MIT Sloan School of Management

Sloan Working Paper 4212-01
September 2001

A FOUR-FACETED KNOWLEDGE-BASED APPROACH FOR SURMOUNTING BORDERS

Amar Gupta

A Four-Faceted Knowledge-Based Approach for Surmounting Borders

Amar Gupta
Co-Director
Productivity From Information Technology (PROFIT) Initiative
MIT Sloan School of Management
Room E40-198, MIT
1 Amherst Street
Cambridge, MA 02139, USA
Telephone: (617) 253-8906
Fax: (617) 253-8839
email: agupta@mit.edu

Abstract

While the evolving high bandwidth information highways provide the infrastructure for attaining "physical connectivity" across computing resources and information systems, the "on/off" ramps to such highways are still at a primitive stage. Huge manual effort is currently expended to develop knowledge-based paradigms that can effectively transcend national borders as well as other types of borders. This paper examines the prevailing situation from four perspectives: (i) knowledge acquisition, which deals with the issue of nationwide applications that are still paper-intensive; (ii) knowledge discovery, which deals with the issue of mining of huge amounts of historical and current information in numerical, textual, and other formats; (iii) knowledge management, which focuses on aspects for which dominant standards and procedures prevail at the national level, but not at the international level; and (iv) knowledge dissemination, which deals with extracting knowledge that is tailored to the needs to each user. Unlike current approaches that tend to focus on one aspect only, an integrated approach that attaches appropriate weightage to each of the four facets is emphasized in this paper.

Keywords

Knowledge Management; Knowledge Discovery; Knowledge Acquisition; Knowledge Dissemination; Knowledge Based Framework; Trans-national Transformations

INTRODUCTION

Today's world is characterized by evolving islands of knowledge. On one side, one sees small and modest-sized knowledge management systems being created in organizations, primarily in developed countries. On the other, many traditional paper-based systems that deal with applications at local, provincial, and national level have remained virtually untouched by recent advances.

In order to create an integrated archipelago of knowledge-based assets, one needs to look at new paradigms that can help to effectively transcend national borders, corporate borders, cultural borders, functional borders, and other types of borders, and provide the material needed to address the

individual needs of an increasingly diverse set of users of such systems. By providing effective "on-off ramps" to the emerging information highways, the goal is to drastically enhance drastically the ability to make effective use of large volumes of information obtained from disparate sources (each with its own set of underlying meanings and assumptions), by transforming automatically the incoming streams of information to the desired meaning (or context) needed for a particular job or function in a particular nation or organization. These sets of information need to be complemented by ones from traditional systems; then, through the use of new knowledge discovery techniques, one can establish very sophisticated transborder knowledge assets. (The term "data" is used in this paper to include numerical data, textual data, pictorial data, other types of data and any combinations thereof; the same applies to the term "information").

OPPORTUNITIES RELATING TO AUTOMATED TRANSFORMATIONS ACROSS BORDERS

As high bandwidth information highways continue to evolve, they will provide the infrastructure for attaining "physical connectivity" across computing resources and information systems located in different countries and organizations. However, the "on/off" ramps to such highways are still at a primitive stage, and huge manual effort is currently expended in providing "logical connectivity," especially with the rapidly increasing volumes and diversity of information exchanged and the growing needs of users located in different nations (and organizations) of the world.

In virtually all situations involving one or more natural borders or manmade borders (such as more than one organization, or more than one function in an organization, or even within a single function in a decentralized organization), the meaning of data acquired from a source environment is different from that needed or expected in the receiver's environment. This problem becomes even more acute when one deals with users in multiple countries. Let us illustrate this problem using a specific market that involves users in multiple countries: the stock market. The concept of stock markets evolved to assist individual investors to buy and sell shares of companies. Stock markets

generally follow common business practices within a country, which may change over time. For example, shares in the US were earlier quoted in dollars and fractions of dollars (like 16 and one-half; or 32 and a quarter), but are now quoted in decimal form (16.50 and 32.25, for example). Apart from these temporal (or time-based) differences, stock market practices vary very significantly across national borders.

In Spain, the first stock market was established in 1831. For most of the time since its inception, shares have been quoted on the Spanish stock market relative to of their nominal book values. So, if this percentage exceeded 100, then the share is trading above its nominal book value and vice versa. Individuals familiar with the Spanish stock market would argue that this percentage value provides a good first order reference for comparing the performance of two different companies, unlike the US stock market where no equivalent first order comparison techniques exists.

When one thinks of converting share prices across national borders, one usually thinks of differences in currencies. But as the above example shows, there are many other types of barriers that need to be surmounted. The nominal book value of shares of all companies must be known before one can reconcile the two market practices described above. And there are temporal issues too. The stock exchanges in Spain moved from relative terms to absolute terms (measured in pesetas, the Spanish currency) in 1998, and more recently to the European currency, Euro.

In the above example, the problem of reconciling dissimilar business practices was partially reduced over time when several countries in Europe decided to adopt a common currency and more unified business practices. Note that such simplifications occurred for only a fraction of the stock markets across the world. And the stock market is just one aspect of the commercial arena. Manual techniques are grossly inadequate for reconciling differences across national borders and other types of borders, because of the types of differences involved, the breadth of domains involved, the number of countries involved, and especially by the time constraints imposed on resolving the differences if one wishes to make meaningful trans-border transactions in the electronic era.

The type of problem described above is currently addressed either at the source-end, or at the receiver-end, and sometimes partially at both ends. For example, users of data from stock exchanges

around the world must modify the incoming data streams to match their needs and context; in such cases, the transformations occur at the receiver end as shown in Figure 1. As an example of source-end transformations, when a multifunctional corporation fills income-tax returns in different countries, the payee organization must access the information held in its various information systems, and modify the data to conform to the context mandated by the tax authorities of each of the concerned nations. This is shown in Figure 2. Finally, as an example of the third case, the source may connect the data to a more commonly accepted standard (US dollar, for example), and the receiver then reconverts the data to meet his/her needs.

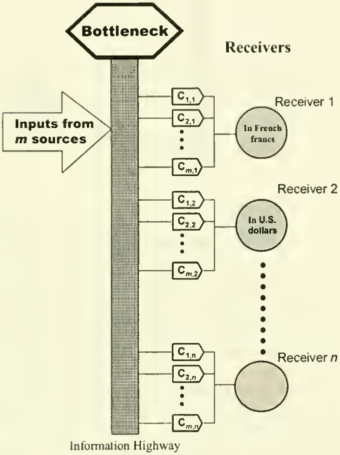

Figure 1: Receiver-End Transformations

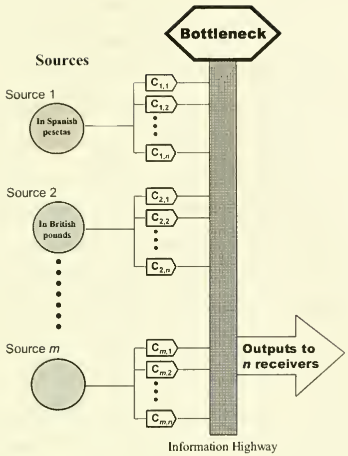

Figure 2: Source-End Transformations

In a global economy, one is witnessing an increasing number of information sources (say "m") and information users (say "n"). With the growing practice of utilizing information from all possible sources, any of the sources can be connected to any of the receivers. Accordingly, a total of "m x n" different transformation scenarios are theoretically involved. The product-form of this relationship emphasizes the fact that the aggregate effort involved is increasing much faster than the growth rate in either the number of sources or the number of receivers. This product form relationship leads to a limiting situation where no additional complexity can be incorporated. As highlighted in Figures 1 and 2, S_1, S_2, . . ., S_m represent the "m" sources, R_1, R_2, . . . , R_n represent the "n" receivers, and the different "C" represent the conversions performed at either end, or at both ends. Incorporating (or revising) a single additional receiver implies writing (or modifying) potentially "m"

new conversion routines, and serves as a growing barrier to the inclusion of new users and new sources (Gupta and Madnick, 1995).

The problem of effective knowledge management across heterogeneous systems is closely related to the problem of integration of islands of disparate information systems. Early research on the latter problem at MIT's Composite Information System Laboratory (now a component of MIT's Productivity From Information Technology "PROFIT" initiative) highlighted needs at three levels: strategic, organizational, and technical. Strategic connectivity involves clear delineation of benefits and costs associated with enhanced connectivity in a multiorganizational environment; organizational connectivity encompasses the processes of making controlled changes in complex organizational environments; and technical connectivity includes mitigating hurdles at both physical and logical levels.

If one bridges two islands of information assets directly, or incorporates a converter between each source-receiver pair, then one is adopting a two-schema approach (see (Gupta, 1989), for example). If there are more sources and receivers, one can adopt either the interfacing approach or the integration process. In the interfacing approach, the total number of converters needed is equal to the number of sources of information multiplied by the number of receivers of information, since any source could be connected to any receiver. In the integration approach or the three-scheme approach, a new global schema is employed that incorporates information about all the participating information systems in a single schema, which provides a single basis for all transformations. Such a global schema can be created and updated in acceptable periods of time when the number of constituents is small, and when someone possesses some degree of control on the constituents. We find neither of these requirements is met when dealing with most transborder applications.

Under the aegis of contracts awarded by the Defense Advance Research Agency of the US Government and other sponsors, researchers have developed new architectures that can significantly reduce the effort involved in mapping knowledge constructs across national borders, organizational borders, and other types of borders (Reddy, Siegel, and Gupta, 1993), (Garcia-Molina et al, 1997) and (Knoblock et al, 2001). In one approach, the underlying "context" for each type of information can be

defined in explicit terms that may include parameters such as the meaning, definition, and context of data, as well as the source for the data, the security classification and the date/time of creation of those data. The differences between the context of the source and the context of the receiver are mitigated through the use of shared ontologies (or vocabularies), which are expected to evolve over time on a domain specific basis. The ontology for the share-market could contain the names of different stocks, the names of the stock markets, and the details of quoting practices as they have evolved over time. Using such ontology, one can "mediate" between vastly different business practices that exist across national borders, both with respect to queries and updates (Reddy and Gupta, 1994), (Reddy et al 1994), (Reddy, Uma, and Gupta, 1998).

While the shared ontology contains a list of different kinds of data that can be represented in a given database, the conversion libraries are utilized to convert data to the context of the receiver. In the case of the income-tax example, the salary data may need to be converted from weekly/monthly format to annual basis or from Japanese Yens to U.S. dollars. A more complicated scenario involves income tax payable based on the number of days spent by a particular employee in the particular nation, which requires access to employee's travel vouchers. In the case of the example involving data feeds from multiple stock exchanges, the conversion libraries will handle currency conversions and the "unit of quote" conversions, taking the appropriate date/time into consideration given the dynamic nature of such conversion parameters, so that an individual user is able to make decisions quickly without having to perform any manual transformations.

The context mediator acts as a "super" query-handler and handles both queries and the initiation of requests to the shared ontology and the conversion routines. A typical sequence involves six steps: a receiver makes a request for data; the context mediator interprets the query based on the receiver's source context; the context mediator issues a modified query to the source based on the context of the source; the raw information is received from the source; the information is converted to the context of the receiver; and the information is provided to the receiver.

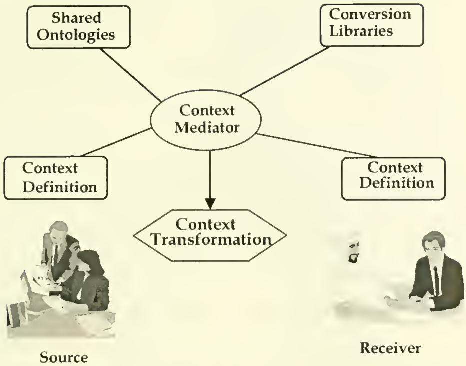

Figure 3: Context Interchange Architecture for Reducing Effort for Mapping Knowledge Constructs Across Borders

The above approach has been utilized to transcend several types of borders including those related to national stock market idiosyncrasies, diverse measuring systems (e.g., the British measuring system still used in the US versus the metric system now used by most countries), and national currencies (e.g., British pounds versus the Italian Liras). This approach falls under the broad category of "federated systems", in which the different information assets retain their original structure and autonomy, and a new "unified" picture is developed to enable individuals from other operating environment to gain quick insights into information from diverse countries and organizations.

OPPORTUNITIES RELATING TO TRADITIONAL SYSTEMS

Major day-to-day applications are handled at the local level, the regional level, the national level, especially by government agencies around the world. This has resulted in a large number of stand-alone systems; these latter systems serve as foundations for providing inputs to evolving transborder knowledge management systems.

Consider the financial industry, in particular, one major aspect of the banking industry that deals with processing bank checks (or "cheques" in some countries). Each nation has its own way of processing bank checks. In the U.S., for example, nationwide check processing and check clearance occur under the aegis of the Federal Reserve Board. If a person holding an account with a bank located in the state of Massachusetts gives a check to a dealer located in the state of California, the dealer presents the check to a local bank located in the latter state. A person at that local bank actually reads the courtesy amount (the amount in figures), and keys the same to create an imprint on the bottom line of the check. Thereafter, the check is physically transported to a designated branch of the Federal Reserve Board, which may be located 300 miles away from the first bank. Next, the check is sent by an air courier to the branch of the Federal Reserve Board located in Boston, which gives it to the lead bank, who in turn gives the check it to the concerned local bank in Massachusetts; the latter bank collects the particular check, stacks it into a pile, and mails all the checks together to the account holder once a month. The process of truncation, encouraged by the Federal Reserve Board, involves converting the paper check into an electronic equivalent; however, this improvement has so far impacted only the last step of this cumbersome process. The total cost of the check processing cycle today exceeds $1.25 per check in the US.

About 66 billion checks are processed each year in the US alone. The global total is estimated at around 200 billion checks per year. Further, checks, which go across national borders, require special processing and involve still higher costs today. As one looks towards global systems, one can propose that a totally new paperless system implemented, in which all financial transactions are performed purely through electronic means. However, such a solution will not appeal to many nations; for example, in the US, consumers are used to writing checks in order to get the benefit of "float", and checks constitute the dominant mechanism for making payments for many types of transactions.

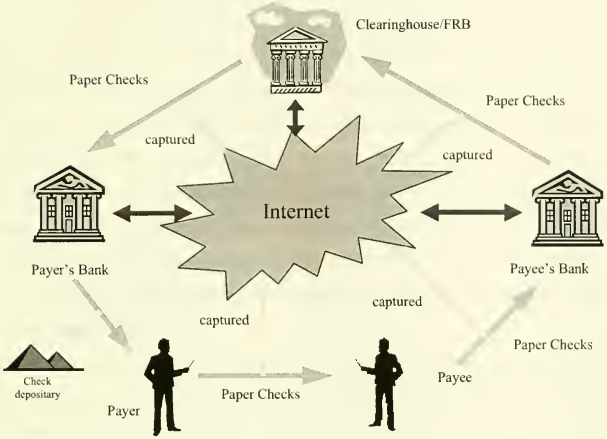

Figure 4: Architecture for Global Automated Approach

Consider an alternative scenario. Imagine that the individual user is still able to write checks. Once the check is given to a dealer or to a local bank, the check is converted into electronic form and the paper copy retained for auditing purposes only for a limited time (Bunke and Wang, 1997). Through a multinational global check clearance infrastructure, the check is processed, irrespective of whether the transaction is domestic or international in nature. If it is international, the knowledge for triggering appropriate sets of national systems is maintained on a continuous basis. Based on the above vision, the concept of "Check Anywhere" has been developed. A check presented anywhere in the world, even in a remote village in India or Brazil, can be scanned, and the courtesy amount read using a neural network based approach. The scanned image, as well as the numerical value of the check, is transmitted using a secure briefcase concept, via the Internet (see Figure 4), to the bank on which the check was issued, bypassing the Federal Reserve Bank in the US and equivalent institutions in other countries (Nagendraprasad, Sparks, and Gupta, 1993), (Agarwal et al, 1996), (Hussein et al,

1999), and (Nagendraprasad et al, 2001). This concept can drastically reduce the cost and time involved in check processing, as well as eliminate many types of frauds and problems arising from dissimilar idiosyncrasies of check-processing infrastructures of different countries around the world. The concept demonstration prototype for the above vision utilizes new technologies for reading characters, for transferring information over the web with high security and privacy, and for taking photographic images of the checks at high speeds. The same hardware and software can be utilized to handle diverse financial, legal, and other types of documents in a quick and reliable manner across organizational and national borders, leading to the notion of "Image Anywhere". The use of such transborder systems can reduce global check processing costs by billions of dollars each year. And there are many other intraorganizational and interorganizational opportunities of similar nature, where existing borders need to be lifted.

Information from systems of the above type will be increasingly utilized in new knowledge management endeavors, particularly in areas related to transborder commerce, distance education, and national/regional security.

OPPORTUNITIES RELATED TO KNOWLEDGE DISCOVERY

The availability of information from multiple sources, located in different countries, competing companies, and challenging environments presents new opportunities. First, one can cross-validate knowledge, and even enhance its overall quality and breadth to derive strategic advantage. Second, one can undertake Knowledge Discovery endeavors to attain new insights into underlying similarities and differences.

Artificial neural networks mimic the broad parallelism that characterizes the human brain. Emerging neural network based data mining techniques produce results that are increasingly outperforming the solutions provided by the best human domain experts. The human mind is good at analyzing problems that involve a few dimensions; when there are 40-50 underlying variables, human beings come up with local optima, not global optima. This is where artificial neural networks and emerging Knowledge Discovery paradigms come in (Fayyad et al, 1996) and (Dhar and Stein, 1997).

In one example involving Inventory Management of over 5000 different items, each sold via one of 2000 outlets located at geographically dispersed sites, researchers were able to reduce the total inventory levels from $1 billion to $500 million using such a neural network based knowledge discovery approach (Bansal, Vadhavkar, and Gupta, 1998) and (Reyes-Aldasoro et al, 1998).

The emerging techniques for knowledge discovery can be applied to situations where the raw inputs are in virtually any format: numerical, textual, pictorial, audio, video, or specialized ones. When one is dealing with inputs coming across borders of different types, one has to incorporate knowledge discovery tools with appropriate preprocessing and post processing modules to overcome issues of physical and logical connectivity.

In one current application, material from chat groups around the world is being analyzed, processed and distilled to create knowledge on alternative medicine (Gopal and Gupta, 2001). Patients, and sometimes their friends and relatives, write in about their symptoms, the medicines they have taken and the frequencies thereof, and the subsequent short-term and long-term effects on them. Since the material on the chat groups is in free-format, natural language techniques have been employed to extract key concepts. And since the inputs come from different countries, where the same medicine (or symptoms) may be called by different names, semantic mapping techniques have been incorporated to mitigate such trans-national issues. Through a combination of artificial intelligence and neural network techniques, one is able to take over 10,000 messages and create a one-page finished report whose quality surpasses the best report that human domain experts have been able to produce in the past. Plus, the automated approach minimizes the problem of human bias, in spite of the fact that the raw material is coming from a much larger and diverse set of individuals than ever before.

One cautionary message should be included here. When raw inputs have significant structure, traditional techniques are able to cope with such inputs. When there is less structure, new neural network based approaches are gradually becoming sophisticated enough to be able to cope with them. However, when there is no structure at all, these neural network techniques fail too. In one controlled experiment, the level of the signal (useful inputs) was varied as compared to the level of noise

(erroneous inputs); this ratio is called the signal-to-noise ratio. When this ratio was above a particular threshold, neural networks demonstrated good learning ability. As this ratio was lowered further, traditional approaches failed first; as it was lowered further, neural network based approaches failed too (Nagendraprasad et al, 2001).

The Knowledge Discovery process is generally performed in three phases: a training phase where neural networks are "trained" with real inputs; a testing phase where results obtained through the experimental neural network are matched against results from the real environment; if there is significant match, then the network is deemed to have "learned" the real environment. Otherwise, additional work is needed. After adequate "learning", the knowledge discovery module can be deployed in a production environment as shown in Figure 5.

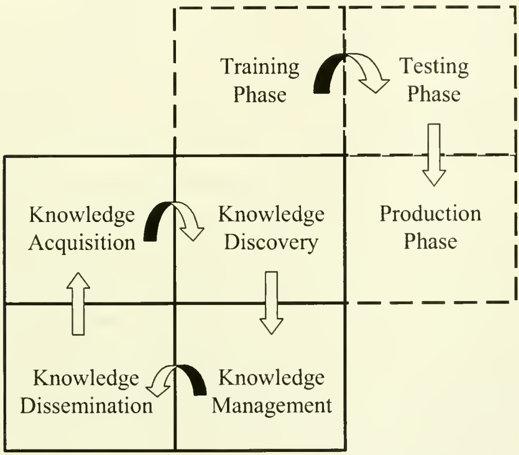

Figure 5: Knowledge Discovery Facet Involves Three Key Phases: Training Phase; Testing Phase; and Production Phase

MULTIFACETED FRAMEWORK

The ideas discussed in the preceding sections can be broadly classified into four categories as follows:

Knowledge Acquisition, or tapping traditional systems to provide accurate and comprehensive material for the new knowledge-based systems;

Knowledge Discovery or automated mining of numerical, textual, pictorial, audio, video, and other types of information, to capture underlying knowledge, either on a one-time basis or on a continuous basis;

Knowledge Management to deal with different types of heterogeneities which invariably exist when inputs have to cross-over borders of different types (national, organizational, departmental, and other); and

Knowledge Dissemination to extract, customize, and direct knowledge to appropriate departments and users, based on their individual needs.

Of the above, the areas of Knowledge Acquisition and Knowledge Management are comparatively more advanced. The area of Knowledge Discovery is now witnessing great interest, and one has just started to do serious work in the Knowledge Dissemination area.

Virtually all organizations are currently focusing on only one, or sometimes two, of the four facets. In Figure 6, a small subset of our sponsor organizations is shown, along with their respective areas of interest. The World Bank, for example, has focused heavily on the Knowledge Management aspect, whereas large commercial companies (codenamed Steelcorp, Bankcorp, and Medicorp) have emphasized the Knowledge Discovery aspect only. In "Steelcorp", one looked at over 40 different parameters to help predict future temperatures within the blast furnace; knowing this temperature is vital to the overall productivity of the blast furnace. In "Bankcorp", a knowledge discovery approach was developed to optimize the inventory of bank securities, partially through prediction of expected prices. And in "Medicorp", one analyzed a large retail distribution environment to reduce inventory

levels. However, in each of these cases, relatively little time was devoted to the other facets of the Knowledge-based paradigm.

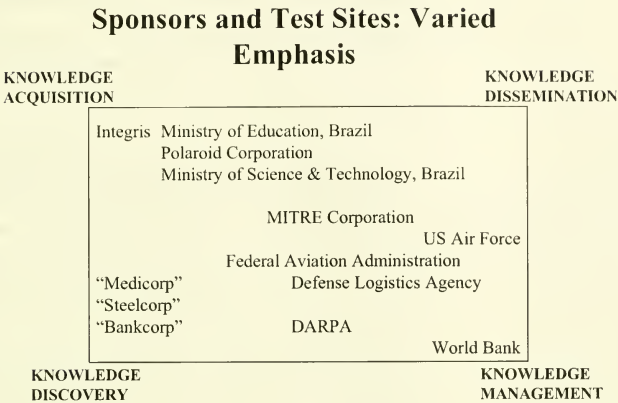

Figure 6: Different Organizations Emphasize Different Facets: Few Adopt Multifaceted

The greatest benefits can be attained through the adoption of a multi-faceted approach that accords adequate weightage to each of the four parallel sets of opportunities. We are currently working with MITRE, a federally funded research and development center of the US government, to help design and develop the next generation integrated command and control system. In this endeavor, all the four aspects are being kept in view, right from the beginning. We have utilized a similar approach with several other multinational organizations to conceptualize knowledge-based systems to address diverse needs of different levels of the hierarchy, as well as different departments and subsidiaries located in different parts of the world.

ACKNOWLEDGEMENT

A significant part of research work described in this paper has been conducted by current and former team members of the Productivity from Information Technology (PROFIT) Initiative at MIT Sloan School of Management. The author sincerely acknowledges their help and support, especially that of Dr. Rafael Palacios, as well as the assistance provided by an array of research sponsors spread over four continents of the world.

REFERENCES

Agarwal, A., Hussein, K., Wang, P., and Gupta, A. (1996) "Detection of Courtesy Amount Block on Bank Checks" *Journal of Electronic Imaging*, Vol. 5(2), pp. 214-224.

Bansal, K., Vadhavkar, S., and Gupta, A. (1998) "Neural Networks Based Forecasting Techniques for Inventory Control Applications", *Data Mining and Knowledge Discovery*, Vol. 2, pp. 97-102.

Bunke, H., and Wang, P.S.P. (1997), *Handbook of Character Recognition and Document Image Analysis*, (Ed.), World Scientific Press.

Casey, R., and Lecolinet, E. (1996), "A Survey of methods and strategies in character segmentation", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18 (7) pp. 690-706.

Dhar, V., and Stein, R. (1997), Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996), *Advances in Knowledge Discovery and Data Mining*, MIT Press.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., and Widom, J. (1997), "The TSIMMIS Approach to Mediation: Data Models and Languages", *Journal of Intelligent Information Systems.*

Gopal, S., and Gupta, A., "Knowledge Extraction From Chat Groups", under preparation.

Gupta, A. (1989) Integration of Information Systems: Bridging Heterogeneous Databases (editor), IEEE Press, New York, NY.

Gupta, A., and Madnick, S. E. (1995) "Context Interchange: Research in Using Knowledge about Data to Integrate Disparate Sources", *PROFIT Working Paper #95-29*, MIT Sloan School of Management, Cambridge, USA.

Hussein, K., Agrawal, A., Wang, P., Gupta, A. (1999) "A Knowledge Based Segmentation Algorithm for Enhanced Recognition of Handwritten Courtesy Amount", *International Journal of Pattern Recognition,* Vol. 32 (2).

Knoblock, C.A., Minton, S., Ambite, J.L., Ashish, N., Muslea, I., Philpot, A.G., and Tejada, S. (2001) "The Ariadne Approach to Web-based Information Integration", *International Journal on Cooperative Information Systems (IJCIS)*, 10 (1/2), pp 145-169.

Nagendraprasad, M.V., Sparks, P.L., and Gupta, A. (1993) "A Heuristic Multi-stage Algorithm for Segmenting Simply Connected Handwritten Numerals", *Heuristics: The Journal of Knowledge Engineering & Technology*, Vol. 6, No. 4, pp 16-26.

Nagendraprasad, M.V., Gupta, A., Shanmugasundaram, J., and Vadhavkar, S. (2001) "Temporal Data Mining", *PROFIT Working Paper*, MIT Sloan School of Management, Cambridge, USA.

Reddy, M.P, Siegel, M., and Gupta, A. (1993) "Towards an Active Schema Integration Architecture for Heterogeneous Database Systems", *Proceedings of International Conference on Research Issues on Data Engineering: Interoperability in Multidatabase Systems*, Vienna, Austria.

Reddy, M.P., and Gupta, A. (1994) "Context Interchange: A Lattice Based Approach", *Knowledge-Based Systems*, Butterworth-Heinemann, Oxford, England.

Reddy, M.P., Prasad, B.E., Reddy, P.G., and Gupta, A. (1994) "A Methodology for Integration of Heterogeneous Databases", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 6, No. 6.

Reddy, M.P., Prasad, B.E., and Gupta, A. (1995) "Formulating Global Integrity Constraints During Derivation of Global Schema", *Data and Knowledge Engineering*, Vol. 16, 1995, pp. 241-268.

Reddy, M.P., Uma, B., and Gupta, A. (1998) "Query Optimization in Heterogeneous Distributed Information Environment", *Heuristics: The Journal of Intelligent Technologies*, Vol. 10, No. 2, pp. 67-90.

Reyes-Aldasoro, C. C., Ganguly, A., Lemus, G., Gupta, A. (1998) "A hybrid model based on dynamic programming, neural networks, and surrogate value for inventory optimisation applications", *Journal of the Operational Research Society*, Vol. 49, pp. 1-10.