Volume 2 ▪ Issue 2 ▪ May 2011

Editor-in-Chief
Professor Walid Aref

INTERNATIONAL JOURNAL OF DATA ENGINEERING (IJDE)

ISSN : 2180-1274
Publication Frequency: 6 Issues / Year

CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF DATA ENGINEERING (IJDE)

VOLUME 2, ISSUE 2, 2011

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 2180-1274

International Journal of Computer Science and Security is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJDE Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF DATA ENGINEERING (IJDE)

Book: Volume 2, Issue 2, May 2011
Publishing Date: 31-05-2011
ISSN (Online): 2180-1274

IJDE Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2011

EDITORIAL PREFACE

This is second issue of volume two of the International Journal of Data Engineering (IJDE). IJDE is an International refereed journal for publication of current research in Data Engineering technologies. IJDE publishes research papers dealing primarily with the technological aspects of Data Engineering in new and emerging technologies. Publications of IJDE are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJDE is Annotation and Data Curation, Data Engineering, Data Mining and Knowledge Discovery, Query Processing in Databases and Semantic Web etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJDE appears in more focused issues. Besides normal publications, IJDE intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJDE is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJDE as one of the top International journal in Data Engineering, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Data Engineering fields.

IJDE editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJDE provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts..

Editorial Board Members
International Journal of Data Engineering (IJDE)

EDITORIAL BOARD

Dr. Andrey Balmin
IBM Almaden Research Center
United States of America

Dr. Rishi R. Sinha
Microsoft Corporation
United States of America

Dr. Qiong Luo
Hong Kong University of Science and Technology
China

Dr. Thanaa M. Ghanem
University of St. Thomas
United States of America

Dr. Ravi Ramamurthy
Microsoft Research
United States of America

TABLE OF CONTENTS

Volume 2, Issue 2, May 2011

Pages

An Organizational Memory and Knowledge System (OMKS): Building Modern Decision Support Systems

Jon Blue *jonblue@udel.edu*
University of Delaware
College of Business and Economics
Department of Accounting and MIS
42 Amstel Avenue, 212 Purnell Hall
Newark, DE 19716, USA

Francis Kofi Andoh-Baidoo *andohbaidoof@utpa.edu*
College of Business Systems Administration,
Computer Information Systems and Quantitative Methods,
University of Texas – Pan American,
Edinburg, TX 78539, USA

Babajide Osatuyi *osatuyi@njit.edu*
Department of Information Systems,
New Jersey Institute of Technology
College of Computing Sciences, University Heights,
Newark, NJ 07102, USA

Abstract

Many organizations employ data warehouses and knowledge management systems for decision support activities and management of organizational knowledge. The integration of decision support systems and knowledge management systems has been found to provide benefits. In this paper, we present an approach for building modern decision support systems that integrates the data warehouses and organizational memory information systems processes to support enterprise decision making. This approach includes the use of Scenarios for capturing tacit knowledge and ontology for organizing the diverse data and knowledge sources, as well as presenting common understanding of concepts among the organizational members. The proposed approach, which we call Organizational Memory and Knowledge System approach, is expected to enhance organizational decision-making and organizational learning.

Keywords: DSS, Decision Support Systems, Decision Making, Tacit Knowledge, Organizational Learning, Organizational Memory, Knowledge Management, Data Warehousing, Ontology, Scenarios, Metadata

1. INTRODUCTION

Individuals in organizations have to make on-going decisions. In view of the complexities of decision-making, organizations provide Decision Support Systems (DSS) to enhance the decision making process. Hence, DSS are critical in the daily operations of organizations.

Data warehousing has become an integral component of modern DSS.The data warehousing infrastructure enables businesses to extract, cleanse, and store vast amounts of corporate data from operational systems which, when queried, can provide answers to the questions posed by decision makers. On-Line Analytical Processing (OLAP) is used heavily by Data Warehouses to make aggregate queries to answer domain specific questions. In addition, OLAP, data mining, and other knowledge discovery techniques are used to establish relationships existing in the data that reside in the warehouse repository. These relationships are used to create, access and reuse knowledge that supports decision-making[HYPERLINK \l "DOL98" 1].

Modern DSS require that various types of knowledge are captured and used in decision-making2]. More so, the importance of tacit knowledge in the knowledge management domain has received great attention [HYPERLINK \l "Bus00" 3]. Tacit knowledge is valuable organizational knowledge, whichresides in the minds of individuals. These persons build up their knowledge by working on the job over extended periods of time. The organization is limited in its ability to leverage this expertise as much as this knowledge remains personal to the individual.

Repeated studies have documented that tacit knowledge can be articulated, captured, and represented 4], [HYPERLINK \l "Gra97" 5], 6], [HYPERLINK \l "Rag96" 7], 8], [HYPERLINK \l "Gol90" 9], 10]. However, the traditional data warehouse does not possess capabilities to acquire, store, and use tacit knowledge[HYPERLINK \l "Bus00" 3], 11], [HYPERLINK \l "YuN99" 12].Researchers have indicated that integrating knowledge processes and decision processes would enhance decision-making13][HYPERLINK \l "Nem02" 11].

In this paper, we present an approach for modern decision support systems that combine features of data warehouses and organizational memory information systems to form an Organizational Memory and Knowledge System (OMKS) that supports enterprise decision-making.This approach includes capturing tacit knowledge and uses ontology as the metadata for organizing the diverse data and knowledge sources, as well as presenting common understanding of concepts among the organization'smembers. This approach is expected to enhance organizational decision-making and organizational learning. Other benefits of integrated decision support system such as a data warehouse and an organizational memory information system include real-time adaptive decision support, support of knowledge management activities, facilitation of knowledge discovery and efficient ways of building organizational memory 13]. We discuss the theoretical foundation for the approach and explain how it meets the requirements for the foundational data warehouse architecture and organizational memory information systems.
The organization of the paper is as follows. In section 2, we review the decision making process, data warehousing,Organizational Memory Information Systems (OMIS), and the knowledge conversion processes. Following, we present and describe our proposed approach of modern DSS that integrates functional features of data warehousing and OMIS. This is done using Scenarios to facilitate the acquisition, storage, use and sharing of tacit knowledge and Ontology for metadata specifications. In Section 4 we discuss the implication of this novel approach for practice and research. Finally, we conclude the paper with suggested future research directions.

2. THEORETICAL BACKGROUND

We draw on several theoretical foundations in proposing the approach for building modern decision support systems. In the following, we present the diverse theoretical foundations that include the decision-making and decision support systems, data warehouse, organizational memory information systems, integrated process management, and the knowledge spiral.

2.1 Decision Making and Decision Support Systems

The works of Gorry and Scott Morton [HYPERLINK \l "Gor71" 14], 15],[HYPERLINK \l "Bon81" 16],17],[HYPERLINK \l "Spr82" 18], define the conceptual and theoretical foundations of decision support systems (DSS). Coined by Gorry and Scott Morton 15], decision support systems are motivated by decision making, as opposed to systems supporting problem or opportunity identification, intelligence gathering, performance monitoring, communications, and other activities supporting organizational or individual performance. Using Simon's [HYPERLINK \l "Sim55" 19]classical four phases of the decision-making process (intelligence, design, choice, and implementation and control), a typical DSS concentrates on the design and choice phases20].

Modern DSS are, however, called to support all the phases of decision-making process [HYPERLINK \l "Bol02" 13]. Data Warehouses, Knowledge Management Systems, and Organizational Memory Information Systems are forms of DSS that help decision makers in the decision making process. Recently these systems have received greater attention.

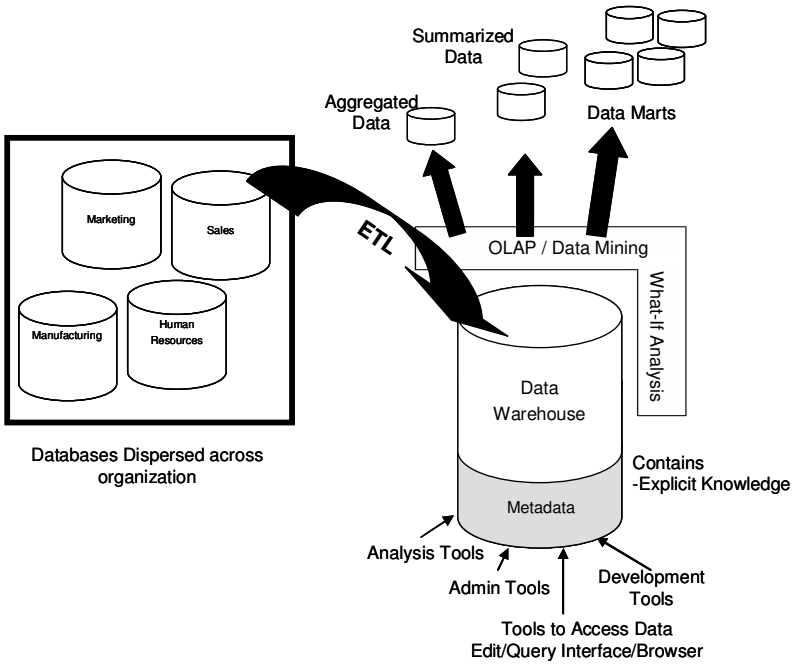


FIGURE 1 : Typical Data Warehouse

2.2 Data Warehousing

Bill Inmon 21]defines Data Warehousing as "… a subject-oriented, integrated, time-variant, and non volatile collection of data in support of management's decision-making process" (p. 1).As can be seen in Figure 1, Extraction, Transformation, and Loading (ETL) tools are used to extract transactional data from diverse sources and transform and load these datainto the warehouse repository. OLAP tools are then used to aggregate data to answer queries. Data mining and other knowledge discovery tools are used to establish relationships that have not been specified in the data sources. These relationships are used to create knowledgeto support decision-making.

A critical element of the data warehouse architecture is metadata management. Metadata defines a consistent description of the data types coming from the different data sources. It provides comprehensive information such as the data sources, definitions of the data warehouse schema, dimensional hierarchies, and user profiles. A metadata repository is used to manage and store all of the metadata associated with the warehouse.Also, the repository allows the sharing of metadata among tools and processes for the design, use, operation, and administration of a warehouse [HYPERLINK \l "Sta99" 22].

2.3 Knowledge Management and the Knowledge Spiral 23]

Knowledge Management

The field ofKnowledge Management continues to grow and stems from the realization that an organization cannot afford to lose knowledge as individuals leave. Knowledge management is not a product or solution that can be bought or sold; it is a process that is implemented over time [HYPERLINK \l "Ben98" 24],

Knowledge comes in two forms: explicit and tacit 25]. Explicit knowledge is systematic and can be expressed formally as language, rules, objects, symbols, or equations. Thus, explicit knowledge is communicable as mathematical models, universal principles, or written procedures[HYPERLINK \l "Nem02" 11].

Tacit knowledge includes the beliefs, perspectives, and mental models ingrained in a person's mind. This type of knowledge is hard to transfer or verbalize because it cannot be broken down into specific rules.

However, many authors have purported that this type of knowledge can be articulated, captured, and represented 4], [HYPERLINK \l "Gra97" 5], 6], [HYPERLINK \l "Rag96" 7],8], [HYPERLINK \l "Pyl81" 10].

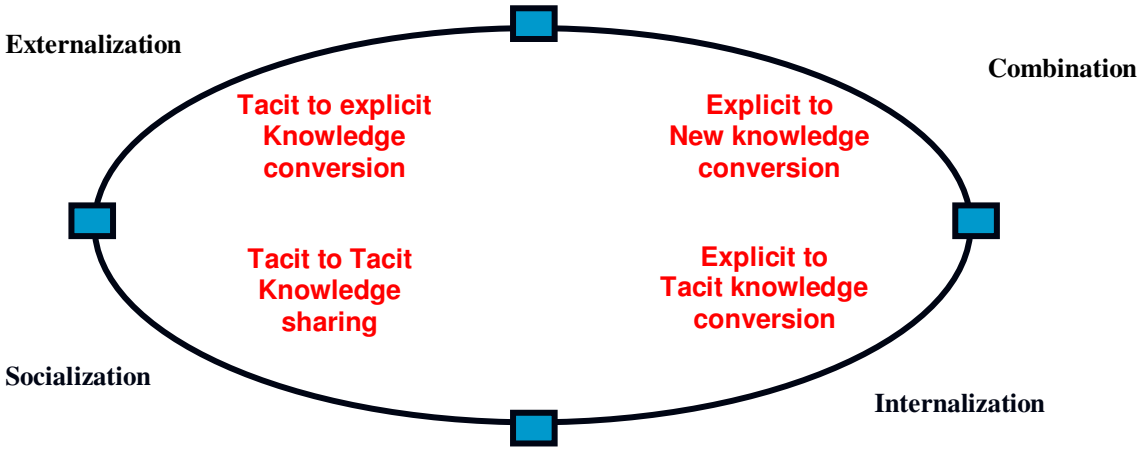


FIGURE 2 : The Knowledge Spiral 23])

Nonaka and Takeuchi [HYPERLINK \l "Non95" 23] assert that new knowledge is created through the synergistic relationship and interplay between tacit and explicit knowledge. This concept, depicted with the Knowledge Spiral (Figure 2), has four spokes: (1) Externalization (conversion of tacit knowledge to explicit knowledge), (2) Socialization (sharing tacit knowledge); (3) Combination (conversion of explicit knowledge to new knowledge); and (4) Internalization (learning new knowledge and conversion of explicit knowledge to tacit knowledge).

The Knowledge Spiral
Externalization involves the conversion of tacit knowledge to explicit knowledge. It allows the explicit specification of tacit knowledge. Socialization is sharing tacit knowledge, i.e. an employee shares their tacit knowledge with other employees during social meetings. Combination is the knowledge conversion step where explicit knowledge is converted to new knowledge. New knowledge is learnt during the Internalization stage. In this process, explicit knowledge is converted to implicit (tacit) knowledge.

2.4 Organizational Memory Information Systems
It has been discussed that an Organizational Memory Information Systems (OMIS) architecture can be more effective in assisting in decision support because it additionally fully supports organizational learning26], [HYPERLINK \l "Lia03" 27].An OMIS is an integrated knowledge based information system with culture, history, business process, and human memory attributes28].

An OMIS is expected to bring knowledge from the past to bear on future activities that would enhanceorganizational responsiveness and effectiveness[HYPERLINK \l "Ste95" 29]. Walsh and Ungson30] proposed that organizational memory occurs in five retention facilities: individuals, culture, structures, transformations (processes such as production and personnel lifecycles), ecology (the physical setting of a workplace), and structures (the roles to which individuals are assigned).

Atwood [HYPERLINK \l "Atw02" 31] presents applications of OMIS in various domains including corporate environments and governmental settings. Hackbarth proposes that direct activities related to

experiences and observations must be stored by an OMIS in a suitable format to match individual cognitive orientations and value systems. These activities refer to decision making, organizing, leading, designing, controlling, communicating, planning, motivating, and other management processes.
Heijst, Spek, and Kruizinga 32] suggest that OMISfacilitates organizational learning in three ways: individual learning, learning through direct communication, and learning using a knowledge repository.

Atwood [HYPERLINK \l "Atw02" 31] suggests three challenges facing OMIS. These challenges include managing informal as well as formal knowledge, motivating knowledge works to generate (for submittal into an OMIS) and using the knowledge in the system, and systems development practices. Having a process view of the knowledge management and data warehouse design, deployment and use can address those challenges. For instance organizations should be cognizant of the processes that the codified knowledge supports as the knowledge has a high tendency to lose its process perspective when stored in the knowledge storage in the OMKS. Similarly, it is critical that both knowledge and its context are captured in the OMKS because information and knowledge are useful only when the context of that knowledge is known.

2.5 Integrated Process Management: Integrating Data Warehousing and Organizational Memory Systems

This research follows the integration process management (IPM) approach prescribed in Choi, Song, Park, and Park 33]. IPM seeks to "provide the theories, techniques, and methodologies to integrate processes and to support design, analysis, automation and management of process knowledge"(p. 86).

The applicability of the IPM to the integration of data warehouse and Organizational memory systems is that both have been characterized as processes [HYPERLINK \l "Cho04" 34], 2]. In fact, knowledge management is considered as a business process [HYPERLINK \l "Ros01" 35].

A process view of knowledge dictates that knowledge management systems and for that matter organizational memory systems take particular focus on the processes that deal with the creation, the sharing, and the distribution of knowledge. Bolloju et al. 13] also claim that knowledge management and decision support are interdependent and propose an approach for integrating those processes for building modern decision support systems.

We contribute to the discussions on building effective modern decision support systems. While we argue for integration of the data warehouse and organizational memory processes, we use a different approach. We propose the use of scenarios to capture tacit knowledge and ontology to standardize the data and knowledge in the knowledge systems developed to support decision-making and for facilitating the sharing of knowledge across the organization.

3. PROPOSED APPROACH FOR MODERN DECISION SUPPORT SYSTEMS

Data warehouses and OMIS support decision making in organizations. We propose the integration of functional features of data warehousing and organizational memory information systems to produce modern DSS that provide more effective support to decision makers and enhance organization learning. To enable this integration, we present an approach that involves the use of Scenarios [HYPERLINK \l "YuN00" 36] to assist in the acquisition of tacit knowledge from workers in the organization and the use of ontology-based metadata. The approach also incorporates ideas of 23]knowledge conversion process in their Knowledge Spiral. Knowledge has been recognized as animportant critical component of the knowledge management process and modern DSS by otherresearchers (e.g., [HYPERLINK \l "Bol02" 13], 11]). Our proposed approach describes: (1) data and knowledge acquisition processes including the use of Scenarios, (2) ontology-based metadata development and maintenance, and (3) knowledge conversion processes. The remainder of this sectionprovides a detailed description of the components and dynamics in the OMKS approach depicted inFigure 3.

3.1 Data and Knowledge Acquisition Processes

One of the functional benefits of our approach is the accommodation of diverse knowledge sources: explicit knowledge from organizational databases and tacit knowledge from individual decision makers in

the organization (see Figure 3). The traditional data warehouse architecture sources it information and data from transactional processing systems. Thus data and knowledge are formally captured and stored in databases, the Internet and transactional data and informally in interpersonal networks, informal common references, and discourse in professional communities that are relevant for decision-making [HYPERLINK \l "Nem02" 11]. ETL tools and other data acquisition modules used in the data warehousing environments would be used in the extraction, transformation, and loading of the data from these diverse sources into the OMKS.

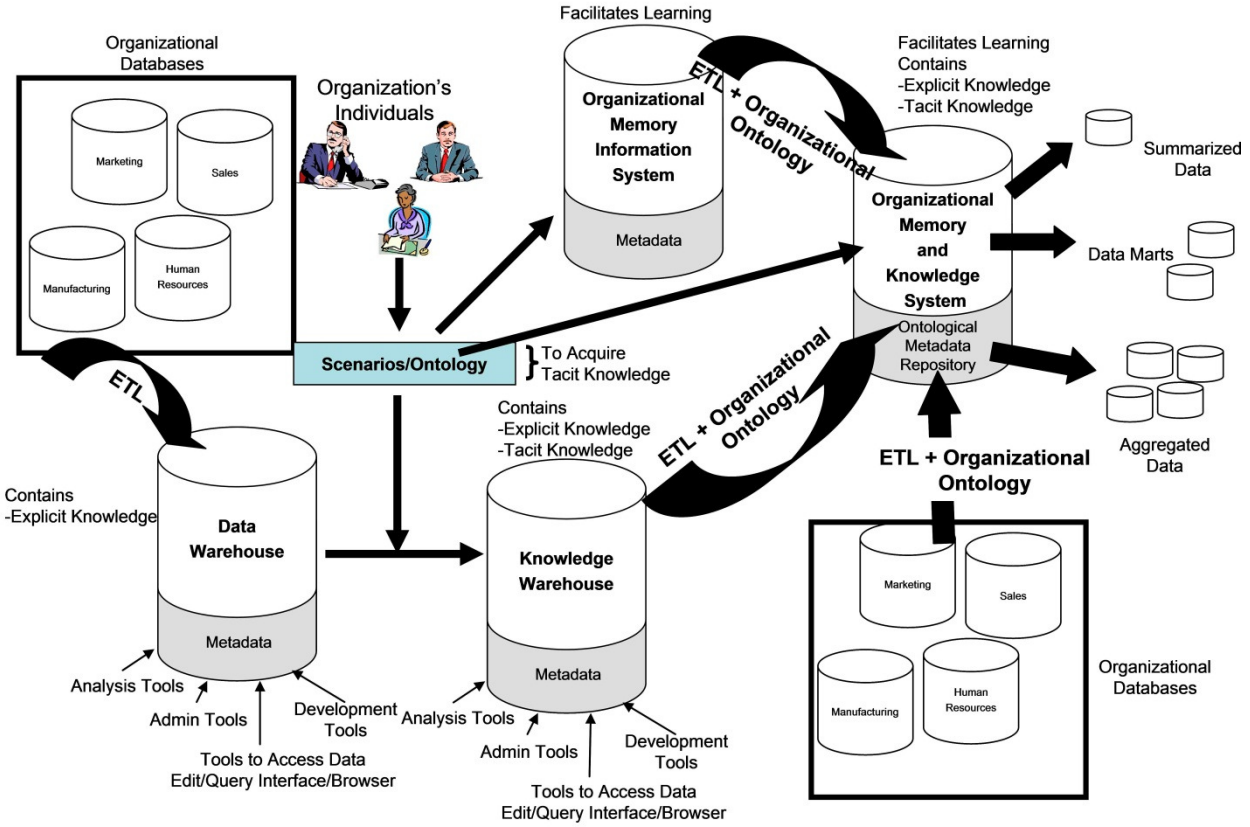


Figure 3 Organizational Memory and Knowledge System

.

Knowledge workers in the organization would be involved with the acquisition of data and information and the assessment of the validity of such data for their suitability as knowledge models for supporting decision making.

OLAP and other knowledge discovery tools would be employed to create data and knowledge models. In addition, tacit knowledge would be captured into the OMKS using Scenarios and ontology. The capture of such knowledge and its conversion in the OMKS is elaborated in a subsequent section. Thus, one issue that needs critical attention, and is therefore discussed later in this paper, is that unlike the data warehousing environment, the OMKS requires an ontology-based metadata for the management of the metadata that are associated with the diverse data and knowledge sources.

Scenarios and Tacit Knowledge Acquisition

Scenarios

A scenario, as described by Yu-N and Abidi **36]**, is a customized, goal-oriented narration or description of a situation, with mention of actors, events, outputs, and environmental parameters. Similarly, a scenario

can be considered an ordered set of interactions between partners, usuallybetween a system and a set of actors external to the system for generating some output.

Tacit knowledge acquisition using Scenarios

Yu-N and Abidi [HYPERLINK \l "YuN001" 37] use ontology with Scenarios to standardize how tacit knowledge is acquired from multiple healthcare practitioners. Scenarios have also been used in other information technology areas such as: human-computer interaction, software requirements engineering and system requirements38]. In each case, Scenarios are useful for capturing tacit knowledge from experts about some business activities or processes.

By placing domain experts into contextual and realistic situations, Scenarioscan capture tacit knowledge from these employees. Scenarios are typically custom designed to reflect atypical problems[HYPERLINK \l "YuN00" 36] and Scenarios can be used to play out what-if situations 39]. This facilitates the capture intentions of domain experts in response to atypical situations [HYPERLINK \l "YuN001" 37]. This is useful because the instinctive aspects of tacit knowledge are best captured while domain experts are actively using their mental models to solve realistic problems. Scenarios enable domain experts to reflect on potential occurrences, opportunities, or risks and therefore facilitate the detection of capable solutions and reactions to cope with the corresponding situations and provide an outlook on future activities 40]. Domain experts are presented with hypothetical or atypical situations in their domains, which allows for the acquisition of tacit knowledge by recording the expert's problem-solving responses. As such, scenario components aim at addressing concrete business situations and make intelligent use of them, in order to drive and reason about decisions without having to expend valuable resources in true trial-and-error ventures [HYPERLINK \l "Kav96" 41]. A further benefit is that they can formalize possible organizational goals held within domain experts' tacit knowledge 41].

Yu-N and Abidi [HYPERLINK \l "YuN00" 36] explain why using Scenarios is a viable method to capture tacit knowledge. The authors' strategy is ground in the assumption that domain experts' knowledge can best be explicated by provoking them to solve typical problems. This is done by repetitively giving domain experts hypothetical scenarios that pertain to typical/novel problems. Then, the domain experts are observed and their tacit knowledge-based problem-solving methodology and procedures are analyzed. As Yu-N and Abidi (p. 2) state, "…the proposed problem-specific scenario presents domain experts the implicit opportunity to introspect their expertise and knowledge in order to address the given problem, to explore their 'mental models' pertaining to the problem situation and solution, and finally to apply their skills and intuitive decision making capabilities. This sequence, allows tacit knowledge to be 'challenged', explicated, captured and finally to be stored."

Ontology and Scenarios

Benjamins et al. 24] describe ontology as a common and shared understanding of some domain that is capable of being communicated across people and systems. Ontologies are applicable to many domains. For instance, van Elst and Abecker [HYPERLINK \l "van02" 42] indicate that ontologies have been used in areas such as agent based computations, distributed information systems, expert systems, and knowledge management. Further, van Elst and Abecker succinctly cite the benefits of using ontologies as "… the major purpose of ontologies is to enable communication and knowledge reuse between different actors interested in the same, shared domain of discourse by finding an explicit agreement on common ontological commitments which basically means having the same understanding of a shared vocabulary…" (p. 357).

Yu-N and Abidi 37] argue that ontology can be used to enforce standardization given that tacit knowledge is deemed to be hierarchical in structure and personal in nature. They also suggest that ontologies hold great potential in facilitating the acquisition of tacit knowledge through the use of Scenarios. In the scenario-based infrastructure, ontologies are used as a means to achieve a defined taxonomy of knowledge items and a standard (conceptual) vocabulary for defining Scenarios to achieve knowledge standardization. In this environment, ontology refers to a specification of a conceptualization that aims to describe concepts and the relationships between entities that share knowledge. The flow of events and structure suggested by the scenario also assist in providing a basis for tacit knowledge capture, which is congruent with the taxonomical nature of ontology [HYPERLINK \l "Gli00" 43].

3.2 Ontology-based Metadata

Metadata design and management is an important process in the OMKS. In the traditional data warehouse environment, metadata may reside in various sources and need to be integrated to ensure consistency and uniform access of data. In the proposed integrated approach, the ontology serves as the global metadata for managing the definition of the data warehousing schema and schema from other data and knowledge sources. Traditionally, different experts, even within a single domain, use different formats in their communications. Issues of data heterogeneity and semantic heterogeneity need to be addressed with respect to the OMKS. Data heterogeneity refers to differences among local definitions, such as attribute types, formats, or precision; and semantic heterogeneity describes the differences or similarities in the meaning of local data. It is noted that two schema elements in two local data sources can have the same intended meaning but different names or two schema elements in two data sources might be named identically, while their intended meanings are incompatible. Hence, these discrepancies need to be addressed during schema integration 44].

While schemas (used in the databases and data warehousing) are mainly concerned with organizing data, ontologies are concerned with the understanding of the members of the community, which helps to reduce ambiguity in communication.Hakimpour and Geppert[HYPERLINK \l "Hak01" 44] present an approach that uses formal ontologies to derive global schemas. Ontologies pertaining to local schemas are checked for similarities. Knowledge about data semantics is then used to resolve semantic heterogeneity in the global schema. Hence, formal ontology can help solve heterogeneity problems.

Several studies in Organizational Memory Information System have used ontology 27], [HYPERLINK \l "She03" 45], 26].Additionally, van Elst and Abecker [HYPERLINK \l "van02" 42] provide a framework for organizational memory technology to support vertical and horizontal scalability. This framework provides means of understanding the issues of integrating organizational memories in distributed environments. The ontology-based metadata specifies validated sets of vocabulary that is consistent among the diverse sources and maintained by the organization's knowledge worker. New vocabulary from new knowledge is used to modify the systems and is communicated among the organization's workers. Each piece of data, information or knowledge has its own data source. As all these sources are captured into the integrated system, inconsistencies may occur. The ontology-based metadata therefore represents a common global metadata that manages all the other metadata associated with the diverse data, information and knowledge sources and also provides explicit semantics. Ittherefore presents a source–independence vocabulary for the domain that the OMKS supports. The ontology-based metadata also facilitates the sharing of metadata among the diverse decision technologies or tools that are present in the OMKS and other processes for the design, use, and administration of the OMKS. It has been recognized that ontology-based metadata enhances organizational members' accessibility to domain knowledge 1].

The Organizational Ontology module should interface with the Data and Knowledge Extraction/Acquisition modules (See Figure 2). The knowledge to be stored in the OMKS must be formally represented and "is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist is some area of interest and the relationships that hold among them" ([HYPERLINK \l "Gen87" 46] qtd. In 47], p. 1). This conceptualization represents the real world and is an abstract and simplified view. The ontology forces the explicit specification of this conceptualization and ensures that information is stored consistently in the OMKS. Given that schema definitions are based on ontology definitions, and vice versa, a symbiotic relationship is constructed between the two.

Knowledge Conversion Activities in Organization Memory and Knowledge System

We have discussed how Scenarios can be used to capture tacit knowledge into the OMKS. This knowledge needs to be made explicit in the system and used to enhance organizational learning and improve the organizational decision through Nonaka's knowledge conversion model. Hence tacit knowledge held in the human mind, and within the shared community/professional memory, will also be managed in the OMKS.

3.3 Knowledge Sources

A knowledge base will store all the different knowledge that are generated by the OLAP and knowledge discovery tools as well as the tacit knowledge that is captured and represented in the OMKS. These knowledge structures serve as sources for enhancing organizational learning and as a basis for enhancing the performance of the ontology-based metadata. New knowledge from these knowledge bases will be used to modify or change the vocabulary and other properties of the metadata. Not only will new knowledge support organizational decision-making, it will also enhance organizational learning because previously learnt knowledge may be modified by the new knowledge. In the following we describe how OMKS supports the knowledge conversion cycle.

Externalization

The Scenarios facilitates the tacit knowledge acquisition from experts from diverse disciplines within the organization. It also enables standardization of the knowledge process that is performed [HYPERLINK \l "YuN00" 36]. Not only does Scenarios capture the instincts of domain experts, they offer an ease of communication and understanding by allowing the domain experts to react from within their own frames of reference and points of view. A scenario is a rich tool, which provides valuable information based on experimenting-in-action on practical cases. Domain experts are able to reason against their experiential knowledge in what is ultimately a sterile environment.

The OMKS can enhance tacit to explicit knowledge by using mathematical models. The knowledge worker can produce mathematical models that reflect tacit knowledge that has been built up over many years. Such models can be stored as explicit mathematical inequalities, as graphs of arc descriptions, or as a canonical model formulation with links to relational tables in the DSS 11]. Brainstorming also provides a potential medium for tacit to explicit knowledge conversion. Output from the brainstorming sessions can be captured into the OMKS and shared among decision makers.

Socialization

The Ontology is a common vocabulary for communication among the organization's employees. The shared understanding then serves as the basis for expressing knowledge contents and for sharing and managing knowledge in the organization. It creates a community of individuals that are likely to embrace collaboration using common knowledge that they share. Further, the sharing of tacit knowledge can happen using tasks such as digitized filming of physical demonstrations[HYPERLINK \l "Nem02" 11]. These digitized films are stored for viewing at anytime by anyone in the organization. The films may also include verbal explanations that explain the process. Kinematics, a form of artificial intelligence, can also be used where an individual is suited with probes and a system records the movements of the person. Busch and Richards 3] indicate that people tend to be reliant on electronic and formal information impeding sharing of knowledge through socialization. The Knowledge Management literature suggests that sharing of tacit knowledge through socialization is effective in small groups [HYPERLINK \l "von00" 48].

Combination

The Ontology reconfigures the explicit knowledge in the OMKS. Through the daily interaction with the OMKS, employees perform their duties using the explicit knowledge that has been captured from diverse knowledge and data sources. Further, Text mining tools (AI-based data mining) on the output from the brainstorming sessions is a representation of this step 49][HYPERLINK \l "Kup95" 50]. This provides key words and extracts the appropriate statistical information in a textual document. A set of rules and guidelines are part of the input to the tool for it to appropriately mine the data. The new knowledge are fed into the OMKS to create, delete or modify existing knowledge in the Ontology's metadata which is distributed among the diverse knowledge to revising metadata that exists in these knowledge sources. This then creates opportunity for the organization's members to revise their understanding of some of the processes and activities that support decision-making.

Internalization

As members of the community gain better understanding of how they can improve their work activities through the shared knowledge from the OMKS, they gain new tacit knowledge about their work.New knowledge is learnt during the Internalization stage. Explicit knowledge is converted to implicit knowledge.

Explicit to implicit knowledge conversion occurs with a modification of a knowledge worker's internal mental model. This modification can occur after discovering new relationships. The OMKS becomes support systems as the knowledge workers validate the new knowledge that has been created.

4. DISCUSSION

In this paper, we have looked at how knowledge management systems such as Data Warehouses can be integrated with organizational memory systems to provide enhanced services in the decision-making environments. Knowledge management technologies are expected to create innovation by supporting the following activities: externalization, internalization, intermediation and cognition 51]. According to the author, Externalization is the process of capturing knowledge repositories and matching them to other knowledge repositories. Internalization seeks to match bodies of knowledge to a particular user's need to know (transfer of explicit knowledge). Intermediation matches the knowledge seeker with knowledge by focusing on tacit knowledge or experience level in the organization. Cognition is the function of business systems to make decisions based on available knowledge by matching knowledge with firm processes. Clearly, three of these activities are directly related to the knowledge spiral that the OMKS seeks to support.

Marakas[HYPERLINK \l "Mar03" 52] identified over 30 different design and construction approaches to decision support methods and systems. Of these many different approaches, none have been considered the best. Most of the DSS development processes are very distinct and project specific. There has been proposed DSS methodologies, such as: 53], [HYPERLINK \l "Mar82" 54], 18], and [HYPERLINK \l "Bla79" 55]; cycles, such as: 56], [HYPERLINK \l "Kee78" 15]; processes, such as: 57], and[HYPERLINK \l "Cou79" 58] and guidelines for such areas as end-use computing. Differently, the approach that we present, OMKS, introduces an approach that is general and applicable across multiple organizations and contexts.

In this section, we discuss how the proposed approach can influence research and practice in the areas of knowledge management, organizational learning and decision support systems. We also present some of the benefits for organizations that employ the approach. Nevertheless, there are issues whenever information technologies are used to support organizational decision-making. We therefore highlight some of these issues and how organizations may alleviate problems that they may face.

Like previous work (e.g., 29], [HYPERLINK \l "Bol02" 13], 11], [HYPERLINK \l "Cho04" 34]), our research presented here only presents an approach, but not the actual implementation. Actual implementation of the approach is beyond the scope of the current work. However, in the following, we demonstrate the validity of the approach in terms of how the process approach meets requirements for such systems and how the approach is based on prior theoretical research.

Our approach seeks to take features of data warehouse and organizational memory systems. Hence, the base requirements for such systems should be met. Thus, we demonstrate how our approach meets both the technical and theoretical requirements for data warehousing11], organizational memory information systems [HYPERLINK \l "Ste95" 29], 31] and business process integration [HYPERLINK \l "Cho04" 34]. Atwood 31]notes that OMIS as presented by Stein and Zwass [HYPERLINK \l "Ste95" 29] hasapplication in the real world. Hence, our description of OMIS in this paper refers to systems that have use in the real world[1]. In proposing an effective-based integrative framework for OMIS, Stein and Zwass prescribes the following meta design requirements (goals) and meta designs (attributes of the artifacts) of four layers of such systems: integrative subsystem, adaptive subsystem, goal attainment subsystem, and pattern maintenance subsystem.

The integrative system enables organization's internal knowledge including technical issues, designs, past decisions and projects to be made explicit. Our approach emphasizes the ability to transform knowledge and make knowledge explicit for reuse by organizational actors across both space and time,

[1] Atwood [HYPERLINK \l "Atw02" 31] provides more detailed examples of how OMIS have been used in an academic setting, a government setting, and businesses such as Anderson Consulting and insurance industries.

which is the meta requirement for this subsystem as prescribed by Stein and Zwass 29]. The meta requirements for the adaptive subsystem include activities “to recognize, capture, organize, and distribute knowledge about the environment to the appropriate organizational actors [HYPERLINK \l "Ste95" 29]. OMKS facilitates these activities. The meta requirements of the goal attainment subsystem are to assist organizational actors design, store, evaluate and modify goals. The OLAP capabilities presented in the OMKS addresses this requirement by enabling the decision maker define key performance indicators in the OLAP environment and to modify, evaluate the specific goals as often as possible. The meta requirements of the pattern maintenance subsystem deal with the values, attitudes and norms of the organizational actors. The use of ontologies enhances understanding of the different organizational actors standardizing practices and norms. These ontologies also enables data from different sources be integrated in such a way that they have consistent meaning for the different actors and therefore enable organization build its culture and value among the diverse actors.

By explaining the various processes in the data warehousing activities such as extraction, transformation and loading of source data into the warehouse and knowledge management as process of creating, storing, sharing and reuse of knowledge, we show the applicability of the integrated process management theoretical concept in the proposed approach34]. Finally, we have presented details of how the traditional extraction, transformation and loading of the data warehouse architecture will progress in the proposed approach.

4.1 Implications for Research

Presented in this paper is an approach that integrates the functions of a typical data warehouse and organizational memory information systems. This approach uses an ontological metadata repository to assist in the structuring of data and facilitating learning, as well as introduces Scenarios as a process to capture tacit knowledge. Researchers in knowledge management, data warehousing, organizational learning, and decision support may use this approach as a way to review how the construction of DSS may be enhanced by looking specifically at the design and maintenance of the components of the OMKS presented in this paper.Systems integration issues are another areas that researchers may want to develop more knowledge.

4.2 Implications for Practice

Realizing all of the benefits of anOrganizational Memory and Knowledge System remains forthcoming because while the approach presented in this paper offers a process and technological means to realizing this end, much of the very important human aspects, which are very important parts of the system, are still evolving. One of the critical issues is the issue of motivating workers to contribute to the tacit knowledge capture and acquisition. Just as we have explained how OMIS have been used in the real world, we believe that building an integrated OMKS have far more reaching benefits. This is because all the processes that involve the building and use of traditional data warehouses and knowledge management systems would be joined; enabling diverse processes to be captured, stored, shared and used from a single unified system.

5. CONCLUSION

The data warehouse is a major component of modern DSS. Others have suggested that theintegration of DSS and knowledge management systems processes can enhance decision-making. In this paper, we have presented an approach for integrating functional features of data warehousing and organizational memory information system to enhance decision-making and organizational learning. We have discussed how Scenarios can be employed to facilitate the acquisition and representation of tacit knowledge into the Organizational Memory and Knowledge System and use ontologies as a metadata for managing other metadata from the diverse sources. The metadata also provides effective standardized semantics for the organization’s members to contribute, use, and learn from the knowledge in the OMKS.

There has been an extensive amount of research on ontology and its application to knowledge management. The use ofontology has been proven to be effective and efficient in this task. The inclusion of *Scenarios* in the solution is justified by appealing to the research and studies that have been presented to show the effectiveness of *Scenarios* in acquiring tacit knowledge.

Future research would involve proving the efficacy of the Knowledge Warehouse architecture and its ability to facilitate tacit knowledge acquisition and schematic integration. We argue that using ontologies is an effective means for tacit knowledge acquisition, standardization, presentation, and storage. This paper is a step in that direction.

6. REFERENCES

[1] D. O'Leary. "Using AI in knowledge management: Knowledge Bases and ontologies." IEEE Intelligent Systems, vol. 13, pp. 34-39, 1998.

[2] M. Alavi and D. Leidner. "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues." MIS Quarterly, vol. 25, no. 1, pp. 107-136, March 2001.

[3] P. Busch and D. Richards. "Mapping Tacit Knowledge Flows within Organization X." in Proc. 12th Australasian Conference on Information Systems, Coffs Harbour, Australia, 2001, pp. 85-94.

[4] P. Busch, D. Richards, and C. Dampney. "Mapping Tacit Knowledge Flows in Organisation X." in Proc. 11th Australasian Conference on Information Systems, Brisbane, Australia, 2000.

[5] E. Grant and M. Gregory. "Tacit knowledge, the life cycle and international manufacturing transfer." Technology Analysis & Strategic Management, vol. 9, no. 2, pp. 149-161, 1997.

[6] I. Nonaka, H. Takeuchi, and K. Umemoto. "A theory of organizational knowledge creation." International Journal of Technology Management, vol. 11, no. 7/8, pp. 833-845, 2009.

[7] S. Raghuram. "Knowledge creation in the telework context." International Journal of Technology Management, vol. 11, no. 7/8, pp. 859-870, 1996.

[8] J. Howells. "Tacit Knowledge and Technology Transfer." ESRC Centre for Business Research and Judge Institute of Management Studies, University of Cambridge, United Kingdom, Working paper 16, 1995.

[9] G. Goldman. "The tacit dimension of clinical judgment." The Yale Journal of Biology and Medicine, vol. 63, no. 1, pp. 47-61, 1990.

[10] Z. Pylyshyn. "The imagery debate: Analogue media versus tacit knowledge," in Readings in Cognitive Science: A perspective from Psychology and Artificial Intelligence, A. Collins and E. Smith, Eds. San Mateo, California, USA: Morgan Kaufman, 1981.

[11] H. Nemati, D. Steiger, L. Iyer, and R. Hershel. "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing." Decision Support Systems, vol. 33, no. 2, pp. 143-161, 2002.

[12] C. Yu-N and S. Abidi. "Tacit Knowledge Creation in a Knowledge Management Context," in Proc. 4th International Symposium on Computer and Information Sciences., 1999.

[13] N. Bolloju, M. Khalifa, and E. Turban. "Integrating Knowledge Management into enterprise environments for the Next Generation." Decision Support Systems, vol. 33, pp. 163-176, 2002.

[14] G. Gorry and M. Scott Morton. "A Framework for Management Information Systems." Sloan Management Review, vol. 13, no. 1, pp. 55-70, 1971.

[15] P. Keen and M. Scott Morton.DSS: An Organizational Perspective. Reading, MA, USA: Addison-

Wesley, 1978.

[16] R. Bonczek, C. Holsapple, and A. Whinston. "A Generalized Decision Support Systems Using Predicate Calculus and Network Data Base Management." Operations Research, vol. 29, no. 2, pp. 263-281, 1981.

[17] M. Ginzberg and E. Stohr. "Decision Support Systems: Issues and perspectives," in Decision Support Systems, M. Ginzberg, R. Reitman, and E. Stohr, Eds. North-Holland, Amsterdam: 1982, pp. 9-32.

[18] R. Sprague Jr. and E. Carlson.Building Effective Decision Support Systems. Englewood Cliffs, N.J., USA: Prentice-Hall, 1982.

[19] H. Simon. "A Behavioral Model of Rational choice." Quarterly Journal of Economics, vol. 69, pp. 99-118, 1955.

[20] B. Fazlollahi and R. Vahidov. "A Method for Generation of Alternatives by Decision Support Systems." Journal of Management Information Systems, vol. 18, no. 2, pp. 229-250, 2001.

[21] W. Inmon. "What is a Data Warehouse?. " Prism Tech Topic, vol. 1, no. 1, 1995.

[22] M. Staudt, A. Vaduva, and T. Vetterli. "The Role of Metadata for Data Warehousing." University of Zurich, Technical Paper 1999.

[23] I. Nonaka and H. Takeuchi.The Knowledge-Creating Company, How Japanese companies manage the dynamics of innovation. New York, USA: Oxford University Press, 1995.

[24] V. Benjamins, D. Fensel, and A. Perez. "Knowledge Management through Ontologies," in Proc. 2nd International Conference on Practical Aspects of Knowledge Management (PAKM 98), Basel, Switzerland., 1998.

[25] M. Polanyi.The Tacit Dimension. London, United Kingdom: Routledge and Kegan Paul, 1966.

[26] J. Vasconcelos, F. Gouveia, and C. Kimble. "An Organizational Memory Information System using Ontologies," in Proc. Third Conference of the Associacao Portuguesa de Sistemas de Informacao, Coimbra, Portugal, 2002.

[27] S.-H Liao. "Knowledge Management Technologies and Applications – Literature Review from 1995 to 2002." Expert Systems With Applications, vol. 25, no. 2, pp. 155-164, 2003.

[28] G. Hackbarth. "The Impact of Organizational Memory on IT Systems," in Proc. Fourth Americas Conference on Information Systems, 1998, pp. 588-590.

[29] E. Stein and V. Zwass. "Actualizing Organizational Memory with Information Systems." Information Systems Research, vol. 6, no. 2, pp. 85-117, 1995.

[30] J. Walsh and G. Ungson. "Organizational Memory." Academy of Management Review, vol. 16, no. 1, pp. 57-91, 1991.

[31] M. Atwood. "Organizational Memory Systems: Challenges for Information Technology," in Proceedings of 35th Hawaii International Conference on System Sciences, Big Island, HI, 2002, pp. 1-9.

[32] G. Heijst, R. Spek, and E. Kruizinga. "Corporate memories as a tool for knowledge management." Expert Systems With Applications, vol. 13, no. 1, pp. 41-54, 1997.

[33] I. Choi, M. Song, C. Park, and N. Park. "An XML-based process definition language for integrated process management." Computers in Industry, vol. 50, no. 1, pp. 85-102, 2003.

[34] I. Choi, J. Jung, and M. Song. "A framework for the integration of knowledge management and business process management." International Journal of Innovation and Learning, vol. 1, no. 4, pp. 399-408, 2004.

[35] M. Roseman. "Integrated knowledge and process management." B-HERT News, pp. 24-26, 2001.

[36] C. Yu-N and S. Abidi. "A Scenarios Mediated Approach for Tacit Knowledge Acquisition and Crystallisation: Towards Higher Return-On-Knowledge and Experience," in Proc. Third International Conference on Practical Aspects of Knowledge Management (PAKM2000), Basel, Switzerland, 2000.

[37] C. Yu-N and S. Abidi. "An Ontological-Based Scenario Composer for Knowledge Acquisition in Intelligent Systems," in Proc.IEEE Region Ten Conference (TENCON 2000), Kuala Lumpar, Malaysia, 2000.

[38] C. Potts, K. Takahashi, and A. Anton. "Inquiry-based scenario analysis of systems requirements." Georgia Tech, Atlanta, GA, Technical Report GIT-CC-94/14, 1994.

[39] R. Clarke, D. Filippidou, P. Kardasis, P. Loucopoulos, and R. Scott. "Integrating the Management of Discovered and Developed Knowledge," in Proc. Panhellenic Conference on New Information Technology, NIT'98, Athens, Greece, 1998.

[40] G. Neumann and M. Strembeck. "A scenario-driven role engineering process for functional RBAC roles," in Proc.7th ACM Symposium on Access Control Models and Technologies, Monterey, California, 2002, pp. 33-42.

[41] E. Kavalki, P. Loucopoulos, and D. Filippidou. "Using Scenarios to Systematically Support Goal-Directed Elaboration for Information System Requirements." Information Systems Engineering Group, Department of Computation, UMIST, Manchester, Technical Report ISE-96-1, 1996.

[42] L. van Elst and A. Abecker. "Ontologies for Information Management: Balancing Formality, Stability, and Sharing Scope." Expert Systems With Applications, vol. 23, pp. 357-366, 2002.

[43] M. Glinz. "Improving the Quality of Requirements with Scenarios," in Proc. Second World Congress for Software Quality, Yokohama, Japan, 2000, pp. 55-60.

[44] A. Hakimpour and A. Geppert. "Resolving semantic heterogeneity in schema integration," in International Conference on Formal Ontology in Information Systems (FOIS), 2001, pp. 297-308.

[45] A. Sheth. "Semantic Meta Data for Enterprise Information Integration." DM Review Magazine, 2003.

[46] M. Genesereth and N. Nilsson.Logical Foundations of Artificial Intelligence, M. Genesereth and N. Nilsson, Eds. San Mateo, CA: Morgan Kaufmann Publishers, 1987.

[47] T. Gruber. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." Stanford University Knowledge Systems Laboratory, Stanford, CA, Technical Report 1993.

[48] G. von Krough, K. Ichijo, and I. Nonaka.Enabling Knowledge Creations: How to Unlock the Mystery of Tacit Knowledge and Release the Power of Innovation. New York, USA: Oxford, 2000.

[49] D.-H. Jang and S. Myaeng. "Development of a document summarization system for effective information services," in Proc. RIAO 97 Conference: Computer-Assisted Information Searching on Internet, Montreal, Canada, 1997, pp. 101-111.

[50] J. Kupiec, J. Pedersen, and F. Chen. "A trainable document summarizer," in Proc.SIGIR-95: 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 68-73.

[51] C. Frappaolo. "What's in a name?," KMWorld, pp. 18-19, 1998.

[52] G. Marakas.Decision support systems in the 21st century. Upper Saddle River, USA: Prentice Hall, 2003.

[53] K. Sexena. "Decision support engineering: a DSS develoopment methodology," in Proc.24th Annual Hawaii International Conference on System Sciences, Los Alamitos, CA, 1991.

[54] M. Martin. "Determining Information Requirements for DSS." Journal of Systems Management, pp. 14-21, December 1982.

[55] R. Blanning. "The functions of a decision support system." Information and Management, pp. 71-96, September 1979.

[56] A. Sage.Decision support systems engineering. New York, USA: Wiley, 1991.

[57] C. Stabell. "A Decision-Oriented Approach to Building DSS," in Building Decision Support, J.L. Bennett, Ed. Reading, MA, USA: Addison-Wesley, 1983, pp. 221-260.

[58] J.-C. Courbon, J. Drageof, and J. Tomasi. "L'approche évolutive," Informatique et Gestion, vol. 103, Janvier-Février 1979.

Face Emotion Analysis Using Gabor Features In Image Database for Crime Investigation

Lt. Dr. S. Santhosh Baboo, Reader *santhos2001@sify.com*
P.G & Research, Dept. of Computer Application,
D.G. Vaishnav College,
Chennai-106. INDIA

V.S. Manjula *manjusunil.vs@gmail.com*
Research Scholar, Dept. of Computer Science & Engineering,
Bharathiar University,
Coimbatore-641 046. INDIA

Abstract

The face is the most extraordinary communicator, which plays an important role in interpersonal relations and Human Machine Interaction. Facial expressions play an important role wherever humans interact with computers and human beings to communicate their emotions and intentions. Facial expressions, and other gestures, convey non-verbal communication cues in face-to-face interactions. In this paper we have developed an algorithm which is capable of identifying a person's facial expression and categorize them as happiness, sadness, surprise and neutral. Our approach is based on local binary patterns for representing face images. In our project we use training sets for faces and non faces to train the machine in identifying the face images exactly. Facial expression classification is based on Principle Component Analysis. In our project, we have developed methods for face tracking and expression identification from the face image input. Applying the facial expression recognition algorithm, the developed software is capable of processing faces and recognizing the person's facial expression. The system analyses the face and determines the expression by comparing the image with the training sets in the database. We have followed PCA and neural networks in analyzing and identifying the facial expressions.

Keywords: Facial Expressions, Human Machine Interaction, Training Sets, Faces and Non Faces, Principal Component Analysis, Expression Recognition, Neural networks.

1. INTRODUCTION

Face expressions play a communicative role in the interpersonal relationships. Computer recognition of human face identity is the most fundamental problem in the field of pattern analysis. Emotion analysis in man-machine interaction system is designed to detect human face in an image and analyze the facial emotion or expression of the face. This helps in improving the interaction between the human and the machine. The machines can thereby understand the man's reaction and act accordingly. This reduces the human work hours. For example, robots can be used as a class tutor, pet robots, CU animators and so on.. We identify facial expressions not only to express our emotions, but also to provide important communicative cues during social interaction, such as our level of interest, our desire to take a speaking turn and continuous feedback signaling or understanding of the information conveyed. Support Vector Algorithm is well suited for this task as high dimensionality does not affect the Gabor Representations. The main disadvantage of the system is that it is very expensive to implement and maintain. Any changes to be upgraded in the system needs a change in the algorithm which is very sensitive and difficult; hence our developed system will be the best solution to overcome the above mentioned disadvantages.

In this paper, we propose a complete face expression recognition system by combining the face tracking and face expression identifying algorithm. The system automatically detects and extracts the human face from the background based on a combination of a retrainable neural network structure. In this system, the computer is trained with the various emotions of the face and when given an input, the computer detects the emotion by comparing the co-ordinates of the expression with that of the training examples and produces the output. Principle Component Analysis algorithm is the one being used in this system to detect various emotions based on the coordinates of the training sample given to the system.

2. RELATED WORK

Pantic & Rothkrantz [4] identify three basic problems a facial expression analysis approach needs to deal with: face detection in a facial image or image sequence, facial expression data extraction and facial expression classification. Most previous systems assume presence of a full frontal face view in the image or the image sequence being analyzed, yielding some knowledge of the global face location. To give the exact location of the face, Viola & Jones [5] use the Adaboost algorithm to exhaustively pass a search sub-window over the image at multiple scales for rapid face detection.

Essa & Pentland [6] perform spatial and temporal filtering together with thresholding to extract motion blobs from image sequences. To detect presence of a face, these blobs are then evaluated using the eigenfaces method [7] via principal component analysis (PCA) to calculate the distance of the observed region from a face space of 128 sample images. To perform data extraction, Littlewort et al. [8] use a bank of 40 Gabor wavelet filters at different scales and orientations to perform convolution. They thus extract a “jet” of magnitudes of complex valued responses at different locations in a lattice imposed on an image, as proposed in [9]. Essa & Pentland [6] extend their face detection approach to extract the positions of prominent facial features using eigenfeatures and PCA by calculating the distance of an image from a feature space given a set of sample images via FFT and a local energy computation.

Cohn et al. [10] first manually localize feature points in the first frame of an image sequence and then use hierarchical optical flow to track the motion of small windows surrounding these points across frames. The displacement vectors for each landmark between the initial and the peak frame represent the extracted expression information. In the final step of expression analysis, expressions are classified according to some scheme. The most prevalent approaches are based on the existence of six basic emotions (anger, disgust, fear, joy, sorrow and surprise) as argued by Ekman [11] and the Facial Action Coding System (FACS), developed by Ekman and Friesen [12], which codes expressions as a combination of 44 facial movements called Action Units. While much progress has been made in automatically classifying according to FACS [13], a fully automated FACS based approach for video has yet to be developed.

Dailey et al. [14] use a six unit, single layer neural network to classify into the six basic emotion categories given Gabor jets extracted from static images. Essa & Pentland [6] calculate ideal motion energy templates for each expression category and take the euclidean norm of the difference between the observed motion energy in a sequence of images and each motion energy template as a similarity metric. Littlewort et al. [8] preprocess image sequences image-by-image to train two stages of support vector machines from Gabor filter jets. Cohn et al. [10] apply separate discriminant functions and variance-covariance matrices to different facial regions and use feature displacements as predictors for classification.

2.1. Principle Component Analysis

Facial expression classification was based on Principle Component Analysis. The Principle Component Analysis (PCA) is one of the most successful techniques that have been used in image recognition and compression. PCA is a statistical method under the broad title of factor analysis. The purpose of PCA is to reduce the large dimensionality of the data space (observed variables) to the smaller intrinsic dimensionality of feature space (independent variables), which

are needed to describe the data economically. [3] The main trend in feature extraction has been representing the data in a lower dimensional space computed through a linear or non-linear transformation satisfying certain properties.

The jobs which PCA can do are prediction, redundancy removal, feature extraction, data compression, etc. Because PCA is a classical technique which can do something in the linear domain, applications having linear models are suitable, such as signal processing, image processing, system and control theory, communications, etc. [3]

Given an s-dimensional vector representation of each face in a training set of images, Principal Component Analysis (PCA) tends to find a t-dimensional subspace whose basis vectors correspond to the maximum variance direction in the original image space. This new subspace is normally lower dimensional. If the image elements are considered as random variables, the PCA basis vectors are defined as eigenvectors of the scatter matrix. PCA selects features important for class representation.

The main idea of using PCA for face recognition is to express the large 1-D vector of pixels constructed from 2-D facial image into the compact principle components of the feature space. This can be called eigenspace projection. Eigenspace is calculated by identifying the eigenvectors of the covariance matrix derived from a set of facial images(vectors). [3] we implement a neural network to classify face images based on its computed PCA features.

Methodology

Let $\{X_1, X_2,\ldots X_n\}$, $\mathbf{x} \in \Re^n$ be N samples from L classes $\{\omega_1, \omega_2, \ldots\ldots \omega_L\}$, and $p(\mathbf{x})$ their mixture distribution. In a sequel, it is assumed that a priori probabilities $P(\omega i)$,$I = 1,2,...,L$, are known.
Consider $\mathbf{m}$ and $\mathbf{\Sigma}$ denote mean vector and covariance matrix of samples, respectively. PCA algorithm can be used to find a subspace whose basis vectors correspond to the maximum variance directions in the original n dimensional space. PCA subspace can be used for presentation of data with minimum error in reconstruction of original data. Let $\mathbf{\Phi}^p PCA$ denote a linear $n\times p$ transformation matrix that maps the original n dimensional space onto a p dimensional feature subspace where $p < n$. The new feature vectors,

$$Y_i = (\mathbf{\Phi}^p PCA)^t * i,\ I = 1,2,\ldots N \quad \text{---} \quad (1)$$

It is easily proved that if the columns of $\mathbf{\Phi}^p PCA$ are the
eigenvectors of the covariance matrix corresponding to its p largest eigenvalues in decreasing order, the optimum feature space for the representation of data is achieved. The covariance matrix can be estimated by:

$$\Sigma = \left(\sum_{i=1}^{N} (x_i - m)(x_i - m)^t\right) / (N - 1) \quad \text{------} \quad (2)$$

Where m in (2) can be estimated by:

$$M_k = m_{k-1} + \eta\,(x_k - m_{k-1}) \quad \text{-------} \quad (3)$$

Where mk is estimation of mean value at k-th iteration and xk is a the k-th input image. PCA is a technique to extract features effective for representing data such that the average reconstruction error is minimized. In the other word, PCA algorithm can be used to find a subspace whose basis vectors correspond to the maximum variance directions in the original n dimensional space. PCA transfer function is composed of significant eigenvectors of covariance matrix. The following equation can be used for incremental estimation of covariance matrix :

$$\Sigma_k = \Sigma_{k-1} + \eta_k(X_k X_k^t - E_{k-1}) \quad \text{-------} \quad (4)$$

Where $\boldsymbol{\Sigma}k$ is the estimation of the covariance matrix at k-th iteration, $\mathbf{x_k}$ is the incoming input vector and η_k is the learning rate.

The emotion categories are : Happiness, sadness, surprise, disgust, fear, anger, neutral. In this project the emotions such as Happiness, sadness, Surprise and neutral is taken into account for human emotion reorganization.

2.2 Training Sets

The ensemble of input-desired response pairs used to train the system. The system is provided with various examples and is trained to give a response to each of the provided examples. The input given to the system is compared with the examples provided. If it finds a similar example then the output is produced based on the example's response. This method is a kind of learning by examples.

Our system will perform according to the number of training sets we provide. The accuracy of our system depends on the number of training sets we provide. If the training samples are higher the performance of the system is accurate. Sample training sets are shown in the fig[1] below,

FIGURE 1: Samples of training set images

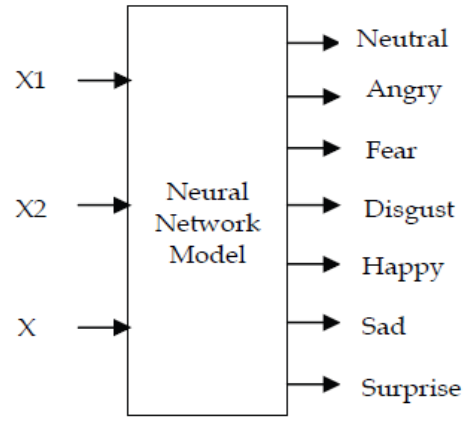
2.3. Neural Networks

Recently, there has been a high level of interest in applying artificial neural network for solving many problems**.** The application of neural network gives easier solution to complex problems such as in determining the facial expression**.** Each emotion has its own range of optimized values for lip and eye. In some cases an emotion range can overlap with other emotion range. This is experienced due to the closeness of the optimized feature values. For example, in Table, X1 of Sad and Dislike are close to each other. These values are the mean values computed from a range of values. It has been found that the ranges of feature values of X1 for Sad and dislike overlap with each other. Such overlap is also found in X for Angry and Happy. A level of intelligence has to be used to identify and classify emotions even when such overlaps occur.

Emotions	Manually Computed Mean Value (in pixels)			Optimized Mean Value by GA (in pixels)		
	b1	b2	b	X1	X2	X
Neutral	38	41	21	34.2644	35.2531	19.6188
Fear	25	41	16	23.0287	36.9529	14.7024
Happy	25	48	16	21.5929	43.4742	15.0393
Sad	33	34	19	30.9104	28.5235	16.9633
Angry	25	34	16	24.2781	30.8381	15.4120
Dislike	35	29	13	31.3409	21.6276	12.8353
Surprise	43	57	17	42.6892	55.5180	16.0701

FIGURE 2: Optimized Value of three features

A feed forward neural network is proposed to classify the emotions based on optimized ranges of 3-D data of top lip, bottom lip and eye. The optimized values of the 3-D data are given as inputs to the network as shown in Figure. The network is considered to be of two different models where the first model comes with a structure of 3 input neurons, 1 hidden layer of 20 neurons and 3 output neurons (denoted by (3x20x3)) and the other model with a structure of (3x20x7). The output of (3x20x3) is a 3-bit binary word indicating the seven emotional states. The output (Oi, i=1, 2,.., 7) of (3x20x7) is of mutually exclusive binary bit representing an emotion. The networks with each of the above listed input sizes are trained using a back-propagation training algorithm. A set of suitably chosen learning parameters is indicated in Table. A typical “cumulative error versus epoch” characteristic of the training of NN models as in Figure 10 ensures the convergence of the network performances. The training is carried out for 10 trials in each case by reshuffling input data within the same network model. The time and epoch details are given in Table 3 which also indicates the maximum and minimum epoch required for converging to the test-tolerance.



(b) NN of (3x20x7) structure

Neural Network Structures – (a & b)

Hidden neurons: 20	Learning rate: 0.0001	Activation function: (1/ (1+e-x))
Momentum factor: 0.9	No. of samples: 70	Maximum no. of epoch: 1000
Testing tolerance: 0.1	Training tolerance: 0.0001	No. of trained samples: 50

NN structure	Epoch (in 10 trials)			Training Time (sec) (in 10 trials)			Classification % (in 10 trials)	
	Min	Max	Mean	Min	Max	Mean	Range	Mean
3x20x3	105	323	225	2.18	7.32	5.02	75.71 - 90.00	83.57
3x20x7	71	811	294	3.43	39.25	13.76	81.42 -91.42	85.13

FIGURE 3: Details of Neural Network Classification using Neural Networks

2.4. Feature Extraction

A feature extraction method can now to be applied to the edge detected images. Three feature extraction methods are considered and their capabilities are compared in order to adopt one that is suitable for the proposed face emotion recognition problem. They are projection profile, contour profile and moments (Nagarajan et al., 2006).

Implementation Overview

For PCA to work properly, we have to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have subtracted, and all the y values have subtracted from them. This produces a data set whose mean is zero as shown in the fig [2].

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

FIGURE 2:

The next step is to calculate the Co-Variance Matrix. Since the data is 2 dimensional, the covariance matrix will be 2 x 2. We will just give you the result as

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

2.5 Eigen Vectors and Eigen Faces

Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues for this matrix. These are rather important, as they tell us useful information about our data. Here are the eigenvectors and eigenvalues,

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

It is important to notice that these eigenvectors are both unit eigenvectors that is their lengths are both 1. This is very important for PCA. Most maths packages, when asked for eigenvectors, will give you unit eigenvectors. It turns out that the eigenvector with the highest eigen value is the principle component of the data set. In our example, the eigenvector with the largest eigen value is the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions.

In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigen value, highest to lowest. This gives the components in order of significance. If we leave out some components, the final data set will have less dimensions than the original. To be precise, if we originally have n dimensions in our data, and so calculate n eigenvectors and eigen values, and then choose only the first p eigenvectors, then the final data set has only p dimensions. This is the final step in PCA is choosing the components (eigenvectors) that we wish to keep in our data and form a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Final Data=RowFeatureVector x RowDataAdjust

where RowFeatureVector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and RowDataAdjust is the mean-adjusted data transposed, ie. the data items are in each column, with each row holding a separate dimension. Final Data is the final data set, with data items in columns, and dimensions along rows.

The implementation chart is as in Fig [4],

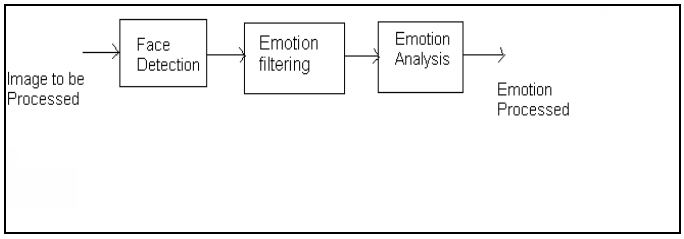


FIGURE 4: Implementation Chart

2.6 Face Detection

Face Detection follows the Eigen face approach. Once the eigenfaces have been computed, several types of decision can be made depending on the application. The steps involved in this approach are mentioned below: (1) Acquire initial set of face images (the training set). (2) Calculate Eigenvector from the training set keeping only M images (face space). (3) Calculate the corresponding distribution in the face space. The process is explained in the following flow diagram Fig [5],

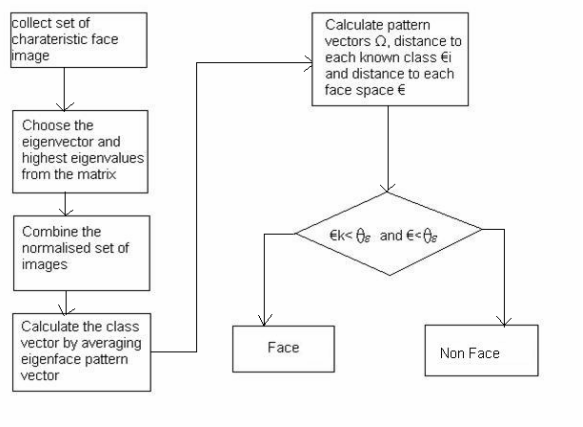


FIGURE 5: Flow Diagram

2.7 Emotion Filtering

Emotion Filtering is where in the detected image is projected over the various average faces such as Happy, Neutral, Sad, and Surprised. The filters provide the output values of the given input by comparing the input with the training samples present in the system. The input is compared to all the training sets of various emotions. The values obtained after the comparison in this module is provided to the analyzer.

2.8 Emotion Analyzer

This is the decision making module. The output of the emotion filter is passed into the emotion analyzer. There are threshold values set for each emotions. For example,

- If output value is equal to the threshold value, then it is considered as neutral
- If output value is greater than the threshold value, then it is considered as happy
- If output value is less than the threshold value, then it is considered as sad

Thus the emotion analyzer makes the decision based on the threshold values provided. By adding new threshold values one add a new emotion to the system.

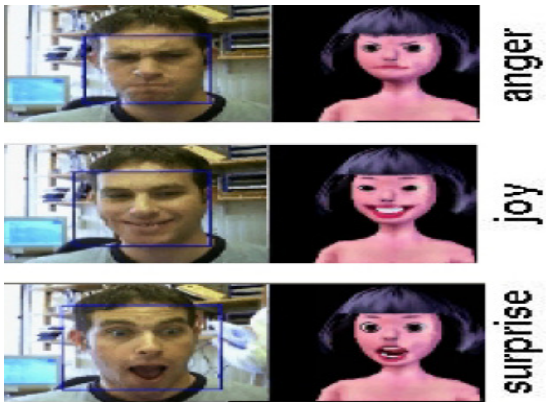


FIGURE 6: Sample Emotions

In our project, we can have nearly 300 images with which our software is completely trained. Some samples of the training set images are shown in the following figure (7). We have developed this system using Dotnet language. In our software we have developed a form with various input and functionality objects.

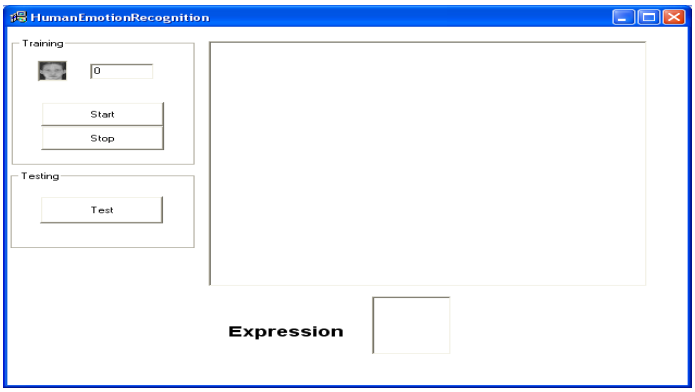


FIGURE 7

In the first phase our system runs the training sets by which system training is over. In the second phase we can stop the training and in the third phase we continue with the testing module. In the testing phase our system will start showing the emotions of the images continuously as in the following figure (8).

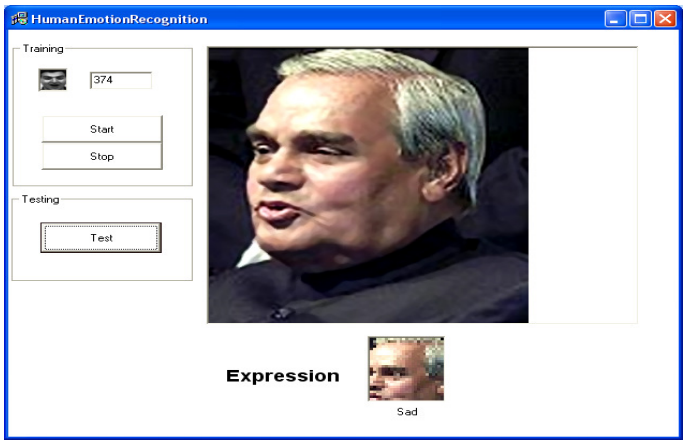


FIGURE 8

The expression in the above image seems to be sad, and the result of the system is also sad. Thus it is proven that our system is able to identify the emotions in the images accurately.

3. EXPERIMENTAL RESULTS

Face segmentation and extraction results were obtained for more than 300 face images, of non-uniform background, with excellent performance. The experimental results for the recognition stage presented here, were obtained using the face database. Comparisons were accomplished for all methods using face images. In our system, all the input images are getting converted to the fixed size pixels before going for the training session. However, this recognition stage has the great advantages of working in the compressed domain and being capable for multimedia and content-based retrieval applications.

Face detection and recognition can also be done using Markov random fields, Guided Particle Swarm Optimization algorithm and Regularized Discriminant Analysis based Boosting algorithm. In Markov Random Fields method the eye and mouth expressions are considered for finding the emotion, using the edge detection techniques. This kind of emotion analysis will not be perfect because the persons actual mood cannot be identified just by his eyes and mouth. In the Guided Particle Swarm optimization method video clips of the user showing various emotions are used, the video clip is segmented and then converted into digital form to identify the emotion. This is a long and tedious process. Our proposed system was implemented using C# programming

language under .NET development framework. The implemented project has two modes, the training mode and the detection mode. The training mode is completed once if all the images are trained to the system. In the detection mode the system starts identifying the emotions of all the images which will be 97% accurate. The following figure shows the comparison of recognition rates.

Methods	Recognition Rates
Markov Random Fields	95.91
Guided Particle Swarm Optimization	93.21
Regularized Discriminant Analysis based Booting	96.51
Our Proposed method	97.76

FIGURE 9

Comparison of facial expression recognition using image database

The above results are shown as comparison graphs in the following Fig [11],

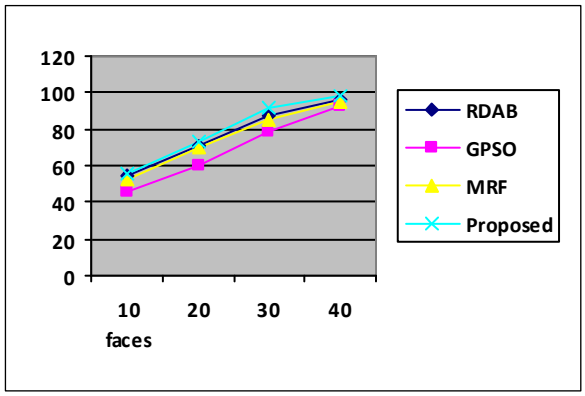


FIGURE 10

Here, we have taken number of training samples in the x-axis and recognition rate on the y-axis. The accuracy of emotion recognition is high in our proposed method compared to the other two algorithms.

4. CONCLUSION

In this paper we have presented an approach to expression recognition in the images. This emotion analysis system implemented using PCA algorithm is used in detection of human emotions in the images at low cost with good performance. This system is designed to recognize expressions in human faces using the average values calculated from the training samples. We evaluated our system in terms of accuracy for a variety of images and found that our system was able to identify the face images and evaluate the expressions accurately from the images.

REFERENCES

[1] M. Pantic, L.J.M. Rothkrantz, “Automatic Analysis of Facial Expressions”: The State of the Art, *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 22, No. 12, pp. 1424-1445, 2000

[2] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, “Classifying Facial Actions", IEEE *Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 974-989, 1999

[3] Face Recognition using Principle Component Analysis -Kyungnam Kim, Department of Computer Science, University of Maryland, and College Park, MD 20742, USA.

[4] M. Pantic, L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions": The State of the Art, IEEE *Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1424-1445, 2000

[5] P. Viola and M. Jones. Robust real-time object detection. Technical Report 2001/01, Compaq Cambridge Research Lab, 2001.

[6] I. Essa and A. Pentland. Coding, analysis, *I*nterpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[7] M. Turk and A. Pentland. "Eigenfaces for recognition". *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[8] G. Littlewort, I. Fasel, M. Stewart Bartlett, and J. R. Movellan. Fully automatic coding of basic expressions from video. Technical Report 2002.03, UCSD INC MPLab, 2002.

[9] M. Lades, J. C. Vorbr¨uggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. W¨urtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[10] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.

[11] P. Ekman. Basic emotions. In T. Dalgleish and T. Power, editors, *The Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd., 1999.

[12] P. Ekman and W. Friesen. "*Facial Action Coding System" (FACS): Manual*. Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[13] Y.-L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for Facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[14] M. N. Dailey, G. W. Cottrell, and R. Adolphs. A six-unit network is all you need to discover happiness. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, Mahwah, NJ, USA, 2000. Erlbaum.

[15] Nagarajan R., S. Yaacob, P. Pandiyan, M. Karthigayan, M. Khalid & S. H. Amin. (2006). Real Time Marking Inspection Scheme for Semiconductor Industries. *International Journal of Advance Manufacturing Technology (IJAMT)*, ISSN No (online): 1433-3015.

[16] Karthigayan. M, Mohammed Rizon, Sazali Yaacob & R. Nagarajan,. (2006). An Edge Detection and Feature Extraction Method Suitable for Face Emotion Detection under Uneven Lighting, *The 2nd Regional Conference on Artificial Life and Robotics (AROB'06)*, July 14 - July 15, Hatyai, Thailand.

[17] Karthigayan M, M.Rizon, S.Yaacob and R. Nagarajan. (2007). On the Application of Lip features in Classifying the Human Emotions, *International Symposium on Artificial life and Robotics,* Japan.

A Performance Based Transposition Algorithm for Frequent Itemsets Generation

Sanjeev Kumar Sharma *spd50020@gmail.com*
Research Scholar
Devi Ahilya University, Takshashila Campus,
Khandwa Road, Indore (M.P.) India

Ugrasen Suman *ugrasen123@yahoo.com*
Associate Professor
Devi Ahilya University,Takshashila Campus,
Khandwa Road Indore (M.P.) India

Abstract

Association Rule Mining (ARM) technique is used to discover the interesting association or correlation among a large set of data items. It plays an important role in generating frequent itemsets from large databases. Many industries are interested in developing the association rules from their databases due to continuous retrieval and storage of huge amount of data. The discovery of interesting association relationship among business transaction records in many business decision making process such as catalog decision, cross-marketing, and loss-leader analysis. It is also used to extract hidden knowledge from large datasets. The ARM algorithms such as Apriori, FP-Growth requires repeated scans over the entire database. All the input/output overheads that are being generated during repeated scanning the entire database decrease the performance of CPU, memory and I/O overheads. In this paper, we have proposed a Performance Based Transposition Algorithm (PBTA) for frequent itemsets generation. We will compare proposed algorithm with Apriori and FP Growth algorithms for frequent itemsets generation. The CPU and I/O overhead can be reduced in our proposed algorithm and it is much faster than other ARM algorithms.

Keywords: Data Mining, Association Rule Mining (ARM), Association rules.

1. INTRODUCTION

There are several organizations in the mainstream of business, industry, and the public sector, which store huge amount of data containing their transaction information online and offline. Such data may contain hidden information that can be used by an organization's decision makers to improve the overall profit. The efficient transformation of these data into beneficial information is thus a key requirement for success in these organizations. Data mining techniques are heavily used to search information and relationships that would be hidden in transaction data. There are various techniques of data mining such as clustering, classification, pattern recognition, correlation, and Association Rule Mining (ARM). The ARM is most important data mining technique that is used to extract hidden information from large datasets. In ARM algorithms, association rules are used to identify relationships among a set of items in database. These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather based on co-occurrence of the data items.

The association rules are firstly introduced and subsequently implemented for the generation of frequent itemsets from the large databases [1],[2]. Association rules identify the set of items that are most often purchased with another set of items. For example, an association rule may state that 75% of customers who bought items A and B also bought C and D. The main task of every ARM is to discover the sets of items that frequently appear together called frequent itemsets.

ARM has been used for a variety of applications such as banking, insurance, medicine, website navigation analysis etc.

Frequent itemset can be produced from discovering useful patterns in customer's transaction databases. Suppose T ={t_1,t_2,t_3…..t_n} is a customer's transaction database, which is a sequence of transactions where each transaction is an itemset ($t_i \subseteq T$). Let J ={i_1,i_2,…..i_n} be a set of items, and D is a task relevant data**,** which can be a set of database transactions where each transaction T is a set of items such that $T \subseteq J$. Each transaction is associated with identifier called TID. Let A be a set of items and the transaction T is said to contain A if and only if $A \subseteq T$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transaction in D that contain $A \cup B$ (i.e. both A and B). This is taken to be the probability $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transaction in D containing A that also contain B. This is taken to be the conditional probability P(B|A). Therefore, the Support($A \Rightarrow B$) = $P(A \cup B)$ and Confidence($A \Rightarrow B$) = P(B|A). Those rules that satisfy both minimum support threshold and minimum confidence threshold are called strong. The values for support and confidence have to occur between 0% and 100%. The problem of mining association rules is to generate all rules that have support and confidence greater than some user specified minimum support and minimum confidence thresholds, respectively. This problem can be decomposed into the following sub-problems: i). All itemsets that have support above the user specified minimum support are generated. These itemsets are called the large itemsets. ii). For each large itemset, all the rules that have minimum confidence are generated as follows: for a large itemset X and any $Y \subset X$, if support(X)/support(X - Y) ≥ minimum- confidence, then the rule $X - Y \rightarrow Y$ is a valid rule.

There are various algorithm of ARM such as Apriori, FP-growth, Eclat etc. The most important algorithm of ARM is Apriori, which is not only influenced the association rule mining community, but it has affected other data mining fields as well [3]. Association rule and frequent itemset mining has become now a widely research area and hence, faster and faster algorithms have been presented. Numerous of them are Apriori based algorithms or Apriori modifications. Those who adapted Apriori as a basic search strategy, tended to adapt the whole set of procedures and data structures as well [4],[5],[6],[7]. Since the scheme of this important algorithm was not only used in basic association rules mining, but also used in other data mining fields such as hierarchical association rules [8],[9],[10], association rules maintenance [11],[12],[13] , sequential pattern mining [14], episode mining [15] and functional dependency discovery [16],[17] etc. Basically, ARM algorithms are defined into two categories; namely, algorithms respectively with candidate generation and algorithms without candidate generation. In the first category, those algorithms which are similar to Apriori algorithm for candidate generation are considered. Eclat may also be considered in the first category [9]. In the second category, the FP-Growth algorithm is the best–known algorithm. Table-1, defines the comparison among these three algorithms [3].

Algorithm	Scan	Data Structures
Apriori	M+1	HashTable &Tree
Eclat	M+1	HashTable &Tree
FP-Growth	2	PrefixTree

TABLE 1: Comparison of Algorithms

The main drawback of above discussed algorithms given above is the repeated scans of large database. This may be a cause of decrement in CPU performance, memory and increment in I/O overheads. The performance and efficiency of ARM algorithms mainly depend on three factors; namely candidate sets generated, data structure used and details of implementations [18]. In this paper we have proposed a Performance Based Transposition Algorithm (PBTA) which uses these three factors. Transactional database is considered as a two dimension array which works on boolean value dataset. The main difference between proposed algorithm and other algorithms is that instead of using transactional array in its natural form, our algorithm uses transpose of array i.e. rows and columns of array are interchanged. The advantage of using transposed array

is to calculate support count for particular item. There is no need to repeatedly scan array. Only by finding the row sum of the array will give the required support count for particular item, which ultimately results in increased efficiency of the algorithm. In the first pass of PBTA, we will receive all the support count value for the 1-itemset. Mining of association rules is a field of data mining that has received a lot of attention in recent years.

The rest of this Paper is organized as follows. In Section 2, we will explain the Apriori algorithm through association rules mining. Section 3 introduces our proposed PBTA algorithm with an illustration and compare with other algorithms. Experimental results are shown in Section 4. The concluding remarks are discussed in Section 5.

2. APRIORI ALGORITHM

ARM is one of the promising techniques of data mining to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. There are several ARM algorithms such as Apriori, FP-Growth, Eclat. The Apriori algorithm is also called the level-wise algorithm to find all of the frequent sets, which uses the downward closure property. The advantage of the algorithm is that before reading the database at every level, it prunes many of the sets which are unlikely to be frequent sets by using the Apriori property, which states that all nonempty subsets of frequent sets must also be frequent. This property belongs to a special category of properties called anti-monotone in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well. Using the downward closure property and the Apriori property the algorithm works as follows. The first pass of the algorithm counts the number of single item occurrences to determine the L_1 or single member frequent itemsets. Each subsequent pass, K, consists of two phases. First, the frequent itemsets L_k-1 found in the $(k-1)^{th}$ pass are used to generate the candidate itemsets C_k, using the Apriori candidate generation algorithm. Therefore, the database is scanned and the support of the candidates in C_k is determined to ensure that C_k itemsets are frequent itemsets [19].

Pass 1
1. Generate the candidate itemsets in C_1
2. Save the frequent itemsets in L_1

Pass k

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}
   a) Join L_{k-1} p with L_{k-1} q, as follows:
      insert into C_k
      select p.item1, p.item2, . . . , p.item$_{k-1}$, q.item$_{k-1}$
      from L_{k-1} p, L_{k-1}q
      where p.item1 = q.item1, . . . p.item$_{k-2}$ = q.item$_{k-2}$,
      p.item$_{k-1}$ < q.item$_{k-1}$
   b) Generate all (k-1)-subsets from the candidate itemsets in C_k
   c) Prune all candidate itemsets from C_k where some (k-1)-subset of the candidate itemset is not in the frequent itemset L_k-1
2. Scan the transaction database to determine the support for each candidate itemset in C_k
3. Save the frequent itemsets in L_k

We will use Apriori algorithm for ARM as a basic search strategy in our proposed algorithm. The proposed algorithm will adapt the whole set of procedures of Apriori but the data structure will be different. Also, the proposed algorithm will use the transposition of transactional database as data structures.

3. PERFORMANCE BASED TRANSPOSITION ALGORITHM (PBTA)

In Apriori algorithm, discovery of association rules require repeated passes over the entire database to determine the commonly occurring set of data items. Therefore, if the size of disk and database is large, then the rate of input/output (I/O) overhead to scan the entire database may be very high. We have proposed Performance Based Transposition Algorithm (PBTA), which improves the Apriori algorithm for repeated scanning of large databases for frequent itemsets generation. In PBTA, transaction dataset will be used in the transposed form and the description of proposed algorithm is discussed in the following sub-sections.

3.1 Candidate Generation Algorithm

In the candidate generation algorithm, the frequent itemsets are discovered in k-1 passes. If k is the pass number, L_{k-1} is the set of all frequent (k-1) itemsets. C_k is the set of candidate sets of pass k and c denotes the candidate set. $I_1, I_2 \ldots I_k$ are the itemsets[19]. The candidate generation procedure is as follows.

Procedure Gen_candidate_itemsets (L_{k-1})
$C_k = \Phi$
for all itemsets $I_1 \in L_{k-1}$ *do*
for all itemsets $I_2 \in L_{k-1}$ *do*
if $I_1[1] = I_2[1]$ ^ $I_1[2] = I_2[2]$ ^ … ^ $I_1[k-1] < I_2[k-1]$
then $c = I_1[1], I_1[2] \ldots I_1[k-1], I_2[k-1]$
$C_k = C_k \cup \{c\}$

3.2 Pruning Algorithm

The pruning step eliminates some candidate sets which are not found to be frequent.

Procedure Prune(C_k)
for all $c \in C_k$
for all (k-1)-subsets d of c do
if $d \notin L_{k-1}$
then $C_k = C_k - \{c\}$

3.3 PBTA Algorithm Description

The PBTA uses candidate generation and pruning algorithms at every iteration. It moves from level 1 to level k or until no candidate set remains after pruning. The step-by-step procedure of PBTA algorithm is described as follows.

1. Transpose the transactional database
2. Read the database to count the support of C_1 to determine L_1 using sum of rows.
3. L_1= Frequent 1- itemsets and k:= 2
4. While (k-1 ≠ NULL set) do
   Begin
   C_k := Call Gen_candidate_itemsets (L_{k-1})
   Call Prune (C_k)
   for all itemsets $i \in I$ do
   Calculate the support values using dot-multiplication of array;
   L_k := All candidates in C_k with a minimum support;
   K:k+1
   End
5. End of step-4

3.3.1 An Illustaration

Suppose we have a transactional database in which the user transactions from T1 to T5 and items from A1 to A5 are stored in the form of boolean values, which is shown in Table 1. We have assumed that this database can be generated by applying Apriori algorithm for frequent itemsets generation.

.

	A1	A2	A3	A4	A5
T1	1	0	0	0	1
T2	0	1	0	1	0
T3	0	0	0	1	1
T4	0	1	1	0	0
T5	0	0	0	0	1

TABLE 2: Transaction Database

Consider the transpose of transactional database of Table 1 is stored in Table 2 by applying metrics arithmetics that can be used in our proposed algorithm (PBTA). Assume the user-specified minimum support is 40%, and then the steps for generating all frequent item sets in proposed algorithm will be repeated until NULL set is reached. In PBTA, transactional dataset will be used in the transposed form. Therefore, candidate set and frequent itemset generation process will be changed as compared to Apriori algorithm. In the first pass, we will receive L_1.

{A1}→ 1, {A2}→2, {A3}→1, {A4}→2, {A5}→3
L_1 := { {A1}→1, {A2}→2, {A3} →1, {A4} → 2, {A5} → 3}

Then the candidate 2-itemset will be generated by performing dot-multiplication of rows of array, as array consist of boolean values, the resultant cell will be produce in the form of 1. If the corresponding cells of the respective rows have 1, otherwise 0 will be in the resultant cell. In this approach, we will receive a new array consisting of candidate 2-itemsets to get the higher order of itemsets. The above process between rows of array can be performed to find out the results.

A1	1	0	0	0	0
A2	0	1	0	1	0
A3	0	0	0	1	0
A4	0	1	1	0	0
A5	1	0	1	0	1

TABLE 3: Transpose Database of Transaction

In the second pass, where k=2 , the candidate set C_2 becomes
C_2: = {{A1*A2}, {A1*A3}, {A1,*A4}, {A1*A5}, {A2*A3}, {A2*A4},{A2*A5},{A3*A4},{A3*A5},{A4*A5} }
The pruning step does not change C_2 as all subsets are present in C_1.
Read the database to count the support of elements in C_2 to get:
{ {A1*A2}→ 0, {A1*A3}→0, {A1*A4} →0, {A1*A5}→1, {A2*A3}→ 1,
{A2*A4} →1, {A2*A5}→0, {A3*A4}→0, {A3*A5}→0, {A4*A5}→1}} and reduces to
L_2 = { {A1*A5}→1, {A2*A3}→1, {A2*A4}→1, {A4*A5}→1}}

In the third pass where k=3, the candidate generation step proceeds:
In the candidate generation step,

- Using {A1*A5} and {A4*A5} it generates {A1*A4*A5}
- Using {A2*A3} and {A2*A4} it generates {A2*A3*A4}
- Using {A2*A4} and {A4*A5} it generates {A2*A4*A5}

Thus, C_3:= { {A1*A4*A5}, {A2*A3*A4}, {A2*A4*A5}}

The pruning step prunes {1,4,5}, {2,3,4},{2,4,5} as not all subsets of size 2, i.e., {1,4}, {3,4}, {2,5} are not present in L_3.

So $C_{3:}= \Phi$

Hence the total frequent sets becomes $L:= L_1 \cup L_2$.

By comparing both Apriori and proposed algorithm, we found that Apriori algorithm requires multiple passes of the dataset to calculate support count for different itemsets. Therefore, in the case of Apriori, the record pointer moves the order of candidate item set * no of records while in the case of PBTA algorithm, record pointer moves equal to only order of candidate itemsets. For example, if we have to find out support count value for 2-itemset in a dataset having 5 items with 5 records using Apriori algorithm number of time record pointer will be 2*5 i.e. 10 while in case of our proposed algorithm it will be 2 only.

4. EXPERIMENTAL EVALUATIONS

The performance comparison of PBTA with classical frequent pattern-mining algorithms such as Apriori, FP-Growth is presented in this Section. All the experiments are performed on 1.50 Ghz Pentium-iv desktop machine with 256 MB main memory, running on Windows-XP operating system. The program for Apriori, FP-Growth and proposed algorithm PBTA were developed in Java JDK1.5 environment. We report the experimental results on three synthetic boolean datasets with 300K, 500K and 700K records, each having 130 columns. The datasets consists of boolean values are shown in table 3, table 4 and table 5. The performances results of Apriori, FP-growth and PBTA are shown with Fig. 1, Fig. 2 and Fig. 3 with data size 300K, 500K, and 700K represented in the graphical form. The X-axis in these graphs represents the support threshold values while the Y-axis represents the response times (in milliseconds) of the algorithms being evaluated as shown.

	Response Time (in ms)		
Support Count (in %)	**Apriori**	**FP-Growth**	**PBTA**
50	50	40	10
40	100	80	40
30	150	100	150
20	300	230	60
10	500	400	80

TABLE 4: Response Time Comparison of algorithms with 300k database

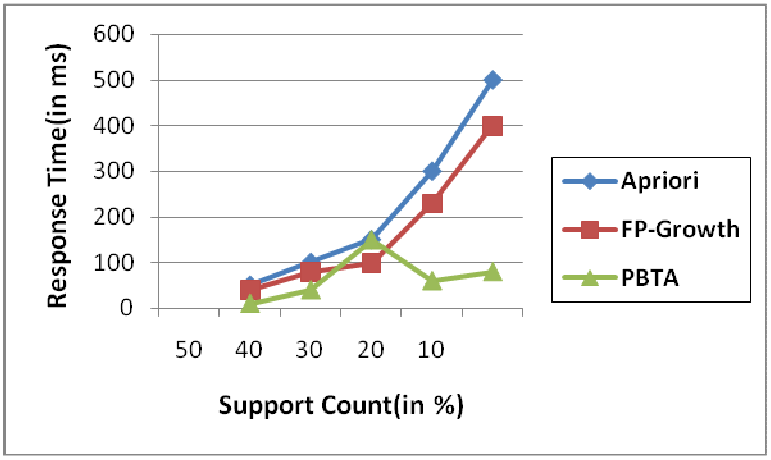


FIGURE 1: Performance analysis of algorithms (300k database)

In the first case, we have considered the transactional database with 300k in size as it is shown in Fig. 1. We have compared the performance of Apriori, FP-Growth with PBTA on the basis of response time. The observation shows that as the support count will be decreased and the response time taken by PBTA is much lesser then Apriori and FP-Growth algorithm.

	Response Time (in ms)		
Support Count (in %)	**Apriori**	**FP-Growth**	**PBTA**
50	84	67	17
40	167	134	67
30	251	167	251
20	501	384	100
10	835	668	134

TABLE 5: Response Time Comparison of algorithms with 500k database

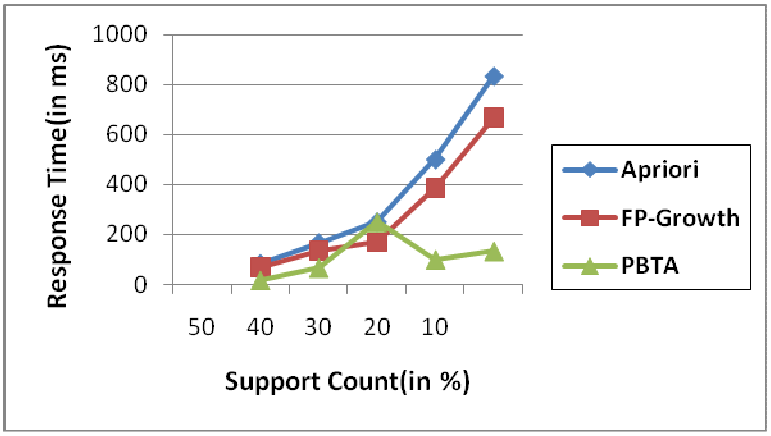


FIGURE 2: Performance analysis of algorithms (500k database)

In the another case, the transactional database with 500k size is considered, which is shown in Fig. 2. Hence, we have observed that as the support count threshold is reduced and the response time taken by PBTA is much lesser then Apriori and FP-Growth algorithm.

	Response Time (in ms)		
Support Count (in %)	**Apriori**	**FP-Growth**	**PBTA**
50	117	93	23
40	233	186	93
30	350	233	350
20	699	536	140
10	1165	932	186

TABLE 6: Response Time Comparison of algorithms with 700k database

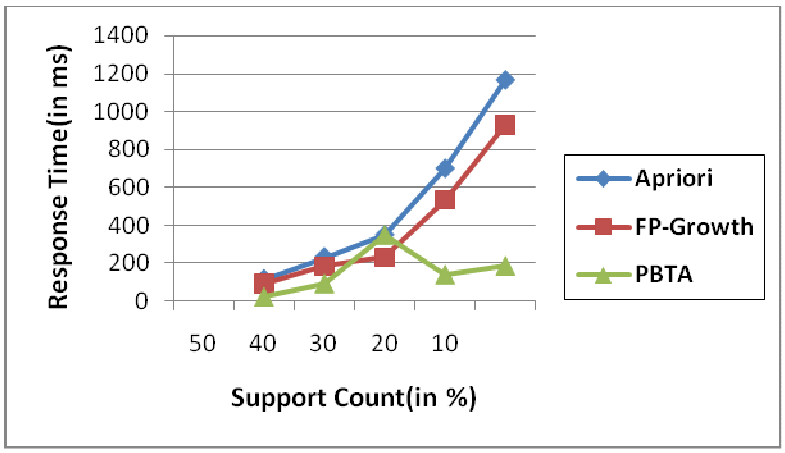


FIGURE 3: Performance analysis of algorithms with 700k database

In the Fig. 3, the transactional database with 700k is used. Here, we observed that in comparison to Apriori and FP-Growth, PBTA will take lesser response time while support threshold is reduced. This PBTA algorithm may be used for extraction of useful frequent hyperlinks or URLs for web recommendation [20].

5. CONCLUSIONS

ARM algorithms are important to discover frequent itemsets and patterns from large databases. In this paper, we have designed a Performance Based Transposition Algorithm (PBTA) for generation of frequent itemsets similar to Apriori algorithm. The proposed algorithm can improve the efficiency of Apriori algorithm and it is observed to be very fast. Our algorithm is not only efficient but also very fast for finding association rules in large databases. The proposed algorithm drastically reduces the I/O overhead associated with Apriori algorithm and retrieval of support of an itemset is quicker as compared to Apriori algorithm. This algorithm may be useful for many real-life database mining scenarios where the data is stored in boolean form. At present this algorithm is implemented for only boolean dataset that can also be extend to make it applicable to all kind of data sets.

6. REFERENCES

[1] R.Agrawal, T. Imielinski, and A.Sawmi, “Mining association rules between sets of items in large databases” in proc. of the ACM SIGMOD Conference on Management of Data, pages (207-216), 1993.

[2] D.W-L.Cheung, S.D.Lee, and B.Kao, “A general incremental technique for maintaining discovered association rules” in Database Systems for advanced Applications, pages 185-194, 1997.

[3] M.H.Margahny and A.A.Mitwaly, “Fast Algorithm for Mining Association Rules” in the conference proceedings of AIML, CICC, pp(36-40) Cairo, Egypt, 19-21 December 2005.

[4] J.S.Park, M-S. Chen and P.S.YU, “An effective hash based algorithm for mining association rules” in M.J. Carey and D.A. Schneider, editors, Proceedings of the 1995 ACM SIG-MOD International Conference on Management of Data, pages 175-186, San Jose, California, 22-25. 1995.

[5] S.Brin, R.Motwani, J.D.Vllman, and S.Tsur, “Dynamic itemset counting and implication rules for market basket data” SIGMOD Record (ACM Special Interest Group on Management of Data), 26(2): 255, 1997.

[6] H.Toivonen, “Sampling large databases for association rules” in the VLDB Journal, pages 134-145, 1996.

[7] A.Sarasere,E. Omiecinsky, and S.Navathe, “An efficient algorithm for mining association rules in large databases” in Proceedings of 21St International Conference on Very Large Databases (VLDB) , Zurich, Switzerland, Also Catch Technical Report No. GIT-CC-95-04 1995.

[8] Y.F.Jiawei Han., ”Discovery of multiple-level association rules from large databases” In the Proceedings of AIML 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt the 21St International Conference on Very Large Databases (VLDB), Zurich, Switzerland, 1995.

[9] Y.Fu., “Discovery of multiple-level rules from large databases”, 1996.

[10] D.W-L.Cheung, J.Han, V.Ng, and C.Y.Wong, “Maintenance of discovered association rules in large databases : An incremental updating techniques” In ICDE, pages 106-114,1996.

[11] D.W-L.Cheung, S.D.Lee, and B.Kao, “A general incremental technique for maintaining discovered association rules” in Database Systems for advanced Applications, pages 185-194, 1997.

[12] S.Thomas, S.Bodadola, K.Alsabti, and S.Ranka, “An efficient algorithm for incremental updation of association rules in large databases” in Proc. KDD'97, Page 263-266, 1997.

[13] N.F.Ayan, A.U. Tansel, and M.E.Arkm, ”An efficient algorithm to update large itemsets with early pruning”, in Knoweldge discovery and Data Mining, pages 287-291,1999.

[14] R.Agrawal and R.Srikant, “Mining sequential patterns”, In P.S.Yu and A.L.P. Chen, editors, Proc.11the Int. Conf. Data engineering. ICDE, pages 3-14. IEEE pages, 6-10, 1995.

[15] H.Mannila, H.Toivonen, and A.I.Verkamo, “Discovering frequent episodes in sequences” in proceedings of the First International Conference on knowledge Discovery and Data mining”, pages 210- 215. AAAI pages, 1995.

[16] Y.Huhtala, J.Karkkainen, P.Pokka, and H.Toivonen, ”TANE: An efficient algorithm for discovering functional and approximate dependencies”,. The computer Journal, 42(2): 100-111, 1999.

[17] Y.Huhtala, J.Kinen, P.Pokka, and H.Toivonen, “Efficient discovery of functional and approximate dependencies using partitions” in ICDE, pages 392- 401, 1998.

[18] F.Bodon, “A Fast Apriori Implementation”, in the Proc.1st IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI2003, Melbourne,FL).CEUR Workshop Proceedings 90, A acheme, Germany 2003.

[19] Akhilesh Tiwari, Rajendra K. Gupta, and Dev Prakash Agrawal, “Cluster Based Partition Approach for Mining Frequent Itemsets” in the International Journal of Computer Science and Network Security(IJCSNS), VOL.9 No.6,pp(191-199) June 2009.

[20] Sanjeev Kumar Sharma, Ugrasen Suman, “A Semantic Enhance Data-Mining Framework for Web Personalization”, In the proceedings of International conference on Data Analysis, Data Quality and Metadata Management(DAMD-2010), pp(49-57),Singapore,2010.

A Naïve Clustering Approach in Travel Time Prediction

Rudra Pratap Deb Nath *rudra@kde.cs.tut.ac.jp*
Department of Computer Science & Engineering
Toyohashi University of Technology
Toyohashi, Aichi, 441-8580, Japan

Nihad Karim Chowdhury *umchowdn@cs.umanitoba.ca*
Department of Computer Science
University of Manitoba
Winnipeg, R3T 2N2, Canada

Masaki Aono *aono@kde.cs.tut.ac.jp*
Department of Computer Science & Engineering
Toyohashi University of Technology
Toyohashi, Aichi, 441-8580, Japan

Abstract

Travel time prediction plays an important role in the research domain of Advanced Traveler Information Systems (ATIS). Clustering approach can be acted as one of the powerful tools to discover hidden knowledge that can easily be applied on historical traffic data to predict accurate travel time. In our proposed Naïve Clustering Approach (NCA), we partition a set of historical traffic data into several groups (also known as clusters) based on travel time, frequency of travel time and velocity for a specific road segment, day group and time group. In each cluster, data objects are similar to one another and are sufficiently different from data objects of other groups. To choose centroid of a cluster, we introduce a new method namely, Cumulative Cloning Average (CCA). For experimental evaluation, comparison is also focused to the forecasting results of other four methods namely, Rule Based method, Naïve Bayesian Classification (NBC) method, Successive Moving Average (SMA) and Chain Average (CA) by using same set of historical travel time estimates. The results depict that the travel time for the study period can be predicted by the proposed strategy with the minimum Mean Absolute Relative Errors (MARE) and Mean Absolute Errors (MAE).

Keywords: Travel Time Prediction, Advanced Traveler Information Systems (ATIS), Naïve Clustering Approach (NCA), Cumulative Cloning Average (CCA) ,Successive Moving Average (SMA), Chain Average (CA), Naïve Bayesian Classification (NBC).

1.INTRODUCTION

In the research area of Intelligent Transportation Systems (ITS), travel time prediction is a very important issue and is becoming increasingly important with the advancement of ATIS [1]. Moreover, information, provided by travel time forecasting, helps traveler to decide whether they should change their routes, travel mode, starting time or even cancel their trip [2]. Therefore, the reliable and accurate travel time prediction on road topology plays an indispensible role in any kind of dynamic route guidance systems such as trip planning, vehicular navigation systems, etc. to fulfill the users' whim. Most importantly, the importance of travel time information is also significant to find the shortest path in terms of time. On top of that, accurate travel time estimation can improve the service quality of delivery industries by delivering products on time.

Predicted travel time information provides the capacity for road users to organize travel schedule pre-trip and en-trip. It helps to save transport operation cost and reduce environmental impacts. As congestion increases on urban freeways, more and more journeys are impacted by delays. Unless a traveler routinely traverses a given route, the extent of possible delays are unknown

before departing on a journey and the uncertainty must be addressed by allocating extra time for traveling. ATISs attempt to reduce the uncertainty by providing the current state of the system and sometimes a prediction of future state. In this context, travel time is an important parameter to report to travelers. Generally, prediction of travel time depends on vehicle speed, traffic flow and occupancy that are extremely sensitive to external event like weather condition and traffic incident [3]. Addressing the uncertainty on road network is also a crucial research issue. Additionally, prediction on uncertain situation is very complex, so it is important to reach optimal accuracy. Yet, the structure of the traffic flow of a specific road network fluctuates based on daily, weekly and occasional events. For example, the traffic structure of weekend may differ from that of weekday [17]. So, time-varying feature of traffic flow is one of the major issues to estimate accurate travel time [12].

In this research, we propose a new clustering way that is able to predict travel time accurately and reliably. Here, we attempt to combine the merits of our previous methods namely NBC [12], Rule based method, SMA and CA [13] by eliminating the shortcomings of those methods. Actually, this is the update version of our most recent research [16]. With the same set of historical traffic data, comparison is also made to evaluate our proposed method. Experimental results show the superiority of our proposed method over other prediction methods namely, NBC, Rule based, SMA and CA.

The remaining portions of this paper are organized as follows: Section 2 introduces some related researches in this field. An outline of our proposed NCA with example is demonstrated in section 3, Section 4 presents a concise experimental evaluation. Finally, the conclusion words and guidelines of future research are discussed in section 5.

2.LITERATURE REVIEW AND MOTIVATION

Nowadays, travel time prediction has emerged as an active and intense research area. So, a healthy amount of researchers have paid their concentration on the accurate travel time prediction. Several methodologies have been developed till date to compute and predict travel time with varying degree of success. A wide-ranging literature review on the topic of travel time prediction is presented in this section.

Park et al [5], [6] proposed Artificial Neural Network (ANN) models for forecasting freeway corridor travel time rather than link travel time. One model used a Kohonen Self Organizing Feature Map (SOFM) whereas other utilized a fuzzy c-means clustering technique for traffic pattern classification. Lint et al [7], [8] proposed a state-space neural network based approach to provide robust travel time predictions in the presence of gaps in traffic data. In [14], Kitaoka et al. developed a new computational method that they called the “Three-Range Composite Prediction Method” to realize optional dynamic route guidance and arrival travel time prediction with the TOYOTA G-BOOK telematic service. Kwon et al [9] proposed linear regression method to predict travel time.

A linear predictor consisting of a linear combination of the current times and the historical means of the travel times was proposed by Rice et al [10]. They proposed a method to predict the time that would be needed to traverse a given time in the future. Wu et al [3] applied support vector regression (SVR) for travel time predictions and compared its results to other baseline travel-time prediction methods using real highway traffic data. Most recent research in this field was proposed by Erick et al [11]. They investigated a switching model consisting of two linear predictors for travel time prediction. UI et al. [15] investigated an approach based on pattern matching which had relied on historic data patterns for estimating future travel times.

An efficient method for predicting travel time by using NBC was proposed by Lee et al [12] which had also been scalable to road networks with arbitrary travel routes. The main idea of NBC was that it would give probable velocity level for any road segment based on historical traffic data. It was shown from experiments that NBC could reduce MARE significantly rather than the other predictors. Another effective rule-based method was proposed by Chang et al [17] in where they had considered vehicle’s current road information, day time and week day information to extract

best suited decision rule. In [13], we formulated two completely new methods, namely SMA and CA that were based on moving average. In that research, we eliminated the drawbacks of conventional moving average approach such as unwanted fluctuation in data set. These methods were also scalable to large network with arbitrary travel routes. Moreover, both methods were less expensive in terms of computational time. Consequently, it was revealed that these proposed methods can reduce error significantly, compared with existing methods.

The prediction of travel time has been received an increasing attention in recent years that urges many researchers to motivate themselves in the research of travel time forecasting. Besides, travel time estimation and prediction form an integral part of any ATIS and ITS. In NBC and rule-based methods, a whole day and velocity of vehicle are divided into several groups in an effective and efficient manner. Moreover, in rule-based method, authors also concentrate on week days. But, the calculation of velocity level for a particular route enhances the complexity. Furthermore, it emphasizes only on those data that have high probability i.e. it doesn't take all data in consideration. In rule-based method, road information, day time and week day information are taken into account to carry out rule generation process. Generated rules are used to predict velocity class. As they generate some fixed rules, so it is unable to address uncertain situation. On the other hand, SMA and CA compute all data and are not based on probability theory. Although, SMA and CA provide an almost accurate travel time, those are unable to find uncertain data from the available traffic data. Clustering is one of the powerful leading data mining tools for discovering hidden knowledge that can be applied in the large historical traffic data set. To address the uncertain situation, and predict travel time more accurately, we propose NCA. In this study, our attempt to eliminate the shortcoming of NBC, rule-based, SMA and CA as well as combining their facilities. The key challenges of this research are to increase prediction accuracy and to address uncertain situation. On top of that, proposed method can also be scalable to large network with arbitrary travel routes. To clarify our method, the complete scenario of our method is presented in the next section.

3. PROPOSED NAÏVE CLUSTERING APPROACH

Cluster analysis or clustering is an assignment of separating the set of observations into subset. A cluster is therefore a collection of objects which are similar between themselves and are dissimilar to the objects belonging to other clusters. From available clustering techniques, partitioning and hierarchical clustering ways are popular and effective. In our research, we emphasize on partitioning clustering. For its simplicity and speed, K-means clustering, one of the partitioning clustering techniques is a better candidate to run on large data set. The procedure of K-means follows a simple and easy way to classify a given data set through a certain number of clusters (assume K clusters) fixed a priori. The main concept is to define K centroids, one for each cluster. The main disadvantage of K-means clustering is that it doesn't yield the same result with each run, since the resulting clusters depend on the initial random assignment. In contrast, we formulate our approach in a cunning way so that it eliminates all shortcomings of traditional K-means algorithm. Our algorithm can automatically determine the number of clusters without the intervention of users i.e. no fixed K clusters. Apart from it, we incorporate a technique so that centroids of different clusters maintain a sufficient difference by placing them as much as possible far away from each other. Furthermore, the initial random assignment problem is also handled. In addition, to re-estimate the centroid from a cluster, we introduce a new method, namely Cumulative Cloning Average which is described in section 3.1.

At first, an origin with start time, day and destination is provided by user. A route may consist of several road segments from origin to destination. Initially, we apply our NCA on the data set of the first road segment to calculate the end time of first road segment which in turn becomes the start time of the next road segment. Finally, applying successive repetition approximate travel time from origin to destination can be measured.

3.1. Cumulative Cloning Average (CCA)

To re-estimate a suitable centroid from available data of a cluster, a new method has been proposed. For the better understandability of the reader, CCA method with an appropriate example is presented, here.

Let $t = (t_1, t_2, ...t_n)$ be the data set where n is the number of elements in that set. The value of $\tau[i, j]$ gives the desired result for $t_i, t_{i+1}, ...t_j$ where $1 \le i \le j \le n$. Finally, the value of $\tau[1, n]$ indicates the CCA of the data set, *t*. CCA can be mathematically defined by following formula:

$$\tau[i, j] = \begin{cases} t_i & \text{if } i = j \\ \dfrac{\sum_{k=1}^{i+1} \tau[k,\ (j-i)+k-1]}{(i+1)} & \text{if } i < j \end{cases} \quad (1)$$

3.1.1. CCA with Example

A set of data with five elements is given below i.e. n=5, here. So, let's see how CCA works.

Sample Data $(t_1, t_2, t_3, t_4, t_5)$: 5, 3, 5, 4, 2

Total Sample Data (n):5

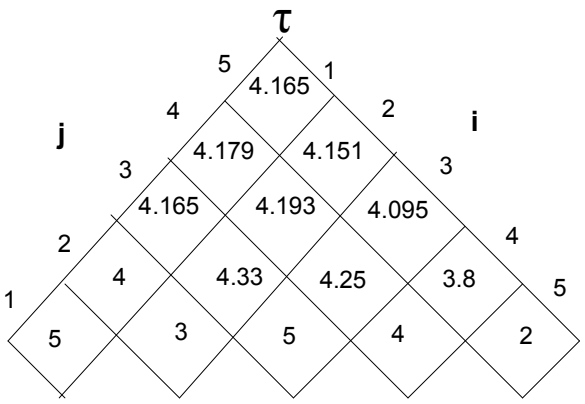


FIGURE 1: τ table for Cumulative Cloning Average

The τ table is used for storing the value of $\tau[i, j]$. Figure 1 illustrates CCA method on a sample data set where n =5. When i=j, then the value of $\tau[i, j] = t_i$. Using equation 1, we can calculate the value of $\tau[2,4]$ as $\frac{\tau[1,2] + \tau[2,3] + \tau[3,4]}{3} = 4.193$. Therefore, CCA of this data set travel would be 4 after applying round-off operation.

3.2. Definition of Time Group and Day group

The road environment of the same road network for running vehicles on the different time periods of a day is different. In NBC, the whole day time is separated into several groups according to the time. In our research, we also use their time grouping table which is illustrated in Table 1 [12].

Start_time_range	Time_group	Start_time_range	Time_group
06:01~10:00	1	16:01~18:00	6
10:01~11:00	2	18:01~22:00	7
11:01~12:00	3	22:01~00:00	8
12:01~14:00	4	00:01~06:00	9
14:01~16:00	5		

TABLE 1: Time group definition

Name of Day_ group	Symbol
Holiday	HD
Day_Before_holiday	BD
Remaining_day	RD

TABLE 2: Definition of Day group

Vehicle_ID	Road_ID	Time_group	Start_time	End_time	Travel_time (min)	Velocity (km/min)	Day_group
1	1	6	16:50	16:57	7	1.8725	RD
2	1	6	17:20	17:31	11	1.1916	RD
3	1	6	17:43	17:56	13	1.0082	RD
4	1	6	16:02	16:11	9	1.456	RD
5	1	6	16:16	16:32	16	0.8192	RD
6	1	6	16:05	16:18	13	1.0082	RD
7	1	6	17:03	17:10	7	1.8725	RD
8	1	6	17:11	17:18	7	1.8725	RD
9	1	6	17:35	17:46	11	1.1916	RD
10	1	6	16:09	16:16	7	1.8725	RD

TABLE 3: Sample historical traffic data

If a vehicle starts from any road segment between 16:01 and 18:00, its *Time_group* will be 6. The traffic flow of road network also depends on holiday, before holiday and remaining day. For our convenience, we group all national holiday and week holiday into holiday group. The previous day of holiday is also crucial in traffic structure. So, we put them in another category and the remaining days are kept in another group. For example, if it is Saturday, the group will be HD. Table 2 exhibits the day group definition. Table 3 illustrates the sample snapshot of historical traffic data for any road segment. Each record of the table contains seven attributes. The value of *Time_group* is calculated from the *Start_time. Travel_time* is the difference from *End_time* to *Start_time.* Dividing length of road segment by *Travel_time, Velocity* is measured.

To calculate approximate travel time for any road segment, we introduce NCA in the following section with appropriate example.

3.3. Procedure of Naïve Clustering Approach

When start time, day and the destination are given, our algorithm extracts the related data from the large data set according to the time_group, day_group, and road segment. Then the following step by step procedure is executed to predict the travel time of that road segment.

PROCEDURE

Step 1: Frequency for each travel time is measured by counting the repetition of that travel time in different records.
Step 2: Define Prediction relation that contains three attributes namely *Frequency*, *Travel_time* and *Velocity*. Each record of Prediction relation must contain distinct travel time.
Step 3: Find the greatest value from the *Frequency* attribute (f_{max}). A tuple $P\ (x_p\ ,\ y_p\ ,\ z_p)$ is chosen as centroid of a cluster, where x_p is the maximum frequency, y_p is the corresponding travel_time associated with x_p and z_p is the velocity associated with travel_time y_p. If two or more tuples contain the greatest value then make those tuples as the centroids, each for one cluster. Hence, we get a set of centroids, P where each centroid has maximum frequency.
Step 4: Compare each tuple T_i (x_i, y_i, z_i) of relation *Prediction* with the selected each centroid P_k (x_p, y_p, z_p) by using the following formula:

$$COST\ (P_k, T_i) = |x_p - x_i| + |y_p - y_i| + |z_p - z_i| \quad (2)$$

Where, subscript k, is the centroid number and can be ranged from 1 to n, depends on the duplication of frequency number. Choose tuple Q_k (x_q, y_q, z_q) as the centroid of another cluster, where *COST* (P_k, Q_k) is maximum. In this way, we also get another set of centroids, Q. Now, to select the final centroids, we perform intersection operation i.e. P ∩ Q. So, the number of tuples or elements in (P ∩ Q) set is the total cluster number.

Step 5: Build clusters where the centroid of each cluster is the distinct element of (P ∩ Q) set.

Step 6: Define the cluster memberships of tuples by assigning them to the nearest cluster representative tuple. The cost is given by Eq.2.

Step 7: Re-estimate the cluster centre by assuming the memberships found above are correct. To re-estimate we use our CCA method which has been illustrated in section 3.1.

Step 8: Step 6 and Step 7 are repeated until no change in clusters

Step 9: After complete preparation of clusters, desired predicted time is calculated separately for each cluster by using the following formula:

$$\tau_r = \frac{\sum_{i=1}^{N} f_i * t_i}{\sum_{i=1}^{N} f_i} \quad (3)$$

Where τ_r is the travel time obtained from r-th cluster, N is the total number of tuple in associated cluster, f_i is the *Frequency* of the i-th tuple, and t_i is the *Travel_time* of the i-th tuple.

Step 10: If the number of elements of (P ∩ Q) is R i.e. |P ∩ Q| =R, then the final predicted approximate travel time, T for the road segment of the specific time group and day group can be defined by following formula:

$$T = \frac{\sum_{i=1}^{R} \tau_i}{R} \quad (4)$$

3.4. Explanation of NCA method with example

Considering the sample historical traffic data of Table 3 that contains data for Road_id =1, Time_group=6 and day_group=RD. Steps of NCA procedure are explained below:

Step 1: There are 10 records in Table 2 where *Road_id* and *Time_group* and day_group are common. First step of NCA reveals to find the frequency of each distinct travel time. If we observe Table 3, then we find that the frequency of *Travel_time* 7 is four (4) because the number of repetition of *Travel_time* 7 in different records is four. Similarly, frequencies of *Travel_time* 16,9,13, and 11 are 1, 1, 2, and 2 respectively.

Step 2: *Prediction* relation is illustrated in Table 4. Each tuple in relation has three attributes namely *Frequency*, *Travel_time* and *Velocity*. The relation also reveals that it contains only those tuples that have distinct travel time.

Frequency	Travel_time(min)	Velocity (km/min)	Frequency	Travel_time(min)	Velocity (km/min)
1	16	0.8192	2	11	1.1916
1	9	1.456	4	7	1.8725
2	13	1.0082			

TABLE 4: *Prediction* relation of Table 2.

Step 3: The Frequency column of relation *Prediction* represents that the maximum value of it is 4. No more than one tuple contain the highest frequency. So, only one member in P set that is the tuple $P\ (x_p, y_p, z_p)$ = (4, 7, 1.8725).
Step 4: Table 5 calculates the cost of each tuple T_i (x_i, y_i, z_i) from the seed of *P Set* by using Eq.2

Frequency	Travel_time (min)	Velocity (km/min)	Distance from (4,7,1.8725)
1	16	0.8192	| 4-1 | + | 7-16 | + | 1.8725 - 0.8192 | = 3 + 9 + 1.0533 = **13.0533**
1	9	1.456	3 + 2 + 0.4165 = 5.4165
2	13	1.0082	2 + 6 + 0.8643 = 8.8643
2	11	1.1916	2 + 4 + 0.6809 = 6.6809
4	7	1.8725	0

TABLE 5: Comparison of each tuple with the centroid of *P set*

The maximum cost **(13.0553)** from centroid (4, 7, 1.8725) is marked as block in the Distance column of Table 5. So, the tuple $Q\ (x_q, y_q, z_q)$ = (1, 16, 0.8192) is selected as the centroid of Q *Set* As Set P has only one element, Set Q also contains only one element. Here, $|P \cap Q| = 2$.
Step 5: Two clusters are built where the centroid of *Cluster1* is the tuple $P\ (x_p, y_p, z_p)$ = (4, 7, 1.8725) and that of *Cluster2* is the tuple $Q\ (x_q, y_q, z_q)$ = (1, 16, 0.8192).
Step 6: Table 6 decides the cluster memberships of tuples by assigning them to the nearest cluster representative tuple. The numbers marked as block indicate the lowest cost comparison to other. Eq.2 is also used to find cost. 1st scenario of both clusters is shown in Table 7.

Freq-uency	Travel _time (min)	Velocity (km/min)	Distance from *Cluster1 centroid* (4,7,1.8725)	Distance from *Cluster2 centroid* (1,16,0.8192)
1	16	0.8192	3 + 9 + 1.0533 = 13.053	**0**
1	9	1.456	3 + 2 + 0.4165 = **5.4165**	0+7+0.6368=7.6368
2	13	1.0082	2 + 6 + 0.8643 = 8.8643	1+3+0.189=**4.189**
2	11	1.1916	2 + 4 + 0.6809 = 6.6809	1+5+0.3724=**6.3724**
4	7	1.8725	**0**	3+9+1.0533=13.0533

TABLE 6: Deciding cluster memberships.

	Frequency	Travel_time(min)	Velocity(km/min)
Cluster1	4	7	1.8725
	1	9	1.456
Cluster2	1	16	0.8192
	2	13	1.0082
	2	11	1.1916

TABLE 7: 1st scenario of both clusters with their members.

Step 7: Re-estimating of new centroid for each cluster. We calculate the new centroid for each cluster by using CCA (Eq. 1) method separately for frequency, travel_time and velocity.

New centroid for *Cluster1 using CCA*
$P_1\ (x_p, y_p, z_p)$ = (2.5, 8, 1.664).

New centroid for *Cluster2 using CCA*
$Q_1\ (x_q, y_q, z_q)$ = (1.55, 13.91, 0.96).

Step 8: Repetition of Step 6 with new centroids of both clusters. Blocking numbers indicate lowest cost comparing to other. Detail description illustrates in Table 8.

Frequency	Travel_time (min)	Velocity (km/min)	Distance from *Cluster1 new centroid* (2.5,8,1.664)	Distance from *Cluster2 new centroid* (1.55,13.91,0.96)
1	16	0.8192	1.5+8+0.84=10.34	0.55+2.09**+**0.14=**2.7**
1	9	1.456	1.5+1+0.208=**2.708**	0.55+4.91+0.49=5.95
2	13	1.0082	0.5+5+0.655=6.155	0.45+0.91+0.048=1**.408**
2	11	1.1916	0.5+3+0.4724=3.9724	0.45+2.91+0.23=**3.589**
4	7	1.8725	1.5+1+0.2085=**2.7085**	2.45+6.91+0.912=10.27

TABLE 8: Deciding cluster memberships with new centroids.

Re-estimating the cluster memberships from Table 8, 2nd scenario of both clusters has been represented in Table 9.

	Frequency	Travel_time(min)	Velocity(km/min)
Cluster1	4	7	1.8725
	1	9	1.456
Cluster2	1	16	0.8192
	2	13	1.0082
	2	11	1.1916

TABLE 9: 2nd scenario of both clusters with new centroids.

After repetition of step 7 we get that the most recent centroids of *Cluster1* P_{new1} (x_p, y_p, z_p) *and Cluster2* Q_{new2} (x_q, y_q, z_q) *are* (2.5, 8, 1.664) and (1.55, 13.9, 0.96) respectively. The most recent centroids of both clusters are similar to the 2nd most recent centroids. So, the need of repetition of step 6 and step 7 again and again are unnecessary. Table 8 shows the final clusters.

Step 9: By using Eq. 3, desired travel time from *Cluster1* and *Cluster2 can* be measured

Expected Travel Time from *Cluster1*

Here, N=2

So, τ_1 = (4*7+1*9)/ (4+1)
= (28 +9) / 5
= 37/5
= 7.4

Expected Travel Time from *Cluster2*

Here, N=3

So, τ_2 = (1*16+2*13+2*11) / (1+2+2)
= (16+26+22) / 5
= 64/5
= 12.8

So, expected travel time from *Cluster1,* τ1 =7 min (applying round operation) and expected travel time from *Cluster2,* τ2 =13 min (applying round operation)

Step 10: The final approximate travel time, T (for *Road_id=1*, Day_group=RD *Time_ group= 6)* is predicted by using Eq. 4 such as the simple arithmetic mean of τ_1 and τ_2. So, the final approximate travel time is T= ((7+13)/2) min =10 min.

4. PERFORMANCE ANALYSIS

4.1. Data Set Description

To measure the performance of different predictors, a real data set is used in our research. The data set generator is based on real traffic situation in Pusan city, South Korea. GPS sensor is used to collect real traffic delay for building this well–organized PNU generator. Traffic pattern of Pusan city was extracted from this data. According to this traffic pattern, generator simulates and generates trajectory data which almost same as real data. User interface of PNU (Punsan National University) is shown in the following figure 2.

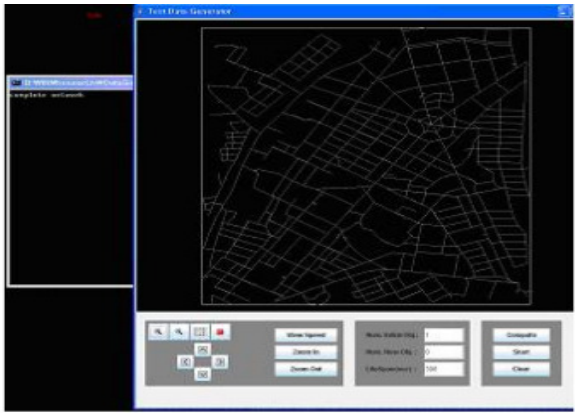


FIGURE 2: User interface of PNU trajectory data generator

By using this generator, 167,669 trajectories are generated. Every trajectory may compose of several road segments. The period of real traffic data covers both week days and weekends, and both peak hours and non-peak hours. This data organization format sufficiently reflects real traffic situations. For computing easily and efficiently and accurate evaluation of performance of the algorithms, data is divided into two categories, namely training data and test data sets. 365 days traffic data are used as training data set and 30 days traffic data are used as testing data set. Data from 365 training days are used for fitting the model. However, 30 days test data are used to measure prediction performance for all methods.

4.2. Comparison of Prediction Accuracy

The prediction error indices, Mean Absolute Relative Error (MARE) and Mean Absolute Error (MAE) are used to compare the accuracy among all prediction methods. MARE is the simplest & well-known method for measuring overall error in travel time prediction. MARE measures the magnitude of the relative error over the desired time range. The MARE is measured by the following formula:

$$MARE = \frac{1}{N}\sum_{t}\frac{|\, x(t) - x^{*}(t)\,|}{x(t)} \tag{5}$$

where $x(t)$ is the observation value; $x^{*}(t)$ is the predicted value and *N* is the number of samples.

On the other hand, the Mean Absolute Error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The MAE is a common measure of forecast error in time series analysis. This error measurement is defined as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\, x(t) - x*(t)\,| = \frac{1}{n}\sum_{t=1}^{n}|\, e(t)\,| \tag{6}$$

As the name suggests, the mean absolute error is an average of the absolute errors
e(t) = x(t) – x*(t), where x(t) is the prediction and x*(t) is the true value. In equation (6), n is the number of samples.In experimental evaluation, proposed methods are tested against other predictors like NBC, Rule-based, SMA and CA. In this section, mean relative absolute error (MRAE) and mean absolute error (MAE) among all travel time predictors are investigated. Prediction errors of all predictors from 8 AM to 6 PM are examined. There are 11 test cases are evaluated between 8 AM to 6 PM. The line chart shown in figure 3 illustrates relative performance of all travel time predictors according to MARE. From the overall point of view, proposed method performs much better than NBC, SMA, CA and Rule based methods. In case of NCA method, it is shown that eight test cases exhibit errors less than 0.40.

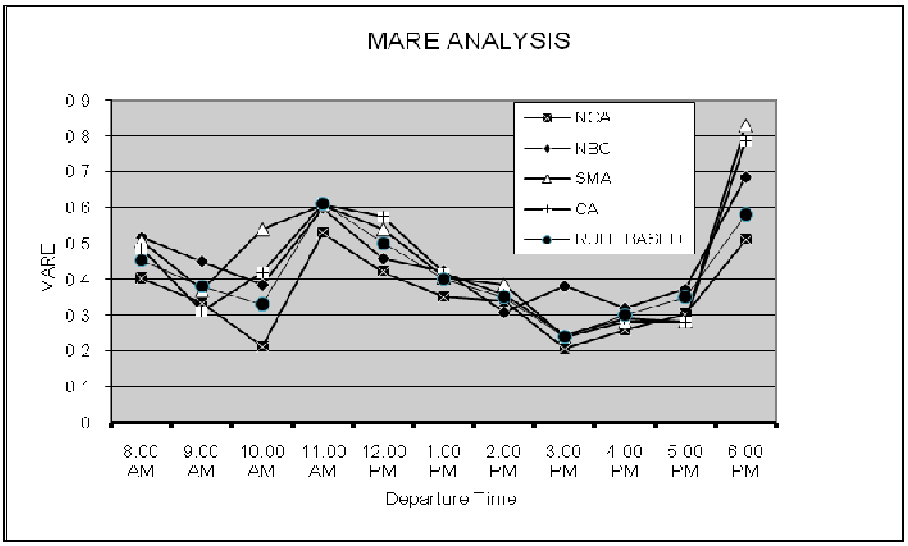


FIGURE 3: MARE of each method during different time interval.

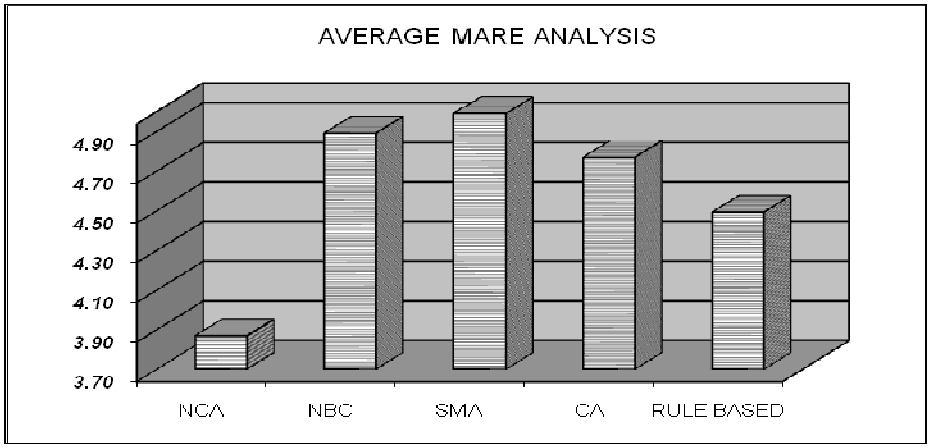


FIGURE 4: Summarized MARE of each prediction method.

Summarized MARE for different methods are shown in figure 4. Summarized MARE of NCA, NBC, SMA, CA and Rule based methods are 3.8692, 4.891, 4.9902, 4.769 and 4.493 respectively. Hence, our method reduces MARE from NBC, SMA, CA and Rule based methods by 20.89 %, 22.4%, 19%, and 14% respectively.

MAE of different methods during different time interval are shown in figure 5. In major cases, our method outperforms other methods in most of the cases. Figure 6 displays that the summarized MAE of NCA, NBC, SMA, CA and Rule based methods are 2.9601, 3.0727, 3.1648, 3.2173 and 3.24 respectively and our method reduces MAE from NBC, SMA, CA and Rule based method by 3.66%, 6.4%, 8% and 8.63% respectively.

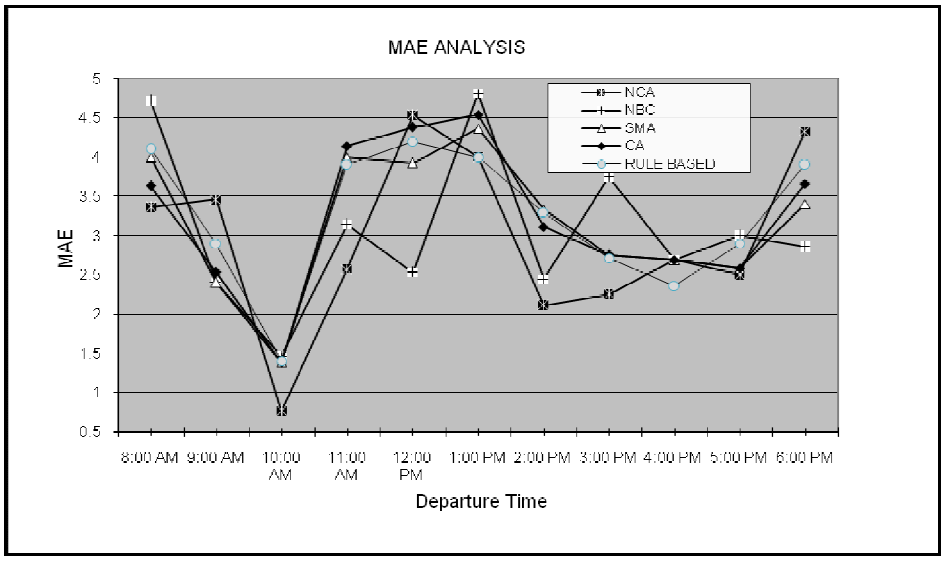


FIGURE 5: MAE of each method during different time interval.

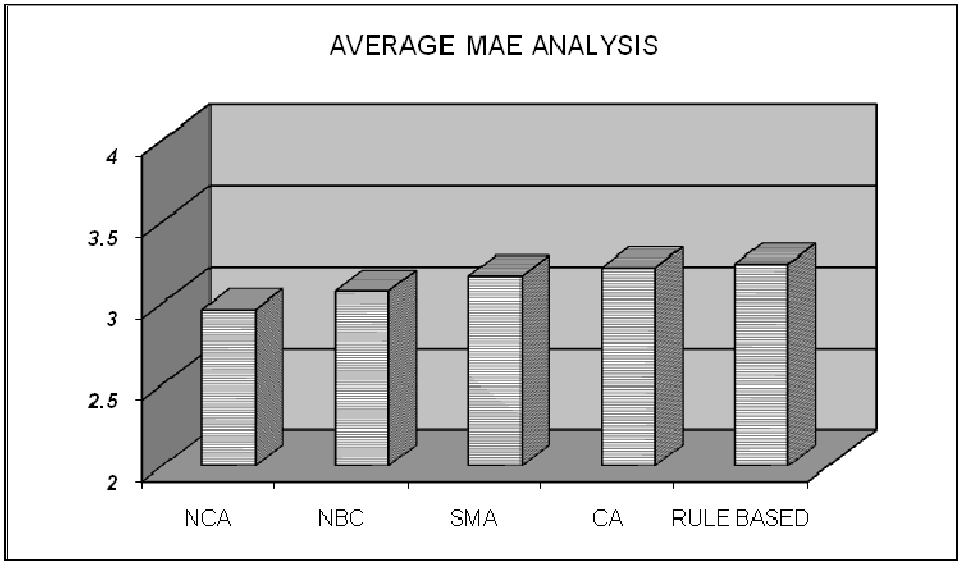


FIGURE 6: Summarized MAE of each method .

5. CONCLUSION

In this research, we focus an effective and efficient method to predict travel time more accurately. From the performance analysis portion, we can easily conclude that our method significantly reduces errors comparing with other methods. We also formulate our method in a cunning way so that we can eliminate so called partitioning problems. The centroids of the clusters are placed in a cunning manner so that they maintain as much as possible far way from each other. The superiority of our method is that the more the historical data set increases, the more the predictor is able to predict accurately. In our future plan, we will extend our NCA approach considering not only time and day but also seasonal event. The relationship between the length of roadways and accuracy of the prediction will also be tried to focus. Most importantly, analysis of our NCA will be extended with respect to real field data.

ACKNOWLEDGEMENTS
We would like to thank Prof. Jae-Woo Chang and Prof. Ki-Joune for providing us the PNU (Pusan National University) trajectory data generator.

6.REFERENCES

[1] M. Chen and S. Chien. "Dynamic freeway travel time prediction using probe vehicle data: Link-based vs. Path-based". J. of Transportation Research Record, TRB Paper No. 01-2887, Washington, D.C. 2001

[2] C. H. Wei and Y. Lee. "Development of Freeway Travel Time Forecasting Models by Integrating Different Sources of Traffic Data". IEEE Transactions on Vehicular Technology. Vol. 56, 2007

[3] W. Chun-Hsin, W. Chia-Chen, S. Da-Chun, C, Ming-Hua and H. Jan-Ming. "Travel Time Prediction with Support Vector Regression". IEEE Intelligent Transportation Systems Conference, 2003

[4] J. Kwon and K. Petty. "A travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes". Transportation Research Board 84th Annual Meeting, Washington, D.C. 2005

[5] D. Park and L. Rilett. "Forecasting multiple-period freeway link travel times using modular neural networks". J. of Transportation Research Record, vol. 1617, pp.163-170. 1998

[6] D. Park and L. Rilett. "Spectral basis neural networks for real-time travel time forecasting". J. of Transport Engineering, vol. 125(6), pp.515-523, (1999)

[7] J. W. C. V. Lint, S. P. Hoogenoorn and H. J. V. Zuylen. "Towards a Robust Framework for Freeway Travel Time Prediction: Experiments with Simple Imputation and State-Space Neural Networks". Presented at 82 Annual Meeting of the Transportation Research Board, Washington ,D.C., 2003

[8] J. W. C. V. Lint, S. P. Hoogenoorn and H. J. V. Zuylen. "Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks". In Transportation Research Record: Journal of the Transportation Research Board, No. 1811, TRB, National Research Council, Washington, D.C., pp. 30-39. 2002

[9] J. Kwon, B. Coifman and P. J. Bickel. "Day-to-day travel time trends and travel time predicti-on from loop detector data". J. of Transportation Research Record, No. 1717, TRB, National Research Council, Washington, D.C., pp. 120-129. 2000

[10] J. Rice and E. Van Zwet. "A simple and effective method for predicting travel times on freeways". In: IEEE Trans. Intelligent Transport Systems, vol. 5, no. 3, pp. 200-207, 2004

[11] J. Schmitt Erick and H. Jula. "On the Limitations of Linear Models in Predicting Travel Times". In: IEEE Intelligent Transportation Systems Conference, 2007

[12] H. Lee, N. K. Chowdhury and J. Chang. "A New Travel Time Prediction Method for Intellige-nt Transportation System". In: International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, LNAI 5177, pp: 473-483, 2008

[13] N. K. Chowdhury, R. P. D. Nath, H. Lee and J. Chang. "Development of an Effective Travel Time Prediction Method using Modified Moving Average Approach". 13th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Part 1. LNAI 5711, pp: 130-138 2009

[14] H. Kitaoka, T. Shiga, H. Mori, E. Teramoto and T. Inoguchi. "Development of a Travel Time Prediction Method for the TOYOTA G-BOOK Telematics service". R & D Review of TOYOTA CRDL vol. 41 no.4 ,2006

[15] S. Ul, I. Bajwa and M. Kuwahara, "A Travel Time Prediction Method Based on Pattern Matching Technique". In proceedings of the 21st ARRB and 11th REAAA Conference. Transport. Vermont South, Victoria 3133, ZZ N/A Australia.2003.

[16] R. P. D. Nath, H. Lee, N. K. Chowdhury and J. Chang. "Modified K-means Clustering for Travel Time Prediction Based on Historical Traffic Data". 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Part 1. LNAI 6276, pp: 511-521, 2010.

[17] J. Chang, N. K. Chowdhury and H. Lee. "New travel time prediction algorithms for intelligent transportation systems". Journal of intelligent and fuzzy systems, vol.21, pp: 5-7, 2010.

Mining of Prevalent Ailments in a Health Database Using Fp-Growth Algorithm

Onashoga, S. A. *bookyy2k@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Sodiya, A. S. *sinaronke@yahoo.co.uk*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Akinwale, A. T. *aatakinwale@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Falola, O. E. *wunmi3005@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Abstract

Health databases are characterised by large number of attributes such as personal biological and diagnosis information, health history, prescription, billing information and so on. The increasing need for providing enhanced medical system has necessitated the need for adopting an efficient data mining technique for extracting hidden and useful information from health database. In the past, many data mining algorithms such as Apriori, Eclat, H-Mine have been developed with deficiency in time-space trade off. In this work, an enhanced FP-growth frequent pattern mining algorithm coined FP-Ail is applied to students' health database with a view to provide information about prevalent ailments and suggestions for managing the identified ailments. FP-Ail is tested on a student's health database of a tertiary institution in Nigeria and the results obtained could be used by the management of the health centre for enhanced strategic decision making about health care. FP-Ail also provides the possibility to refine the minimum support threshold interactively, and to see the changes instantly.

Keywords: FP-Ail, Frequent Pattern, Health Database, Knowledge.

1. INTRODUCTION

A good number of domains of life such as the scientific institutions, government agencies and businesses have dedicated a good part of their resources to collecting and storing data, of which only a minute amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analysed effectively. The primary reason is that the original effect to create a data set is often focused on issues such as storage efficiency, it does not include a plan for how the data will eventually be used and analysed for decision making purposes. Thus, establishing the fact that we are drowning in data but starving for knowledge.

Data mining is defined as the process of extracting trends or patterns from data in a large database and carefully and accurately transforms them into useful and understandable information [2].

Frequent pattern mining has been a constantly addressed factor in the field of data mining as a result of its great and promising applicability at mining association [2], causality [9], sequential patterns [3], just to mention a few.

1.1 Frequent Pattern Mining

Let I be a set of items. A set $X = \{i_1, \ldots, i_k\} \subseteq I$ is called an itemset, or a k-itemset if it contains k items. A transaction over I is a couple T = (tid , I) where tid is the transaction identifier and I is an itemset. A transaction T = (tid , I) is said to support an itemset $X \subseteq I$, if $X \subseteq I$. A transaction database D over I is a set of transactions over I.

A number of algorithms have been proposed for mining frequent patterns in a large database, these include apriori algorithm, pattern growth methods, such as FP-growth [6] and tree projection [7] e.t.c. In a transaction database, a frequent set would be a set of items that co-occur frequently in the database. A pattern growth algorithm, FP-growth, reported to be an order of magnitude faster than apriori algorithm was proposed [6]. In this work, an enhanced FP-growth algorithm is used to mine students' health database by compressing the whole database in a compact manner

This paper is organised as follows: Section 2 has the related works, while section 3 discusses the modified algorithm. Section 4 has the implementation details with the results and section 5 concludes the work with area of further research.

2. RELATED WORKS

[4] describes a scalable, distributed software architecture that is suitable for managing continuous activity data streams generated from body sensor network. The system, when applied to healthcare, helps in taking care of patients' well-being through continuous and intelligent monitoring. The objective was achieved through observation of frequent patterns of the inherent structures of the concerned patients.

CLOTELE, a pattern growth algorithm for mining closed frequent calling patterns of a telecommunication database from a telecommunication provider was proposed in [5]. The knowledge obtained is useful for telecommunication network operators in order to make crucial decisions.

E-CAST used a dynamic threshold and indicated that the cleaning step of the original CAST algorithm may be unnecessary, in which the threshold value was computed at the beginning of each new cluster was introduced [1] . The knowledge gained could be used for the analysis of gene expression data.

[8] presented a systematic approach for expressing and optimizing frequent itemsets queries that involve complex conditions across multiple datasets. This work provided an important step towards building an integrated, powerful and efficient KDDMS which provides support for complex queries on multiple datasets in a KDDMS.(Knowledge Discovery and Data Mining System).

The benefits of performing episodic mining of health data which is a method of compressing transactional set health care episodes, that are standardised medical practise are clearly highlighted in [9]. The benefits include preprocessing data to some temporal principle that is clinically meaningful. It allows for filtering irrelevant attributes that will not be included in data analyses.

3. METHODOLOGY

3.1 Procedure for Mining Frequent Ailments Pattern

Records of different forms such as Patients' billing information, Staff Routine, Patients' health information are kept on daily basis in the health domain which could aid strategic decision making for better health care and profit maximization, if it is well exploited. The procedure used in this research for mining frequent ailments patterns involve the following stages:

1. Data Collection: At this stage, a secondary database of students' health information is extracted from the health management system.

2. Data Cleaning: In order to extract useful frequent ailment pattern, data preprocessing and data cleaning are needed. This stage identifies the most relevant/fundamentally required attributes for mining and also deals with outliers.
3. Pattern Discovery: After the data cleaning stage, the data mining algorithm to discover frequent ailment pattern is designed (section 3.2).
4. Deduction: this involves the comprehensibility of the discovered pattern i.e the conclusion drawn from the discovered knowledge.

StudentID	Gender	Level	Diagnosis	Abode	Dept.	…
A4	M	400	Malaria	Hostel	Maths	…
E3	F	300	Typhoid	Town	ABG	…
F6	F	300	HepatitisA	Town	Stat	…
D9	M	400	HepatitisB	Hostel	Home_ Sc.	…
A4	M	400	Cough	Hostel	Maths	…
…	…	…	…	…	…	…

TABLE 1: An example of selected health database

3.2 Algorithm Design – FP-Ail

This section discusses the FP-Ail algorithm.
Algorithm: Mining frequent ailments pattern, an FP-growth based approach.
Input: Valid (User, Password) Authentication
π (D), σ_{abs} // π (D) is the projection on the database and σ_{abs} is the minimum support.
Output: F(π (D), σ_{abs}) //Frequent ailment patterns.
Methods: The algorithm is given as below:

```
        Login (Username, Password)
        If( Login () is successful )
                then Select π (D)
        Else re-Login()
end if
H= {} // frequent-1 itemsets
TDB={} // Transaction Database
F(π (D), σabs )={}
// create TDB
for all i ϵ π (D) do
    TDB={i, {j}} where  set j corresponds to i = tid
// Prune (Delete infrequent itemsets)
        for distinct j ϵ TDB do
                get Supp (j) // get count of j
            if Supp (j) >= σabs
                H={j, Supp(j)}// get frequent-1 itemsets
             end if
        end for
        for all (tid, X) ϵ π (D) with j ϵ X do
                K= sort(tid(i), Xi ) in Support descending order
        // Depth first recursion
                Compute F[k] //frequent iemsets
        F(π (D), σabs ) = {H U F[k]}
```

FIGURE 1: Fp-Ail Algorithm

3.3 An Illustrative Example for Mining Frequent Ailment patterns

Consider Table 1 for illustration. Suppose the frequent ailments are to be mined based on two attributes: students' ID and diagnosis

Tid	Diagnosis (Itemsets)
A4	m, c, t, d
E3	m, t, u, h, d, t
F6	m, h, d, c, t, m
D9	m, h, c, d

TABLE 2: Transaction Database, TDB

Step 1**-** Compute frequent patterns

The database, from which the transaction database (TDB) is extracted, is allowed to be queried by an authenticated user. This algorithm ensures strict denial to an invalid user as the health domain is always meticulous over privacy issue. However, the user has been saved the stress of the need to write SQL queries by just having to select the required attributes from a populated combo box component of the application. Table 1 is used for illustration, in order to reduce the processing time, the second attribute, diagnosis, is encoded with the first letter of each item in the transaction i.e Malaria= m, Cough= c e.t.c. The TDB generated based on the query is scanned in order to compute the support of each itemset, after which the infrequent itemsets are eliminated based on the given minimum support threshold. (Note: the support is the absolute occurrence of items)

Itemset	Support
C	3
D	4
H	3
M	5
T	4
U	1

TABLE 3: TDB showing item support counts

Note: From the above, u is not frequent and is thus eliminated. (Min. support (σ_{abs}) is set to 3)

Tid	Diagnosis (Itemsets)
A4	m, t, d, c
E3	m, t, t, d, h
F6	m, m, t, d, c, h
D9	m, d, c, h

TABLE 4: TDB sorted in support descending order

Step 2: Use FP-Ail algorithm to mine FP-Ail tree

FP-Ail tree has the information of the whole database, so the algorithm mines this tree and not the database.

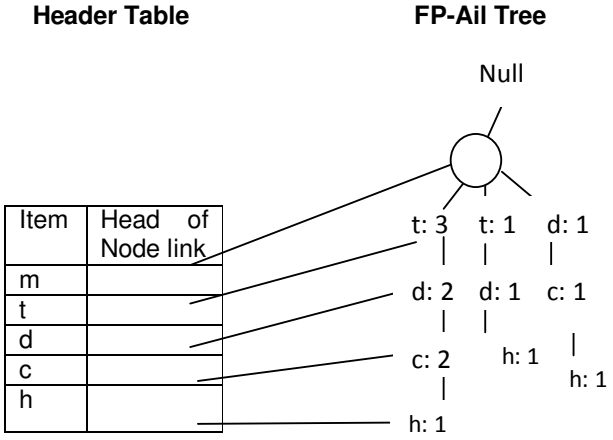
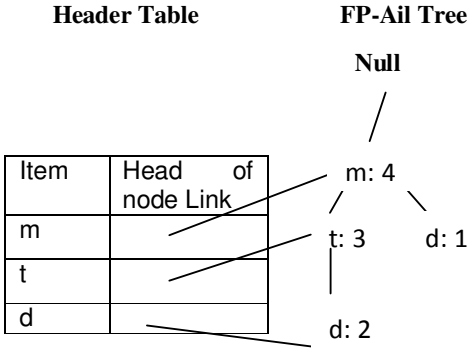


FIGURE 2: The Header Table and FP-Ail Tree

Starting from h, for each frequent-1 itemsets, construct its conditional pattern base. A conditional pattern base for an itemset contains the transactions that end with that itemset.
i). Item h's conditional pattern base is: {m:2, t:1, d:1, c:1},{m:1, t:2, d:1},{m:1, d:1, c:1}.

Note: In this conditional pattern base, c occurs only twice and is thus eliminated.

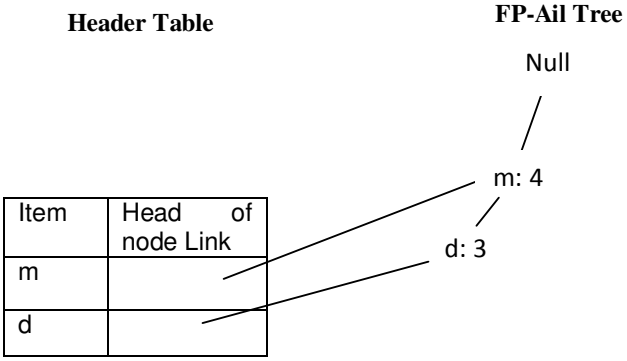
The conditional FP-Ail tree is constructed thus;



The frequent patterns generated for item h is as given below:

{mdh: 3},{dh: 3}, {mh: 3}.

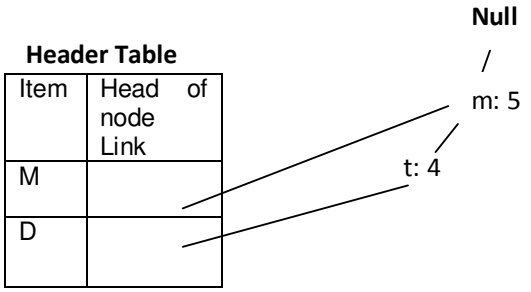
ii). Item c's conditional pattern base is: {m: 1, t: 1, d: 1}, {m: 2, t: 1, d: 1}, {m:1, d:1}.
Note: Items t occurs only twice and is thus eliminated.
The conditional FP-Ail tree is constructed thus:



The generated frequent itemsets is as follows: {mdc: 3}{dc: 3}{mc: 3}

iii). Items d's conditional pattern base is {m: 1, t: 1}, {m: 1, t: 2}, {m: 2, t: 1}, {m: 1}

The conditional FP-Ail tree is constructed thus;



The generated itemsets is as follows: {mtd: 3}, {md: 4}, {td: 3}.

iv). Items t's conditional pattern base is given as: {m: 1}, {m: 1}, {m: 2} and the generated itemsets is {mt: 3}.

Combined with the frequent-1 itemsets generated during the first database scan, we have the following frequent patterns: {mtd: 3}, {md: 4}, {td: 3}, {mt: 3}, {mdc: 3}, {dc: 3}, {mc: 3}, {mdh: 3},{dh: 3}, {mh: 3}.

4. IMPLEMENTATION AND RESULTS

The algorithm is experimented on the student health database of 4 different academic sessions, which is of size 600KB, tested on varying minimum support in order to test the flexibility and adaptability of the algorithm. All experiments were carried out on a 733MHz Pentium III PC, with a 1GB RAM size, 100GB HDD running Microsoft Windows Vista. FP-Ail was implemented in Java using NetBeans IDE 6.1 version.

The system designed is so flexible in that it allows different attributes to be selected based on the operational environment and the expected knowledge to be discovered. In particular, during this implementation, three different database is selected from the TDB for mining. These databases are T60I15D100K, T4I15D100K and T10I15D100K where T represents students' ID: 60, different levels of students: 4 and different students' abodes: 10 respectively and I is the number of diagnosis and D represents the number of records in the data base. The student ID is mined against diagnosis.

4.1 Types of Knowledge Discovered

The knowledge to be acquired from health databases can not be over-emphasized. However, in this discourse, from the sequence of diagnosis pattern, the management could discover the most effective therapy, know the most likely ailment a particular patient could have from the health record or determine the most affected group of patients with a particular ailment in order to make strategized and optimal decisions. Table 5 shows examples of patterns extracted from respective databases with their interpretations and several decision that could be taken by the authorities concerned.

	Patterns	Interpretation	Decision
T60I15D100K	{H8:cough, Tuberculosis: 8}	Student with ID H8 is noticed to have been diagnosed with the ailments in that order on 6 occasions.	Student's health should be monitored and may need to be sent home for treatment in order to avoid spread of the diseases.
T4I15D100K	{200L: diarrhoea, malaria: 20}	200 level students were been attacked severally with the identified ailments	The pharmaceutical unit should be stocked with drugs and measures should be taken to reduce or combat the attack.
T10I15D100K	{Asero: diarrhoea, malaria: 18}	Several students living at Asero (an abode outside the campus at Abeokuta, Ogun State, Nigeria) were discovered to have been frequently diagnosed with the listed ailments.	The university management could call the attention of the government in carrying out a health inspection of the area and provide necessary measures.

TABLE 5: Interpretation of the patterns extracted

4.1.1 Prevalent Ailments

We went further to mine the prevalent ailments among students by considering each session starting from 2003/2004. This window (Figure 3) displays the result of the successful query, for example:

SELECT *STUDENTID, DIAGNOSIS*
FROM *TDB*
WHERE *LEVEL = 100*

on the database. The patterns generated are displayed in the lower right end component of the windows displayed. This result would go a long way in assisting the health management make crucial decisions.

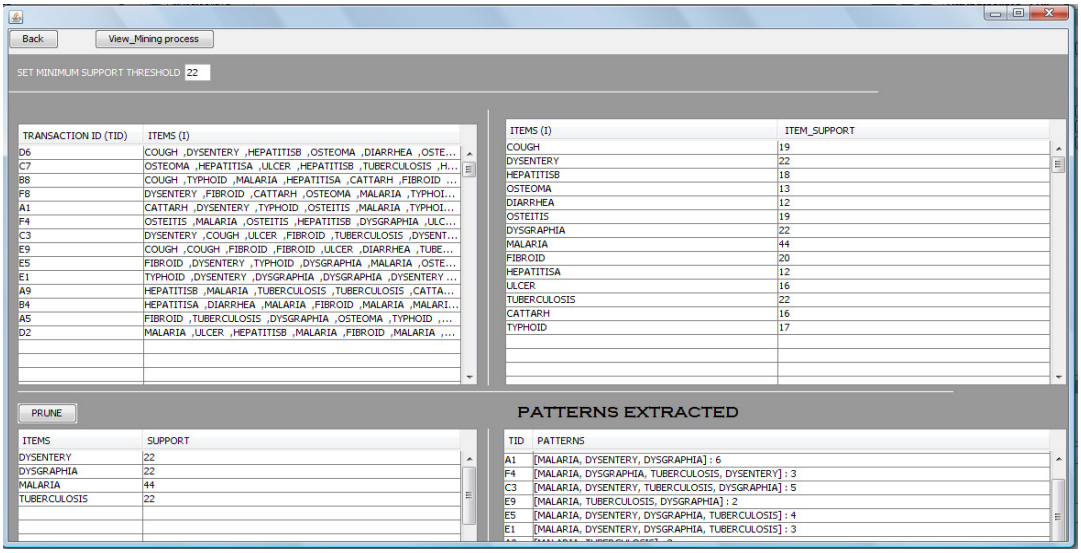


FIGURE 3: Mining View with Patterns generated.

The result as depicted in Table 6 using diagnosis against the year showed that {Malaria, Headache} was rampant among the students in 2003/2004 and 2005/2006. This clearly

shows that the common ailments as reported in the health database is this pattern. In this regards, the management should try and design a measure of reducing stress on the parts of students which could be the cause of headache and find a way of reducing the attack of malaria either by always fumigating the environment at the end of each semester or provide some mosquito repellent tools.

	SUPPORT		
YEAR	*10*	*15*	*20*
2003/2004	{Malaria, Headache, cold, catarrh, diarrhoea, dysentery, hepatitis}	{Malaria, Headache, cold, diarrhoea}	{Malaria, Headache}
2004/2005	{Malaria, diarrhoea, catarrh}	{Malaria, diarrhoea, catarrh}	{Malaria}
2005/2006	{Malaria, Headache, tuberculosis, dysentery, catarrh}	{Malaria, Headache, tuberculosis, catarrh}	{Malaria, Headache, catarrh}
2006/2007	{cold, pains}	{cold, pains}	{ pains}

TABLE 6: Patterns on prevalent ailments with support values

The choice of the minimum support value is critical in many applications, if it is too high, the FP-Ail tree is empty, if it is too low, the number of items in the FP-Ail tree is too high. Therefore FP-Ail provides the possibility to refine the minimum support threshold interactively, and to see the changes instantly. Figure 4 depicts each of these databases in comparison with the changing of support. The size of the patterns generated is inversely proportional to the minimum support.

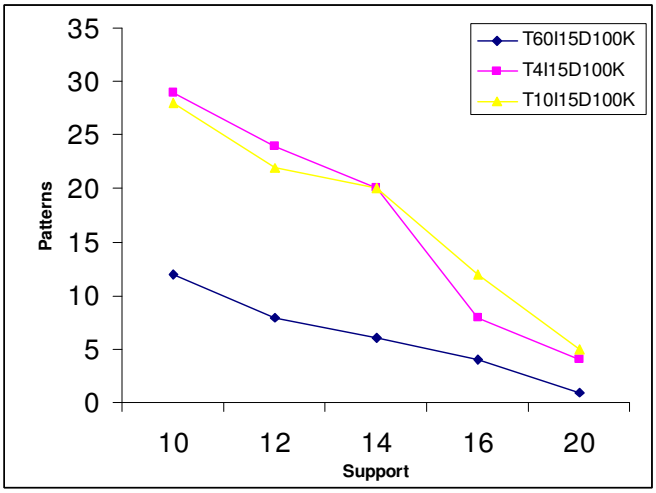


FIGURE 4: Effects of changing support threshold

5. CONCLUSION

In this paper, we have highlighted the advantages of FP-growth as a frequent pattern mining algorithm. It is thus integrated into an undergraduate students' health database coined FP-Ail algorithm for extracting hidden knowledge that could aid strategic decision making in the health unit of any organization. The tool designed is flexible and integrates security based on privacy issues of the health domain. The algorithm is tested and several knowledge is discovered.

ACKNOWLEDGEMENT

Our appreciation goes to the Director of the University Students' Health Centre for giving us assess to the database.

REFERENCES

[1] B. Abdelghani, P. David, C. Yidong, G. Abdel (2004). "E-CAST: A Data Mining Algorithm For Gene Expression Data". *BIOKDD02: Workshop on Data Mining in Bioinformatics with SIGKDD02 Conference.*

[2] R. Agrawal and R. Srikant (1994). "Fast Algorithms for mining association rules", *Proceedings 20th International Conference on Very Large Data Bases, pages 487- 499. Morgan Kaufmann.*

[3] R. Agrawal and R. Srikant (1995). "Mining sequential patterns". *ICDE'95.*

[4] R. Ali, M. ElHelw, L. Atallah, B. Lo, Y. Guang-Zhong. (2008). "Pattern mining for routine behaviour discovery in pervasive healthcare environments". *International Conference on Technology and Applications in Biomedicine, 2008. ITAB 2008.*

[5] S. A. Ibrahim, O. Folorunso, O. B. Ajayi (2005). "Knowledge Discovery of Closed Frequent Calling Patterns in a Telecommunication Database". *Proceedings of the 2005 Information Science and IT Education Joint Conference. Flagstaff, Arizona, USA. June 16-19.*

[6] J. Han, J. Pei, and Y. Yin. (2000). "Mining Frequent Patterns without Candidate Generation". *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00)*

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, & M. C. Hsu (2001). "Prefix-Span: Mining sequential patterns efficiently by prefix-projected pattern growth". *Proceedings 2001International Conference Data Engineering (ICDE'01).*

[8] J. Ruoming, A. Gagan (2004). "A Systematic Approach for Optimizing Complex Mining Tasks on Multiple Databases". *Department of Computer Science and Engineering, Ohio State University, Columbus OH 43210.*

[9] T. Semenova (2003). "Episode-Based Conceptual Mining of Large Health Collections" *Lecture notes in Computer Science, Vol. 2813/2003, Publisher: Springer Berlin / Heidelberg.*

Classification Based on Positive and Negative Association Rules

B.Ramasubbareddy *rsreddyphd@gmail.com*
Associate Professor,Dept.of CSE,
Jyothishmathi Institute of Technology& Science,
Karimnagar505001, India

A.Govardhan *govardhan_cse@yahoo.co.in*
Professor of CSE,
JNTUH College of Engineering,
Nachupally,Karimnagar,505001,India

A.Ramamohanreddy *ramamohansvu@yahoo.com*
Professor of CSE, S.V.University,
Tirupati 517502, India.

Abstract

Association analysis, classification and clustering are three different techniques in data mining. Associative classification is a classification of a new tuple using association rules. It is a combination of association rule mining and classification. In this, we can search for strong associations between frequent patterns and class labels. The main aim of this paper is to improve accuracy of a classifier. The accuracy can be achieved by producing all types of negative class association rules.

Keywords: data Mining, Association Analysis, Classification, Positive and Negative Association Rules.

1. INTRODUCTION

Data mining algorithms aim at discovering knowledge from massive data sets. Association analysis, classification and clustering are three different data mining techniques. The aim of any classification algorithm is to build a classification model given some examples of the classes we are trying to model. The model we obtain can then be used to classify new examples or simply to achieve a better understanding of the available data. Classification generally involves two phases, training and test. In the training phase the rule set is generated from the training data where each rule associates a pattern to a class. In the test phase the generated rule set is used to decide the class that a test data record belongs to. Different approaches have been proposed to build accurate classifiers, for example, naive Bayes classification, Decision trees, and SVMs. Data mining community proposed Association Rule Mining based Classification. This approach is called Associative Classification produces transparent classifier consisting of rules that are straight forward and simple to understand. Associative classification based on association rule mining searches globally for all rules that satisfy minimum support and confidence thresholds. In associative classification the classifier model is composed of a particular set of association rules, in which consequent of each rule is restricted to classification class attribute. Many improvements have been done in associative classification approach in recent studies and experiments thereof show that this approach achieves higher accuracy than traditional approaches.

The traditional associative classification algorithms basically have 3 phases: Rule Generation, Building Classifier and Classification as shown in *Fig.1.* Rule Generation employ the association rule mining technique to search for the frequent patterns containing classification rules. Building Classifier phase tries to remove the redundant rules, organize the useful ones in a reasonable order to form the classifier and the unlabeled data will be classified in the third step. Some experiments done over associative classification algorithms such as CBA [26], CMAR [23] and

MCAR [28] state that the associative classification methods share the features of being more accurate and providing more classification rules.

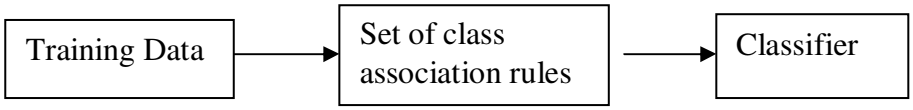


FIGURE 1: Associative Classifier

This paper is structured as follows: section II recalls preliminaries about Association Rules, In Section III, existing methods for associative classification are reviewed. The proposed algorithm is presented in Section IV and V. Section VI contains conclusions and future work.

2. BASIC CONCEPTS AND TERMINOLOGY

This section introduces association rules terminology and some related work on negative association rules and associative classification systems.

2.1 Association Rules

Let I = {i1, i2 ...in} be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with a unique identifier TID. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An association rule is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ has a support s in the transaction set D if s% of the transactions in D contains $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than given thresholds. These rules are called strong rules.

2.2 Negative Association Rules

A *negative association rule* is an implication of the form $X \rightarrow \neg Y$ (or $\neg X \rightarrow Y$ *or* $\neg X \rightarrow \neg Y$), where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \Phi$ (Note that although rule in the form of $\neg X \rightarrow \neg Y$ contains negative elements, it is equivalent to a positive association rule in the form of $Y \rightarrow X$. Therefore it is not considered as a negative association rule. In contrast to positive rules, a negative rule encapsulates relationship between the occurrences of one set of items with the absence of the other set of items. The rule $X \rightarrow \neg Y$ *has* support s*% in* the data set s, if *s %* of transactions in *T* contain itemset *X* while do not contain itemset *Y*. The support of a negative association rule, $supp(X \rightarrow \neg Y)$, is the frequency of occurrence of transactions with item set *X* in the absence of item set *Y*. Let *U* be the set of transactions that contain all items in *X*. The rule $X \rightarrow \neg Y$ holds in the given data set (database) with confidence *c*, if c% of transactions in *U* do not contain item set *Y*. Confidence of negative association rule, $conf(X \rightarrow \neg Y)$, can be calculated with $P(X \neg Y)/P(X)$, where *P(.)* is the probability function. The support and confidence of itemsets are calculated during iterations. However, it is difficult to count the support and confidence of non-existing items in transactions. To avoid counting them directly, we can compute the measures through those of positive rules.

3. RELATED WORK IN ASSOCIATIVE CLASSIFICATION

The problem of AC is to discover a subset of rules with significant supports and high confidences. This subset is then used to build an automated classifier that could be used to predict the classes of previously unseen data. It should be noted that MinSupp and MinConf terms in ARM

(Association Rule Mining) are different than those defined in AC since classes are not considered in ARM, only itemsets occurrences are used for the computation of support and confidence.

The CBA algorithm[26] was one of the first AC(Associative Classification) algorithms that employed an Apriori candidate generation step to find the rules. Classification Based on Associations (CBA) was presented by (Liu et al., 1998) and it uses Apriori candidate generation method (Agrawal and Srikant, 1994) for the rule discovery step. CBA operates in three steps, where in step 1, it discretises continuous attributes before mining starts. In step 2, all frequent ruleitems which pass the MinSupp threshold are found, finally a subset of these that have high confidence are chosen to form the classifier in step3. Due to a problem of generating many rules for the dominant classes or few and sometime no rules for the minority classes, CBA (2) has introduced by (Liu *et al*. 1999), which uses multiple support thresholds for each class based on class frequency in the training data set. Experiment results have shown that CBA (2) outperforms CBA and C4.5 in terms of accuracy.

Classification based on Multiple Association Rules (CMAR)[23] adopts the FP-growth ARM algorithm (Han et al., 2000) for discovering the rules and constructs an FP-tree to mine large databases efficiently (Li et al., 2001). It consists of two phases, rule generation and classification. It adopts a FP- growth algorithm to scan the training data to find the complete set of rules that meet certain support and confidence thresholds. The frequent attributes found in the first scan are sorted in a descending order, i.e. F-list. Then it scans the training data set again to construct an FP-tree. For each tuple in the training data set, attribute values appearing in the F-list are extracted and sorted according to their ordering in the F-list. Experimental results have shown that CMAR is faster than CBA and more accurate than CBA and C4.5. The main drawback documented in CMAR is the need of large memory resources for its training phase.

Classification based on Predictive Association Rules (CPAR)[29] is a greedy method proposed by (Yin and Han, 2003). The algorithm inherits the basic idea of FOIL in rule generation (Cohen,1995) and integrates it with the features of AC.

Multi-class Classification based on Association Rule (MCAR)[28] is the first AC algorithm that used a vertical mining layout approach (Zaki et al.,1997) for finding rules. As it uses vertical layout, the rule discovery method is achieved through simple intersections of the itemsets Tid-lists, where a Tid-list contains the item's transaction identification numbers rather than their actual values. The MCAR algorithm consists of two main phases: rules generation and a classifier builder. In the first phase, the training data set is scanned once to discover the potential rules of size one, and then MCAR intersects the potential rules Tid-lists of size one to find potential rules of size two and so forth. In the second phase, the rules created are used to build a classifier by considering their effectiveness on the training data set. Potential rules that cover a certain number of training objects will be kept in the final classifier. Experimental results have shown that MCAR achieves 2-4% higher accuracy than C4.5, and CBA.

Multi-class, Multi-label Associative Classification (MMAC) [27] algorithm consists of three steps: rules generation, recursive learning and classification. It passes over the training data set in the first step to discover and generate a complete set of rules. Training instances that are associated with the produced rules are discarded. In the second step, MMAC proceeds to discover more rules that pass MinSupp and MinConf from the remaining unclassified instances, until no further potential rules can be found. Finally, rule sets derived during each iteration are merged to form a multi-label classifier that is then evaluated against test data. The distinguishing feature of MMAC is its ability to generate rules with multiple classes from data sets where each data objects is associated with just a single class. This provides decision makers with useful knowledge discarded by other current AC algorithms.

4. FINDING CLASS ASSOCIATION RULES

Apriori-based implementations are efficient but cannot generate all valid positive and negative ARs. In this section, we try to solve that problem without paying too high a price in terms of computational costs. Generating negative class association rules of the form ¬ (= XY)=>C For simplicity, we also limit ourselves to support and confidence to determine the validity of ARs.

Algorithm:

1. **Generating negative class association rules of the form ¬ I(= XY)=>C**
2. **Generate negative class association rules of the form ¬ X →C**
3. **Generate negative class association rules of the form ¬ X ¬ *Y* →C**
4. **Generate negative class association rules of the form ¬ XY →C**

4.1. Finding Positive class Association Rules XY=>C

1. AR← φ;
2.S ← φ;
3.Find L $(P_1)_1$ i.e. Frequent 1-itemsets
4. L(P_1) ← L(P_1)
5. For k=2; L $(P_1)_{k-1}$≠Ø; k++
6.{
7.// Generating C_k
8.for each l_1,l_2 ε L$(P_1)_{k-1}$
9.If($l_1[1]=l_2[1]$^……$l_1[k-2]=l_2[k-2]$^$l_1[k-1]<l_2[k-1]$)
10. C_k=Ck ∪ {{l_1 [1]…….l_1 [k-2],$l_1[k-1]$,$l_2[k-1]$}
11. end if
12. end for
13. // Pruning using Apriori property
14. for each (k-1)- subsets s of l ε C
15. If s ∉ L $(P_1)_{k-1}$
16. $C_k=C_k$ – {l}
17. end if
18. end for
19. //Pruning using Support Count
20. Scan the database and find supp(l) for all l ε C_k
21. S=S ∪ {l with support count}
22. For each l in C_k
23. If supp (l) ≥ ms
24. L $(P_1)_k$=L $(P_1)_k$ ∪ {l}
25. end if
26. end for
27. L (P_1) =L(P_1) ∪ L$(P_1)_k$
28. }
29. end for
30. // Generating Positive Classification Rules of the form l(=XY) => c
31. for each l(=XY) ε L(P_1)
32. for each c ε C
33. If conf (l →c) ≥ mc
34. AR=AR ∪ {l → c}
35. end if
36. end for
37. end for

4.2. Generating negative class association rules of the form ¬ I(= XY)=>C

1. for each l ε L(P1)
2. if 1-supp(l) ≥ ms

3. L(P2)=L(P2) U I
4. end for
5. // Generating Negative Association Rules of the from $\neg$ (XY) => c
6. for each I ε L(P2)
7. for each c ε C
8. If conf($\neg$ I→c)≥ mc
9. AR=AR $\cup$ {$\neg$ I→c}
10. end for
11. end for

4.3. Generating negative class association Rules of the form I(=$\neg$ X$\neg$ Y) => C

1. C(P3)2 = {¬{i1 }¬{i2 }|i1 , i2 ε $L(P_1)_1$, i1 ≠ i2 }
2. for {k = 2; $C(P_3)_k \neq \emptyset$; k + +} do
3. for all I = ¬X¬Y ε $C(P_3)_k$ do
4. if supp(I) ≥ ms then
5. insert I into $L(P_3)_k$
6. else
7. for all i $\notin$ XY do
8. // Generating Candidates
9. Cand ={$\neg$ (XU{i})$\neg$ Y, $\neg$ X($\neg$ Y U{i})}
10. // Pruning Cand
11. for each item in Cand
12. If (X{i} is not in $L(P_1)$ or $\neg X^1 \neg Y^1$ is in L(P3) where $X^1 \subseteq$ X{i} and $Y^1 \subseteq$ Y)
13. Cand= Cand – {XY {i}}
14. $C(P3)_{k+1}$= $C(P3)_{k+1}$ U Cand
15. if Cand ≠ ∅, XY {i} $\notin$ S and
16. (!$\exists$ $I^1 \subseteq$ XY {i}) ($supp(I^1) = 0$) then
17. insert XY {i} into $S(P_3)_{k+1}$
18. end if
19. end for
20. end if
21. end for
22. compute support of itemsets in $S(P_3)_{k+1}$
23. S = S $\cup$ $S(P_3)_{k+1}$
24. end for
25. // Generating Negative class association Rules of the from I(= $\neg$ X$\neg$ Y) => C
26. for each I ε $L(P_3)$
27. for each c ε C
28. If conf(I→c)≥ mc
29. AR=AR $\cup$ {I→c}
30. If conf (I→$\neg$ c) ≥ mc
31. AR=AR $\cup$ {I→$\neg$ c}
32. end for
33. end for

4.4. Generating negative class Association Rules of the form $\neg$ XY=>C

1. $C(P_4)_{1,1}$ = {¬{i_1 }{i_2 }|i_1 , i_2 ε $L(P_1)_1$, $i_1 \neq i_2$ }
2. for {k = 1; $C(P_4)_{k,1} = \emptyset$; k + +} do
3. for {p = 1; $C(P_4)_{k,p} \neq \emptyset$; p + +} do
4. for all I ε $C(P_4)_{k,p}$ do
5. if supp(I) ≥ ms then
6. insert I into $L(P_4)_{k,p}$
7. end if

8. end for
9. //Generating Candidates
10. // I_1 and I_2 are joinable if $I1_1 \neq I_2$, I_1 .negative = I_2 .negative, I_1 .positive and //I_2.positive share the same k – 1 items, and I_1 .positive U I_2 .positive ε $L(P_1)_{p+1}$
11. for all joinable I_1 , I_2 ε $L(P_4)_{k,p}$ do
12. X = I_1 .negative, Y = I_1 .positive U I_2 .positive
13. I = ¬XY
14. if (! $\exists X^1 \subset X$)(supp(¬X^1 Y) ≥ ms) and (! $\exists Y^1 \subset Y$)(supp(¬XY^1) < ms) then insert I into $C(P_4)_{k,p+1}$
15. if XY $\notin$ S and ! $\exists I^1 \subset XY$, supp(I^1) = 0 then
16. insert XY into $S(P_4)_{k,p+1}$
17. end if
18. end if
19. end for
20. compute support of itemsets in $S(P_4)_{k,p+1}$
21. S = S ∪ $S(P_4)_{k,p+1}$
22. end for
23. for all X ε $L(P_1)_{k+1}$, i ε $L(P_1)_1$ do
24. if (! $\exists X^1 \subset X$)(¬X^1 {i} ε $L(P_4)$ then $C(P_4)_{k+1,1} = C(P_4)_{k+1,1} \cup$ ¬X{i
25. end if
26. end for
27. end for
28. // Generating Negative Association Rules of the from ¬ XY=> C
29. for each I ε $L(P_4)$
30. for each c ε C
31. If conf(I→c)≥ mc
32. AR=AR ∪ {I→c}
33. If conf(I→¬ c)≥ mc
34. AR=AR ∪{I->¬ c}
35. end for
36. end for

5. ASSOCIATIVE CLASSIFIER

The set of rules that were generated as discussed in the previous section represent the actual classifier. This categorizer is used to predict to which classes new objects are attached. Given a new object, the classification process searches in this set of rules for those classes that are relevant to the object presented for classification. The set of positive and negative rules discovered as explained in the previous section are ordered by confidence and support. This sorted set of rules represents the associative classifier. This subsection discusses the approach for labeling new objects based on the set of association rules that forms the classifier

Algorithm: CPNAR (Classification based on **P**ositive and **N**egative **A**ssociation **R**ules)

Input: A new object to be classified o;
The associative classifier (AC);
The confidence margin T;
Output: Category attached to the new object

Method:
1. S←φ /* set of rules that match o*/
2. for each r in AC /* the sorted set of rules */
3. If (r⊂o) {count++}
4. S←S U r
5. If(count==1)
6. fr.conf←r.conf /* keep the first rule confidence*/

7. S←S U r
8. else if(r.conf>fr.conf- τ)
9. S←S U r
10. else break
11. Divide S in subsets by category:S_1,S_2, ... S_n
12. for each subset S_1,S_2, ... S_n
13. Sum/substract the confidences of rules and divide by the number of rules in S_k
14. $Score_i$ = Σr.conf/#rules
15. Put the new object in the class that has the highest confidence score
16. o→ c_i, with $score_i$=max{ $score_1$, ……$score_n$}

In the above algorithm (Classification of a new object), a set of applicable rules is selected in the lines 1-8. The set of applicable rules is selected within a confidence margin. The interval of selected rules is between the confidence of the first rule and this confidence minus the confidence margin as checked in line 7. The prediction process is starting at line 10. The applicable set of rules is divided according to the classes in line 10. In lines 11-12 the groups are ordered according to the average confidence per class. In line 13 the classification is made by assigning to the new object the class that has the highest score.

6. EXPERIMENTAL RESULTS

The implementation of our algorithm is a java program. The experiments have been performed using datasets downloaded from UCI machine learning repository. To run the experiments, we have used ten-fold cross validation test to compute the accuracy of the classifier. To discretize the continuous attributes, we have adopted the technique used in CBA. All the experiments are performed on a 600 MHz Pentium PC with 128MB main memory running Microsoft XP. From the table 2, CPNAR algorithm has performed well for Heart, Iris and Zoo datasets when compared to C4.5,CBA,CMAR and CPAR.

DATASET	#ATTS	#CLS	#REC	#Rules Generated
BREST	10	2	699	478
HEART	13	2	270	209
HEPATITIS	19	2	155	87
IRIS	4	3	150	123
ZOO	16	7	101	68

TABLE 1: No. of CARs generated by our algorithm on various UCI ML datasets

DATASET	C4.5	CBA	CMAR	CPAR	CPNAR
BREST	95.0	96.3	**96.4**	96.0	96.6
HEART	80.8	81.9	82.2	82.6	**83.0**
HEPATITIS	80.6	81.8	80.5	**82.6**	82.3
IRIS	95.3	94.7	94	94.7	**95.6**
ZOO	92.2	96.8	97.1	95.1	**97.5**

TABLE 2: Accuracies of Various Classifiers on UCI ML datasets

7. CONCLUSION AND FUTURE WORK

We proposed an algorithm that integrates classification and association rule generation. It mines both positive and negative class association rules. Our method generates positive and negative class association rules with existing support-confidence framework. We conducted experiments on UCI datasets. In future we wish to improve accuracy of our algorithm and then we conduct experiments on some more datasets and compare the performance with other related algorithms.

8. REFERENCES

[1] B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy. Adaptive approaches in mining negative association rules. In Intl. conference on ITFRWP-09, India Dec-2009

[2] B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy. Mining Positive and Negative Association Rules, IEEE ICSE 2010, Hefeai, China, August 2010.

[3] R. Agrawal and R. Srikant. *Fast algorithms for mining associationrules.* In VLDB, Chile, September 1994.

[4] J. Han, J. Pei, and Y. Yin. *Mining frequent patterns* without candidate generation. In SIGMOD, dallas, Texas, 2000.

[5] C. Blakeand C. Merz. UCI repository of machine learning databases.

[6] S.Brin, R. Motwani, and C.Silverstein. *Beyond market* baskets: Generalizing association rules to correlations. In ACM SIGMOD, Tucson, Arizona, 1997.

[7] D. Thiruvady and G. Webb. *Mining negative association rules using grd*. In PAKDD, Sydney, Australia, 2004

[8] Goethals, B., Zaki, M., eds.: *FIMI'03: Workshop on* Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. (2003) http://CEUR-WS.org/Vol-90/.

[9] Teng, W., Hsieh, M., Chen, M.: *On the mining of* substitution rules for statistically dependent items. In: Proc. of ICDM. (2002) 442–449

[10] Tan, P., Kumar, V.: Interestingness measures for association patterns: A perspective.In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining. (2000)

[11] Gourab Kundu, Md. Monirul Islam, Sirajum Munir, Md. Faizul Bari ACN: An Associative Classifier with *Negative Rules* 11th IEEE International Conference on Computational Science and Engineering, 2008.

[12] Brin,S., Motwani,R. and Silverstein,C., " *Beyond Market* Baskets: Generalizing Association Rules to *Correlations,*" Proc. ACM SIGMOD Conf., pp.265-276, May 1997.

[13] Chris Cornelis, peng Yan, Xing Zhang, Guoqing Chen: Mining Positive and Negative Association Rules from Large Databases , IEEE conference 2006.

[14] M.L. Antonie and O.R. Za¨ıane, "*Mining Positive and* Negative Association Rules: an Approach for Confined *Rules"*, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2004, pp 27–38.

[15] Savasere, A., Omiecinski,E., Navathe, S.: *Mining for* Strong negative associations in a large data base of *customer transactions.* In: Proc. of ICDE. (1998) 494- 502..

[16] Wu, X., Zhang, C., Zhang, S.: efficient m*ining both positive and negative association rules*. ACM Transactions on Information Systems, Vol. 22, No.3, July 2004,Pages 381-405.

[17] Wu, X., Zhang, C., Zhang, S.: *Mining both positive and* negative association rules.In: Proc. of ICML. (2002) 658–665

[18] Yuan,X., Buckles, B.,Yuan, Z.,Zhang, J.:*Mining* Negative Association Rules. In: Proc. of ISCC. (2002) 623-629.

[19] Honglei Zhu, Zhigang Xu: *An Effective Algorithm for* Mining Positive and Negative Association Rules. International Conference on Computer Science and Software Engineering 2008.

[20] Pradip Kumar Bala:A Technique for Mining Negative *Association Rules .* Proceedings of the 2nd Bangalore Annual Compute Conference (2009).

[21] Data Mining: Concepts and Techniques Jiawei Han, Micheline Kamber

[22] Quinlan, J. 1993 C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann

[23] Li, W., Han, J. & Pei, J. 2001 CMAR: Accurate and efficient classification based on multiple-class association rule. In Proceedings of the International Conference on Data Mining (ICDM'01), San Jose, CA, pp. 369–376

[24] Dong, G., Zhang, X., Wong, L. & Li, J. 1999 CAEP: Classification by aggregating emerging patterns. In Proceedings of the 2nd Imitational Conference on Discovery Science. Tokyo, Japan: Springer Verlag, pp. 30–42.

[25] Antonie, M. & Zaïane, O. 2004 An associative classifier based on positive and negative rules. In Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Paris, France: ACM Press, pp. 64–69

[26] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'98)*, pages 80–86, New York City, NY, August 1998.

[27] Thabtah, F., Cowling, P. & Peng, Y. 2004 MMAC: A new multi-class, multi-label associative classification approach. In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, pp. 217–224.

[28] Thabtah, F., Cowling, P. & Peng, Y. 2005 MCAR: Multi-class Classification based on Association Rule approach. In Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications,Cairo, Egypt, pp. 1–7.

[29] Yin, X. & Han, J. 2003 CPAR: Classification based on predictive association rule. In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, pp. 369–376. B. Liu, W. Hsu, &Y. Ma, "*Integrating classification and association rule mining*", Proceeding of KDD'98, 1998, pp. 80-86.

[30] B.Ramasubbareddy, A.Govardhan, A.Ramamohanreddy, An Approach for Mining Positive and Negative Association Rules, Second International Joint Journal Conference in Computer, Electronics and Electrical, CEE 2010

[31] B.Ramasubbareddy,A.Govardhan,A.Ramamohanreddy, Mining Indirect Association between Itemsets, proceedings of Intl conference on Advances in Information Technology and Mobile Communication-AIM-2011 published by Springer LNCS , April 21-22, 2011, Nagapur, Maharastra, India

[32] B.Ramasubbareddy, A.Govardhan, and A.Ramamohanreddy Mining Indirect Positive and Negative Association Rules, Intl Conference on Advances in Computing and Communications, July 22-24 2011, Kochi, India

INSTRUCTIONS TO CONTRIBUTORS

Data Engineering refers to the use of data engineering techniques and methodologies in the design, development and assessment of computer systems for different computing platforms and application environments. With the proliferation of the different forms of data and its rich semantics, the need for sophisticated techniques has resulted an in-depth content processing, engineering analysis, indexing, learning, mining, searching, management, and retrieval of data.

International Journal of Data Engineering (IJDE) is a peer reviewed scientific journal for sharing and exchanging research and results to problems encountered in today's data engineering societies. IJDE especially encourage submissions that make efforts (1) to expose practitioners to the most recent research results, tools, and practices in data engineering topics; (2) to raise awareness in the research community of the data engineering problems that arise in practice; (3) to promote the exchange of data & information engineering technologies and experiences among researchers and practitioners; and (4) to identify new issues and directions for future research and development in the data & information engineering fields. IJDE is a peer review journal that targets researchers and practitioners working on data engineering and data management.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 2, 2011, IJDE appears in more focused issues. Besides normal publications, IJDE intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJDE LIST OF TOPICS

The realm of International Journal of Data Engineering (IJDE) extends, but not limited, to the following:

- Approximation and Uncertainty in Databases and Pro
- Autonomic Databases
- Data Engineering
- Data Engineering Algorithms
- Data Engineering for Ubiquitous Mobile Distributed
- Data Engineering Models
- Data Integration
- Data Mining and Knowledge Discovery
- Data Ontologies
- Data Privacy and Security
- Data Query Optimization in Databases
- Data Streams and Sensor Networks
- Data Warehousing
- Database Tuning
- Database User Interfaces and Information Visualiza
- Knowledge Technologies
- Metadata Management and Semantic Interoperability
- OLAP and Data Grids
- Personalized Databases
- Query Processing in Databases
- Scientific Biomedical and Other Advanced Database
- Semantic Web
- Social Information Management
- Spatial Temporal

CALL FOR PAPERS

Volume: 2 - **Issue:** 4 - July 2011

i. Paper Submission: July 31, 2011 **ii. Author Notification:** September 01, 2011

iii. Issue Publication: September / October 2011

CONTACT INFORMATION

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2011
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org