 diverse areas of AI were presented at KBCS-2000. These papers are outlined in this report. Out of this broad panorama, a few themes

seem to show up vividly. Three of them are singled out for special focus in this report, namely natural language processing for machine translation, information retrieval and data mining for Internet and human interface by speech and handwritten script. The conference also included a few invited talks by IT experts, which displayed visions and a thought-provoking panel discussion on data mining, a much-hyped area. The gist of the talks and the panel discussion, not included in the conference proceedings, is provided here.

2 About KBCS-2000

2.1 General

The conference is a bi-annual event, and KBCS-2000 is the third one in the current series. More than 170 delegates spanning across 11 countries attended. Among the delegates, scientists from government and companies, academicians from various organizations, research program advisers and directors were present. Students from IEEE and other organizations also benefited from this high-level conference. About 50 research papers, chosen by an international panel of referees, were presented in 15 technical sessions running mostly in parallel. In addition, there were invited talks by well-known professionals and a panel discussion on "Has Data Mining Proved its Worth?". The first day of the conference was devoted to the pre-conference tutorials on "Web Intelligence: Goals and Challenges" and "Lexical Semantics and Knowledge Representation: Some

* Reproduced from the Asian Technology Information Program (ATIP01.011) with permission.

Challenges, Formalisms and Tools". About 50 participants attended each tutorial. The International Federation for Information Processing (IFIP) and the Computer Society of India sponsored the conference. It was also supported by the software industry (see the Weblinks and Contacts section for details). Some sponsors exhibited a few relevant software products.

2.2 Scope

The theme of the conference was AI. Within this, it was fairly broad based. The stated aim of the conference was to act as a forum for exchange of ideas, complementary and conflicting, within the vast spectrum of AI. Accordingly, the papers chosen represented a wide number of areas, and often presented alternate methodologies for achieving identical goals. However, there was a distinct slant toward thrust areas, such as information retrieval and natural language processing (NLP), that have assumed great significance in the wake of the Internet revolution. Also observable was the marginal predominance of technology over pure theory in the research papers presented. The areas covered by this conference are as follows:

- character recognition
- data mining
- intelligent agents
- information retrieval
- intelligent tutoring system
- knowledge engineering
- natural language processing
- neural networks
- planning and scheduling
- reasoning
- robotics
- soft computing
- speech processing

Section 3 gives an overview of the presentations. Sections 4 through 6 describe the theme topics of the conference in some detail.

2.3 Participation

About 170 delegates attended the conference from 11 countries, namely the US, Japan, the United Kingdom, Germany, Spain, France, Russia, China, Korea, Australia and India. More than half of the participants were from the host country, India. All three invited speakers are experts of Indian origin, now settled in U.S. The participants were predominantly research scientists in various areas of AI. The authors of the papers were also mostly affiliated with either academic institutions (Departments of Computer Science, Information Technology, Automation, etc.) or with applied research laboratories. Response from industry, however, was not insignificant, and several were conference sponsors. Their delegates followed the R&D presentations keenly, presumably with an eye to future collaborations. There were product demonstrations, and a few paper presentations by the delegates from Indian IT organizations. The Government of India, too, has displayed interest by the participation of a few very senior level officials from the Ministry of Information Technology (including the secretary) to the conference. As far as the host country, India, is concerned, the best AI research laboratories were represented in the conference. The IT houses which were sponsors and participants are the leading ones in India. As for the overseas participation, the conference did not succeed in attracting the best names in AI. Nevertheless, the works presented by the overseas authors were interesting and thought provoking without exception.

2.4 Significance

AI has established itself as an important field of Science and Technology over the last two decades. The growth was uneven and often unpredictable: particularly in the areas requiring formalization of common sense knowledge, e.g., vision, speech and machine translation. Interest from industry and governments, however, remains strong for the simple fact that the stakes are too high to ignore this field. With falling prices of hardware and availability of improved resources, attention in computer and IT is now turning towards more demanding tasks ranging from providing better human interfaces (such as speech input-output, handwritten script interface, natural language processing) to solving difficult practical problems (such as resource scheduling, process control,

machine translation). Techniques from fuzzy logic, neural networks and case-based reasoning have been deployed to provide practical solutions in a number of areas. With the Internet boom, automatic classification and extraction of information and data have emerged as important needs.

For Asian countries, AI has some additional roles to play. For example, countries such as India, China and Japan have unique scripts and languages and need special interfaces, e.g., for speech or hand written characters, to draw their native populations into the IT revolution. On the other hand, the developing nations need, and are in acute shortage of, knowledge of advance technology. Knowledge based computer systems can also help to bridge this gap in a big way. Under this backdrop, KBCS-2000 assumes a special significance. Its avowed aim of acting as a forum for exchange of notes between AI scientists across a wide range of disciplines has, no doubt, been reasonably well met.

Probably more importantly, it also acted as a meeting ground for researchers and academics with those from government and industry. For the latter, who are willing to put money into AI, but who are not sure where or when to put it, KBCS-2000 presents a broad view of the scenario and provides many important details, too. The results presented in the conference were not earth shaking. However, they were often significant and worth taking note of, particularly in the Asian technological context. Areas of AI, such as speech, script and NLP, are population-specific anyway, and need indigenous R&D. The subsequent sections of this report attempt to present the conference in this context. It presents a panorama, then zooms on some landmarks, and finally concludes with a few pointers.

3 Presentation Overview

As mentioned earlier, this conference presented a broad view of the AI R&D scenario, and was not sharply focused on any one or two areas. However, a few trends naturally predominated and are highlighted in the following sections. NLP is an area where a great deal of effort currently is being applied. A majority of the NLP efforts seem to be directed toward MT. The conference presentations toward this direction are summarized in Section 4. With the explosion of the Internet, applying intelligence for information retrieval has assumed

prime importance. Section 5 describes the presentations on information retrieval and data mining, which were mainly oriented toward Internet usage. Providing input/output modes other than via English-type keyboard-monitors for computers is an important issue, particularly for Asian countries with languages and scripts strikingly different from English. Section 6 sums up the presentations related to alternate human interfaces to computers, through speech and hand written scripts.

Needless to say, there were many other interesting sessions and presentations. In robotics, three presentations from the Bhabha Atomic Research Centre (BARC), India, mainly focused on the advanced problems of collision avoidance and robot path finding, which are very important for application areas, such as hazardous material handling. Another paper threw light on the design of a PC-based four finger robotic hand.

In intelligent agents, VK Rai has claimed to have resolved the famous principal-agent problem. F. Macia-Perez, Spain, proposed a mobile agent-based model for a network node regeneration system that will be useful for coping with the ever-increasing size and complexity of present communication networks. K.R.K. Murthy proposed network distributed collaboration' (NDC), particularly for the dynamic ensemble of information activity infrastructure (IAI). This is expected to optimize usage in web personalization, information dissemination and gathering and in web mining.

In intelligent tutoring, VS Kumar presented a system that supports the helpers in a helpdesk system by suggesting which help scheme may be useful for providing assistance in each given situation. The system is operational in an academic campus. P. Ravi Prakash presented 'Vyasa', a system that automatically generates multiple-choice questions for a computerized test.

In reasoning, G Vijaya Kumari presented interval duration network (INDU), a proposed extension of interval algebra for modeling qualitative information about interval and duration in a single binary constraint network. Michael Emmerich, Germany, dealt with knowledge integration for thermodynamic flowsheet synthesis and optimization, and presented a technique to represent knowledge about interval domains of state variables in chemical plant networks.

For neural networks, MS Puranik presented an

application for automatic trajectory plotting of an object under radar surveillance. His claim is that this AI-based approach will be superior to conventional pattern matching methodology. P Bhattacharyya discussed the applicability of a back-propagation algorithm for off-line signature verification. In planning and scheduling, S Kambhampati made an excellent presentation on 'AltAlt', a planner that successfully combines the advantages of two competing paradigms, namely graphplan and heuristic state search [PS1]. Sudeshna Sarkar presented a logic PLTL-B that embeds CTL subformulae within a past linear temporal logic (PLTL) framework. In knowledge engineering, GK Palishkar presented a formal framework to represent fuzzy knowledge. KRK Murthy proposed concept network (CNW), a directed acyclic graphical structure for context sensitive document representation with the aim of information filtering. Neither of the two soft computing papers published in the proceedings was presented in the conference.

4 Natural Language Processing for Machine Translation

NLP is related to a number of AI applications. Probably the most important is MT, and that was reflected in KBCS-2000, where most of the NLP presentations had direct or indirect relevance for MT. Fully-automated general purpose high-quality machine translation (FGH-MT) systems are cherished goals, which have yet to be realized anywhere. The creative aspects of translation and the global knowledge associated with human translation cannot easily be conferred to a machine. As of now, MT is, therefore, applied either semi-automatically or in a domain-specific manner. Such efforts are still important, particularly in view of the enormous amount of information available on the web. Manual translation is a skilled and time-consuming job. However, if a machine can take some load from a human translator, substantial time can be saved. This approach is quite important for local newspapers and TV channels, which receive their input mostly through networks. India is a multi-lingual country, and considerable efforts are underway to develop MT between English and Indian languages, and between Indian languages. It is comparable to the situation in Europe, where MT between English and continental languages, or between the continental

languages are important issues.

1. In his presentation, Rajeev Sangal described 'Anusaaraka', a type of MT system that makes text in one Indian language accessible through another. The machine presents an image of the source text in a language close to the target language. Such systems have been built between Hindi, the main Indian language, and five regional languages, and are available for use through E-mail. The system presented is semi-automatic. While a person is better in resolving ambiguities by using world knowledge, a computer can better handle the language processing part. The authors claim that their approach of building language access in Anusaarakas allows rapid development of systems by separating the analysis based on language and that requiring world knowledge. They take the view that language encodes information, and the information can be extracted and re-expressed in the target language by enhancing the target language with additional notation. They claim that a user, after some training, learns to read and understand the text in this 'new dialect' of the target language. The output may be post-edited by a trained user to make it grammatically correct and stylistically better. Anusaarakas follows the principle of substitutivity and reversibility of the string produced, which implies preservation of information while transiting from one language to another. The Ministry of Information and Technology, Government of India, under their Technology Development for Indian Language (TDIL) program, initially funded the project. A project for systems between English and Hindi is being supported by Satyam Computers Pvt. Ltd., and will be available as "free" open-source software under GPL.
2. Another MT system 'Anubaad', for translating news items from English to an Indian language, was presented by Sivaji Bandopadhyay. The paper described the prototype development of a knowledge driven generalized example based system that translates short single program news items from English to an Indian language. News headlines are translated using knowledge bases and example structures. The sentences in the body of the news items are generally

translated by analysis and synthesis. Word sense disambiguation is done at various levels. Some semantic categories are associated with words to identify the inflections to be attached with corresponding words in the target language, as well as to identify the context in the sentence. Context identification is also done by the recognition of idiomatic expressions and using context templates for each word. The knowledge bases include bilingual dictionaries for names of places, month, date, news agencies, dignitaries, political parties, organizations, festivals, etc., as well as directories and tables for synthesis in the target language. Translation examples of news headlines, both specific and general, are stored in the example base. The example base also includes grammatical phrases in the source language and their corresponding mapping in the target language. The methodology is claimed to be extendable to other Indian languages.

3. Durgesh Rao presented a human-aided MT system for translating English news to Hindi. Using world knowledge and heuristics, the topic of the news story is identified. The processed text is analyzed and tagging of the parts of speech is done. The text is then transferred into a case-frame like structure. Generation of target language is then achieved by using the parameterized templates from the case-frame structures and the bi-lingual lexicon. The authors presented a practical framework for syntactic transfer of compound-complex sentences from English to Hindi in the context of a transfer based Machine Assisted Translation (MAT) system. The analysis is based on the linguistic instruction of the authors, backed up by evidence from a real-life corpus, and ongoing work on a practical MAT system. The description of the framework is based on a template-like representation. The most important component of the framework is the mapping of finite, as well as non-finite, verb groups, in order to cover simple, as well as compound-complex, sentences. The authors have adopted a pragmatic approach by combining the best of human linguistic intuitions about how to solve various issues involved, with statistical evidence that helps them in prioritizing the most important

issues. In order to have firm analysis, authors use a representative parallel corpus in English and Hindi and carry out syntactic transfer, taking a structure representation of the source text, and mapping it using a heuristic rule based structure that is appropriate for the target language. The framework presented was informally analyzed and found to be adequate to handle the range of sentences encountered in the parallel corpus. The authors have implemented it for mono-finite sentences, and are in the process of extending it to cover multi-finite and compound-complex sentences. It is claimed that this framework would be fairly intuitive, easy to implement, and may be maintained without very elaborate linguistic knowledge.

4. G Vazquez, Spain, presented a paper detailing a few lexical semantic mismatches between Spanish and English from the point of view of MT. A verb here is assumed to be pivotal for translation, and the mismatches formulated are accordingly verb-centered. The mismatches presented are of the following broad categories:
 - (a) the verb in one language doesn't have a match, with the same meaning, in the other (lexical-semantic mismatch),
 - (b) a 'construction' in one language does not have a match in the other (lexical semantic compositional mismatch),
 - (c) the way the verb interacts with its arguments are different in the two languages (argument structure mismatch), and
 - (d) the 'events' associated with equivalent items in the two languages are different (event structure mismatch).

The paper also presented a way to resolve such mismatches, for possible MT application, by using 'conceptual transfers'. An example of resolving a mismatch, based on diathesis alterations, was provided.

5 Information Retrieval and Data Mining for the Internet

Efficiency in storage and retrieval of information and data has assumed great importance in the wake

of the explosion of the Internet. The many related jobs of classification, indexing, structuring, and so on, cannot be done entirely by humans. Current trends are to apply more and more intelligence for information retrieval. Often we do not have enough idea about the data patterns, particularly when the scenario changes rapidly. Data Mining is the concept that attempts to dig into the data, discover hidden patterns and create the framework automatically. Naturally, KBCS-2000 included a number of information retrieval and data mining papers. Most of them were inclined toward Internet applications. The following subsections summarize a few significant presentations.

1. Vivek Balaraman discussed the possibility and advantages of using case-based reasoning (CBR) for intelligent retrieval of data. Locating the relevant document over large data volumes, especially in the presence of noise and inexactness, is a problem faced by server application domains. Standard tools in such a domain depend on exact correspondence between the query and the search record, leading to brittle, unintelligent systems. By viewing the data volumes as case bases, all the advantages of case-based retrieval, such as intelligent search, high precision and recall, are available to the application users. CBR is a computational model of instance- or precedence-based problem solving, and has been employed in many commercially-available software packages as a complementary approach to reasoning. CBR does not look for exact correspondence between the input problem and the stored cases. Considerable work in CBR was focused on defining what similarity means in different domains. A CBR tool 'Consult' has been developed by Tata Consultancy Services, India, using the K-Nearest Neighbor (KNN) algorithm to determine the similarity between cases. The 'inexact matching' capability of Consult is exploited for domains that require intelligent search over data sets containing noise and incomplete input information. For this paper, the existing matching algorithm in Consult was modified to cater to large and differently structured data sets. This new implementation tries to reduce the retrieval time, while retaining the power of the search to handle incomplete/noisy information by

applying a graded retreat approach, i.e., to begin with a high precision, unforgiving search and gradually relax the criterion to increase the search space during various stages. It is claimed that intelligent directory assistance and skill-based personnel selection type of applications will greatly benefit from this approach that provides improved search power and quality of retrieval.

2. A Kannan presented a system that provides a knowledge based search technique with an effective user interface for searching research papers from the Internet. At first level, user-supplied keywords are used to get relevant URLs from the Internet via conventional search engines. Then the system explores each and every URL to check whether it corresponds to a relevant research paper, by applying heuristic search techniques. If there happens to be a research paper, certain algorithms and production rules are applied to determine how far that paper meets the requirement of the user, and the results are stored in the database. The system provides an effective query language interface. It provides past searches and citation lists, which can be used to streamline searches. The author points out that conventional searching tools on the Internet support broad and general queries, and produce a large volume of irrelevant documents and consume inordinate time. The author claims that a rule-based search helps the researcher to select focused areas and easily go to deeper levels. Lack of natural language interface is another limitation of current search engines, which this system has overcome. The system provides the user with three interfaces: a natural language interface, an extended structural query language (ESQL) interface with rule extension and a keyword-based interface. It makes use of the standard search engines for searching the Web. The URLs collected are again explored and filtered to provide the specific URLs to the users, along with some information about their contents. For this purpose, the system uses a set of event-condition-action (ECA) rules. Moreover, various search results are stored on a local database so that the user can search through it before the actual search on the Web, thus, saving substantial time.

3. Dave Shachi presented a scheme for knowledge extraction from text for Hindi, the main language in India. Knowledge extraction from documents is important for Internet-related activities, such as multilingual access to the interface, information retrieval, text summarization or automatic linking of hypertext. NLP techniques are usually employed. The presented work represents knowledge extracted from documents in an electronic language called universal networking language (UNL) expressions. Sentence information is represented in UNL as a hyper-graph having concepts as nodes and relations as areas. Concepts are represented as character strings that provide further information about the node where the concept is being used in the specific sentence. The 'Enconverter', a language independent parser, is used as the analysis tool for Hindi, and makes use of the UW-Hindi lexicon and syntactico-semantic rules for the language. The Enconverter scans the input string from left to right, and all matched morphemes with the same starting characters are retrieved from the dictionary. The lexicon contains information about every word, such as parts of speech, the possible semantic role it can play, its gender and other information, and maps universal words into Hindi words. The paper describes morphological analysis, objective morphology, verb morphology and their usage. It gives an idea of which rules are applied under what conditions. The author claims to have implemented a system to handle almost all the relation labels given in the UNL specifications and all sentences including single, clausal, interrogative and imperative types. This system currently can handle 3000 rules, and the lexicon size is about 60,000.
4. Lipika Dey presented a rough set based filtering scheme for customized information retrieval, which intends to achieve search efficiency by user-centric approach and user profiling. User behavior can change over time, hence the user profiles should be updateable with constant monitoring of user actions. The authors propose a rough-set based filtering scheme to build domain specific customized filters at the user end, instead of the present practice of server-based filtering which does not take care of user interests. A tool 'Extractor', that extracts user interests by analyzing user reactions, initially creates the domain dependent filtering agents, called "sieves". The user reactions are analyzed using a rough set based methodology that exploits the structure of web documents and combines user intelligence to generate a user profile. The profile, once available, is used to filter unwanted documents. The paper details the design of the filtering scheme, how to extract user interest, computation of rank of a document, development of information systems using rough sets, document classification, etc. The results obtained by using the proposed scheme were also presented. It was suggested that users could design sieves at their end to encode their interests within a particular domain in the form of document structures. The authors argue that sieves cannot be generalized, since even within the same domain, different users can rate the same document differently.
5. Hiranmoy Ghosh presented a novel approach for knowledge based query interpretation for multimedia retrieval. In this approach, a conceptual query is interpreted as a set of expected media patterns, which are in turn used for feature based retrieval. The authors claim that this approach is well suited for open and extensible retrieval systems, where incompleteness of the domain knowledge does not permit logical deduction of conceptual objects from observed media. The authors also introduce a generic query model that combines expressiveness of conceptual, as well as feature based specifications, and described a new domain knowledge representation and reasoning algorithms for interpreting a generic query as a logical combination of expected media patterns. A content-based retrieval system requires semantic interpretation of the contents of multimedia documents with some domain specific knowledge. Most contemporary retrieval systems use a deductive approach or probabilistic logic. In this system, observed media patterns are interpreted as conceptual objects to be compared with a query. Previously, the authors had proposed an open-agent based architecture for multimedia retrieval where different system components can be

independently developed and dynamically plugged into the system. This architecture allows rescue of feature-based retrieval capabilities in conceptual queries. In this paper, the focus is on the interpretation of semantic queries with domain knowledge. The authors' contributions include:

- a) a new model that combines expressiveness of semantic and feature based specifications,
 - b) a knowledge description language that can be used to describe the domain knowledge required for query interpretation in any application domain, and
 - c) an inferencing algorithm that can reason with the query model and the domain representation to derive the search specifications for the feature based retrieval engine.
6. Llango Krishnamurthy presented a method for the handling of video data, and proposed a novel content-based retrieval model namely, the Entity relationship diagram-based model (ERDBM). The data model handles the semantics at the various data granularity levels within a chosen framework. The query model parses the natural language query into query form, which is then matched with the video database using a three-layered similarity-matching model. Retrieval of structured records from traditional databases is based on the values of entries in the fields. However, multimedia databases include unstructured data, hence retrieval is possible only after looking into the contents of the data. On the proposed ERDBM model, video objects are handled at various data granularity levels. The paper provides details about the ERDBM model, a new CBVR model, the ERDBM query model, the similarity-matching model and the algorithm, which has design choices at various layers. The proposed ERDBM is NLP oriented. The authors claimed that they found a fuzzy model more suitable than a probabilistic model at the top layer, and observed that weights associated with the feature values do not influence the effectiveness of the system at the middle layer.

7. Sudeshna Sarkar presented a paper on automatic generation of 'Taxonomy', i.e., classification scheme, for information repository. With the surfeit of information documents must be efficiently organized in the repositories for efficient storage and retrieval. This paper presents an algorithm 'Chain' (clustering with hierarchical and incremental approach) that automatically evaluates taxonomy. For each document, a set of key phrases is first automatically extracted. These are used to group the documents by unsupervised and incremental clustering. Subcategories are created when some threshold is exceeded. The algorithm provides for dynamic re-evaluation of taxonomy with the addition/deletion of documents in the repository and permits multiple links to a class. The algorithm was tested and it was found that, due to the lack of semantic knowledge, it could not do better than a human in initial taxonomy generation. However, it can be useful if the initial clustering is done semi-automatically, and can be effectively used for subsequent on-line upgrading of taxonomy.

8. V Uma Maheswari presented a paper on mining web usage graphs. This paper considers the mining of web usage graphs as a classification problem, and defines a new metric, based on cumulative graphs containing positive and negative example sessions. How this matrix is useful in classification was illustrated through experimental results. It is shown that:

- (a) cumulative graphs are built efficiently,
- (b) these graphs can be built incrementally and are scalable to very large data sets,
- (c) entire training data set need not be brought into memory, and
- (d) a test session can be classified in time that is proportional to the size of the test session and cumulative graphs.

The results obtained from the experiment showed that 94 percent of the test examples are classified correctly.

9. Chen Ning from PRC presented a paper on clustering data with mixed data types. Clustering has been recognized as an

attractive task of data mining and knowledge discovery in databases. Most clustering algorithms are designed for databases having a single data type. This paper discusses the similarity of different data types, and presents a reasonable definition of similarity measure among records with mixed type, based on geometric adjacency and information gain.

The author proposes a grid-based clustering algorithm, GCMD, by finding the connected components for data with mixed types. Clustering is an important task of data mining for identifying distribution patterns in databases. However, none of the clustering problems deal with mixed data types, while real-life data often involves mixed types. A reasonable definition of similarity measure among records with mixed data types, based on geometric adjacency and information gain, was presented in this paper. A grid-based clustering algorithm is proposed for finding connected components for a database with mixed data types so that it can become insensitive to the order of input data, noise and 'outliers'.

10. Nitesh V Chawla presented 'Smote': Synthetic Minority Over-Sampling Technique for data mining. He described classifiers to deal with imbalanced datasets by under sampling the majority class combined with over sampling of the minority class, with resultant better classifier performance. This new method of over sampling involves creation of synthetic minority class examples, by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k-minority class nearest neighbors. Results with application of combined Smote and under sampling shows better performance than plain understanding. This has been tested and confirmed over a variety of data sets.
11. MN Murty presented a technique for mining multiple databases for inter-database association rules. Mining of association rules are based on a large collection of data type transactions. In reality, data may be stored in several databases with an implicit association among various parts of this data. There may be salary information of employees in an 'Employee' database and goods purchased in the 'Customer' database. Associations

between these two databases, generated by association rules, are called inter-database association rules. The present novel approach uses a knowledge-guided process (in the form of a semantic network) to generate association rules between sets of values. The following points are achieved:

- (a) data structures to find associations between two sets of values, which semantically partition the resultant database, are generated,
- (b) such a partition helps in generating large set of values by scanning the resultant database only once, and
- (c) in a single step, it is possible to generate the association rules directly from the large set of values.

The above methodology is quite efficient, as it needs only a single scanning step, thus saving time and total computational effort.

12. DK Subramanian presented a paper on And-Or (AO) concept taxonomies for mining generalized association rules. The And-Or taxonomy is a new taxonomy (traditional approach uses only an 'is-a' taxonomy). Only one database scan is more economical than the minimally two database scans of all earlier approaches. The proposed AO taxonomy is based on items and concepts/functions. It is sound, as the scheme uses knowledge for mining only those concepts that are of interest to the user. It is interactive, as the user can adjust a controllable parameter called level of interestingness. The proposed AO single scan, used to generate all interesting generalized association rules between seemingly unrelated concepts and associations, is interesting. The AO single scan generates meaningful association rules using less computational resources than previous approaches.

6 Human Interface by Speech and Handwritten Script

An important area of AI is the development of human computer interfaces, so that users have more freedom in the ways they can interact with computers. Speech and handwritten script are two modes that can widen the scope of computer usage

substantially. This is particularly true for Asian countries, such as China, Japan and India, with their special scripts and languages, and with large populations uncomfortable with an English keyboard-monitor environment. The conference had two papers on speech processing and two on handwritten character recognition, all from India. The relevant technologies in India fall marginally short of the world level, but are fast catching up. The efforts are significant in view of the large number of potential beneficiaries, and considering that there is no alternative to indigenous development in these areas.

1. Text-to-speech systems (TTS) can support useful applications, such as access of computerized information or reading of e-mail over a telephone. An 'Indian English' TTS was presented and demonstrated by A Sen. This reads out English text stored in a computer with pronunciation and accents acceptable to native Indian users who know English. The technology utilized is 'formant' synthesis, with a standard source-filter speech production model. The model parameters are varied, using indigenously developed rules suitable for Indian speech sounds. A front-end Indian English text analyzer provides pronunciation and accent information to the formant synthesizer. The text analyzer uses a small Indian English phonetic dictionary, applies improvised techniques of morphological analysis with alternative root hypotheses and as a fall-back option, and uses letter-to-phoneme rules appropriate to Indian English. There is a special provision to detect Indian names and pronounce them appropriately. The synthesizer has a very flexible structure for improvement through research. The TTS is being integrated with demonstration systems for e-mail and web page readers.
2. Coding of 'transparent audio', which consists of speech plus other sounds, is important in view of storage and transmission of audio on the web. Standard LPC-type coders use a speech production model for coding, and hence are not suitable for this purpose. An algorithm for compressing transparent audio, using wavelet representation of the signal, was presented by PS Sathidevi. An analysis-synthesis method was used to optimize the coder coefficients by minimizing

the error through iterations. The framework of the coder takes the subjective criteria of human perception, e.g., critical frequency bands, into account. Encoding in this algorithm is slow, whereas decoding is fast. Hence, the method is suitable to deliver stored messages over the Internet. A subjective evaluation test of the algorithm was done, and the results were found to be acceptable for a wide variety of sounds.

3. A 'writing pad' for feeding input through handwritten script in two Indian languages (viz. Kannada and Tamil) was presented and demonstrated by R. Srinivasa Rao Kunte. Indian scripts have superscripts and subscripts. To enter text through English-like keyboards, techniques such as compose characters are often used, but none are really natural. This alternative mode for input through a graphic tablet type of device is expected to make text entry easier for a lay user. The technique can be extended to develop small note-pad type computers in Indian languages. The script recognition system is on-line, i.e., it must have timing information. It uses wavelet features and a feedforward neural net classifier with a single hidden layer. The system was trained to recognize even English letters and alphabets interspersed between Indian scripts. It can also be extended to other Indian scripts, with suitable training. The only restriction is that the characters are to be written in a non-cursive way. The scripts of the two languages selected are non-cursive by nature. The accuracy of recognition is currently about 95 percent.
4. Another paper presented an off-line handwritten character (numeral) recognition system by SK Jayanti. An off-line character recognizer does not need time information and is suitable to read scanned images. While the basic idea is not new, the methodology, i.e., application of Fuzzy approach to solve the problem, is novel and promising.

7 What the Experts Say

The conference included three invited talks. Noted IT professionals delivered two talks, and an eminent academic delivered one. There was also a panel

discussion with experts in the field. All these, in short, present a vision, show the way the things are likely to be in future and indicate how to direct immediate activities toward the new frameworks that are going to emerge. A summary of the three invited talks and the panel discussions are presented here with that purpose in mind.

1. Dr R Uthurusamy, GM Research Labs., USA, delivered the inaugural lecture on emerging technologies, and set the tone of the conference. He spoke about the impact of emerging IT and advances in web-enabled environments. He felt that these rapid changes have made a tremendous difference in the computing setups in the (particularly U.S.) automobile industry. He illustrated his point with changes in business models with the rapidly-changing IT landscape, and presented case studies to show its potential benefits, when innovative automobile applications were deployed with AI based techniques. He mentioned that the spending of the automobile industry on IT R&D has increased over the years, and presented some figures of R&D expenditures of large companies, such as Ford and GM as illustrations. He commented that GM's new policy is to give customers 'what they need', a change from the earlier policy of 'give them what they want'. He emphasized cost reductions, rapid deployment of new technology, use of COTS, and predicted that the next big thing in the auto industry will be widespread use of robotics technology and intelligent assist devices (IAD). He also felt that the present Internet environment, which is data centric, would become people centric in the coming future.
2. Dr Ramesh Jain, CEO, PRAJA Inc., USA, delivered an invited talk on digital experience. He mentioned that in the rapid evolution of IT, we have moved from data processing to information processing, and the next move would be to experience processing. He defined experience as a direct observation and participation in an event, and believes that the digital experience will be the next natural major step in technology evolution, and added that contextual search and inference are keys to the development of a realistic experience. He explained that people's desire to communicate their experience has led to

major inventions, starting from spoken language, writing paper, print media, telegraph, telephone, radio, TV, recording media and digital processing, to the present multimedia Internet. He examined the role of knowledge search, perception systems and visualization approaches to realize a digital experience. He believes that this will impact every aspect of our society, including education, business and health care. He said that the current Internet technology represents the world's largest library of documents, and searching documents on the web is transitioning from simple keyword searches to updates and alerts, media search, contextual search, to immersive search. He pointed out that people want to know about every event happening anywhere in the world instantaneously, without waiting to logon to the net. Web searches, therefore, should be event driven. He predicted that creation of real experience by capture from any real world source (via text, tactile, video, etc.) based on event model and ability to transmit it through any device will be desired by users. He further predicted evolution of content management systems into experience management systems in the future.

3. Dr Subbarao Kambhampati, Dept. of Computer Science and Engineering, Arizona State University, USA, delivered an invited talk on Integrating Planning and Scheduling: Status and Prospects. Many real-world decision-making tasks need the ability to synthesize courses of action, and schedule them optimally in the presence of resource limitations and changing mission goals. For most of their over-30-year history, planning and scheduling have been seen as related problems that are solved by separate types of techniques. The speaker pointed out that planning, which can be informally characterized as action selection and causal reasoning, has been the domain of plan synthesis techniques in AI, while scheduling, which involves action sequencing and resource allocation for a pre-synthesized set of plans, has been addressed by operations research and (more recently) constraint satisfaction communities. He cautioned that despite the fact that most real-world tasks involve aspects of both planning and scheduling, no

integrated approaches to planning and scheduling have thus far been proven practical. He mentioned that one of the main obstacles for integrating planning and scheduling has been the fact that the plan synthesis problem has in general been much harder to deal with computationally. A related issue is that while plan synthesis relied on predominantly search-based methods, the most promising scheduling approaches relied on compilation approaches, wherein the scheduling problem is compiled into an instance of either a constraint satisfaction problem (CSP), or a mixed integer/linear programming problem (MILP). He said that recent developments in plan synthesis have lead to significant scale-up in plan-synthesis capabilities, as well as made clear the connections between planning and other combinatorial workhorses, such as CSP and IP. These developments have significantly improved the prospects for effective integration of planning and scheduling phases. The speaker first surveyed the state-of-the-art in plan and schedule synthesis in AI, and then approaches for integrating them were discussed. Special attention was paid to his RealPlan project at Arizona State University that proposes a loosely coupled architecture for integrating planning and scheduling.

4. Panel Discussion. The conference panel discussion provided a fertile ground for open discussion on the pros and cons of an emerging technique. The panel topic "Has data mining proved its worth?" reflected the potential opportunities and daunting challenges that confront researchers in bringing AI applications to the commercial market place. Dr. PVS Rao, Head, R&D Unit, Tata Infotech, chaired the panel discussion. The other panelists included Dr Sunita Sarawagi, Dept. of IT, IIT, Mumbai, Dr Sarabjot Singh Anand, CEO, MINEit, USA and Dr S Sengupta, Tata Infotech. Data mining has been a promising area of interest to EDP managers, database users and AI researchers. The panelists tried to determine if these expectations were fulfilled or not. However, as might be expected, the panelists were divided in their views and expressed differing opinions. Dr. Anand gave examples of successful deployment of data mining in

telecom, fraud detection, fault analysis in manufacturing, health care, etc., and argued that data mining has to a great extent delivered on its promises. Dr. Sunita disagreed and pointed out the problem of scalability, lack of penetration in common applications and high cost as major reasons for the failure of AI concepts. She felt that data mining has been used only in large applications where cost is not an issue, and stressed that the projected promise of AI and usefulness has not been realized. Dr. Sengupta, on the other hand, cited a few examples to show that on certain applications data mining has been useful. It was felt that while data warehousing and Internet technology have reached public consciousness, the same is not true for data mining. The general opinion was that getting the right kind of answers through data mining software requires a great deal of human-machine interaction, as well as domain specific knowledge, and the systems are not yet particularly user friendly. However, there was general agreement among all participants that even though full fledged data mining applications have yet not reached wide spread popularity, the technique has been already employed in many software products.

8 Conclusion

KBCS-2000 achieved reasonable success in its effort to act as a broad forum for exchange of ideas among AI researchers, government and industries. This should help to enrich the ideas of the researcher community, and bridge the gap between science and technology in the field of AI. The conference also displayed some technologies emerging in Asian countries, including Japan, India and China, which are not far behind world standards. Despite uncertain immediate outcomes, governments and industries the world over are not averse to sponsoring AI projects and activities, and this displays future vision. It is now becoming increasingly clear that to keep up the tempo of the ongoing IT revolution, intelligence must be applied to computer applications.

The Internet has emerged as a decisive global technology. Users, however, will be choked by the information, if intelligence is not applied to search engines. A large number of KBCS-2000 papers

addressed this problem. Another important issue is to allow users access to computers in more natural ways. Future computers must support multiple languages, and alternate modes of interaction, such as speech and hand written script. Presentations in speech, script and NLP sections pointed toward this direction, and are of particular significance for Asian countries, such as Japan, India and China, with languages and scripts strikingly dissimilar from English. MT was the core issue of NLP. Even a partial success in this effort can result in considerable time and expenditure savings in areas such as media and publications. Several other areas, such as intelligent agents and robotics also show significant potential, and are surely to be in the focus of attention in the near future.

In the field of AI, a clear gap exists between the technological expectations and reality. A conference such as KBCS helps people in the labs and people within companies to understand each other's position more precisely. The situation is such that even the partial solutions that AI can provide to practical problems can be of immense benefit. Prudent ventures are therefore possible and necessary. Proper appraisal of the situation and periodic reviews are prerequisites of such efforts, and KBCS-2000 helped this in a limited way.

There were many interesting proposals and theoretical studies among the presentations. There was also a very noticeable trend to apply the results, even if those gave only partial solutions. Rajeev Sangal's semi-automatic MT scheme is a case in point, and indicates the course of events to come, where the respective strengths of man and machine will be utilized and integrated into intelligent systems. There was a fairly good level of representation in the conference from the host country, India. Unfortunately, despite the best efforts of the organizers, the quality and level of overseas participation has not really increased over the years. This is an aspect where more attention needs to be paid. Japan is a first-world country in knowledge based computer system R&D. It is important that the conference attracts more representation from that country, so that lessons can be learned from their experiences. Overall, KBCS-2000 was a rewarding event for all concerned with computer and information technology in general, and was more so for the AI community, in particular.

9 Weblinks and Contacts

9.1 Sponsors

- International Federation for Information Processing [IFIP]
<http://www.ifip.or.at>
- Computer Society of India
<http://www.csi-india.org>

In co-operation with:

- Association of Computing Machinery
<http://www.acm.org/chapters/bombay>

It was also supported by:

- Tata Infotech Limited
- Tata Consultancy Services
- Tata Research Development & Design Centre
- Tata Consultancy Services
- Intel Asia Electronics Inc.

[The proceedings of the conference was published by Allied Publishers, New Delhi and copies are available from them. A few copies are also available at NCST. Please write to Editor, Vivek if you are interested.]

Vivek

A Quarterly in
Artificial Intelligence

Volume 14, Number 1
January 2001

Editor:

S Ramani

Associate Editors:

KSR Anjaneyulu

M Sasikumar

Editorial Board

Aravind Joshi, *University of Pennsylvania, USA*

S Arunkumar, *Indian Institute of Technology,
Bombay, India*

Amitava Bagchi, *IIM Calcutta, India*

Margaret A Boden, *The University of Sussex, UK*

Nick Cercone, *University of Regina, Canada*

R Chandrasekar, *Microsoft Corporation, USA*

BB Chaudhuri, *Indian Statistical Institute,
Calcutta, India*

Shri K Goyal, *GTE Laboratories Inc., USA*

JR Isaac, *NIIT, Delhi, India*

RA Kowalski, *Imperial College of Science &
Technology, London, UK*

HN Mahabala, *Infosys, Bangalore, India*

M Narasimha Murty, *Indian Institute of Science,
Bangalore, India*

R Narasimhan, *CMC Ltd, Bangalore, India*

Arun K Pande, *Tata Infotech Ltd, Bombay, India*

PVS Rao, *Tata Infotech, Navi Mumbai, India*

Patrick Saint-Dizier, *Universite Paul Sabatier,
France*

Rajeev Sangal, *Indian Institute of Information
Technology, Hyderabad, India*

Ramasamy Uthurusamy, *General Motors
Corporation, USA*

LM Vidyasagar, *Tata Consultancy Services,
Secunderabad, India*

Publisher:

Truptee C Shah

Typeset at NCST using \LaTeX .

The cover depicts a thousand-petalled lotus, a traditional Indian symbol of knowledge. Today we find a parallel to this idea in AI, in the belief that knowledge is multi-faceted, and is too complex to be described by a few simple rules. The cover was designed by RK Joshi of IDC, IIT Bombay, and produced using Indo-GKS.

*Vivek (ISSN 0970-8618) is published by the
National Centre for Software Technology
Gulmohar Cross Road No. 9, Juhu
BOMBAY 400 049, India.*

Telephone : +91 (22) 620 1606

Fax : +91 (22) 621 0139

Email : vivek@saathi.ncst.ernet.in

Annual Subscription Rates (for Volume 12):

*Individual: Rs 95 (India), Rs 125 (India, by regd post),
\$20 (abroad, by regd air mail)*

*Institution: Rs 380 (India, by regd post), \$40
(abroad, by regd air mail)*

Please direct editorial material to the Editor, Vivek at the address given above. Articles should conform to the format described in "Information for Authors", published in Volume 13, Number 1, January 2000. Copies are available on request. Queries related to subscription and advertisements may be directed to Ms Shah at the same address.

Copyright NCST, 2001