ISSN 1913-8989 (Print)
ISSN 1913-8997 (Online)

COMPUTER AND INFORMATION SCIENCE

Vol. 2, No. 4
November 2009

Canadian Center of Science and Education

Editorial Board

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Contents

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Contents

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

A Weighted-Density Connected Dominating Set Data Gathering Algorithm for Wireless Sensor Networks

Larry King

Clemson University, Clemson, SC 29634, USA

E-mail: larryfking3@gmail.com

Natarajan Meghanathan (Corresponding author)

Department of Computer Science, Jackson State University

P. O. Box 18839, 1400 John R. Lynch Street, Jackson, MS 39217, USA

Tel: 01-601-979-3661 E-mail: natarajan.meghanathan@jsums.edu

This research is funded by the U.S. National Science Foundation through grant (CNS-0851646) entitled: "REU Site: Undergraduate Research Program in Wireless Ad hoc Networks and Sensor Networks."

Abstract

We propose a weighted-density connected dominating set (wDCDS) based data gathering algorithm for wireless sensor networks. The wDCDS is constructed using the weighted-density of a sensor node, which is defined as the product of the number of neighbors available for the node and the fraction of the initially supplied energy available for the node. A data gathering tree (wDCDS-DG tree) rooted at the wDCDS Leader (the node with the largest available energy) is formed by considering only the nodes in the wDCDS as the intermediate nodes of the tree. The leader node forwards the aggregated data packet to the sink. The wDCDS and wDCDS-DG tree are dynamically reconstructed for each round of data gathering. Simulation studies reveal that the wDCDS-DG tree yields a significantly larger network lifetime, lower delay and lower energy consumption per round compared to the density-only CDS and energy-only CDS based data gathering trees.

Keywords: Connected Dominating Set, Density, Energy, Data Gathering Tree

1. Introduction

A wireless sensor network is a distributed system of smart sensor nodes that collect data about the ambient environment and propagate the data back to one or more control centers called 'sinks', which access the data. A sensor node typically has limited battery charge, computing capability and memory capacity. The transmission range of a sensor node is the distance within which the signals emanating from the node can be received with appreciable signal strength. Wireless sensor networks have limited bandwidth as the sensor nodes within the transmission range of each other share the communication medium. As the sink is normally fixed and located far away from the sensor network field, direct transfer of the collected data to the sink is not a viable solution from both energy as well as bandwidth point of view. This motivates the need for data gathering algorithms that can be effectively and efficiently run at the sensor nodes to combine the data and send only the aggregated data (that is a representative of the data collected from all the sensor nodes) to the sink. Throughout this paper, the terms 'data aggregation', 'data fusion' and 'data gathering' are used interchangeably. They mean the same.

Data gathering algorithms typically proceed in rounds, wherein during each round, data from all the sensor nodes are collected and aggregated, and then forwarded to the sink. The communication structures normally used for such data aggregation are clusters (Heinzelman, Chandrakasan & Balakrishnan, 2004), grid (Luo, Ye, Cheng, Lu & Zhang, 2005), chain (Lindsey, Raghavendra & Sivalingam, 2002), connected dominating set (CDS) (Meghanathan, 2009) and trees (Lindsey, Raghavendra & Sivalingam, 2001). Meghanathan (2009) proposed an energy-based algorithm to construct the CDS (called the ECDS) of the underlying sensor network for every round of communication and also to construct a data gathering tree (ECDS-DG tree) based on the ECDS. The ECDS strategy prefers to include nodes that have a relatively high energy as intermediate nodes of the data gathering tree (i.e., used for data collection, aggregation and forwarding);

whereas, nodes with relatively lower energy are used only as leaf nodes of the tree (i.e., for data collection and forwarding only). Simulation studies (Meghanathan, 2009) revealed that the ECDS based algorithm yields larger network lifetime than that observed with the classical algorithms such as the cluster-based LEACH (Low-Energy Adaptive Clustering Hierarchy; Heinzelman, Chandrakasan & Balakrishnan, 2004) and the chain-based PEGASIS (Power-Efficient Gathering in Sensor Information Systems; Lindsey, Raghavendra & Sivalingam, 2002). Throughout this paper, we define the network lifetime as the time of first node failure due to exhaustion of battery charge.

Even though, the ECDS-based algorithm leads to a significant improvement in the network lifetime, the algorithm does not intend to minimize the number of nodes that form the CDS. We conjecture (this has been also vindicated in our simulations) that the overall delay and energy consumption could be further lowered by attempting to reduce the number of nodes that are part of the CDS. This necessitates the need to consider the number of neighbors per node also as a criterion to construct the CDS. In this paper, we explore the use of a weighted-density based approach to construct the CDS (referred to as wDCDS) of the underlying sensor network for every round of data communication and to construct a data gathering tree based on the wDCDS. The weighted-density of a sensor node is defined as the product of the number of neighbors of the node and the fraction of the initially supplied energy that is available for the node. The wDCDS algorithm prefers nodes with a relatively larger weighted-density for inclusion in the CDS. As illustrated in the simulation results, the wDCDS approach not only minimizes the delay and energy consumption, it also yields a larger network lifetime compared to the ECDS based approach.

In addition to the wDCDS approach, we also explore the use of a density-only approach for constructing the CDS (referred to as the DCDS). For constructing the DCDS, we prefer nodes with a larger number of neighbors over nodes with relatively less number of neighbors. Since the number of neighbors for a sensor node does not change with time (we consider the network is functional only until the time of first node failure), the DCDS remains the same for every round of communication. Hence, nodes with more neighbors are likely to get exhausted earlier, leading to a lower network lifetime. The DCDS algorithm is merely used as a baseline to illustrate how much the network lifetime for a data gathering algorithm could be improved by considering both the density as well as the energy of the nodes.

Once a CDS (based on the density, energy or the weighted-density) is determined for a particular round, the principle behind the construction of the data gathering tree is common to all the three data gathering algorithms. The Leader node is the node with the largest available energy and it is responsible for transmitting the aggregated data packet to the sink. A data gathering tree rooted at the Leader node is formed by considering only the nodes in the CDS as the intermediate nodes of the tree. The non-CDS nodes form the leaf nodes of the tree. The upstream node of an intermediate CDS node in the data gathering tree is the closest CDS node that is also relatively closer to the Leader node.

The rest of the paper is organized as follows: Section 2 reviews the well-known data gathering algorithms based on different communication structures. Section 3 presents the algorithm to construct the wDCDS and the wDCDS-DG tree. Section 4 presents an overview of the algorithm to construct the DCDS-DG and ECDS-DG trees. Section 5 presents a simulation study of the three data gathering trees and discusses the performance results observed. Section 6 concludes the paper.

2. Literature Review

We now briefly review the data gathering algorithms proposed based on different communication structures. The algorithms discussed are: LEACH (clusters), PEGASIS (chain), Improved PEGASIS for CDMA systems (trees), ECDS-DG (connected dominating set) and TTDD (grid).

2.1 Low Energy Adaptive Clustering Hierarchy (LEACH)

The basic idea behind the LEACH algorithm (Heinzelman, Chandrakasan & Balakrishnan, 2004) is that the whole network is divided into clusters and a cluster-head is used for each. A sensor selects the cluster-head closer to it and directly transmits the data to the cluster-head. The cluster-head aggregates all the data collected in its cluster and transmits the aggregated data to the sink. The role of the high-energy consuming cluster-head position is rotated among all the sensor nodes in the network. If P is the percentage of nodes that can be cluster heads, LEACH ensures that a sensor node is elected as cluster head exactly once within every $1/P$ rounds of data communication. The optimal number of clusters (i.e., the cluster-heads) that will reduce the overall energy consumption in a given system depends on several parameters, such as the network topology, the percentage of cluster heads and the relative costs of computation and communication.

2.2 Power-Efficient Gathering in Sensor Information Systems (PEGASIS)

PEGASIS (Lindsey, Raghavendra & Sivalingam, 2002) forms a chain of the sensor nodes and uses this chain as the basis for data aggregation. The chain is formed using a greedy approach, starting from the node farthest to the sink. The nearest node to this node is included as the next node in the chain. This procedure is continued until all the nodes are included in the chain. A node can be in the chain at only one position. For every round, the leader node is responsible for forwarding the aggregated data to the sink. Once the leader node is selected and notified by the sink node, each node

on both sides of the chain (with respect to the leader node), receives and transmits the aggregated data to the next node in the chain, until the data reaches the leader node.

2.3 Improved PEGASIS for CDMA Systems

PEGASIS incurs a huge delay, especially for Time Division Multiple Access (TDMA) systems, as data moves across the complete chain of sensor nodes, one node at a time, before transmitted to the sink. An improved version of PEGASIS (Lindsey, Raghavendra & Sivalingam, 2001) has been proposed for Code Division Multiple Access (CDMA) systems, where there can be simultaneous communication between any pair of nodes if each node is assigned unique CDMA code and each node knows the CDMA code for communication with every other node. The chain formed using the greedy distance-based heuristic is still used as the basis for data aggregation. A round of data aggregation and transmission is comprised of log N levels, where N is the number of nodes in the network. Each node transmits the data to a close neighbor in a given level of the hierarchy. Nodes that receive data at a given level are the only nodes that rise to the next level. In order to lower the delay, data is aggregated simultaneously using as many pairs as possible at each level.

2.4 Energy-aware Connected Dominating Set-based Data Gathering (ECDS-DG) Algorithm

The ECDS-DG algorithm (Meghanathan, 2009) for every round of data communication works as follows: The leader node for a round is the node with the largest available energy. The leader node becomes the root node of the data gathering tree. An energy-aware connected dominating set (ECDS) is constructed and the leader node is the first node to be included in the ECDS. Only nodes that have a relatively larger available energy are included in the ECDS. Every node in the network is either in the ECDS or is a neighbor of a node in the ECDS. Once the ECDS construction phase is completed, the ECDS-DG tree is constructed. The leaf nodes of the ECDS-DG tree are the non-ECDS nodes. The ECDS nodes form the intermediate nodes of the ECDS-DG tree. The upstream node of a leaf node is the closest ECDS node. The upstream node of an ECDS node i is another ECDS node j such that j is closer to i and also relatively closer to the leader node compared to i. The leader node could itself be the upstream node for certain intermediate ECDS nodes. An intermediate ECDS node aggregates data collected from all of its downstream leaf nodes and downstream ECDS nodes, if any exists, and forwards the aggregated data to its upstream ECDS node in the ECDS-DG tree. This procedure is repeated until the aggregated data reaches the leader node, which then aggregates the data received with its own data and forwards the aggregated data to the sink.

2.5 Two-Tier Data Dissemination (TTDD) Algorithm

The TTDD algorithm (Luo, Ye, Cheng, Lu & Zhang, 2005) lets each source sensor node of the data to proactively construct a grid structure such that the sensor nodes at the grid points (called dissemination nodes) forward the data from the source to the sink node. The sink node within a grid, issues a query for the data and the query is routed by the sensors within the grid to the dissemination node for the grid. The query is further propagated only by the dissemination nodes and the source now responds back through the reverse path of the dissemination nodes. Considerable overhead would be involved in establishing the grid structure for each source sensor node. The dissemination nodes at the grid points are bound to run out of battery power quickly. A variant of TTDD called the Energy Efficient Data Dissemination (EEDD) algorithm (Zhou, Xiang and Wang, 2006) divides the entire sensor field into virtual grids of size $R_{trans}/2\sqrt{2}$, where R_{trans} is the transmission range of a sensor node. Each grid has a grid head, most likely to be the node with the largest energy among the nodes in the grid. The grid heads are responsible for forwarding the data from the source node to the sink. The grid heads have to be frequently changed in order to maintain fairness for each sensor node. As a result, more latency will be incurred in propagating the data from a source to the sink.

3. Weighted-Density based Connected Dominating Set and Data Gathering Algorithm

A Connected Dominating Set (CDS) of a graph is the connected sub-graph of G such that for every vertex v in G, v is either in the CDS or is a neighbor of a node in the CDS. The problem of determining a Minimum Connected Dominating Set (MCDS) is critical for data aggregation in wireless sensor networks because a MCDS helps to aggregate data from all the nodes in the network such that each node is involved in only one transmission and reception and the data gets aggregated at a minimum number of intermediate nodes before being forwarded to the sink. The problem of determining a MCDS is NP-complete (Cormen, 2001).

3.1 Assumptions

(a) We assume all the sensor nodes are CDMA (Code Division Multiple Access) enabled so that we can achieve parallel communication between any pair of sensor nodes as and when desired. Such an assumption has also been made in other well-known data aggregation algorithms such as LEACH and PEGASIS.

(b) We assume that the underlying network graph that is used to obtain the CDS is a unit disk graph constructed assuming each sensor node has a fixed transmission range R. A link exists between two nodes in the unit disk

graph if and only if the physical distance between the two nodes is less than or equal to *R*. This assumption prevents the non-CDS leaf nodes of the data gathering tree from incurring higher transmission costs.

(c) We assume that a sensor node can do transmission power control, if desired. In other words, a sensor node can communicate with any other node (even to nodes outside the transmission range) in the network. Transmission power control is done only by the nodes that are part of the CDS (i.e., the intermediate nodes of the data gathering tree) and not by the non-CDS leaf nodes.

(d) For data aggregation, we assume that every upstream node uses a particular CDMA code to communicate with all its immediate downstream nodes. The upstream node broadcasts a time schedule for data transmission to all its immediate downstream nodes. A downstream node sends its data to the upstream node according to the slots provided in the time schedule. Note that such TDMA (Time Division Multiple Access) – based communication between every upstream node and its immediate downstream nodes can occur in parallel using the unique CDMA codes chosen by each of the upstream nodes.

3.2 wDCDS-DG Algorithm

The algorithm (pseudo code in Figure 1) is executed for each round of data aggregation. The weighted-density of a sensor node (equation (1)) at any time instant is the product of the number of neighbors of the sensor node and the fraction of the initial energy currently available at the node during that time instant. The first node (*Start Node* in the pseudo code of Figure 1) to be included in the wDCDS is the node with the largest weighted-density. If more than one node has the largest weighted-density, the tie is broken arbitrarily. The algorithm is executed in two stages: In the first stage, we form a CDS (called wDCDS) of the entire network by preferring nodes with relatively larger weighted-density to be part of the wDCDS. In the second stage, we form a data aggregation tree rooted at the Leader node and by involving only nodes that are part of the wDCDS. The sensor node that has the largest available energy during a round is selected as the Leader node for the round. The pseudo code of the algorithm is illustrated in Figure 1.

$$Weighted-density(u) = \#Neighbors(u) * \frac{Energy(u)}{Initial-Energy} \quad \cdots\cdots\cdots\cdots \quad \cdots\cdots\cdots\cdots \quad (1)$$

Note that the term *Energy(u)/Initial-Energy* in the above equation represents the fraction of the initial energy currently available at node *u*.

We maintain four data structures:

(i) wDCDS-List – includes all the nodes that are part of the wDCDS

(ii) Uncovered-Nodes-List – includes all the nodes that are not covered by a node in the wDCDS-List

(iii) Covered-Nodes-List – includes nodes that are either in the wDCDS-List or covered by a node in wDCDS-List

(iv) Priority-Queue – includes nodes that are in the Covered-Nodes-List and are probable candidates for addition to the wDCDS-List. This list is sorted in the decreasing order of the weighted-density of the nodes. A dequeue operation on the queue returns the node with the largest weighted-density.

The wDCDS is primarily constructed as follows: The *Start Node* is the first node to be added to the wDCDS-List. As a result of this, all the neighbors of the Start Node are said to be covered: removed from the Uncovered-Nodes-List and added to the Covered-Nodes-List and to the Priority-Queue. If both the Uncovered-Nodes-List and the Priority-Queue are not empty, we dequeue the Priority-Queue to extract a node *s* that has the largest weighted-density and is not yet in the wDCDS-List. If there is at least one neighbor node *u* of node *s* that is yet to be covered, all such nodes *u* are removed from the Uncovered-Nodes-List and added to the Covered-Nodes-List and to the Priority-Queue; node *s* is also added to the wDCDS-List. If all neighbors of node *s* are already covered, then node *s* is not added to the wDCDS-List. The above procedure is repeated until the Uncovered-Nodes-List gets empty or the Priority-Queue gets empty. If the Uncovered-Nodes-List gets empty, then all the nodes in the network are covered. If the Priority-Queue gets empty and the Uncovered-Nodes-List has at least one node, then the underlying network is considered to be disconnected.

Note that the Start Node and the Leader node are selected with different criteria. The reason is that as the Leader node has to spend relatively larger amount energy to transmit the aggregated data to the faraway sink, it is better to rotate this role among the nodes in the wDCDS. Also, note that as the Start Node for every round is randomly chosen among the nodes that have the largest weighted-density, the constituent nodes of the wDCDS are more likely to be different for different rounds. Both these strategies are aimed at increasing the network lifetime.

Once the wDCDS is formed, the wDCDS-DG tree is constructed as follows: The upstream node for every non-wDCDS leaf node *v* is the wDCDS node that is closest to *v*. Once the upstream nodes for all the leaf nodes are determined, we determine the upstream-nodes for all the wDCDS nodes, except the Leader. The upstream node for a wDCDS node *v* is the closest wDCDS node *u* that is also relatively closer to the Leader node. The downstream node list of the wDCDS

node is updated as and when the node is selected as the upstream node of a leaf node or another wDCDS node. Figure 7 illustrates the working of the wDCDS-DG algorithm through an example.

3.3 Delay of the wDCDS-DG Tree

The delay per round for data aggregation in the wDCDS-DG tree is computed in terms of the maximum number of time units incurred at the Leader node to collect and aggregate data from all of its immediate downstream nodes. As communication between an upstream node and its downstream nodes in the wDCDS-DG tree occur sequentially using TDMA, every intermediate node in the tree has to wait to collect the data from all of its immediate downstream nodes before aggregating and forwarding the data to its upstream node. The overall delay (time units) for data aggregation and transmission is one plus the delay incurred at the Leader node.

4. Density-only and Energy-only based Connected Dominating Set and Data Gathering Algorithms

The Density-based Connected Dominating Set and Data Gathering (DCDS-DG) tree algorithm uses a heuristic, called the *d-MCDS* heuristic, proposed by Meghanathan and Farago (2008), to determine the Density-based Connected Dominating Set (DCDS). The pseudo code is illustrated in Figure 2. We first include in the DCDS, the *Start Node*, which is the node with the largest number of neighbors. If there is a tie, one of the contending nodes is randomly chosen and included in the DCDS. The heuristic proceeds further by only considering the nodes that have the maximum number of uncovered neighbors. Once the DCDS is formed, the Leader node is selected to be the node with the largest energy level among the nodes in the DCDS. The DCDS-DG tree is constructed similar to the procedure adopted for a wDCDS-DG tree.

The algorithm to determine the Energy-based Connected Dominating Set (ECDS) and the ECDS-based Data Gathering (ECDS-DG) tree is similar to that of the wDCDS-DG algorithm except that the Start Node is the node with the largest energy level and the Priority-Queue is maintained in the decreasing order of the energy level of the nodes. The Leader node of a round is the node with the largest available energy among the nodes in the ECDS. For more information on ECDS-DG, the interested reader is referred to (Meghanathan, 2009).

5. Simulations

We evaluated the performance wDCDS-DG, DCDS-DG and ECDS-DG in a discrete event simulator developed in Java. The DCDS-DG and wDCDS-DG algorithms are newly implemented for this research paper and we used the ECDS-DG code recently developed by its author (Meghanathan, 2009). The network considered is a square network of dimensions 100m x 100m. The location of the sink node is varied between the following three values: (0, 0), (50, 50) and (50, 300). The transmission range per node adopted for forming the unit disk graph is varied from 15m to 50m (i.e., 15% to 50% of the side of the square network considered for our simulations). Note that the transmission range limits only the distance between two neighboring nodes in the CDS and the distance between a non-CDS node and a CDS node. However, the CDS-based data aggregation tree (for all the three algorithms studied) may involve an upstream node and downstream node that need not be neighbors in the CDS. In such cases, a sensor node conducts transmission power control (i.e., vary the transmission range) depending on the distance to the receiver node. We assume all the sensor nodes are CDMA-enabled so that there can be simultaneous communication between any pair of sensor nodes.

5.1 Energy Consumption Model

The energy consumption model is the commonly used first order radio model (Rappaport, 2002) that has also been used in several previous work (Heinzelman, Chandrakasan & Balakrishnan, 2004; Lindsey, Raghavendra & Sivalingamm, 2001; Lindsey, Raghavendra & Sivalingam, 2002),. According to this model, the energy expended by a radio to run the transmitter or receiver circuitry is E_{elec} = 50 nJ/bit and $\in_{amp}$ = 100 pJ/bit/m^2 for the transmitter amplifier. The radios are turned off when a node wants to avoid receiving unintended transmissions. The energy lost in transmitting a k-bit message over a distance d is given by: $E_{TX}(k, d) = E_{elec} * k + \in_{amp} * k * d^2$. The energy lost in receiving a k-bit message is $E_{RX}(k) = E_{elec} * k$. The energy lost for data fusion is 5 nJ/ bit/ message.

The energy lost per round is the sum of the energy lost at all the nodes for the transmission, reception and fusion of the data. The leaf nodes of the data gathering tree do not lose energy to receive or fuse the data, but lose energy to transmit data to their upstream CDS node. Every CDS node, including the Leader node, loses energy to receive, aggregate data from each of the intermediate downstream nodes and to forward the aggregated data to its upstream node in the tree. Note that the sink node can be considered as an upstream node of the Leader node.

5.2 Performance Metrics

The performance metrics considered are: (i) Network lifetime, measured as the number of rounds the network sustains before the first sensor node dies due to exhaustion of battery charge, (ii) Delay (in terms of the number of time units) per round of data aggregation and transmission to the sink, (iii) Energy lost per round and (iv) Energy * Delay per round. A lower value for the energy*delay per round for an algorithm indicates that the algorithm very well balances the tradeoff between energy and delay. The results reported in Figures 4 through 6 and 8 through 13 are obtained for 1000

trials of the appropriate algorithms for each value of the transmission range per node. For each trial, the initial energy supplied to every sensor node is 1J. The delay per round is independent of the sink locations, as it is a measure of only the number of time units needed for data aggregation. Hence, the results for delay per round illustrated in Figure 3 are the values obtained for all the sink locations.

5.3 Performance Results and Discussion

Overall, the following observations can be made: When the sink is located far away from the sensor network field, the wDCDS-DG algorithm significantly outperforms both the DCDS-DG and ECDS-DG algorithms for all values of the transmission range per node. When the sink is located either in the corner or in the center of the network field, the wDCDS-DG algorithm performs better than ECDS-DG up to moderate values of the transmission range. As the transmission range per node gets high, the wDCDS-DG algorithm starts performing poor. Nevertheless, the wDCDS-DG algorithm is better than the DCDS-DG algorithm for all values of the transmission range per node.

The DCDS and wDCDS based data gathering trees yield a significantly lower delay (refer Figure 3) compared to that of the ECDS based data gathering tree. As the transmission range per node increases, the difference between the delay incurred per round for a DCDS or wDCDS based data gathering tree and an ECDS based data gathering tree increases. The delay per round for a DCDS or wDCDS based data gathering tree is about 75% and 40% of the delay per round for an ECDS based data gathering tree at transmission range per node values of 15m and 50m respectively. There is no significant difference between the delay per round incurred by the DCDS and wDCDS based data gathering trees.

If the sink is located far away from the sensor network field, the wDCDS-DG tree yields the largest network lifetime (refer Figure 4) for each value of the transmission range per node. This can be also attributed to the relatively lower energy incurred per round of data communication. If the sink is located in the corner or the center of the network field (refer Figures 5 and 6), the wDCDS-DG tree still yields a larger network lifetime, but only up to moderate values of the transmission range. As the transmission range gets high, the network lifetime decreases as the energy consumed per round (refer Figures 8, 9 and 10) increases. Nevertheless, the wDCDS-DG tree incurs the lowest energy consumed per round for all the scenarios. Because of the lowest energy per round and lowest delay per round, the wDCDS-DG tree also incurs the lowest energy*delay per round (refer Figures 11, 12 and 13). This indicates that the weighted-density based approach very well balances the tradeoff between energy and delay compared to the density-only or the energy-only based approaches.

For the DCDS-DG tree, the energy consumed per round (refer Figures 8, 9 and 10) increases from low to a significantly high value as the transmission range per node increases. The density-based approach for determining the CDS has the smallest number of constituent nodes and the corresponding data gathering tree incurs the least height. Because of the involvement of less number of nodes in the data gathering tree, data transmissions from nodes located in the different regions of the network have to go through a relatively longer distance. Nodes that are part of the DCDS are repeatedly exhausted and die earlier, leading to a lower network lifetime. Also, there is a significant increase in the energy consumed per round because of the long-distance transmissions, especially when the transmission range per node gets high. The ECDS based approach incurs higher energy per round, though improves the network lifetime compared to the DCDS-based approach. As a result, most of the nodes in the network lose relatively more energy per round, resulting in premature node failures. The weighted-density based approach achieves the correct balance between the height of the tree and the number of immediate downstream nodes per intermediate node of the tree. Thus, the wDCDS-DG tree incurs the lowest energy*delay per round of data aggregation.

6. Conclusions

The high-level contribution of this paper is the development of a weighted-density based connected dominating set data gathering (wDCDS-DG) tree algorithm for wireless sensor networks. The weighted-density of a node for a round is defined as the product of the number of neighbors of the node and the fraction of the energy currently available at the node during the round. The weighted-density is used as the metric to form the wDCDS. Only nodes that have a relatively larger weighted-density are considered for inclusion in the wDCDS. The wDCDS-DG tree yields a longer network lifetime, lower delay and lower energy consumption per round when compared to a density-only CDS based data gathering (DCDS-DG) tree and an energy-only CDS based data gathering (ECDS-DG) tree. Since the wDCDS-DG tree performs better than the ECDS-DG tree, the wDCDS-DG algorithm can also be considered to perform better than the classical LEACH and PEGASIS algorithms, which are outperformed by ECDS-DG. (Meghanathan, 2009). Simulation results illustrate that when the sink is located far away from the sensor network field (which is more often the case), wDCDS-DG can yield significantly larger lifetime than ECDS-DG. If the sink is located either in the corner or in the center of the network field, then wDCDS-DG yields larger lifetime up to moderate values of transmission range per node and as the transmission range gets high, ECDS-DG starts yielding slightly larger lifetime.

References

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. (2nd ed.) The MIT Press.

Heinzelman, W., Chandrakasan, A., and Balakarishnan, H. (2004). *Energy-Efficient Communication Protocols for Wireless Microsensor Networks*. Paper presented at the Hawaaian International Conference on Systems Science, USA.

Lindsey, S., Raghavendra, C., and Sivalingam, K. M. (2002). Data Gathering Algorithms in Sensor Networks using Energy Metrics. *IEEE Transactions on Parallel and Distributed Systems*, *13* (9), 924-935.

Lindsey, S., Raghavendra, C., and Sivalingam, K. M. (2002). *Data Gathering in Sensor Networks using the Energy*Delay Metric*. Paper presented at the 15th International Parallel and Distributed Processing Symposium, USA.

Luo, H., Ye, F., Cheng, J., Lu, S., and Zhang, L. (2005). TTDD: Two-Tier Data Dissemination in Large-Scale Wireless Sensor Networks. *Wireless Networks*, *11* (1-2), 161-175.

Meghanathan, N. (2009). *An Algorithm to Determine Energy-aware Connected Dominating Set and Data Gathering Tree for Wireless Sensor Networks*. Paper presented at the 2009 International Conference on Wireless Networks, USA.

Meghanathan, N., and Farago, A. (2008). On the Stability of Paths, Steiner Trees and Connected Dominating Sets in Mobile Ad hoc Networks. *Elsevier Ad hoc Networks*, *6* (5), 744 – 769.

Rappaport, T. S. (2002). *Wireless Communications: Principles and Practice*. (2nd ed.). Prentice Hall.

Zhou, Z., Xiang, X., and Wang, X. (2006). *An Energy Efficient Data Dissemination Protocol in Wireless Sensor Networks.* Paper Presented at the International Symposium on a World of Wireless, Mobile and Multimedia Networks, USA.

Input: Graph $G = (V, E)$, where V is the vertex set and E is the edge set

Initial-energy // Initial energy supplied to each node

Auxiliary Variables and Functions:

wDCDS-List, *Covered-Nodes-List*, *Uncovered-Nodes-List*, *Priority-Queue*, *Start Node*, *Leader*

// Weighted-density for a node $u \in V$

weighted-density(u) = #*Neighbors*(u)* (*energy*(u) / *Initial-energy*)

$$StartNode = \left\{ u \middle| \underset{u \in V}{Max} \left[weighted - density(u) \right] \right\}$$

Output: *wDCDS-DG* Tree Information

- *Upstream-Node*(v) for every vertex $v \in V - \{Leader\}$
- *Downstream-Nodes*(u) for every vertex $u \in V$

Initialization:

wDCDS-List = {*Start Node*}; *Priority-Queue* = {*Start Node*}; *Covered-Nodes-List* = {*Start Node*}

Uncovered-Nodes-List = V – {*Start Node*}

Begin *wDCDS-DG* Construction

while (*Uncovered-Nodes-List* ≠ Φ and *Priority-Queue* ≠ Φ) **do**

node s = Dequeue(*Priority-Queue*) // Extracts the node with the largest weighted-density

alreadyCovered = true // to test whether all neighbors of node s have already been covered or not

for all node $u \in$ *Neighbors*(s) **do**

if ($u \in$ *Uncovered-Nodes-List*) **then**

alreadyCovered = false

Uncovered-Nodes-List = *Uncovered-Nodes-List* – {u}

Covered-Nodes-List = *Covered-Nodes-List* U {u}

Priority-Queue = *Priority-Queue* U {u}

end if

end for

if (*alreadyCovered* = false) **then**

wDCDS-List = *wDCDS-List* U {*s*}

end if

end while

$$Leader = \left\{ u | \underset{u \in DCDS-List}{Max} \left[energy(u) \right] \right\}$$

$$\forall v \notin wDCDSList,$$

$$Upstream-Node(v) = \left\{ \bar{u} = u | \underset{u \in wDCDS-List}{Min} \left[dis\tan ce(v,u) \right] \right\}$$

$$DownstreamNodes(\bar{u}) = DownstreamNodes(\bar{u}) \cup \{v\}$$

$$\forall v \in wDCDSList - \{Leader\},$$

$$UpstreamNode(v) = \left\{ \bar{u} = u | \left[\underset{u \in wDCDSList}{Min} \left[dis\tan ce(v,u) \right] \right] AND \left[dis\tan ce(v, Leader) > dis\tan ce(u, Leader \right] \right\}$$

$$DownstreamNodes(\bar{u}) = DownstreamNodes(\bar{u}) \cup \{v\}$$

End *wDCDS-DG* Construction

Figure 1. Pseudo Code for the *wDCDS-DG* Algorithm

Input: Graph $G = (V, E)$, where V is the vertex set and E is the edge set

Auxiliary Variables and Functions:

DCDS-List, *Covered-Nodes-List*, *Uncovered-Nodes-List*, *Priority-Queue*, *StartNode*, *Leader*

// *StartNode* is the node with the largest number of neighbors

Output: *DCDS-DG* Tree Information

- *Upstream-Node*(*v*) for every vertex $v \in V - \{Leader\}$
- *Downstream-Nodes*(*u*) for every vertex $u \in V$

Initialization:

DCDS-List = {*StartNode*}; *Priority-Queue* = {*StartNode*}; *Covered-Nodes-List* = {*StartNode*}

Uncovered-Nodes-List = *V* – {*StartNode*}

Begin *DCDS-DG* Construction

while (*Uncovered-Nodes-List* ≠ Φ and *Priority-Queue* ≠ Φ) **do**

node *s* = Dequeue(*Priority-Queue*) // Extracts the node with the largest number of neighbors that are not yet covered

DCDS-List = *DCDS-List* U {*s*}

for all node $u \in Neighbors(s)$ **do**

if ($u \in$ *Uncovered-Nodes-List*) **then**

Uncovered-Nodes-List = *Uncovered-Nodes-List* – {*u*}

Covered-Nodes-List = *Covered-Nodes-List* U {*u*}

Priority-Queue = *Priority-Queue* U {*u*}

end if

end for

end while

$$Leader = \left\{ u \middle| \underset{u \in DCDS-List}{Max} \left[energy(u) \right] \right\}$$

$\forall v \notin DCDSList$,

$$UpstreamNode(v) = \left\{ \bar{u} = u \middle| \underset{u \in wDCDS-List}{Min} \left[dis\tan ce(v,u) \right] \right\}$$

$$DownstreamNodes(\bar{u}) = DownstreamNodes(\bar{u}) \cup \{v\}$$

$\forall v \in DCDSList - \{Leader\}$,

$$UpstreamNode(v) = \left\{ \bar{u} = u \middle| \left[\underset{u \in wDCDS-List}{Min} \left[dis\tan ce(v,u) \right] \right] AND \left[dis\tan ce(v, Leader) > dis\tan ce(u, Leader) \right] \right\}$$

$$DownstreamNodes(\bar{u}) = DownstreamNodes(\bar{u}) \cup \{v\}$$

End *DCDS-DG* Construction

Figure 2. Pseudo Code for the *DCDS-DG* Algorithm

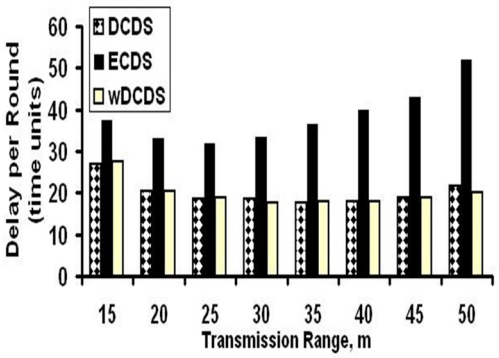


Figure 3. Average Delay per Round

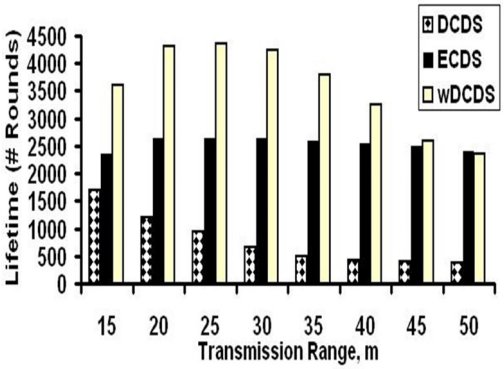


Figure 4. Lifetime per Round [Sink at (50, 300)]

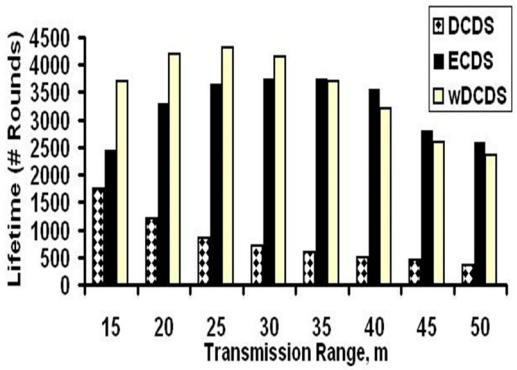


Figure 5. Lifetime per Round [Sink at (50, 50)]

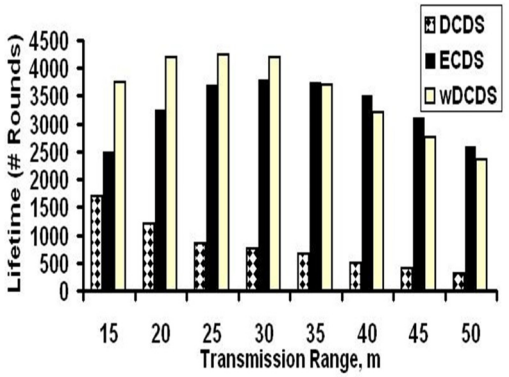
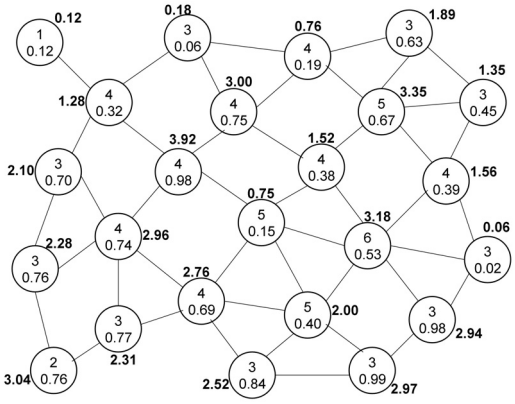
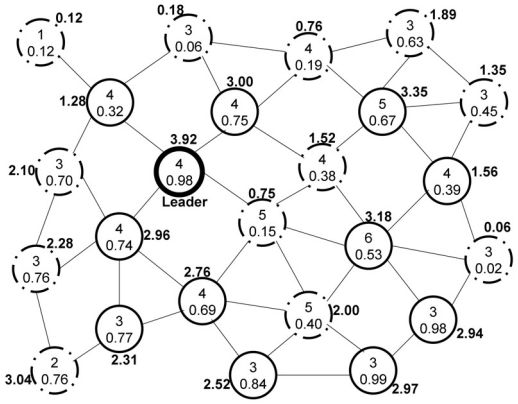


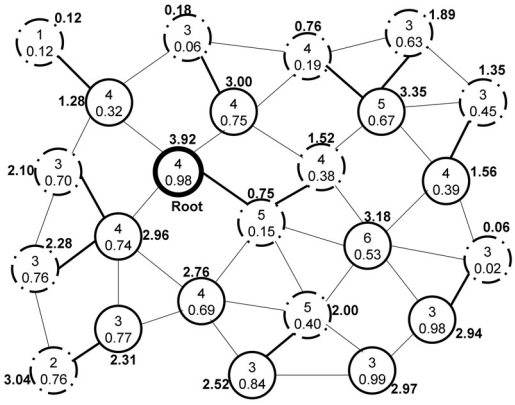
Figure 6. Lifetime per Round [Sink at (0, 0)]



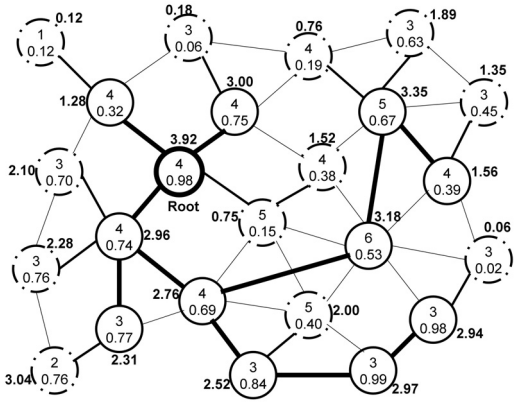
(Initial Network, Before ECDS Construction)



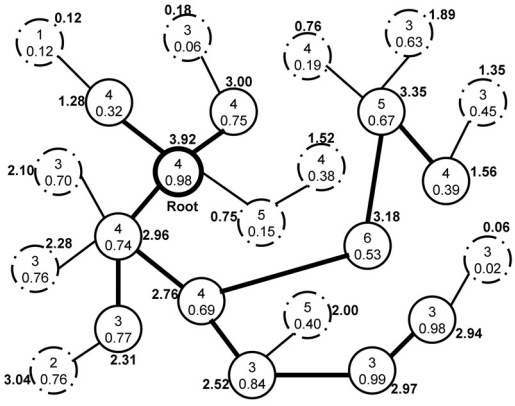
(After ECDS Construction)



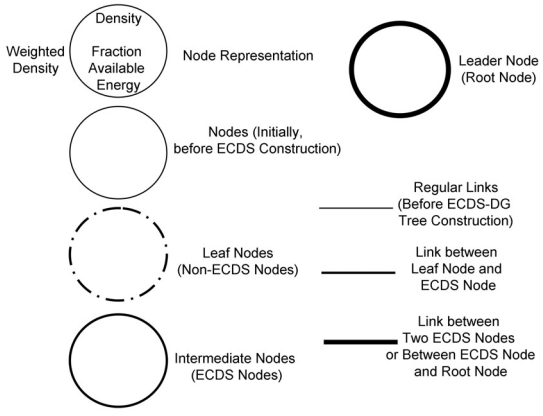
(Formation of Links between Leaf Nodes and ECDS Nodes



(Formation of Links between Two ECDS Nodes or Between an ECDS Node and the Root Node)



ECDS-DG Tree



Legend for the Figures

Figure 7. Example for the Execution of the wDCDS-DG Algorithm

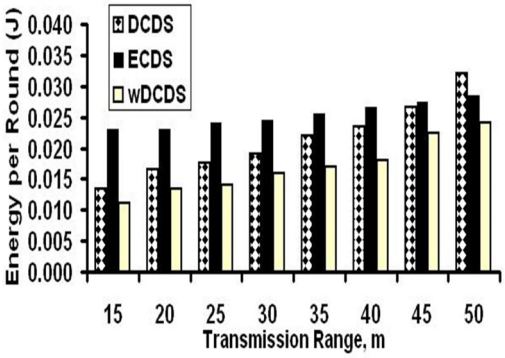


Figure 8. Energy Consumed per Round [Sink at (50, 300)]

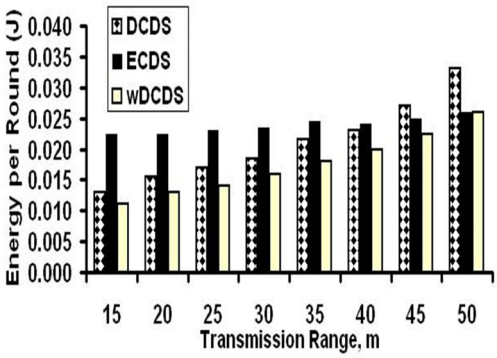


Figure 9. Energy Consumed per Round [Sink at (50, 50)]

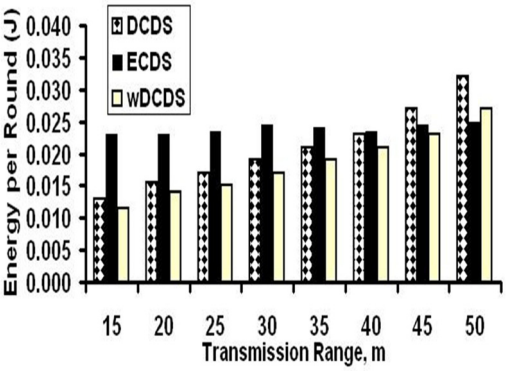


Figure 10. Energy Consumed per Round [Sink at (0, 0)]

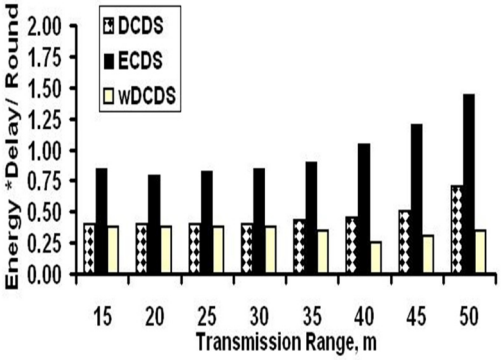


Figure 11. Energy * Delay per Round [Sink at (50, 300)]

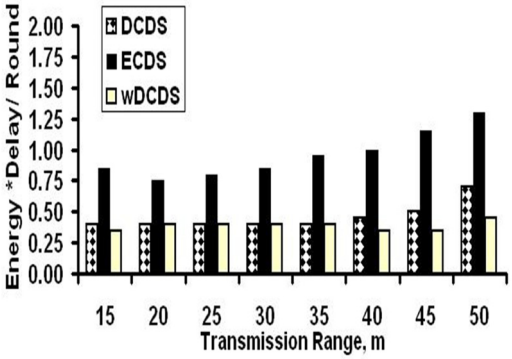


Figure 12. Energy * Delay per Round [Sink at (50, 50)]

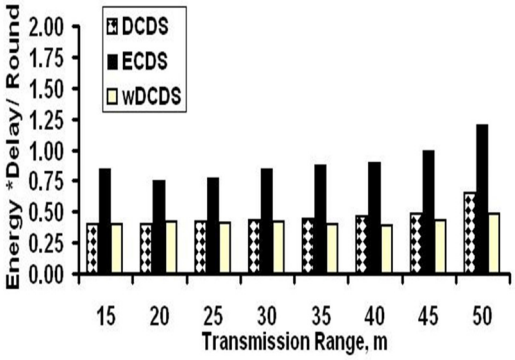


Figure 13. Energy * Delay per Round [Sink at (0, 0)]

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

An Agent-based Online Shopping System in E-commerce

Ziming Zeng
Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China
E-mail: zmzeng1977@163.com

The paper is Supported by the MOE Project of Key Research Institute of Humanities and Social Science in Chinese Universities (NO: 07JJD870220)

Abstract

The paper presents an agent-based shopping system. First, the system can acquire the customer's current needs from system-customer interactions. Then the system integrates built-in expert knowledge and the customer's current needs, and recommends optimal products based on multi-attribute decision method. In order to maintain a semantic conversation with sellers, the commodity ontology is also utilized to support sharable information format and representation. Finally, an experimental prototype based on JADE is developed.

Keywords: E-commerce, Agent, Multi-attribute decision making, Collaborative filtering, Commodity ontology

1. Introduction

The Internet and World Wide Web are becoming an important channel for retail commerce as well as for business to business transactions. It is undeniable that daily life has become convenient with online shopping. People do not drive to a store, do not travel to oversea, they can purchase the commodities and get the services they want. Forrester research, International Data Corp., and Nielsen Media Research have reported that the number of people buying, selling and performing transactions on the Web is increasing at a phenomenal pace. At present, however, the potential of the Internet for transforming commerce is largely unrealized. Electronic purchases are still largely non-automated. So the exponentially increasing information along with the rapid expansion of the business websites causes the problem of information overload. This of courses spends customers too much time on visiting flooding of retail shops on websites to know about the commodities and to survey the relevant commodity information for further comparison.

One way to solve the above problem is to develop intelligent shopping systems to provide personalized information services. The system can interact with customers and capture what they needs, so it provides decision support for them to buy on the Web. Depending on the types of commodities, different kinds of shopping systems should be developed to automate shopping process by assisting customers to have commodity information retrieval and comparison in the massive information environment of the Internet. For the type of commodities that customers buy often, such as food, clothes and books, the shopping system can be developed to acquire a customer's personal preferences by analysing his/her profile information and purchasing records(Lee, J. Lee,J., Podlaseck, M., Schonberg, E., & Hoch, R. 2001). For the commodities such as computers that a customer does not buy often, it is difficult to reason about his/her previous preferences because there is not enough information available about his/her past purchasing record. In addition, the customer may have his/her specific requirements for each single shopping and have inadequate knowledge to evaluate the commodities. In order to automate shopping process of this kind of commodities, the shopping system in this paper is presented, which can provide consultation services and decision support via iterative interaction with customers. Therefore, the system can acquire and analyse a customer's current needs or preferences, then evaluates the candidate commodities within the database to recommend the optimal commodity for him/her.

The paper is organized as follows. In section 2, the shopping process is described. In section 3, the intelligent shopping system is implemented based on the multi-attribute decision making method and consumer-based collaborative approach. Besides, the commodity ontology is established in order to maintain a semantic conversation between the system and seller websites. Finally, the conclusions are drawn in section 4.

2. The analyse of the Intelligent Shopping system

Based on the agent technology, the shopping system integrates knowledge-based decision-making method and

consumer-based collaborative filtering approach to provide decision support for automatic shopping. The shopping process is list as followers and its workflow is shown.

(1) First, the shopping system performs multiple sellers searching task. The commodities can be collected from the sellers by a search engine and stored in the internal commodity database.

(2) After the system gets all the commodities information, it asks the customer answer some qualitative questions to collect his/her needs about the commodities.

(3) After gathering the customer's qualitative needs, the system can obtain the built-in expert knowledge to calculate the optimality of each commodity using multi-attribute decision making method.

(4) Once the currently available commodities have been ranked, the commodity with the top rank will be recommended to the customer as the candidate.

(5) To speed up the shopping process, a consumer-based collaborative filtering approach is used. The approach is based on the similar customer's purchasing record to provide more candidate commodities for the current customer.

3. The implementation of the shopping system based on multi-agent

3.1 System framework

The overall goal here is to analyse a customer's current requirements and to find the most suitable commodity for him/her. To achieve the goal, the system consists of five types of agents that can interact with each other: interface agent, buyer agent, expert agent, evaluation agent and collaboration agent. These agents collaborate with each other by the message delivery mechanism and make the whole system works together. The structure of the system is shown in Figure 1. The detailed functions of each agent in the shopping system are described as follows.

1) Interface agent

The main work of the interface agent is bidirectional communication between the shopping system and customers. In order to collect and analyse the customer's current needs, the interface agent asks him/her some specially designed questions about the commodities. In the shopping system, assuming that the customer does not have enough domain knowledge to answer quantitative questions regarding the technical details about the commodity, the system has to inquire some qualitative ones instead. For example, the system will ask the customer to express his need on the display feature rather than the basic frequency of CPU.

2) Buyer agent

Buyer agent is a mobile agent, which can migrate to the electronic marketplace and search for the commodity information from multiple sellers. When it searches out one seller, it will ask for offers about the commodity from the respective seller. After the buyer agent gets all offers, it will return back and store the commodity information in the internal commodity database. In order to promote the efficiency of searching, it creates a group of child agents and dispatches each to search for the offers of the commodity from the respective seller. These child agents perform parallel searching, so buyer agent should supervise the running state of each child agent and coordinate task distribution among them.

3) Expert agent

As is indicated, an important issue in the design of the system is how to use the expertise to provide the knowledge-based decision support. The expert agent provides the communication interface with human experts, by which the experts can embed their personal knowledge into the system and give a score of a commodity in each qualitative need defined before. With the expert agent, the system can collects opinions from different experts to give more objective suggestions. Then the expert agent will convert them into a specially designed internal form for knowledge representation. However, human experts seldom reach exactly the same conclusions. They may give different scores of the same commodity in the same qualitative need since their preferences are different. In order to resolve this problem, the system synthesizes all the expert's opinions and assigns the same weights for them in the system implementation. In this way, the expert agent can transfer each commodity to a rank form and calculate its optimality accordingly.

4) Evaluation agent

The evaluation agent is an important component of the online shopping system. After receiving the offers of all commodities from the sellers, the evaluation agent will have comparison mechanism to evaluate each commodity in order to make the best possible selection of all the supplied commodities. Since shopping is not just searching for a lower price commodity. There is something else that should be taken into considerations like quality, reliability, brand, service, etc. In the system, the multi-attribute decision making method (Barbuceanu, M., Lo, W. 2000)(Keeney, R. L., Raiffa, H. 1993) is applied to evaluate commodities considering multi-attributes of the commodities. Based on the multi-attribute evaluation model, the evaluation agent calculates the utility value of each commodity and selects one

that has maximal utility value as the recommended commodity. Its mathematical model can be described:

Supposing $C=\{c_1,c_2,...,c_m\}$ as the vector of the commodities information that has been gathered on Internet, $A=\{a_1,a_2,...,a_n\}$ as the qualitative feature vector of the commodities, the utility value of the commodity $c_i(1\le i\le m)$ about the attribute $a_j(1\le j\le n)$ can be denoted as $f_{ij}=f_j(c_{ij})$, which represents the relative performance of the commodity c_j in the qualitative feature i. Therefore, the decision matrix that consists of $m\times n$ f_{ij} can be denoted as:

$$F=\begin{bmatrix} f_{11} & f_{12} & ... & f_{1n} \\ f_{21} & f_{22} & ... & f_{2n} \\ ... & ... & ... & ... \\ f_{m1} & f_{m2} & ... & f_{mn} \end{bmatrix}=(f_{ij})_{m\times n} \tag{1}$$

In order to facilitate mutual reference between the multi- attributes easily, the decision matrix should be normalized, which can be followed by formula (2):

$$f_{ij}^{'}=\frac{f_{ij}}{\sqrt{\sum_{i=1}^{m}(f_{ij})^2}} \tag{2}$$

After normalizing the decision matrix, the value of $f_{ij}^{'}$ is limited in [0,1]. Then the evaluation agent can calculate the utility value of each commodity based on the formula (3).

$$U(c_i)=\sum_{j=1}^{n}\omega_j f_{ij}^{'} \tag{3}$$

In the formula (3), $U(c_i)$ is the utility value of the commodity $c_i(1\le i\le m)\cdot\omega_j$ is the weight of the qualitative feature $j(1\le j\le n)$, which means the customer's current requirement in this qualitative feature and $\sum_{j=1}^{n}\omega_j$ =1. After calculating the utility of all the commodities, the evaluation agent will select one that has maximal utility value as the recommended commodity. Finally, the evaluation agent submits the recommended commodity to the customer via interface agent. The whole computing process is performed by the evaluation agent automatically.

5) Collaboration agent

As indicated before, the user-system interaction is an important factor in achieving optimal recommendation. During the interaction, the consumer can give more feedback to the system by updating his/her current needs until the consumer is satisfied with the shopping result. However, the frequent user-system interactions inevitably take time. In the system, collaboration agent is designed to reduce the time of user-system interaction. The collaboration agent is based on the consumer-based collaboration approach(Zeng Chun, Xing Chun-Xiao et al. 2004), which first compares the need pattern of the current customer to the ones previously recorded and then system recommends the commodities selected by the similar consumers to the current customer.

The qualitative need pattern of a customer can be defined as a vector $W=(\omega_1,\omega_2,...,\omega_n)$, in which $\omega_i(1\le i\le n)$ means the preference score of customer's qualitative need in the feature dimension i, and n is the number of qualitative need feature. The collaboration can acquire the need pattern of previous customers easily by accessing the web log database of the system as shown in table 1.

The collaboration agent uses the correlation coefficient of *Pearson*, which compares the current customer's need pattern with the ones of the previous customers, and then calculate the similarities between the current customer and all the previous customers. Its mathematic model can be expressed as follows:

Supposing the need pattern of a current customer a as the vector: $W_a=(\omega_{a,1},\omega_{a,2},...,\omega_{a,n})$, the need pattern of a previous customer b as the vector: $W_b=(\omega_{b,1},\omega_{b,2},...,\omega_{b,n})$. So the similarities $Sim(a,b)$ between two need pattern can be calculated as formula (4):

$$Sim(a,b)=\frac{\sum_{j=1}^{n}(\omega_{a,j}-\overline{\omega}_a)(\omega_{b,j}-\overline{\omega}_b)}{\sqrt{\sum_{j=1}^{n}(\omega_{a,j}-\overline{\omega}_a)^2}\sqrt{\sum_{j=1}^{n}(\omega_{b,j}-\overline{\omega}_b)^2}} \tag{4}$$

In the formula (4), $\omega_{a,j}$ and $\omega_{b,j}$ represent the preference score of qualitative need feature j that the current

customer a and previous customer b give respectively, while $\overline{\omega}_a$ and $\overline{\omega}_b$ represent the average score of all the features that the current customer a and previous customer b give respectively.

Through the similarity calculation of qualitative need features, the collaboration agent can search out the most similar need pattern for the current customer from the web log database. The system then predicts that what the current customer is targeting may be the commodities that most similar previous customer finally purchased. Hence, the system also recommends commodities derived form the collaborative filtering approach described above to the current customer, in addition to the optimal commodity provided by the evaluation agent. In this way, the system gives the customer more choice space and a customer can share experiences from previous customers. On the other hand, the number of iterations of user-system interaction can thus be reduced, and the system can work even more efficiently.

3.2 Commodity ontology

The shopping system should gather commodities information from multiple sellers, however, it is difficult to exchange information between the shopping system and the sellers because of the different commodity data format in database and representation. In order to maintain a semantic conversation between the shopping system and sellers, there should be a common language to support shared data format and representation about the commodities information. This is established by means of an ontology, which contains the main concepts owning to the domain we are dealing with. In addition to this information, the ontology also includes attributes, values, relations between concepts and axioms so that consistency checking and inferences are done (Yan H., Schreiber G., et al. 1997). Therefore, the main ontological entity in the prototype system developed in the work is the concept, but the use of other ontological entities such as attributes is also possible in the model in order to provide the system with powerful representation capabilities.

In this example, commodity ontology show how a computer is composed by several elements: monitor, keyboard, mouse, processor, etc, which can be described as follows using OWL languages (Deborah LM, Frank VH. 2004):

```
<owl:Class rdf:ID="computer">
  <rdf: subClassOf rdf:resource="#Product" />
  <rdf:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="hasTradeMark" />
       <owl: hasValue rdf:resource="#IBM"/>
      </owl:Restriction>
  </rdf:subClassOf>
  <rdf:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasModel" />
      <owl: cardinality
        rdf:datatype="&xsd;nonNegativeInteger"> 1
      </owl:cardinality>
      <owl: hasValue rdf:resource="#CompaqEvoD220"/>
    </owl:Restriction>
  </rdf:subClassOf>
  <rdf:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasGuaranty"/>
      <owl: cardinality
        rdf:datatype="&xsd;nonNegativeInteger"> 1
      </owl:cardinality>
      <owl: hasValue rdf:resource= 24/>
     </owl:Restriction>
  </rdf:subClassOf>
```

```
  <rdf:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasOrderPrice" />
      <owl: minCardinality
        rdf:datatype="&xsd;nonNegativeInteger"> 1
      </owl:minCardinality>
      <owl: hasValue rdf:resource= 6399.00/>
    </owl:Restriction>
  </rdf:subClass>
</owl:Class>
```

In this case, in addition to the concepts taking part in the semantic relation under question, the relation will have a name with the relation type and eventually some other properties associated to that relation.

3.3 Web application

With the purpose of applying intelligent agents to the e-commerce system, JADE platform should be integrated into the Web application.

At first, the environment initialization is needed in order to start working with JADE. This process can be implemented by AgentLoader, which reads configuration files and creates the AMS and DF agents. Then AMS and DF agents provide white/yellow pages services respectively. On the one hand, DF provides a yellow pages service to the other agents in the system, which executes the tasks of agent registration and lookup. When the buyer or seller agent is created, it should be registered in AMS. On the other hand, DF is responsible for monitoring the life cycle of each agent and tracing the behaviour of it. In this way, all the agents in the system can be effectively managed and their inter-communication will be facilitated well.

The design of the system is based on the Apache Struts Web Application Framework and can be implemented with Java technology.

4. Conclusions

In this paper, I have indicated the need to automate shopping process on Internet and provide more personalized information services for customers. Therefore, developing intelligent shopping system is a promising way to achieve this goal. In the work, I present a multi-agent system to provide shopping service for the commodities that a consumer does not buy frequently. The system integrates built-in expert knowledge and the customer's current needs, and recommends optimal products based on multi-attribute decision making method. To reduce the effort of system-customer interactions, the system utilizes customer-based collaboration filtering approach to recommend the products. Besides, in order to maintain a semantic conversation with sellers, the commodity ontology is also utilized to support sharable information format and representation. A prototype of the system is implemented using the Java Agent Development Framework (JADE). The result shows that the system performs efficiently and can help customers save enormous time for Internet shopping. My future work will be focused on developing some security mechanisms to provide security services for the system.

References

Barbuceanu, M., Lo, W. (2000). "A Multi-Attribute Utility Theoretic Negotiation Architecture for Electronic Commerce," *Proc. 4th Int. Conf. On Autonomous Agent,* 2000.

Deborah LM, Frank VH. (2004). *Owl web ontology language overview*, Stanford University, USA. URL: http://www.w3.org/TR/owl-features/, 2004.

Keeney, R. L., Raiffa, H. (1993). *Decision with Multiple Objectives*, Cambridge University Press, 1993.

Lee, J. Lee,J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising[J]. *Data Mining and Knowledge Discovery*, 2001, 5(1-2): 59-84.

Yan H., Schreiber G., et al. (1997). Using explicit ontologies in KBS development[J]. *International Journal of Human Computer studies*, 1997, 45(2-3): 183-192.

Zeng Chun, Xing Chun-Xiao et al. (2004). Similarity Measure and Instance Selection for Collaborative Filtering [J]. *International Journal of Electronic Commerce*, 2004, 8(4): 115-129.

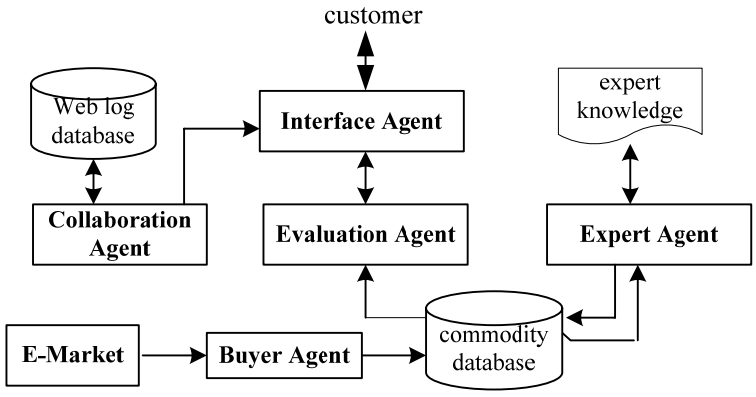


Figure 1. The architecture of the shopping system

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Calora: A Software to Simulate Calcium Diffusion

Gaurav Gupta (Corresponding author)

Department of Mathematics

Maulana Azad National Institute of Technology

Bhopal – 462051, India

Tel: 91-99939-03363 E-mail: gaurav_gaurav2@yahoo.co.in

Shivendra Tewari

Department of Mathematics

Maulana Azad National Institute of Technology

Bhopal – 462051, India

Tel: 91-94257-27127 E-mail: shivendra.tewari@rediffmail.com

K.R. Pardasani

Department of Mathematics

Maulana Azad National Institute of Technology

Bhopal – 462051, India

Tel: 91-94253-58308 E-mail: kamalrajp@rediffmail.com

Abstract

Calcium is a vital second messenger which regulates processes in almost all cell types like myocytes, hepatocytes, oocytes,neurons etc. Due to spatiotemporal limitation of imaging techniques, it is not possible to determine calcium concentration at macroscopic level. For this purpose mathematical modeling of calcium dynamics provides a suitable alternative. Several mathematical models governing intracellular Ca^{2+} dynamics have been proposed thus far. Here in this paper author has developed a software to simulate two distinct approximations, namely, Excess buffering approximation (EBA), and Rapid Buffering approximation (RBA). The software is based on oracle 10g and follows an algorithm based on an explicit Eulers method. The software contains all the important functionalities, like Multiple graphs, Spatio-temporal zoom, desired Ca^{2+} conc. value retrieval at specific space & time etc., to simulate the approximations for all important biophysical constants, to have a graphical representation of a simulated model and to do comparative analysis among the mathematical models.

Keywords: Ca^{2+} Diffusion, Oracle, Partial differential equation

1. Introduction

Calcium ions (Ca^{2+}) impact nearly every aspect of cellular life as they are fundamental important component of cellular signal transduction. It helps in regulating cell growth and differentiation. (Lodish, Berk, Zipursky, Matsudaira, Baltimore, Darnell, 2000). The divalent cation calcium (Ca^{2+}) is also used by cells as a second messenger to control many cellular processes including muscle contraction, secretion, metabolism, neuronal excitability, cell proliferation, and cell death (Simpson, Challiss & Nahorski, 1995).

Neuronal activity can lead to marked increases in the concentration of cytosolic calcium. Calcium binds to calmodulin and stimulates the activity of a variety of enzymes, including calcium-calmodulin kinases and calcium-sensitive adenylate cyclases. These enzymes transduce the calcium signal and effect short-term biological responses, such as the modification of synaptic proteins and long-lasting neuronal responses that require changes in gene expression. The calcium levels outside cells are 10 000 times higher than free intracellular Ca^{2+}. However, free $[Ca^{2+}]_i$ is the physiologically active form of calcium (Rasmussen, 1988). The level of free intracellular calcium ($[Ca^{2+}]_i$) is regulated and maintained as low as (~100 nM) through the action of a number of binding proteins and ion exchange mechanisms.

Each cell has a unique set of Ca^{2+} signals to control its function. Ca^{2+} signal transduction is based on rises in free cytosolic Ca^{2+}concentration. Ca^{2+} can flow from the extracellular space or be released from intracellular stores. The endoplasmic reticulum (ER) is a major site for sequestered Ca^{2+} ions.

Calcium concentrations are strongly buffered in living cells. Buffer site concentrations have been estimated to be in the range of 100-300, μM in the cytoplasm and significantly higher in the endoplasmic reticulum (ER) (Allbritton, Meyer & Stryer, 1992; Michalak, Milner, Burns & Opas, 1992). Certain proteins of the cytoplasm and organelles act as buffers by binding Ca^{2+}.Which binds most of the Ca^{2+} in a cell (up to 99%). Depending on their diffusion characteristics, buffers are considered as mobile or immobile. Moreover, buffers can control intracellular Calcium dynamics by tuning spatial coupling of channel clusters. That determines the average period of oscillations due to wave nucleation (Falcke, 2003; Lukyanenko & Gyorke, 1999).

Two theories have been used to simplify the system of reaction-diffusion equations governing calcium diffusion. One is excess buffer approximation (EBA) which assumes that mobile buffer is present in excess and cannot be saturated. The other is rapid buffer approximation (RBA), which assumes that calcium binding to buffer is rapid compared to calcium diffusion rate. In the present work an attempt has been made to develop software which simulates Ca^{2+} diffusion under both type of approximation.

EBA is appropriate when the saturability of mobile buffer is negligible and when there is significant saturability of mobile buffer and buffer kinetics are fast relative to Ca^{2+} diffusion then RBA is appropriate (Tewari,Tewari & Pardasani, 2009) . In, our software, we have a single interface via which users can switch to either of the approximation. Further, the software is developed using Oracle 10g to provide handling of large databases generated by model simulation.

2. Mathematical Formulation

Author has assumed that only Ca^{2+} ions and buffers are present inside the cytosol. Assuming the following reaction equation between Ca^{2+} and buffers,

$$[Ca^{2+}]+[B_j] \underset{k^-}{\overset{k^+}{\rightleftharpoons}} [CaB_j] \tag{1}$$

Where $[B_j]$ and $[CaB_j]$ are free and bound buffer respectively and 'j' is an index over buffer species. If we further assume that the cytosolic medium is homogenous and is isotropic then, Calcium dynamics in presence of a point source can be framed in the form of following equations (Neher 1986; Smith 1996; Smith, Dai, Miura & Sherman 2000):

$$\frac{\partial[Ca^{2+}]}{\partial t} = D_{Ca}\nabla^2[Ca^{2+}]+\sum_j R_j + \delta\sigma(r) \tag{2}$$

$$\frac{\partial[B_j]}{\partial t} = D_{B_j}\nabla^2 B_j + R_j \tag{3}$$

$$\frac{\partial[CaB_j]}{\partial t} = D_{CaB_j}\nabla^2[CaB_j]-R_j \tag{4}$$

Where, the Laplacian operator written in spherical symmetry,

$$\nabla^2 = \frac{\partial^2}{\partial r^2}+\frac{2}{r}\frac{\partial}{\partial r}$$

$$R_j = -k_j^+[B_j][Ca^{2+}]+k_j^-[CaB_j] \tag{5}$$

D_{Ca}, D_{Bj}, D_{CaBj} are diffusion coefficient of free calcium, free buffer, and Ca^{2+} bound buffer, respectively; k_j^+ and k_j^- are association and dissociation rate constants for buffer 'j', respectively. For stationary, immobile buffers or fixed buffers $D_{Bj} = D_{CaBj} = 0$.

For boundary conditions, we assume a point source of Ca^{2+} at origin and a fixed background Ca^{2+} concentration. There is no source for buffer, and the buffer is assumed to be in equilibrium with Ca^{2+} far from the source. The corresponding equation is (Smith *et al.,* 2000)

$$\lim_{r\to 0}\left\{-2\pi r^2 D_{Ca}\frac{d[Ca^{2+}]}{dr}\right\} = \sigma, \qquad \lim_{r\to\infty}\left[Ca^{2+}\right] = \left[Ca^{2+}\right]_\infty, \tag{6}$$

$$\lim_{r\to 0}\left\{-2\pi r^2 D_{B_j}\frac{d[B_j]}{dr}\right\} = 0, \qquad \lim_{r\to\infty}[B_j] = [B_j]_\infty \tag{7}$$

$$\lim_{r\to 0}\left\{-2\pi r^2 D_{CaB_j}\frac{d[CaB_j]}{dr}\right\} = 0, \qquad \lim_{r\to\infty}\left[CaB_j\right] = [CaB_j]_{\infty,} \tag{8}$$

$$\left[B_j\right]_\infty = \frac{K_j[B_j]_T}{K_j + [Ca^{2+}]_\infty}, \qquad \left[CaB_j\right]_\infty = \frac{\left[Ca^{2+}\right]_\infty [B_j]_T}{K_j + [Ca^{2+}]_\infty}, \tag{9}$$

$K_j = k_j^- / k_j^+$ is dissociation constant for buffer j, and $[B_j]_T = [B_j] + [CaB_j]$ total concentration profile for each buffer.

If we assume that the diffusion constant of each mobile buffer is not affected by the binding of Ca^{2+} (that is, $D_{Bj} = D_{CaBj}$), and also assume that, $[B_j]_T$ is initially uniform, then $[B_j]_T$ will remain uniform for all time. Thus we can write (Smith 1996; Smith *et al.,* 2000)

$$R_j = -k_j^+[Ca^{2+}][B_j] + k_j^-([B_j]_T - [B_j])$$

Finally, we restrict consideration to unsteady state i.e. to the domain of high Ca^{2+}, done in microsecond near sources. Note that the fixed buffer, while indispensable for the time dependent PDE solution, have no influence on steady states because $R_j = 0$ if $D_{Bj} = D_{CaBj} = 0$. For ease, we assume a single mobile buffer species, [B], resulting in following boundary-value problem (Wagner & Keizer, 1994; Smith, Wagner, & Keizer, 1996)

$$\frac{\partial[Ca^{2+}]}{\partial t} = D_{Ca}\nabla^2[Ca^{2+}] - k^+[B][Ca^{2+}] + k^-([B]_T - [B]), \tag{10}$$

$$\frac{\partial[B]}{\partial t} = D_B\nabla^2[B] - k^+[B][Ca^{2+}] + k^-([B]_T - [B]), \tag{11}$$

With these boundaries condition (Smith *et al.,* 2000)

$$\lim_{r\to 0}\left\{-2\pi r^2 D_{Ca}\frac{d[Ca]^{2+}}{dr}\right\} = \sigma, \qquad \lim_{r\to\infty}\left[Ca^{2+}\right] = \left[Ca^{2+}\right]_\infty, \tag{12}$$

$$\lim_{r\to 0}\left\{-2\pi r^2 D_B\frac{d[B]}{dr}\right\} = 0, \qquad \lim_{r\to\infty}[B] = [B]_\infty = \frac{K[B]_T}{K + [Ca^{2+}]_\infty}. \tag{13}$$

Basically there are two approximations which have been proposed in order to simplify equation (1-4),

2.1 The Excess Buffer Approximation (EBA):

In seminal work Neher (1986) made the critical observation that if buffer is present in excess then, free mobile buffer profile is not perturbed by presence of the source. Under this assumption, one can make the approximation that $[B] \approx [B]_\infty$

$$\frac{\partial[Ca^{2+}]}{\partial t} = D_{Ca}\nabla^2[Ca^{2+}] - k^+[B]_\infty[Ca^{2+}] + k^-([B]_T - [B]_\infty). \tag{14}$$

As $t \to \infty$ one can write

$$k^-([B]_T - [B]_\infty) = k^+[B]_\infty[Ca^{2+}]_\infty, \tag{15}$$

$$\frac{\partial\left[Ca^{2+}\right]}{dt} = D_{Ca}\nabla^2[Ca^{2+}] - k^+[B]_\infty([Ca^{2+}] - [Ca^{2+}]_\infty) \tag{16}$$

This linear equation for $[Ca^{2+}]$ at steady state can be solved easily and applying the boundary conditions would yield (Smith 1996),

$$[Ca^{2+}] = \frac{\sigma}{2\pi D_{Ca} r}e^{-r/\lambda} + [Ca^{2+}]_\infty \tag{17}$$

Where λ is the characteristic length constant for the mobile Ca^{2+} buffer given by $\lambda = \sqrt{D_{Ca}/k^+[B]_\infty}$ .This approximation has been shown to be valid when mobile buffer is in high concentration and/or when the source amplitude is small, that is , $\lim_{r\to 0}[B] \approx [B]_\infty$

Initial condition is

$$\left[Ca^{2+}\right]_{t=0} = 0.1\ \mu M, \tag{18}$$

And boundary condition

$$\lim_{r\to 0}\left(-2\pi D_{Ca} r^2 \frac{d\left[Ca^{2+}\right]}{dr}\right) = \sigma_{Ca} \tag{19}$$

$$\lim_{r\to\infty}\left[Ca^{2+}\right] = 0.1\ \mu M \tag{20}$$

2.2 Rapid Buffer Approximation (RBA):

In Rapid Buffering approximation, we assume that the buffering time scales are rapid, reaching equilibrium at each point in space before appreciable diffusion occurs. This means that the buffering is a singular perturbation (Murray, 2001), which we handle by deriving an equation for the local, total concentration of Ca^{2+}, $[Ca^{2+}]_T$:

$$\left[Ca^{2+}\right]_T = \left[Ca^{2+}\right] + \left[CaB_s\right] + \left[CaB_m\right] \tag{21}$$

We find that

$$\frac{\partial\left[Ca^{2+}\right]_T}{\partial t} = D_{Ca}\nabla^2\left[Ca^{2+}\right] + D_{CaB_m}\nabla^2\left[CaB_m\right] \tag{22}$$

This equation does not involve the rapid buffering time scale and, therefore, involves no singular perturbation. It is possible to eliminate $[CaB_m]$ from this equation using the assumption of rapid equilibrium (Wagner and Keizer 1994),

$$\left[CaB_i\right] = \frac{\left[Ca^{2+}\right]\left[B_i\right]_T}{K_i + \left[Ca^{2+}\right]}, \tag{23}$$

Where K_i is the dissociation constant and $[B_j]_T$ is the total concentration of stationary or mobile buffer binding sites. When combined allows $[Ca^{2+}]_T$ to be written in terms of $[Ca^{2+}]$,

$$\left[Ca^{2+}\right]_T = \left[Ca^{2+}\right]\left(1 + \frac{\left[B_s\right]_T}{\left(K_s + \left[Ca^{2+}\right]\right)} + \frac{\left[B_m\right]_T}{\left(K_m + \left[Ca^{2+}\right]\right)}\right) \tag{24}$$

Using equation 21, one can write,

The following transport equation for free Ca^{2+}:

$$\frac{\partial\left[Ca^{2+}\right]}{\partial t} = \beta\left(D_{Ca} + \gamma_m D_{CaB_m}\right)\nabla^2[Ca^{2+}] - \frac{2\beta\gamma_m D_{CaB_m}}{K_m + [Ca^{2+}]}\left(\nabla\left[Ca^{2+}\right]\right)^2, \tag{25}$$

$$\gamma_i = \frac{k_i\left[B\right]_T}{\left(k_i + \left[Ca^{2+}\right]\right)^2},\ \beta = \left(1 + \gamma_m + \gamma_s\right)^{-1}$$

Where, i = m or s. Assuming that D_{CaBm} and D_{Bm} is approximately equal so we can rewrite this equation in this form along with initial boundary condition (Wagner and Keizer 1994; Smith *et al.*, 1996)

$$\frac{\partial\left[Ca^{2+}\right]}{\partial t} = \beta\left(D_{Ca} + \gamma_m D_{Bm}\right)\nabla^2[Ca^{2+}] - \frac{2\beta\gamma_m D_{Bm}}{K_m + [Ca^{2+}]}\left(\nabla\left[Ca^{2+}\right]\right)^2, \tag{26}$$

Initial Condition,

$$\left[Ca^{2+}\right]_{t=0} = 0.1\ \mu M, \tag{27}$$

And boundary condition near the channel is given (Wagner and Keizer, 1994)

$$\lim_{r\to 0}\left\{-2\pi r^2\beta\left[D_{Ca} + \gamma_m D_{Bm}\right]\frac{\partial\left[Ca^{2+}\right]}{\partial r}\right\} = \beta\sigma, \tag{28}$$

In the next section we discuss about the importance of the software and the results obtain by the software.

3. CalOra: The Software

It is Oracle 10g based software in which there is a single interface for both simulations i.e. rapid buffer approximation and excess buffer approximation. The user can switch in between any of the simulations at any time. To start CalOra interface, there is a need for server OC4J (Oracle container for J2EE) instance which has to be started before its interface. The OC4J standalone distribution includes an HTTP(S) server, the required J2EE services, and Web Services capabilities all of which are executed from one Java process. This server comes with the Oracle 10 g setup and is installed along with it. CalOra gives the result in large precision form (the number values after the decimal place) and

this precision number can be changed according to the computational power one needs to spend.

CalOra runs on web browser and has to be configured to run in your machine. Data generated from CalOra can be saved in oracle database for future reference. Saved simulated data can be used to perform comparative studies. Figure-1 shows the main page of CalOra where user enters its choice to proceed with any of the simulations according to his/her need. It is basically software for simulating calcium diffusion obeying excess buffering and Rapid buffering approximations. Help is given in the upper right corner of CalOra where beginners may get help about EBA, RBA module and its limitations as shown in figure 2, so that the user can get familiar with the functioning of our software.

Now, if user selects EBA or RBA, he will move on to the respective module where he can enter different parametric values according to his/her need, but if user does not want to enter any values then the user can proceed with the simulation using the default values stored in the Oracle database, so that naïve user can also perform the simulation without hovering about the parametric values to choose. The user need not to enter all the parametric values and hence can change only few values to get complete results along with graphical representation of their simulation. Input page of EBA and RBA module is given in figure 3 and 4.

In EBA section the user can perform two simulation in one go. How? He can choose two different buffer association rates for two simulations and click the calculate button, to have two different tables corresponding to the two buffer association rates. User can see the simulated calcium concentration values with the help show values buttons adjacent to the table containing $[Ca^{2+}]$ values. The user can see the $[Ca^{2+}]$ values for both the buffer with respect to space and time and can be return back to the input page, using back button, where he/she can again make changes to the parametric values. Further the user can see the results in graphical form with the help of Graph button.

Figure 3 shows the GUI for EBA module. The values shown in the text box along with the text labels are the default parametric values stored in Oracle database. There are there buttons shown namely Calculate, $[Ca^{2+}]$ Values and Graphs. Calculate button can be used, once the user has decided to calculate calcium concentration with his/her own parametric values. These values can be seen with the help of second button $[Ca^{2+}]$ values. These values are shown in a table, where the rows corresponds the space node and the column correspond the time node. From the table, if user wants to know the particular value of calcium concentration on specific space and time node, he/she can use enter query and execute query buttons to see the desired value as shown in figure 5. Graph button takes the user to the graphical representation page where the user can see the simple and logarithmic graphical results of his/her simulation.

Now if we talk about RBA, we can perform comparative study between three different types of buffers namely stationary, endogenous mobile and exogenous as shown in the figure 4.We have taken only one buffer specie per simulation assuming that the other two species are absent. To perform a comparative analysis, the concentration of each single buffer species is taken to be 50 μM. Further the user can enter his/her choice of buffer concentration and buffer dissociation constant. In this module, we have created three tables to store values of three different simulations. In this page, we have created some more buttons like copy1 and copy 2. These two buttons can be used to copy the data in Oracle database for future reference. One another extra button is there i.e. delete all data which is used to delete all data from all tables. This functionality will be used when the user is performing comparative study. [Ca++] Values button has same predefined functionality elaborated in EBA section.

In both of the modules EBA and RBA, there is a common button "Graph". This button is used to show simple linear graph and logarithmic graph. The purpose of having logarithmic graph along with the simple graph is to differentiate between the curves under undistinguishable situations. For this purpose, we have used logarithmic scale on Y axis i.e. Calcium concentration axis which will be shown in the result and discussion section. Further the user can enter any desired value of time to have its corresponding calcium diffusion path. As usual, when the user has finished playing with the software, he can use Quit button to exit from the module.

4. Results and Discussion

For RBA (Rapid Buffer Approximation) and EBA (Excess Buffer Approximation), the numerical values for various parameters used are given in Table 1 and Table 2. With the help of CalOra calcium concentration profiles have been plotted with respect to space which is shown in figure 6 and figure 7.

In figure 6, there are two curves shown in the figure blue (E) represents EGTA and light blue (B) represents BAPTA. In this graph, cytosolic $[Ca^{2+}]$ profile is plotted for two exogenous buffers; both have same affinity for $[Ca^{2+}]$ but different binding rates. 1) EGTA (*Ethylene Glycol-bis(beta-aminoethyl-ether)-N,N,N',N'-Tetra Acetate*) a very slow buffer, 2) BAPTA (*1,2-bis(o-minophenoxy)ethane-N,N,N',N'-tetra acetic acid*) a very fast buffer. Exogenous buffers are used to delay or accelerate the time required to achieve steady state condition. Same thing is also evident from the figure, the results are shown in two scales as to show the difference in between the chelators binding rates and the impact they have on the spatial aspect of $[Ca^{2+}]$ attaining its resting concentration. It is apparent from the log plot that $[Ca^{2+}]$ achieves its resting concentration profile close to the channel 1.001 μm while for EGTA it relaxes around 10.01 μm.

In figure 7, cytosolic calcium profile for three different buffer species, namely stationary, endogenous mobile and

exogenous buffers, under rapid buffering approximation is shown. There are three curves shown in the figure blue (T) represents Troponin-C,light green (E) represents EGTA and light blue (C) represents Calmodulin. The study was done; assuming that only single buffer specie is present at a time i.e. assuming that data to specie are absent.The curve for Troponin-C is closest to the y axis since it is a stationary buffer and its dissociation rate is on the higher side i.e. the rate of calcium dissociation is higher than its binding rate. While the other two curves for EGTA (exogenous) and Calmodulin (endogenous) are in a different league as both of them are mobile and hence affect the path of calcium diffusion as shown in figure 7. The curve for EGTA is the second closest to y axis while the curve for Calmodulin is farthest. It is because of the fact that, the dissociation constant for Calmodulin is greater than that for EGTA. Higher dissociation constant implies that there is more free calcium inside the cytosol and hence the curve for calmoduin is farthest from y axis. But one can argue the fact stated above as the dissociation constant for the Troponin-C is greatest but its curve is steepest. This happens because Troponin-C is a stationary buffer and hence is assumed to have zero diffusion coefficients contributing nothing to calcium diffusion but affecting the time required to achieve steady state nonetheless.

5. Future Scope

Similar software for Ca^{2+} diffusion does exist and one of them is CalC – The Calcium Calculator initiated by Matveev (2002). This project is a large scale project which is under operation since 2002 where as our software is still a stub and is under development process. In future authors would like to extend their work to the extent of this well known work of CalC – The Calcium Calculator. So that present work could prove beneficial in diagnosis of neurodegenerative disorders, such as Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), Huntington's disease (HD) and spinocerebellar ataxias (SCAs), present an enormous medical, social, financial and scientific problem. Recent evidence indicates that neuronal calcium (Ca^{2+}) signaling is abnormal in many of these disorders. Similar, but less severe, changes in neuronal Ca^{2+} signaling occur as a result of the normal aging process. The role of aberrant neuronal Ca^{2+} signaling in the pathogenesis of neurodegenerative disorders is key research area now these days (Bezprozvanny, 2009). With the help of software development we can contribute in this field. So CalOra can be extended up to the level that it would be able in disease diagnosis by simulating complete calcium dynamics of a particular person.

References

Allbritton N., Meyer T., and Stryer L. (1992). Range of messenger action of calcium ion and inositol 1, 4, 5 trisphosphate. *Science.* 258, 1812–1815.

Bezprozvanny I. (2009). Calcium signaling and neurodegenerative diseases. *Trends in Molecular Medicine*, v. 15, pp 89-100.

Falcke M. (2003). On the role of stochastic channel behavior in intracellular Ca^{2+} dynamics, *Biophysical Journal,* 84, 42–56.

Lodish H.,Berk A.,Zipursky S.L.,Matsudaira P.,Baltimore D.,Darnell J.E. (2000). *Molecular Cell Biology* (4th edition), New York, WH Freeman & Company.

Lukyanenko V., Györke S. (1999). Ca^{2+} sparks and Ca^{2+} waves in saponin-permeabilized rat ventricular myocytes, *Journal of Physiology* 521, 575–585.

Michalak M.,Milner R.,Burns K., and Opas M. (1992). Calreticulin, *Journal of Biochemistry*, 285, 681–692.

Matveev V. (2002). CalC Calcium Calculator [Online] Available: http://web.njit.edu/~matveev/calc.html

Murray J.D. (2001). *Mathematical Biology* (3rd edition), New York: Springer-Verlag.

Neher E. (1986). Concentration profiles of intracellular Ca^{2+} in the presence of diffusible chelators. *Experimental Brain Research Series*, 14, 80-96.

Rasmussen. (1988). JE: Calcium and the skin. *Archive of Dermatol*. 124: 443–444.

Simpson P.B., Challiss R.A.J. and Nahorski S.R. (1995). Neuronal Ca^{2+} stores: activation and function. *Trends Neuroscience*, 18, 299–306.

Smith G.D. (1996). Analytical Steady-State Solution to the rapid buffering approximation near an open Ca^{2+} channel, *Biophysical Journal*, 71, 3064- 3072.

Smith G.D., Wagner J., and Keizer J. (1996). Validity of the rapid buffering approximation near a point source of calcium ions, *Biophysical Journal*, 70, 2527-2539.

Smith G.D., Dai L., Miura R. M., and Sherman A. (2000). Asymptotic Analysis of buffered Ca^{2+} diffusion near a point source. *SIAM Journal of Applied of Mathematics*, 61 1816-1838

Tewari V., Tewari S., and Pardasani K.R. (2009). A Model to Study the Effect of Excess buffers and Na+ ions on Ca^{2+}

diffusion in Neuron cell. *International Journal of Biological and Medical Sciences*, 5, 22-27.

Wagner J. and Keizer J. (1994). Effects of rapid buffers on Ca^{21} diffusion and Ca^{21} oscillations. *Biophysical Journal,* 67, 447-456.

Table 1. RBA Approximation: Parameters, Values used

Parameter	Values	*Unit*
Space Initialization	0.001	*Micro meter*
Space Termination	1	*Micro meter*
Space step	100	
No. of time Step	10000	
Initial time	0	*Sec*
Final time	0.1	*Sec*
Initial Condition	0.0000001	*Molar*
Mobile Buffer Association Rate	500	*Micro molar^{-1} sec^{-1}*
Mobile Buffer Dissociation Rate	470	*Sec^{-1}*
Exogenous Buffer Association Rate	1.5	*Micro molar^{-1} sec^{-1}*
Exogenous Buffer Dissociation Rate	0.3	*Sec^{-1}*
Stationary Buffer Association Rate	90	*Micro molar^{-1} sec^{-1}*
Stationary Buffer Dissociation Rate	300	*Sec^{-1}*
Diffusion Coefficient Of Calcium	250	*Micro meter^{2} sec^{-1}*
Diffusion Coefficient Of Mobile	32	*Micro meter^{2} sec^{-1}*
Diffusion Coefficient Of Exogenous	113	*Micro meter^{2} sec^{-1}*
Source	40	*Pico ampere*

Table 2. EBA Approximation: Parameters, Values used

Parameter	Values	*Unit*
Initial time	0	*Sec*
Final time	100	*Sec*
Space Bounds	[0.001, 1]	*Micro meter*
Time Step Length	.01	*Sec*
No. of time step	10000	
No. of space step	40	
Space Step Length	.25	*Micro meter*
Buffer Association Rate	.0015	*Micro molar^{-1} sec^{-1}*
Buffer Concentration	50	*Micro molar*
Diffusion Coefficient	.25	*Micro meter^{2} sec^{-1}*
Initial Condition	.1	*Micro molar*
Source	1	*Pico ampere*

Figure 1. Webpage showing main Page of CalOra where user can switch either of simulation

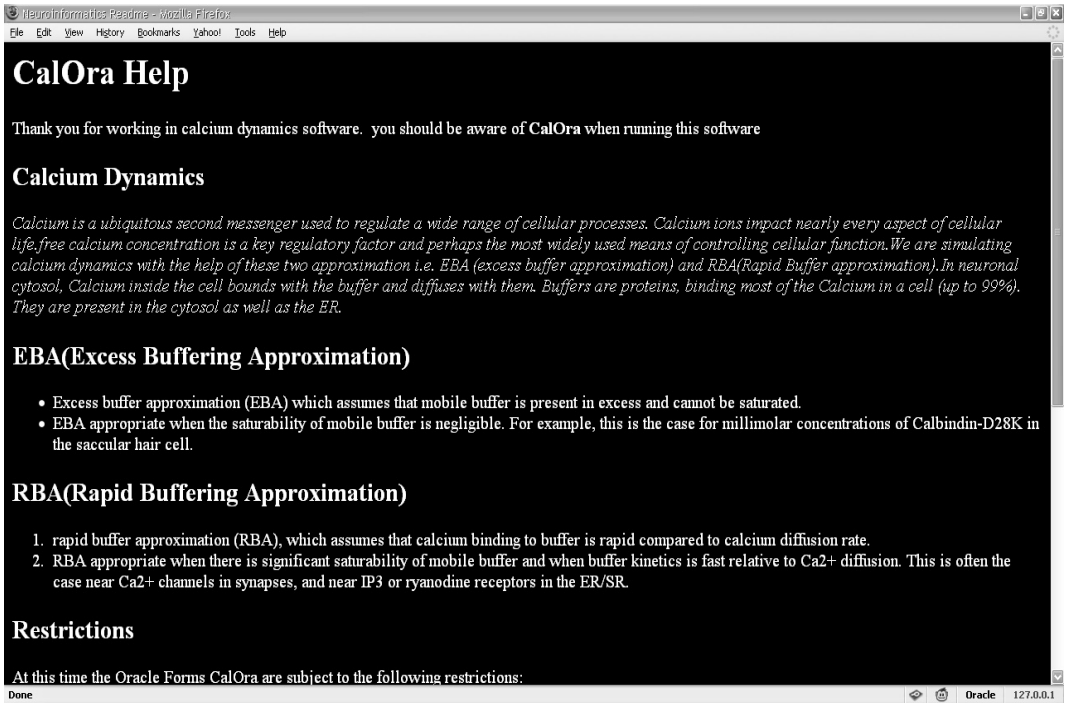


Figure 2. Webpage showing Help section of CalOra where user can aware of its different part

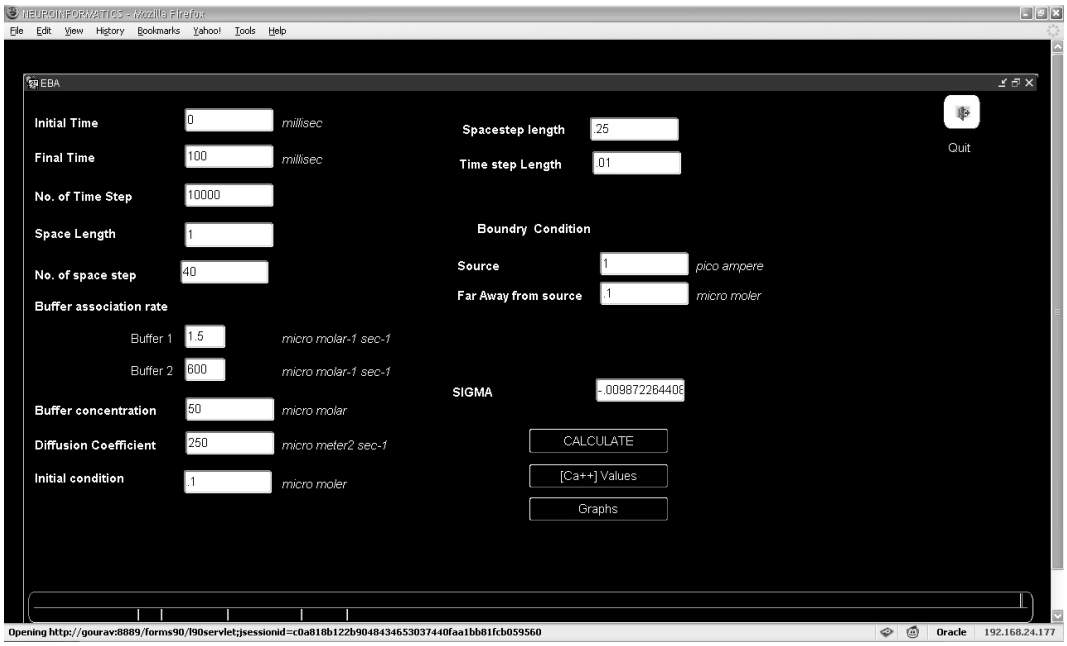


Figure 3. Webpage showing EBA module of CalOra where user can enter different parametric values

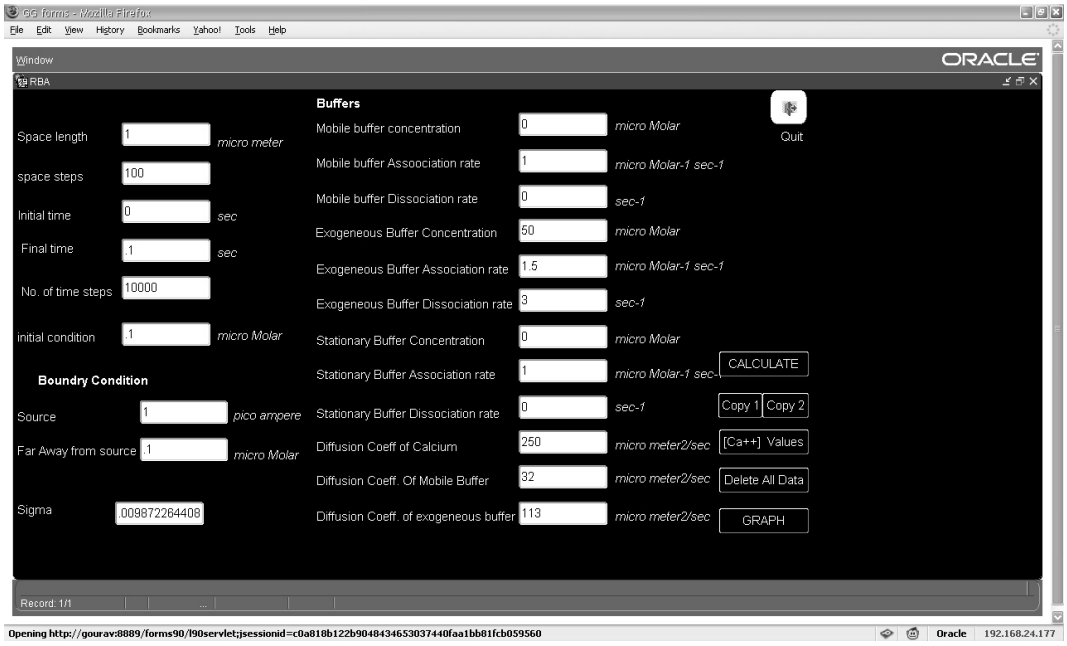


Figure 4. Webpage showing EBA module of CalOra where user can enter different parametric values

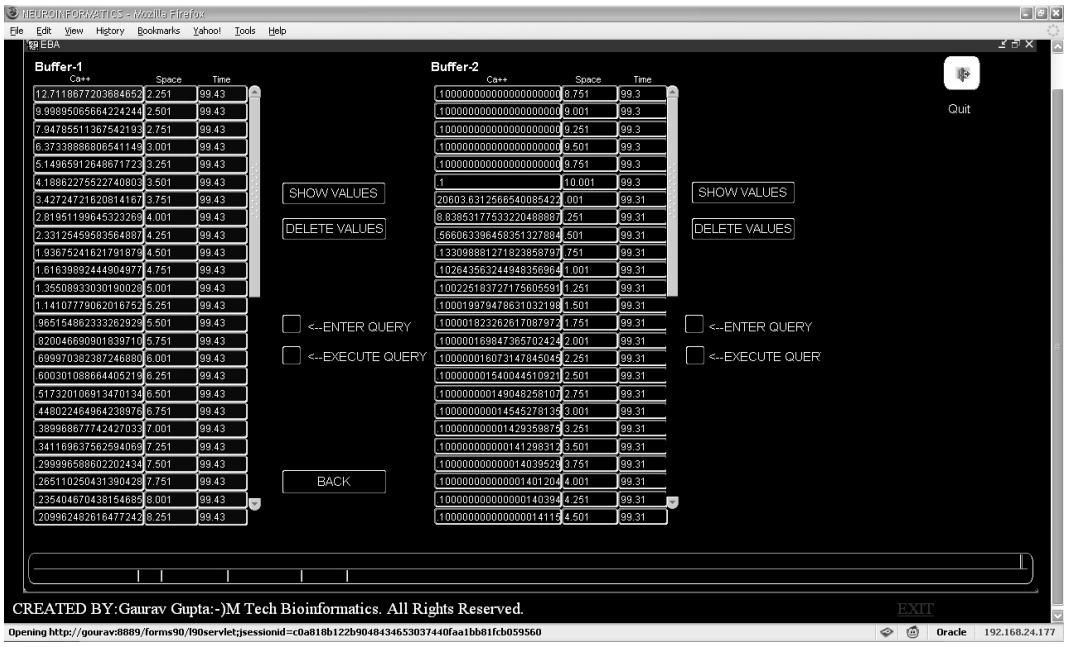


Figure 5. Webpage showing the Ca^{2+} values of buffer1 (EGTA) and buffer2 (BAPTA) after calculation has done with the parametric values shown in fig 3 of EBA

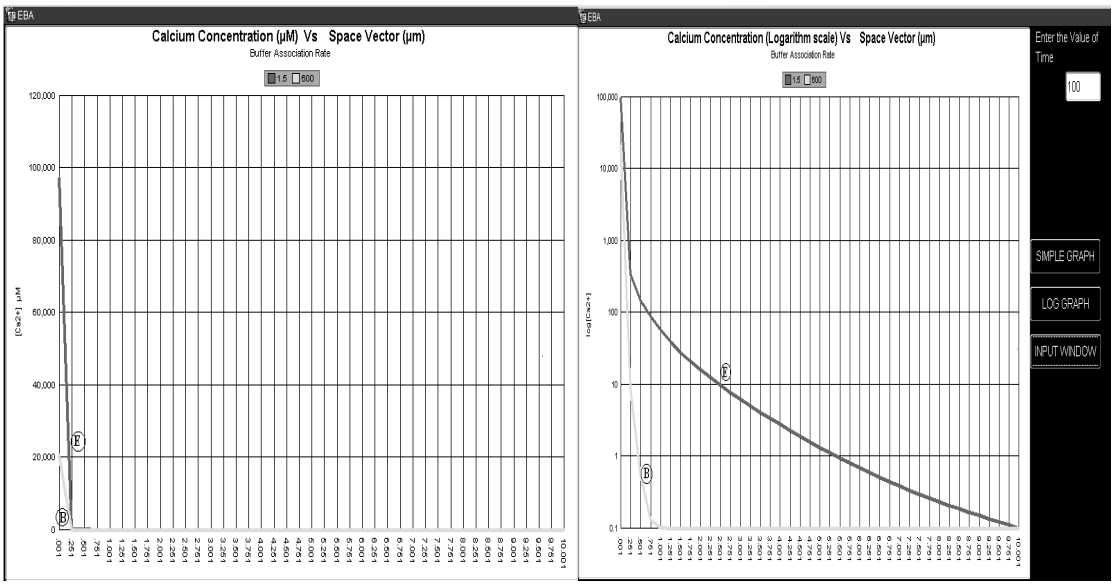


Figure 6. Cropped images showing the results of EBA in the form of simple bar graph and logarithmic graph where (E) & (B) represent EGTA and BAPTA Buffers respectively

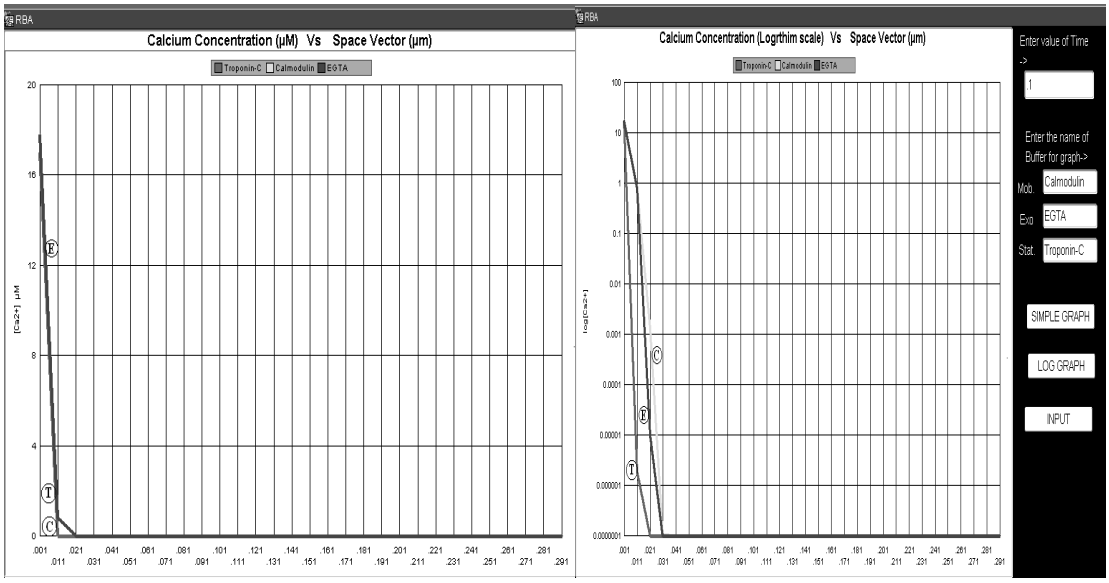


Figure 7. Cropped images showing the results of RBA in the form of simple bar graph and logarithmic graph where (T), (C) and (E) representing the Troponin-C, Calmodulin and EGTA Buffers respectively

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Radial Projection Fourier Transform and its Application for Scene Matching with Rotation Invariance

Lang Su (Corresponding author)

College of Mechatronic Engineering and Automation, National University of Defense Technology

De Ya Road, Changsha 410073, China

E-mail: sulang00@139.com

Zheng Gao

College of Mechatronic Engineering and Automation, National University of Defense Technology

De Ya Road, Changsha 410073, China

E-mail: gaozheng1605@163.com

Abstract

Scene matching is used for image registration in many fields. There are usually translation and an arbitrary unknown rotation angle between reference and template images. The corresponding scene matching algorithm costs far more computing time than that with small rotation angle and translation. Conceptions of generalized vector image, gray-scale image rotation transformation, gray-scale image point transformation, nature of shift invariance and rotation invariance are proposed to form the foundation of this paper. Radial Projection Fourier Transform (RPFT) is proposed and its rotation invariance is formal proved in this paper. It is applied to the Algorithm of Scene Matching with Rotation Invariance (ASMRI).

Calculation on reference and template images can be done separately. Some works can be done before the template images are required. This can improve the matching speed at the cost of more memory.

A program to implement the proposed RPFT and ASMRI based on RPFT is coded by means of Visual C++ 6.0. The results prove that ASMRI based on RPFT is not only rotation invariant, but also more accurate and faster than traditional methods. The program can be carried out with hardware.

Keywords: Radial Projection Fourier Transform, Scene Matching, Rotation Invariance, Discrete Fourier Transform, Generalized Vector Image, Gray-scale Image Point Transform

1. Introduction

Scene matching refers to an image processing technology which can find corresponding region of template images in reference images or find the correlation between them (Zhao, Fengwei, 2002, P.110-113). The images are of the same scene taken at different times, from different perspectives, or by different sensors (Barbara Zitova, 2003, P.977–1000).

In scene matching area, image matching with translation and arbitrary unknown rotation angle is fundamental and important. There have been many researches on rotation invariant image matching, such as Mean Square Difference (MSD) matching algorithm (Pang S N, 2004, P.519-527) based on the minimum average square difference, Normalized Product (NPROD) correlation matching algorithm (Su, kang, 1997, P.1-7) based on regional gray association, etc. However, those methods can perform well only when reference and template images differ in a small rotation angle. To have a better effect for scene matching with rotation of arbitrary unknown angle, Farhan Ullah (Ullah F, 2005, P.201-209) presented orientation codes for rotation invariant template matching, but the method costs much computing time which may leads to less practicability; Wang Jingdong (Wang, Jingdong, 2005, P.6-10) presented a circular projection matching algorithm, but the correct matching probability of this method is not high enough for practice; Sun Bojiao (Sun, Bojiao, 2008, P43-48) presented a matching method based on normalized cross-correlation, but the selection of step has great effect on image registration.

Conceptions of generalized vector image, gray-scale image rotation transformation, gray-scale image point transformation, nature of shift invariance and rotation invariance etc. are proposed to form the foundation of this paper.

Radial Projection Fourier Transform (RPFT) is proposed and its rotation invariance is formal proved in this paper. It is applied to the Algorithm of Scene Matching with Rotation Invariance (ASMRI). It involves eight steps as follows:

step1: extract the edge of original reference image to form reference edge image;

step2: extract Subsets of Suitable Matching Points (SSMP) from reference image which includes original reference image and reference edge image;

step3: calculate Radial Projection Fourier Transform (RPFT) of each point in SSMP of reference image and save it into memory;

step4: extract the edge of original template images to form template edge images;

step5: extract SSMP from template images which includes original template image and template edge image;

step6: calculate RPFT of each point in SSMP of template image;

step7: match the result of MDFT of each RPV utilizing proper comparability measurement to confirm the corresponding points;

step8: calculate the rotation angle between reference and template images.

In most scene matching, the reference image is obtained earlier. Therefore, the computations on reference image can be fully implemented offline. In this paper, the result of the computations on reference image was saved into memory. Then the matching process can be quickly done after reading in the saved data when the template images are acquired, which may greatly improve the matching speed.

This paper is organized as follows. Section 2 describes generalized vector image in detail. Section 3 describes gray-scale image point transformation and its nature of shift invariance and rotate invariance. In section 4, RPFT is proposed and its nature of rotation invariance is formal proved. In section 5, ASMRI based on RPFT is presented. In section 6, experiments of ASMRI based on RPFT are presented and compared with that of tradition methods. Finally, conclusions are drawn in section 7.

2. Generalized Vector Image

In order to interpret RPFT better, the formal definitions of concepts of gray-scale image, gray-scale image shifting transformation, shifting image, gray-scale image rotation transformation, set of rotation transformations of gray-scale image, rotating image, rotation corresponding point, generalized image, N-dimensional generalized vector image, vector image shifting transformation, vector shifted image, vector image rotation transformation and vector rotated image are provided first basing on the describing method of Mathematical Logic and Set Theory (Geng, Suyun, 2002, Shi, Chunyi, 2000)

Definition 1: A *gray-scale image* is a map from R^2 to R, where R is the real number domain. In other words, a gray-scale image is a real function of two variables, it can be noted as $f(x, y)$ or $f : R^2 \to R$. $f(x_0, y_0)$ means the grey value of the pixel with the coordinate of (x_0, y_0). The set of all gray-scale images is noted as GI.

Definition 2: If a map from GI to GI satisfies Proposition 1, the map can be referred to as a *gray-scale image shift transformation* with the shift of (dx, dy). It can be noted as $SHIFT_{dx,dy}$, i.e. $SHIFT_{dx,dy} : GI \to GI$.

Proposition 1: If $g(x, y) = SHIFT_{dx,dy}[f(x, y)]$, then

$$\forall i, j \in R : g(i, j) = f(i + dx, j + dy). \tag{1}$$

Definition 3: If $f(x, y)$ and $g(x, y)$ satisfies

$$g(x, y) = SHIFT_{dx,dy}[f(x, y)], \tag{2}$$

$g(x, y)$ can be noted as a *shifting image* of $f(x, y)$ with the shift of (dx, dy). That is to say, $g(x, y)$ is the result of the shift of (dx, dy) on $f(x, y)$.

Definition 4: If a map from GI to GI satisfies Proposition 2 or 3, the map can be noted as a *gray-scale image rotation transformation* with the rotation of (x_0, y_0) and the angle of α. It can be noted as $ROT_{x_0,y_0,\alpha}$.

Proposition 2: If $g(x, y) = ROT_{x_0,y_0,\alpha}[f(x, y)]$, then $\forall x_1, x_2, y_1, y_2 \in R$,

$$\begin{cases} x_2 - x_0 = (x_1 - x_0)\cos\alpha - (y_1 - y_0)\sin\alpha \\ y_2 - y_0 = (x_1 - x_0)\sin\alpha + (y_1 - y_0)\cos\alpha \end{cases} \to g(x_2, y_2) = f(x_1, y_1). \tag{3}$$

Proposition 3: If $g(x,y) = ROT_{x_0,y_0,\alpha}[f(x,y)]$, then $\forall \rho_1, \rho_2, \theta_1, \theta_2 \in R$,

$$\begin{cases} \rho_2 = \rho_1 \\ \theta_2 = \theta_1 + \alpha \end{cases} \rightarrow g(\rho_2 \cos\theta_2 + x_0, \rho_2 \sin\theta_2 + y_0) = f(\rho_1 \cos\theta_1 + x_0, \rho_1 \sin\theta_1 + y_0). \quad (4)$$

Definition 5: A *set of rotation transformations of gray-scale image* is defined as:

$$ROTSet = \bigcup_{x\in R, y\in R, \alpha\in R} \{ROT_{x,y,\alpha}\}. \quad (5)$$

In other words, set of rotation transformations of gray-scale image is a set of gray-scale image rotation transformations with the arbitrary rotation of (x, y) or with any rotate angle of α.

Definition 6: If $f(x,y)$ and $g(x,y)$ satisfies

$$g(x,y) = ROT_{x_0,y_0,\alpha}[f(x,y)], \quad (6)$$

$g(x,y)$ can be noted as *rotating image* of $f(x,y)$ with the arbitrary rotation of (x_0, y_0) and the rotate angle of α.

Definition 7: If (x_2, y_2) and (x_1, y_1) satisfies

$$\begin{cases} x_2 - x_0 = (x_1 - x_0)\cos\alpha - (y_1 - y_0)\sin\alpha \\ y_2 - y_0 = (x_1 - x_0)\sin\alpha + (y_1 - y_0)\cos\alpha \end{cases}, \quad (7)$$

point (x_2, y_2) can be noted as *rotation corresponding point* of point (x_1, y_1) based on $ROT_{x_0,y_0,\alpha}$.

Definition 8: A *generalized image* is a map from R^2 to S, in which S is a set. Most transformation of image can be described by the conception of generalized image. The form of S can be different in different applications.

Definition 9: *N - dimensional generalized vector image* is a kind of generalized image, "generalized vector image" for short. It is a map from R^2 to V_N, where V_N is N-dimensional Euclidean Vector Space, which is a set of all N-dimensional vectors. For example, a gray-scale image is a one-dimensional generalized vector image; a color image is a three-dimensional generalized vector image; Radial Projection Vector Field expatiated in this paper is a 512-dimensional generalized vector image. That is to say, N-dimensional generalized vector image is a vector function of two variables. It can be noted as $\boldsymbol{f}(x,y)$, i.e. $\boldsymbol{f}: R^2 \rightarrow V_N$.

Definition 10: If a map from VI_N to VI_N satisfies Proposition 4, the map can be called a *vector image shifting transformation* with the shift of (dx, dy).It can also be noted as $\boldsymbol{SHIFT}_{dx,dy}$.

Proposition 4: If $\boldsymbol{g}(x,y) = \boldsymbol{SHIFT}_{dx,dy}[\boldsymbol{f}(x,y)]$, then

$$\forall i, j \in R: \boldsymbol{g}(i,j) = \boldsymbol{f}(i + dx, j + dy). \quad (8)$$

Definition 11: If N-dimensional generalized vector image $\boldsymbol{f}(x,y)$ and $\boldsymbol{g}(x,y)$ satisfies

$$\boldsymbol{g}(x,y) = \boldsymbol{SHIFT}_{dx,dy}[\boldsymbol{f}(x,y)], \quad (9)$$

$\boldsymbol{g}(x,y)$ can be noted as *vector shifted image* of $\boldsymbol{f}(x,y)$ with the shift of (dx, dy).

Definition 12: If a map from VI_N to VI_N satisfies Proposition 5 or 6, the map can be noted as a *vector image rotation transformation* with the shift of (dx, dy).It can be noted as $\boldsymbol{ROT}_{x_0,y_0,\alpha}[\boldsymbol{f}(x,y)]$.

Proposition 5: If $\boldsymbol{g}(x,y) = \boldsymbol{ROT}_{x_0,y_0,\alpha}[\boldsymbol{f}(x,y)]$, then $\forall x_1, x_2, y_1, y_2 \in R$,

$$\begin{cases} x_2 - x_0 = (x_1 - x_0)\cos\alpha - (y_1 - y_0)\sin\alpha \\ y_2 - y_0 = (x_1 - x_0)\sin\alpha + (y_1 - y_0)\cos\alpha \end{cases} \rightarrow \boldsymbol{g}(x_2, y_2) = \boldsymbol{f}(x_1, y_1). \quad (10)$$

Proposition 6: If $\boldsymbol{g}(x,y) = \boldsymbol{ROT}_{x_0,y_0,\alpha}[\boldsymbol{f}(x,y)]$, then $\forall \rho_1, \rho_2, \theta_1, \theta_2 \in R$,

$$\begin{cases} \rho_2 = \rho_1 \\ \theta_2 = \theta_1 + \alpha \end{cases} \rightarrow \boldsymbol{g}(\rho_2 \cos\theta_2 + x_0, \rho_2 \sin\theta_2 + y_0) = \boldsymbol{f}(\rho_1 \cos\theta_1 + x_0, \rho_1 \sin\theta_1 + y_0) \tag{11}$$

Definition 13: If N-dimensional generalized vector image $\boldsymbol{f}(x, y)$ and $\boldsymbol{g}(x, y)$ satisfies

$$\boldsymbol{g}(x, y) = \boldsymbol{ROT}_{x_0, y_0, \alpha}[\boldsymbol{f}(x, y)], \tag{12}$$

$\boldsymbol{g}(x, y)$ can be called *vector rotated image* of $\boldsymbol{f}(x, y)$ with the rotation of (x_0, y_0) and rotate angle of α.

3. Gray-scale image point transformation

The formal concept of gray-scale image point transformations is proposed first, and then its transformation invariance and rotation invariance is defined.

Definition 14: A *gray-scale image point transformation* is a map from GI to VI_N. It can be noted as $GIPT$, i.e. $GIPT : GI \rightarrow VI_N$.

Definition 15: A *set of gray-scale image point transformations* is a set of all gray-scale image point transformations, it can be noted as $GIPTSet$.

Definition 16: A gray-scale image point transformation possesses the nature of *shift invariance*, if the shift of input image leads to the same shift in output image. That is to say,

if $GIPT \in GIPTSet$ is shift invariant, and $\forall f(x, y) \in GI, g(x, y) \in GI, dx \in R, dy \in R$,

$$g(x, y) = SHIFT_{dx, dy}[f(x, y)] \rightarrow GIPT[g(x, y)] = \boldsymbol{SHIFT}_{dx, dy}[GIPT[f(x, y)]]. \tag{13}$$

Definition 17: A gray-scale image point transformation possesses the nature of *rotation invariance*, if the rotation of input image leads to the same rotation in output image. That is to say,

if $GIPT \in GIPTSet$ is rotation invariant , and $\forall f(x, y) \in GI, g(x, y) \in GI, dx \in R, dy \in R$,

$$g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)] \rightarrow GIPT[g(x, y)] = \boldsymbol{ROT}_{x_0, y_0, \alpha}[GIPT[f(x, y)]]. \tag{14}$$

4. Radial Projection Fourier Transform

4.1 The definition of RPFT

The definition of Radial Projection Transform, Radial Projection Vector, and Radial Projection Vector Field is given first, and then Radial Projection Fourier Transform is proposed based on DFT and modulus operator.

With regard to N-dimensional vector $\boldsymbol{x}$, $\boldsymbol{x} = \{\boldsymbol{x}[n]\}_{n=0}^{N-1}$, and the nth element is noted as $\boldsymbol{x}[n]$.

Definition 18: *Radial Projection Vector* (RPV) is a kind of N-dimensional vector. If the rectangular coordinate of point A in grey-level image $f(x, y)$ is (x_0, y_0), then RPV of point A in grey-level image $f(x, y)$ can be noted as $\boldsymbol{RPV}_{f,A}$. $\boldsymbol{RPV}_{f,A}[n]$ is calculated as formula 15.

$$\boldsymbol{RPV}_{f,A}[n] = \sum_{\rho=1}^{\text{radius}} f(\rho \cos\theta + x_0, \rho \sin\theta + y_0), n = 0, 1, ..., N-1 \tag{15}$$

Where $\theta = \frac{2\pi}{N} n$. *radius* is the radius of the rotundity neighborhood of point A.

That is to say, RPV of a pixel point is to project grey-level value of its rotundity neighborhood. According to the method in formula 15, grey-level value of different angle is projected to vector of specifically dimensions.

As is shown in Fig1, (a) is grey-level image $f(x, y)$, in which, a specified point A is pointed by a white line; (b) is Schematic diagram of Radial Projection of point A; (c) is $\boldsymbol{RPV}_{f,A}$.

Definition 19:

Radial Projection Transform (RPT) is a kind of $GIPT$, it belongs to $GIPTSet$.

$$RPT : GI \rightarrow VI_N. \tag{16}$$

RPT: Calculate corresponding RPVs all points in gray-level image.

Radial Projection Vector Field (RPVF) is the result of RPF on each point of gray-level image $f(x, y)$.It can be noted

as RPT_f, or $RPT_f = RPT(f)$. It is a kind of N - dimensional generalized vector image. To any pixel point A in image $f(x, y)$,

$$RPT_f(A) = \boldsymbol{RPV}_{f,A}. \tag{17}$$

It is easy to prove that RPF is shift invariant, but not rotation invariant.

Definition 20:

Modulus Discrete Fourier Transform of vector (MDFT) is a map from N-dimensional Euclidean Space to N-dimensional Euclidean Space, i.e. $MDFT: R^N \to R^N$. (Alanv. Oppenheim, 1998, Guan, Zhizhong, 2004)

To N-dimensional vector $\boldsymbol{x}$, if $MDFT(\boldsymbol{x})=\boldsymbol{XM}$, then

$$\boldsymbol{XM}[k] = \left| \frac{1}{N} \sum_{n=0}^{N-1} x[n]\, W_N^{kn} \right|, \qquad k = 0, 1, 2 \ldots\ldots N-1 \tag{18}$$

where $W_N = \exp[-j2\pi / N]$.

Definition 21:

Point-by-point Modulus Discrete Fourier Transform of vector (PMDFT) is the result of calculating MDFT of each point in a generalized vector image, it is a map from VI_N to VI_N. It can be noted as $PMDFT: VI_N \to VI_N$.

If $\boldsymbol{g} = PMDFT(\boldsymbol{f})$, $\forall (a, b) \in R^2 : \boldsymbol{g}(a, b) = MDFT(\boldsymbol{f}(a, b))$.

Definition 22:

Radial Projection Fourier Transform (RPFT) is the composite map of RPT and PMDFT.

$$GI \underbrace{\xrightarrow{RPT} VI_N \xrightarrow{PMDFT}}_{RPFT} VI_N \tag{19}$$

4.2 Proof of rotation invariance of RPFT

Definition 23:

Circular shift (Alanv. Oppenheim, 1998) on N-dimensional vector $\boldsymbol{x}$ with the shift of m is still a N-dimensional vector, it can be noted as $\tilde{\boldsymbol{x}}_m$, and

$$\forall n \in [0, N-1], \tilde{\boldsymbol{x}}_m[n] = \boldsymbol{x}[(n-m)(\text{mod } N)]. \tag{20}$$

Where m, n is integer, i.e. $m, n \in Z$. That is to say, there is a circular shift with the shift of m between $\tilde{\boldsymbol{x}}_m$ and $\boldsymbol{x}$, i.e. $\tilde{\boldsymbol{x}}_m$ is the *circular shift vector* of $\boldsymbol{x}$ with the shift of m.

Lemma 1:

If $g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)]$, where $\alpha = \frac{2\pi}{N} m, m \in Z$, point A and B is rotation corresponding point on $ROT_{x_0, y_0, \alpha}$, $\boldsymbol{RPV}_{f,A}$ and $\boldsymbol{RPV}_{g,B}$ is corresponding RPV, then there is a circular shift with the shift of m between $\boldsymbol{RPV}_{f,A}$ and $\boldsymbol{RPV}_{g,B}$.

It can be formally described as:

$$g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)] \to (\boldsymbol{x} = \boldsymbol{RPV}_{f,A} \to \tilde{\boldsymbol{x}}_m = \boldsymbol{RPV}_{g,B}) \tag{21}$$

Proof:

According to Definition 6 about rotating image, if $g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)]$,

with regard to corresponding rotation points $A(\rho_1, \theta_1)$ and $B(\rho_2, \theta_2)$,

$$\begin{cases} \rho_2 = \rho_1 \\ \theta_2 = \theta_1 + \alpha \end{cases} \to g(\rho_2 \cos\theta_2 + x_0, \rho_2 \sin\theta_2 + y_0) = f(\rho_1 \cos\theta_1 + x_0, \rho_1 \sin\theta_1 + y_0) \tag{22}$$

According to Definition 15 about RPV, $\forall n \in [0, N-1]$,

$\boldsymbol{RPV}_{f,A}[n] = \sum_{\rho=1}^{\text{radius}} f(\rho\cos\theta + x_0, \rho\sin\theta + y_0)$, $\boldsymbol{RPV}_{g,B}[n] = \sum_{\rho=1}^{\text{radius}} g(\rho\cos\theta + x_0, \rho\sin\theta + y_0)$,

where $\theta = \frac{2\pi}{N}n$.

So, there is a circular shift between $\boldsymbol{RPV}_{f,A}$ and $\boldsymbol{RPV}_{g,B}$, and the shift equals to $\frac{\alpha N}{2\pi}$.

In conclusion, Formula 19 is proved.

Lemma 2:

If there is a circular shift between N-dimensional vector $\boldsymbol{y}$ and $\boldsymbol{x}$, then the result of MDFT of $\boldsymbol{y}$ equal to that of $\boldsymbol{x}$ .It can be formal described as: if $\boldsymbol{XM} = MDFT(\boldsymbol{x})$, $\boldsymbol{YM} = MDFT(\boldsymbol{y})$,

$$\forall m \in [0, N-1],\ \boldsymbol{y} = \tilde{\boldsymbol{x}}_m \rightarrow \forall k \in [0, N-1], \boldsymbol{YM}[k] = \boldsymbol{XM}[k]. \tag{23}$$

Proof:

According to the property of circular shift (Alanv. Oppenheim, 1998),

$\& \forall n \in [0, N-1],\ \boldsymbol{y}[n] = \boldsymbol{x}[(n-m)(\text{mod } N)]$.

$\forall k \in [0, N-1]$,

$$\begin{aligned}
\boldsymbol{YM}[k] &= \left|\sum_{n=0}^{N-1} \boldsymbol{y}[n]W^{kn}\right| = \left|\sum_{n=0}^{N-1} \boldsymbol{x}[(n-m)(\text{mod } N)]W^{kn}\right| \\
&= \left|\sum_{n=0}^{m-1} \boldsymbol{x}[(n-m)(\text{mod } N)]W^{kn}\right| + \left|\sum_{n=m}^{N-1} \boldsymbol{x}[(n-m)(\text{mod } N)]W^{kn}\right| \\
&= \left|\sum_{n=0}^{m-1} \boldsymbol{x}[N+n-m]W^{kn}\right| + \left|\sum_{n=m}^{N-1} \boldsymbol{x}[n-m]W^{kn}\right| \\
&= \left|\sum_{l=N-m}^{N-1} \boldsymbol{x}[l]W^{k(m+l)}\right| + \left|\sum_{l=0}^{N-1-m} \boldsymbol{x}[l]W^{k(m+l)}\right| = \left|\sum_{l=0}^{N-1} \boldsymbol{x}[l]W^{k(m+l)}\right| \\
&= \left|W_N^{mk}\sum_{l=0}^{N-1} \boldsymbol{x}[l]W^{kl}\right| = \left|\sum_{l=0}^{N-1} \boldsymbol{x}[l]W^{kl}\right| = \boldsymbol{XM}[k].
\end{aligned}$$

$$\boldsymbol{YM}[k] = \boldsymbol{XM}[k]. \tag{24}$$

The MDFT of rotation corresponding points can keep invariance.

Theorem 1: Rotation Invariance of RPFT

If $g(x,y) = ROT_{x_0,y_0,\alpha}[f(x,y)]$, and $\alpha = \frac{2\pi}{N}n$, then with regard to rotation corresponding points $A(x_1, y_1)$ and $B(x_2, y_2)$, the RPFT of them can keep invariance.

$$g(x,y) = ROT_{x_0,y_0,\alpha}[f(x,y)] \rightarrow \boldsymbol{RPFT}[g(x,y)] = \boldsymbol{ROT}_{x_0,y_0,\alpha}[\boldsymbol{RPFT}[f(x,y)]] \tag{25}$$

Proof:

Combination at Definition 1-23,

According to formula 19, 20,

$g(x,y) = ROT_{x_0,y_0,\alpha}[f(x,y)] \rightarrow (\boldsymbol{x} = \boldsymbol{RPV}_{f,A} \rightarrow \tilde{\boldsymbol{x}}_m = \boldsymbol{RPV}_{g,B})$.

Calculate the MDFT of RPV of point A and B.

According to formula 22, if $\boldsymbol{XM} = MDFT(\boldsymbol{x})$, $\boldsymbol{YM} = MDFT(\boldsymbol{y})$,

$\forall m \in [0, N-1],\ \boldsymbol{y} = \tilde{\boldsymbol{x}}_m \rightarrow \forall k \in [0, N-1], \boldsymbol{YM}[k] = \boldsymbol{XM}[k]$.

$\boldsymbol{RPFT}[f(x_1, y_1)] = (\boldsymbol{XM}[0], \boldsymbol{XM}[1], ..., \boldsymbol{XM}[N-1])$,

$\boldsymbol{RPFT}[g(x_2, y_2)] = (\boldsymbol{YM}[0], \boldsymbol{YM}[1], ..., \boldsymbol{YM}[N-1])$,

So, $\boldsymbol{RPFT}[g(x_2, y_2)] = \boldsymbol{RPFT}[f(x_1, y_1)]$.

According to the randomicity of A and B, and $g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)]$,

$\boldsymbol{RPFT}[g(x, y)] = \boldsymbol{ROT}_{x_0, y_0, \alpha}[\boldsymbol{RPFT}[f(x, y)]]$.

In conclusion, Theorem 1 is proved.

5. Implement of arithmetic of RPFT

Implement of arithmetic of RPFT consists of three steps.

step1: set up polar coordinates on each point of gray-scale image $f(x, y)$, and uses a reverse transform method [12] to calculate grey level in polar coordinates corresponding to Cartesian coordinates;

step2: calculate RPVF of image $f(x, y)$;

step3: calculate PMDFT of image $f(x, y)$.

Overview of this method is illustrated in Figure 2.

5.1 Inverse transform in polar coordinates

To find the correspondence between Cartesian coordinate system and polar coordinates, the general method is to transform Cartesian coordinates to polar coordinates: calculate the distance and angle (ρ, θ) mapping from the position of given pixel (x, y).

Polar coordinates system (Wang, Qi, 2006, P.6-7):

$$\begin{cases} \rho = \sqrt{(x - x_0)^2 + (y - y_0)^2} \\ \theta = \arctan((y - y_0) / (x - x_0)) \end{cases} \tag{26}$$

where x_0, y_0 refers to the coordinates of a point(noted as K) in Cartesian coordinates which is the origin of polar coordinates.

Combination at the shortcomings of conventional methods, in this paper, reverse polar coordinates transform algorithm (Wang, Qi, 2006, P.6-7) is presented, which is able to overcome the above disadvantages.

Specific ideas are: utilizing polar coordinates of pixel (ρ, θ) to calculate its corresponding value of each axis in the Cartesian coordinates (x, y). That is,

$$\begin{cases} x = \rho \cos\theta + x_0 \\ y = \rho \sin\theta + y_0 \end{cases} \tag{27}$$

Schematic diagram is shown in Figure 3 (Wang, Qi, 2006, P.6-7, Karl C. Walli, 2005, P.8).

Generally, location of corresponding points in Cartesian coordinates has a decimal value, it locates between four pixels. Based on the idea of interpolation (Karl C. Walli, 2005, P.8), the original Cartesian coordinate system can construct out of an "optimal equivalence" of pixels. Its concrete steps are as follows:

1) The reciprocal of distances from the four surrounding pixel points to the actual coordinates can be defined as the weight proportion of each surrounding pixel points respectively. The distances are noted as d_1, d_2, d_3, d_4.

2) Calculate the value of $F(x, y)$ (Karl C. Walli, 2005, P.8)

$$F(x, y) = \frac{F(I(x), I(y)) / d_1 + F(I(x)+1, I(y)) / d_2 + F(I(x), I(y)+1) / d_3 + F(I(x)+1, I(y)+1) / d_4}{1/d_1 + 1/d_2 + 1/d_3 + 1/d_4} \tag{28}$$

where $I(x)$ refers to the maximum integer less than x.

Inverse transform in polar coordinates is shown in Figure 4. (a) shows input image and Cartesian coordinate with the origin of point K (240,300), (b) shows inverse transform in polar coordinates with the origin of K.

After these transformations, each point of polar coordinates has an “optimal” corresponding grey value in Cartesian coordinate, which can guarantee that each polar coordinate point has only a corresponding value.

5.2 Radial Projection Vector

Set up polar coordinates with the origin of arbitrary point A in image $f(x, y)$. According to Definition 18, RPV of point A in $f(x, y)$ as $\boldsymbol{RPV}_{f,A}$ is defined as:

$$\boldsymbol{RPV}_{f,A}[n] = \sum_{\rho=1}^{\text{radius}} f(\rho\cos\theta + x_0, \rho\sin\theta + y_0), \boldsymbol{n} = 0, 1, ..., N-1 \tag{29}$$

where $\theta = \frac{2\pi}{N} n$ and *radius* is the radius of neighborhood area of point A.

512-dimensional vector is adopted to denote the RPV of each point in this paper. Because DFT is designed to meet the precondition that points of signal N satisfies $N = 2^t$. If a RPV is the circular shift vector of another one, RPV of 360 elements can not satisfy the nature of circular shift when $N = 2^t$ after extension.

According to Lemma 1, if there is a rotation of unknown angle between reference image and template image, RPVs of rotation corresponding points differ in circular shift.

$$g(x, y) = ROT_{x_0, y_0, \alpha}[f(x, y)] \rightarrow (\boldsymbol{x} = \boldsymbol{RPV}_{f,A} \rightarrow \tilde{\boldsymbol{x}}_m = \boldsymbol{RPV}_{g,B}) \tag{30}$$

where $m = \frac{\alpha N}{2\pi}$.

As it is shown in Figure 5, (b) is an image of the size 300×250 pixels which is a subscene of an image which only differs in an unknown rotation angle with (a). (c) and (d) is respectively the RPV of rotation corresponding points A and B .As can be seen, there is circular shift between the corresponding RPV of points A and B .

5.3 Modulus Discrete Fourier Transform

The computation of DFT of N-dimensional vector is $O(N^2)$, which could not meet the needs of practical applications. The computation of FFT of N-dimensional vector is $O(N \log_2^N)$ .To speed up the implementation process, FFT is used to implement DFT in the experiment.

According to Lemma 2, if the RPVs differ in a circular shift with the shift of m, the result of MDFT maintain invariance.

As is shown in Figure 6, (a) and (b) is respectively MDFT of RPVs in Figure 5(c) and Figure 5(d).

6. ASMRI based on RPFT

According to rotation invariant of RPFT, it can be applied to ASMRI.

Offline calculation on reference images involves three steps:

step1: extract the edge of original reference image to form reference edge image;

step2: extract SSMP from original reference image and reference edge image respectively;

step3: calculate RPFT on each point from SSMP of reference image and calculate corresponding MDFT.

Calculation after acquiring template images consists of five steps:

step1: extract the edge of original template images to form template edge images;

step2: extract SSMP from template images, select several points from SSMP of original template and template edge images with adjacent coordinate for matching;

step3: calculate RPFT on each point from SSMP of template image and calculate MDFT on each point;

step4: match the result of RPFT of each selected points in template images with that of reference image utilizing proper comparability measurement to confirm the corresponding points;

step5: calculate the rotation angle between reference and template images;

Overview of those steps is illustrated in Figure 7.

6.1 Extraction of edge images

The edge of the images reflects step change of grey level of adjacent pixels. In this paper, the operator of Prewitt with orientation is adopted (Lu, Zongqi, 2006, P192). The result of this processing keeps the form of gray-level images. Templates of Prewitt with orientation are shown in Figure 8.

To each pixel, calculate the gradient of eight orientations, and choose the maximum of them as the output value.

As shown in Figure 9, (b) is the grey level edge image corresponding to (a), and (d) is the grey level edge image corresponding to (c).

6.2 Extraction of SSMP

For some points which are poor in grey level changes with surrounding pixels, the mistake matching probability may be unacceptable. To enhance the matching accuracy and reduce the matching time, SSMP are extracted from images. In this paper, method based on Harris corner extraction method (C.Harris, 1988, P.147-151) is adopted.

Extracting SSMP in template images involves four steps:

1) Filter to each pixel in the image utilizing horizontal and vertical difference operator to confirm I_x and I_y, which stands for horizontal and vertical changes, and calculate matrix $M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{bmatrix}$;

2) Perform Gauss smoothing filter to four elements in matrix M, and update M. Where $Gauss = \exp(-\frac{x^2 + y^2}{2\sigma^2})$;

3) Calculate corner value of every pixel. $Corner\ value = \frac{I_x^2 I_y^2 - (I_x I_y)^2}{I_x^2 + I_y^2}$;

4) Extract corner which satisfies "corner value is not less than 0.6 multiply the maximum corner value of all pixels" and "its corner value is the maximum of its neighboring pixels".

Extracting SSMP in reference images also involves four steps. It is the same as the method above from step 1 to step 3. Instead, step 4 is to extract corners which satisfy "corner value is not less than 0.30 multiply the maximum corner value of all pixels" and "its corner value is not less than 0.60 multiply the maximum of its neighboring pixels".

The steps of extracting SSMP are illustrated in Figure 10. (a) and (b) shows original reference image and template images, (c) and (d) shows reference edge image and template edge images, (e) and (f) shows SSMP extracted from original reference image and template images, (g) and (h) shows SSMP extracted from reference edge images and template edge images.

This can insure that SSMP from template and template edge images can find its corresponding points in reference and reference edge images.

6.3 Determination of corresponding points

After extracting SSMP and calculating their RPVs, matching can be performed to determine corresponding points.

As shown in Figure 6, the value of the both ends of MDFT is much larger than the others. The algorithm is easy to degenerate to be the algorithm of find the point of similar sum of the surrounding grey level, because for MDFT, $\boldsymbol{XM}[0] = \left| \frac{1}{N} \sum_{n=0}^{N-1} \boldsymbol{x}[n] \right|$. Thus $[\boldsymbol{RPFT}[f(x,y)]]$ and $\boldsymbol{RPFT}[g(x,y)]$ should be normalized in matching process, making each result of MDFT have equal or similar impact. $[\boldsymbol{RPFT}[f(x1,y1)]]$ and $\boldsymbol{RPFT}[g(x2,y2)]$ is respectively noted as $\boldsymbol{XM}$ and $\boldsymbol{YM}$. Combined with the method of the similarity of multi-dimensional vector, similarity measure function (Du, Jie, 2007, P11) is defined as:

$$f = \frac{\sum_{k=0}^{n-1} (\frac{\boldsymbol{XM}[k]}{\overline{\boldsymbol{XM}[k]}} - 1) \times (\frac{\boldsymbol{YM}[k]}{\overline{\boldsymbol{YM}[k]}} - 1)}{\sqrt{\sum_{k=0}^{n-1} (\frac{\boldsymbol{XM}[k]}{\overline{\boldsymbol{XM}[k]}} - 1)^2 \sum_{k=0}^{n-1} (\frac{\boldsymbol{YM}[k]}{\overline{\boldsymbol{YM}[k]}} - 1)^2}} \tag{32}$$

where $\overline{\boldsymbol{XM}[k]} = \sum_{k=0}^{jdleft} \boldsymbol{XM}[k] \Big/ jdleft$, $\overline{\boldsymbol{YM}[k]} = \sum_{k=0}^{jdright} \boldsymbol{YM}[k] \Big/ jdright$, $jdleft, jdright$ is respectively the number of points of SSMP in template and reference images.

It is the same principle in determination of corresponding point in reference edge image and template edge images.

6.4 Determination of rotation angle

After acquiring corresponding points M and M' of reference and template images, rotation angle exists between them can be determined.

Calculate corresponding PRV of M, M' separately, noted as $RPV_{f,M}$ and $RPV_{g,M'}$, and then offset $RPV_{g,M'}$ from 0 to 511. According to the offset when relevance of both vectors is the maximum of all, angle of rotation can be determined.

Angle of rotation between reference edge and template edge images can be determined using above methods.

7. Experimental Results

A program to implement the proposed RPFT and ASMRI based on RPFT is coded by means of Visual C++6.0. In this section, three scene matching experiments are presented. For each experiment, reference images, template images, matching accuracy and matching speed are shown. Experiments are done on the PC platform with a CPU of 2.8GHz. In those experiments, *raidus* is 20. The interface of the program is shown in Figure 11.

7.1 Experiment 1: rotation between reference and template images

The first experiment demonstrates robustness of the method to scene matching with rotation between reference and template images. Rotation angle varies from 5° to 355° with interval of 5° during taking photos, totally 71 images. Four pairs of images are considered in this experiment as shown in Figure 12.

7.2 Experiment 2: rotation and grey change between reference and template images

The second experiment demonstrates robustness of the method to scene matching with rotation and grey change between reference and template images. Rotation angle varies from 5° to 355° with interval of 5°. Four pairs of images are considered in this experiment as shown in Figure 13.

7.3 Experiment 3: rotation and Gaussian noise between reference and template images

The third experiment demonstrates robustness of the method to scene matching with rotation and Gaussian noise between reference and template images. Rotation angle varies from 5° to 355° with interval of 5°. Four pairs of images are considered in this experiment as shown in Figure 14.

Four pairs of images are considered in those experiments as shown in Figure 12, 13, 14. (b), (d), (f), (h) is respectively one of the 71 images of corresponding images of (a), (c), (e) and (g). The experimental results are shown respectively in table 1, 2, 3.

As can be seen from Table 1, 2 and 3, matching accuracy rate of this algorithm is higher than the circular projection algorithm (Wang, Jingdong, 2005, 6-10). When the template images have a change in brightness or are interfered by Salt-Pepper noise, matching accuracy rate of circular projection algorithm is not high, while the algorithm proposed in this paper can still achieve a high matching rate.

8. Conclusion

In scene matching, there is usually a difference in rotation angle between reference and template images, which may lead to mistake matching. The algorithm proposed in this paper maintains a very high matching accuracy with arbitrary rotation angle and keep robust with rotation and gray change, rotation and Gaussian noise in template image. The analysis of the reference image have been completed earlier, together with the RPV is only operated by 1D-DFT, so the matching process needs only less computation. Combination at the different character of original image and edge image can enhance the matching accuracy.

In all, the algorithm proposed in this paper possesses the nature of rotation invariance, has a high matching rate, matching speed and strong applicability in scene matching with rotation invariance, it will surely has a wide application.

References

Alanv. Oppenheim, & Ronald W. Schafer, John R. Buck. (1998). *Discrete-time signal processing*, (2nd ed.). (Chapter 6).

An, Ru.(2005). Investigation of image matching algorithm for aircraft navigation based on corner feature. Doctor's degree paper, Nanjing University. 200505. 88.

Barbara Zitova, & Jan Flusser. (2003). Image registration methods: a survey. *Image and Vision Computing*. 21(2003). 977–1000.

C.Harris, & M. Stephens,(1988). ”A combined corner and edge detector”. Alvey Vision Conf. 147-151.

Du, Jie.(2007). Two fast image matching algorithms based on gray value. Master's degree paper, Dalian maritime university. 10-11.

Geng, Suyun, Qu, Wanlin, & Wang, Hanpin. *Tutorial of Discrete Mathmatics*. Press of PeKing University. 2002.06.

Guan, Zhizhong, Xia, Gongge, & Meng, Qiao. (2004). *Signal and linear system*. Higher Education Press, China. 114.

Karl C. Walli. (2005). Multisensor image registration utilizing the LOG filter and FWT. Master's degree paper, Rochester Institute of Technology.

Pang S N, Kim H C, & Kim D,et al. (2004). Prediction of the suitability for image-matching based on self-similarity of vision contents. *Image and Vision Computer*. Vol. 22.5. 519-527.

Shi, Chunyi, Wang, Jiaqin. (2000). *Mathematical Logic and Set Theory*. Press of TsingHua University. 2000.12. (2nd ed.)

Su, kang, Guan, Shiyi, & Liu, Jian et al. (1997). A Practical normalized cross-correlation scene matching, *Journal of Astronautics.* Vol.18.3. 1-7.

Sun, Bojiao, & Zhou, Donghua. (2008). Rotated image registration method based on NCC [J], *Transducer and Microsystems Technologies*. Vol 27.5.

Ullah F, & Kaneko S. (2004). Using orientation codes for rotation-invariant template matching [J]. *Pattern Recognition*. P201-209. Vol.37.2.

Wang, Jingdong, Xu, Yibin, & Shen, Chunlin. (2005). New scene matching method for arbitrary rotation. *Journal of Nanjing University of Aeronautics and Astronautics*. Vol. 37.1. 6-10.

Wang, Qi, Li, Yanjun. (2006). Research on Inverse Log- polar Transformation Based on Sub-pixel Interpolation. *Computer Engineering and Application*. Vol 27. 6-7.

Zhao, Fengwei, Li, Jicheng, & Shen, Zhenkang. (2002). Study of scene matching techniques. *Systems Engineering and Electronics*. Vol. 24.12. 110-113.

Table 1. Summary of the experiment 1

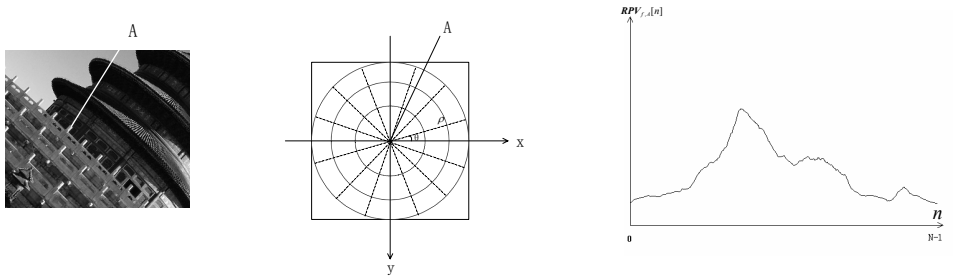
Input	Reference image		(a)	Edge of(a)	(c)	Edge of(c)	(e)	Edge of(e)	(g)	Edge of(g)
	Template image		(b)	Edge of(b)	(d)	Edge of(d)	(f)	Edge of(f)	(h)	Edge of(h)
	Image size	Reference image	600×500		428×321		428×321		250×250	
		Template image	300×250		350×250		269×250		213×212	
Results	Points in SSMP	Reference image	1090	1633	347	303	2455	1234	253	293
		Template image	34	8	27	16	123	81	30	24
	Matching accuracy		97.6%		99.3%		98.1%		99.5%	
	Average Time	Extract SSMP from template images	125 ms		156 ms		141 ms		47 ms	
		Extract template edge images	31 ms		37 ms		31 ms		26 ms	
		Extract SSMP from template-edge images	141 ms		188 ms		171 ms		60 ms	
		RPFT in template and template-edge images	62 ms		63 ms		63 ms		62ms	
		Matching	204 ms		62 ms		328 ms		47ms	
		Determine Rotate angle	162ms		167 ms		171ms		156ms	
		Total time	725 ms		673ms		905ms		398ms	

Table 2. Summary of the experiment 2

Input	Reference image		(a)	Edge of(a)	(c)	Edge of(c)	(e)	Edge of(e)	(g)	Edge of(g)
	Template image		(b)	Edge of(b)	(d)	Edge of(d)	(f)	Edge of(f)	(h)	Edge of(h)
	Image size	Reference image	600×500		428×321		428×321		250×250	
		Template image	300×250		350×250		269×250		213×212	
Results	Points in SSMP	Reference image	1090	1633	347	303	2455	1234	253	293
		Template image	35	8	31	19	128	83	33	25
	Matching accuracy		87.8%		85.3%		86.5%		92.7%	
	Total match time		723 ms		673 ms		897 ms		399ms	

Table 3. Summary of the experiment 3

Input	Reference image		(a)	Edge of(a)	(c)	Edge of(c)	(e)	Edge of(e)	(g)	Edge of(g)
	Template image		(b)	Edge of(b)	(d)	Edge of(d)	(f)	Edge of(f)	(h)	Edge of(h)
	Image size	Reference image	600×500		428×321		428×321		250×250	
		Template image	300×250		350×250		269×250		213×212	
Results	Points in SSMP	Reference image	1090	1633	347	303	2455	1234	253	293
		Template image	50	14	56	32	155	104	31	23
	Matching accuracy		91.6%		92.2%		89.2%		94.1%	
	Total match time		725 ms		680 ms		901 ms		399ms	



(a) grey-level image (b) Radial Projection (c) $\boldsymbol{RPV}_{f,A}$

Figure 1. Schematic diagram of RPV

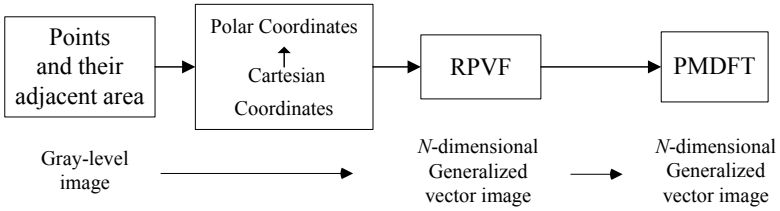


Figure 2. Overview of RPFT

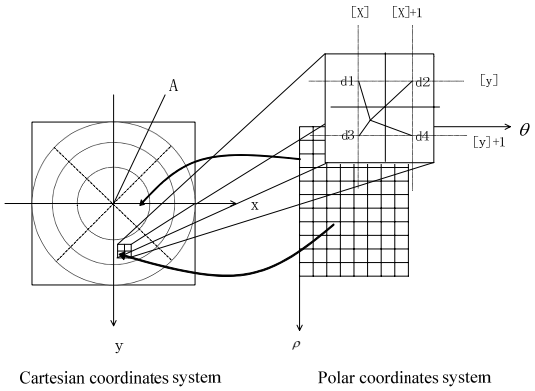
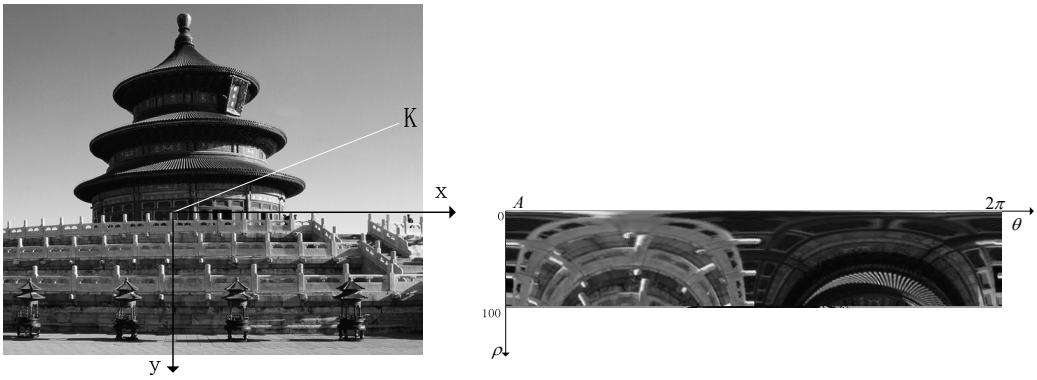
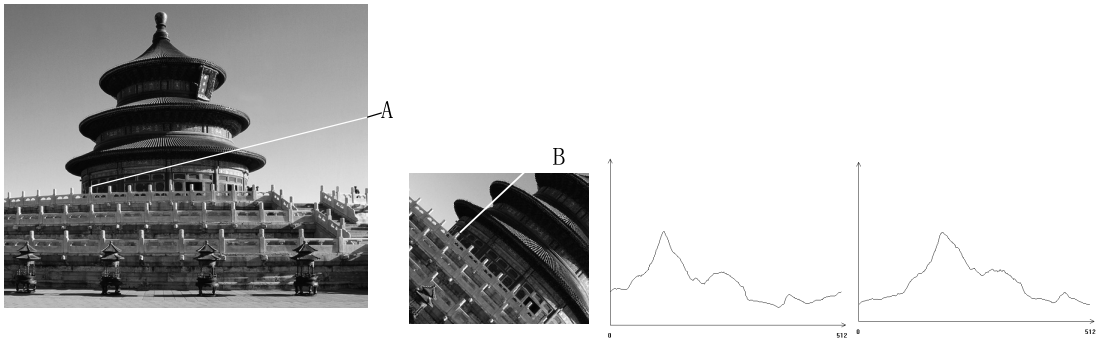


Figure 3. Schematic diagram of inverse transform in polar coordinates



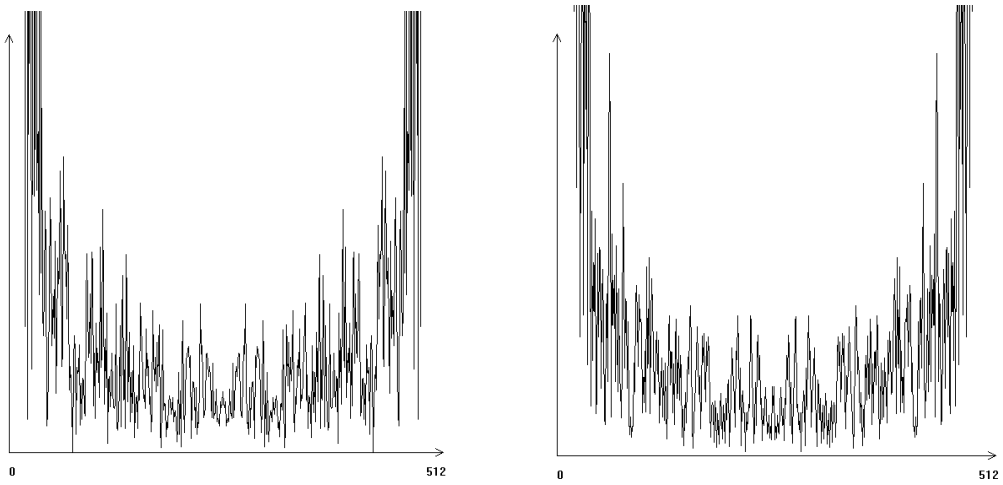
(a) (b)

Figure 4. Inverse transform in polar coordinates



(a) (b) (c) (d)

Figure 5. transform rotation angle to circular shift of RPV



(a) (b)

Figure 6. Rotation invariance of MDFT on RPVs

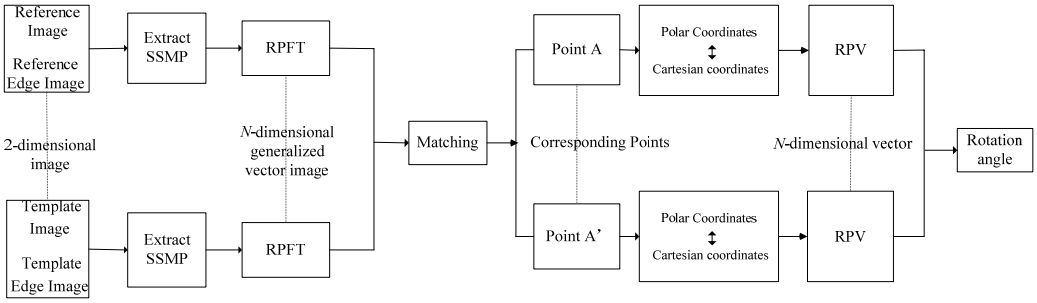


Figure 7. Overview of the ASMRI based on RPFT

1	1	1
1	-2	1
-1	-1	-1

1	1	1
1	-2	-1
1	-1	-1

1	1	-1
1	-2	-1
1	1	-1

1	-1	-1
1	-2	-1
1	1	1

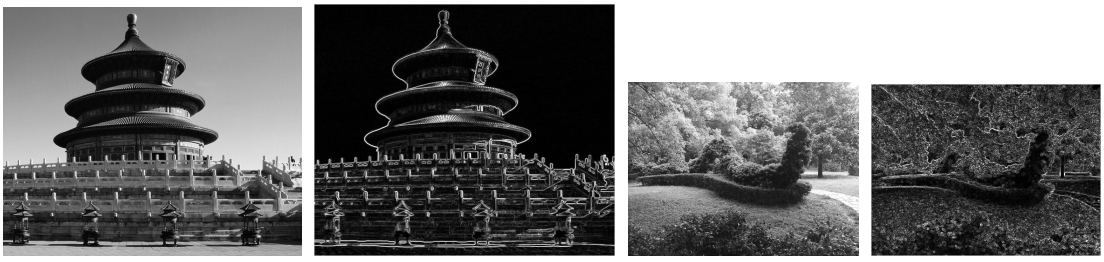
-1	-1	-1
1	-2	1
1	1	1

-1	-1	1
-1	-2	1
1	1	1

-1	1	1
-1	-2	1
-1	1	1

1	1	1
-1	-2	1
-1	-1	1

Figure 8. Templates of Prewitt with orientation



(a) (b) (c) (d)

Figure 9. Extraction of edge images using the operator of Prewitt with orientation

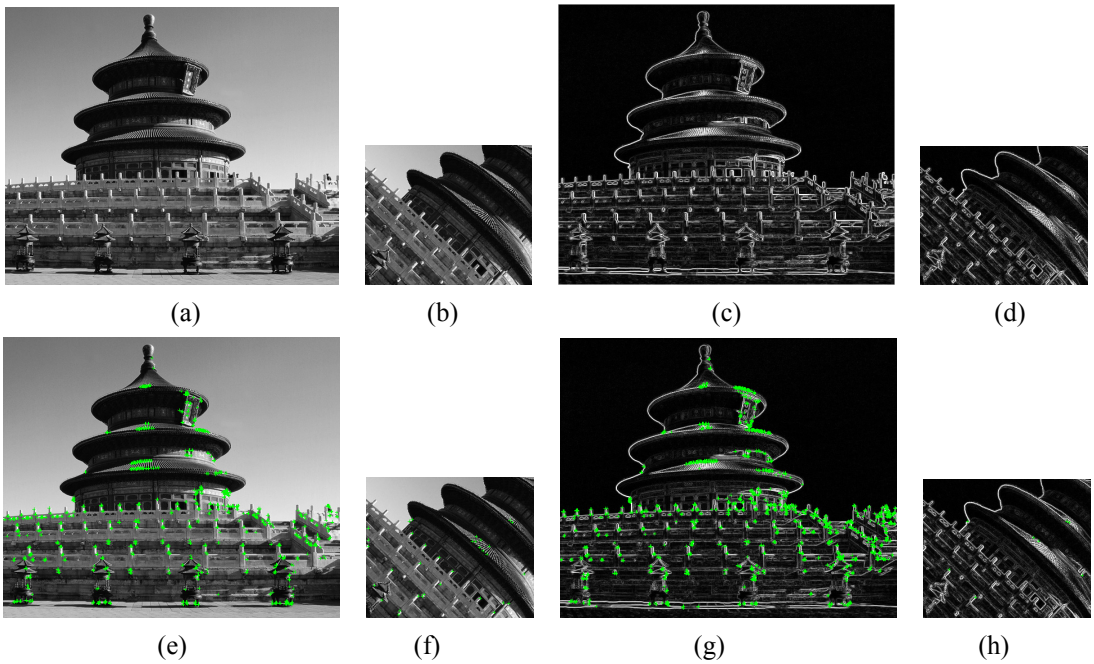


Figure 10. Schematic diagram of steps of extracting SSMP

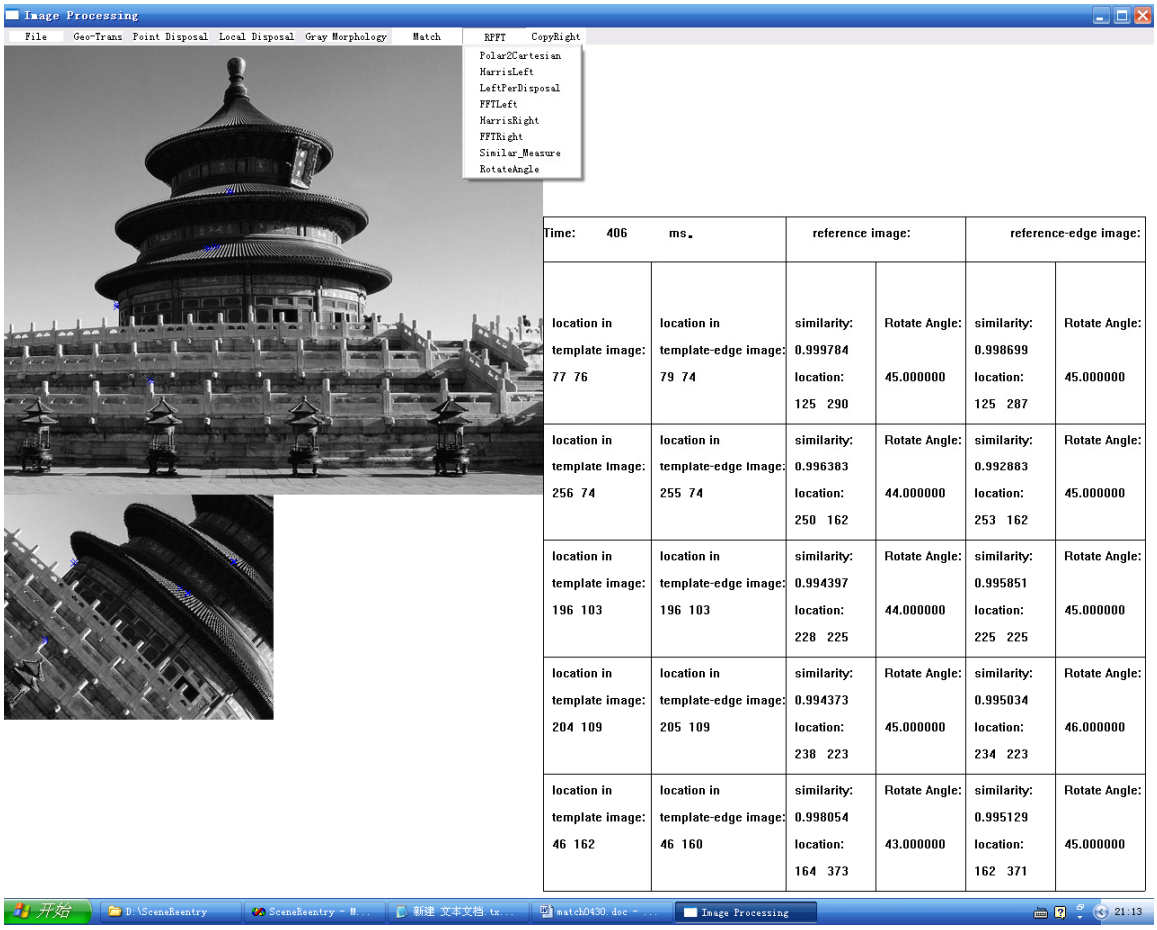
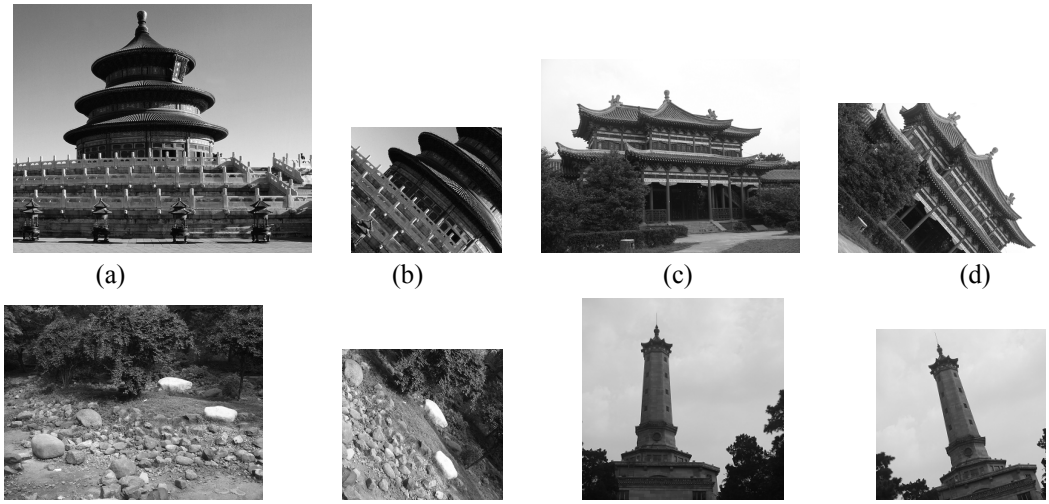


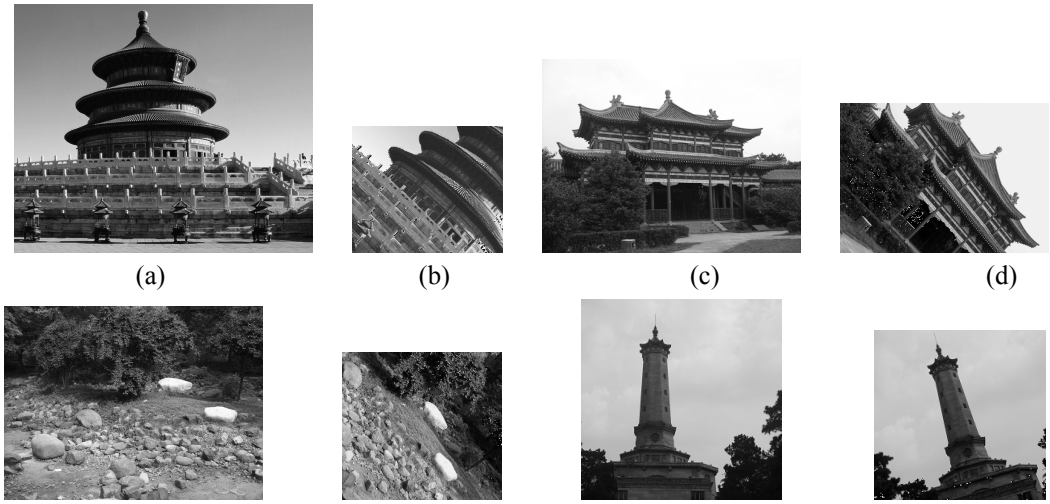
Figure 11. The interface of the program



(a) (b) (c) (d)

(e) (f) (g) (h)

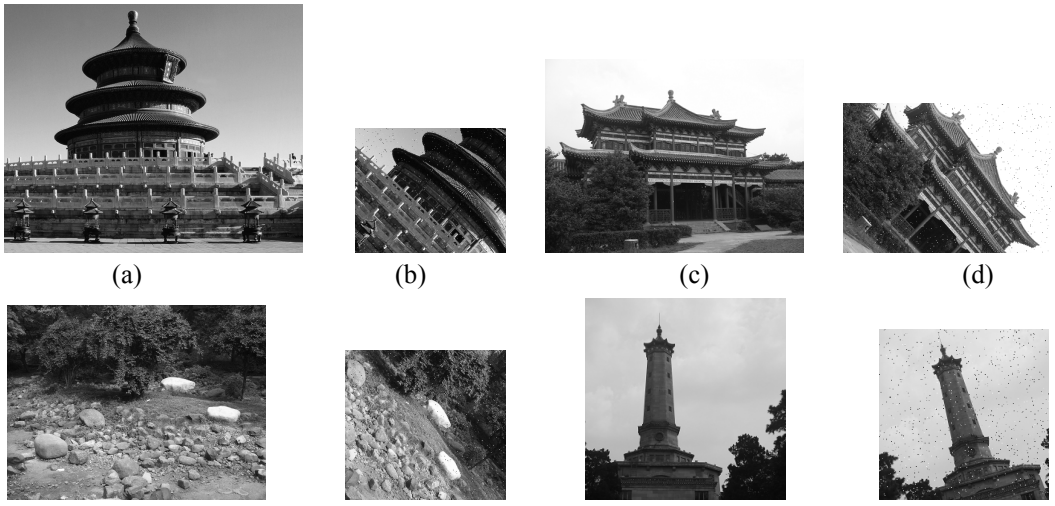
Figure 12. rotation between reference and template images



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 13. rotation and grey change between reference and template images



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 14. rotation and Gaussian noise between reference and template images

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

An Investigation into Methods and Concepts of Qualitative Research in Information System Research

Marzanah A.Jabar (Corresponding author)

Tel: 60-3-8946-6524 E-mail: marzanah@fsktm.upm.edu.my

Fatimah Sidi, Mohd Hasan Selamat, Abd. Azim Abd. Ghani & Hamidah Ibrahim

Faculty of Science Computer and InformationTechnology

University Putra Malaysia (UPM), Serdang

43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

Tel: 60-3-8946-6555 E-mail: {fatimah, hasan, azim, hamidah}@fsktm.upm.edu.my

Abstract

This paper is an initial review of literature, investigating qualitative research, to show its relevance in information system disciplines. Qualitative research involves the use of qualitative data, such as interviews, documents, and participant observation data, to understand and explain social phenomena. Qualitative research can be found in many disciplines and fields, using a variety of approaches, methods and techniques. In Information Systems (IS), there has been a general shift in information system research away from technological to managerial and organizational issues, hence an increasing interest in the application of qualitative research methods. Frequently used methods are the action research, case study, ethnography and grounded theory. Review of each research approaches in qualitative methods, will be discussed. Important considerations in the methods are identified, and cases for each research method are described. Then we will present some benefits and limitations of each method. Based on the result, a framework of an action research was proposed and might be useful in starting a research project in information system using qualitative method.

Keywords: Qualitative research, Information system, Action research, Grounded theory, Case study, Ethnography

1. Introduction

Methodology deals with the methods and principles used in an activity. Research methodology explained on how the research is done, the methods of data collection, materials used, subjects interviewed, or places visited. It details out the account of how and when the research is carried out. It also gives reasons on why a particular method is use, rather than other methods. There are many references in the IS literature of research in the area of Information System, such as the work of Alavi & Carlson (1992), Benbasat & Weber (1996), Olikowski & Iacono (2001), Yin (2002) and Myers (2006). Research in information systems is a relatively new one and without a research tradition that it can claim to be its own. In an awareness to cover other areas of the information systems spectrum, from technical perspectives of systems design and implementation, to social perspectives of the structural and social consequences of information systems, it lead the way to qualitative approaches, and the use of methodologies imported from those fields. Following a general shift in information system research away from technological to managerial and organizational issues, there is an increasing interest in the application of qualitative research methods as highlighted by Mangan (2004) and Singh et.all (2005). Matsuo et.al (2008) has compiled a series of case studies in answering research queries of an experimental learning theory in managing information theory.

The purpose of this article is to provide a perspective on the methods that could be applied to Information Systems disciplines research. This article addresses the usage of qualitative research method in Information System research. This paper is structured as follows. First we will review the research approaches in qualitative methods. Next we highlight the important considerations in the methods are identified, and cases for each research method in qualitative research are described. Then we put forward discussion on the benefits and limitations of each method and proposed a framework for a qualitative research study in IS. Finally we present some conclusions on the research methodologies in IS.

2. Qualitative Research Overview

Research methods can be classified in various ways, Myers (1997) however described that one of the most common

distinctions of research methods is between qualitative and quantitative research methods. Quantitative research methods were originally developed in the natural sciences to study natural phenomena. The focus of quantitative research is objective measures. Data is collected in an objective and replicable manner. Examples of quantitative methods include laboratory experiments, formal methods (e.g. econometrics) and numerical methods such as mathematical modeling. The tools of quantitative research include test performance scores, physiological readings, survey responses and spectrometer readings, Cresswell J. (1994), defined qualitative study as an inquiry process of understanding a social or human problem, based on a complex, holistic picture, formed with words, and reporting in a natural setting. Guba & Lincoln (2000), classified IS research paradigm as positivist, interpretive and engineering as define by Clark (1992) in Table 1. The engineering paradigm in information system is applied here as the result of the development of the application, testing, technology, conceptualization, and prototyping in IS. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions according to Myers (2006) and define with the following characteristics: exploratory, descriptive, emergent, natural setting emphasis on human and qualitative data collection.

2.1 Positivist

Orlikowski and Baroudi (1991) classified IS research as positivist if there was evidence of formal propositions, quantifiable measures of variables, hypothesis testing, and the drawing of inferences about a phenomenon from the sample to a stated population. Yin's (2002) , Straub et. all (2004), and the work of Marzanah (2007) has describe on an approach to qualitative research on case study research which is applicable in information system research.

2.2 Interpretive

According to interpretive philosophy, the study of a phenomena involved researcher attempting to understand the complexities of the social work, which involved qualitative techniques, with the aim to develop a rich and complex understanding of each individual's interpretation of the world as stated in Orlikowski & Baroudi, (1991). It uses qualitative methodological and characterized by a belief in a socially constructed, subjectively-based reality, one that is influenced by culture and history. It still retains the ideals of researcher objectivity, and researcher as passive collector and expert interpreter of data. Interpretive methods of research in IS are "aimed at producing an understanding of the context of the information system, and the process whereby the information system influences and is influenced by the context" Walsham(1993).

2.3 Engineering

Research conducted within the computer science and engineering involves the conception, design and development of in information system using information technology according to Clark, R. (1997). The new technology is designed to intervene in some setting, or to enable some function to be performed, or some aim to be realized. The design is usually based upon a body of theory, and the technology is usually subjected to some form of testing, in order to establish the extent to which it achieves its aims as described and used in the work of Klien et. al (2006) in the use of group support system with respect to technological implication.

3. Qualitative Research Method in Information System

IS research has been the study of processes related to the development of IS applications and the effects that IS applications have on people, particularly in formal settings such as organizations. The importance of IS research until now, has led to a number of different research approaches and methods, usually adapted from other disciplines such as sociology, natural sciences, and business studies. Harvard colloquium on qualitative IS research methods and QualIT conference in Griffith University in Brisbane on November 2005 have highlighted qualitative research, as a distinctive research approach. Qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions. According to Northcutt & McCoy (2004), Myers (2006), and Hesse-Biber & Levy (2006), there are four research methods being used by IS researchers. The research methods are the case study research, ethnography, action research, and grounded theory.

3.1 Case Study

Case study research is the most common qualitative method used in information systems (Alavi and Carlson, 1992). Yin (2002) defines the scope of a case study as an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident. Yin further suggested the following steps techniques for organizing and conducting the case study research. The steps are to determine and define the research questions, to select the cases and determine data gathering and analysis techniques, prepare to collect data, collect data in the field, to evaluate and analyze the data and lastly preparing the report. There are numerous case study research, in the organizational context for the implementation of information systems, to illustrate and investigate theories related to IS and organization.

3.2 Ethnography

This is the research method of anthropology with its emphasis on culture. It is undertaken by observation, interviews and examination of documents. In the research, the researchers observe their collaborators without prejudice or prior assumptions. Ethnography is widely used in the study of information systems in organizations, from the study of the development of information systems (Davies & Nielsen, 1992). Ethnography according to Avison and Myers, (1995) is suited to providing information systems researchers with rich insights into the human, social and organizational aspects of information systems development and application. The goal of ethnographic research is to improve our understanding of human thought and action through interpretation of human actions in context. Basic steps recommended as a general framework for an ethnographic study (Rose et al., 1995), used to conduct an ethnographic study. The steps include preparation to understand, familiarize setting goals and access to observe. Field study to establish rapport with managers and users, observe/interview and collect data. Analysis to compile the collected data, quantify data and compile statistics, preparing report and presenting the findings. Randall, D., et al. (1999), explore the issue of 'legacy' through the use of a long-term empirical investigation into how information technology is employed in a major UK bank. The closeness of their investigation into the day-to-day operations of the bank from the perspectives of individual users (using ethnographic techniques) identifies the embedded nature of the technology and the impact of cultural, organizational, and individual employees 'legacy' on organizational and technical change.

3.3 Action Research

Action research has been promoted and practiced as one way to conduct empirical research within Information System discipline. Information system action research (Davidson, 1998) is applied research to develop a solution that is of practical value to the people with whom the researchers are working, and at the same time to develop theoretical knowledge of value to a research community. According to Baskerville, R. (1999), information system research in has led to a number of different research approaches and methods, adapted from other disciplines such as sociology, natural sciences, and business studies and is often identified by its dual goal of both improving the organization participating in the research project, and the AR practitioner is expected to apply intervention on this environment. Action Research methodology was normally chosen as a research methodology as it provides the research with an inside and working view of the research matter. AR study done is characterized by the researcher applying positive intervention to the organization, while collecting field data about the organization and the effects of the intervention.

3.4 Grounded Theory

Grounded theory is a research method that seeks to develop theory that is grounded in data systematically gathered and analyzed. According to Corbin and Strauss (1990), grounded theory is theory discovery methodology that allows the researcher to develop a theoretical account based on concepts, categories and propositions. There are five phases of grounded theory building: research design, data collection, data ordering, data analysis and literature comparison, and each phase were evaluated against four research quality criteria: construct validity, internal validity, external validity and reliability. Orlikowski, (1993) uses grounded theory research in the findings of an empirical study into two organizations' experiences with the adoption and use of CASE tools over time. The study characterizes the organizations' experiences in terms of processes of incremental or radical organizational change. These findings are used to develop a theoretical framework for conceptualizing the organizational issues around the adoption and use of these tools and issues that have been largely missing from contemporary discussions of CASE tools. Singh et. al (2005) discussed on the challenge of methodological implication of moving from grounded theory to user requirement in IS design.

3.5 Results and Discussion

As research method is a strategy of inquiry to research design and data collection. The choice of research method will influence the way in which the researcher collects data. Specific research methods also imply different skills, assumptions and research practices. The main strength and limitations of each research methods are further discussed in Table 2. According to Benbasat et. al (1996), no single research methodology is better that any other methodology, and in order to ensure the quality of information system research, Clarke (1997) listed the following requirements to be present in an IS research: the research method, applied within the scientific, the interpretive or the engineering tradition, the explication of a body of theory, which in most cases needs to reach back into reference disciplines, and also the extension of the theory. This give rise to the following motivation in conducting qualitative research in IS: spending many hours in the field, collecting extensive data, and trying to gain access, rapport, as to gain an "insider" perspective in natural setting, and doing exploratory studies, where variables cannot be identified, theories are not available to explain behavior of participants or their population of study, and theories need to be developed.

The qualitative research does also present some challenges that the researchers might face in using the method. In grounded theory, the challenges for the researchers are to set aside, as much as possible, theoretical ideas or notions so that the analytic, substantive theory can emerge, the researcher must recognize that this is a systematic approach to

research with specific steps in data analysis. The researcher faces the difficulty of determining when categories are saturated or when the theory is sufficiently detailed. The ethnography is challenging to use for the researchers as the researcher needs to have grounding in cultural anthropology, time to collect data is extensive, involving prolonged time in the field, and there is a possibility to be unable to complete the study or be compromised in the study. In case study research, some of the challenges that the researcher must face is that whether to study a single case or multiple cases. The study of more than one case may dilutes the overall due to the lack of depth. In action research methods, lack of agreed criteria for evaluating action research, further complicates the publication review process, and makes this approach a difficult choice for academics. There is also an issue in both ethical and professional problems. Researchers who do not carefully explain their research orientation may mislead clients who are expecting consulting-type performance, creating an ethical breech regarding informed consent.

In the field of IS a variety of research methodologies has been explored by researchers for different aspects of research study depending on the research focus and application domain of the researchers. Whatever research method to use, there must be some way of assuring the quality of the data collected, and the correctness of interpretation. There is also the need of a framework to guide the effort, and to clarify such methodological details, as it will provide a set of guidelines for a good IS research as suggested by Checkland (1991) and Lau (1997). A framework in Table 3 is proposed and has been used by Marzanah (2007) to guide the effort, clarify methodological details as the role of the researcher, the process of problem diagnosis, the nature of the intervention, the extent of reflection and learning intended, and whether there is new knowledge to be gained in the research. The action research approach enabled us to understand the interaction of social organization and information systems, by introducing changes into these processes and observing the effects of these changes. The action research approach is proposed due to the value of capturing and explaining what is going on in real organization. By using action research, it enabled us to understand the interaction of social organization and information systems, by introducing changes into these processes and observing the effects of these changes. It serves as a checklist with its criteria and questions to assess the quality of the research.

3.6 Conclusions

The qualitative research methodology approach is viewed as significant in IS research due to the value of capturing and explaining what is going on in real organization. It enabled us to understand the interaction of social organization and information systems, the processes and observing the effects of these changes brought forward by IS. A research framework inaction research is proposed as guidance for the research activities to be undertaken to ensure the research objectives are met. The framework would guide the research effort and clarify methodological details of the role of the researcher, the process of problem diagnosis, the real world happening in an organization, the extent of reflection and learning intended, and whether there is new knowledge to be gained.

References

Alavi M. & Carlson P. (1992). A Review of MIS Research and Disciplinary Development. *J. Mngt Inf. Syst*. 8(4), 45-62.

Baskerville, R. (1999). Investigating Information systems with Action Research. *Communications of AIS* Volume 2, Article 19.

Benbasat I. & Weber R. (1996). 'Rethinking Diversity in Information Systems Research' *Info. Sys. Research* 7, 389-399.

Clarke R. (1992). 'Fundamentals of 'Information Systems", September 1992, at http://www.anu.edu.au/people/Roger.Clarke/SOS/ISFundas.html

Clarke R. (1997). Electronic Commerce Definitions. at http://www.anu.edu.au/people/Roger.Clarke/EC/ECDefns.html

Cresswell, J. (1994). Research Design: qualitative and quantitative approaches, Sage: Thousand Oaks.

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. Qualitative Sociology, 13, 3-21.

Davies, L.J. and Nielsen, S. (1992). An Ethnographic Study of Configuration Management and Documentation Practices in an Information Technology Centre, in Kendall, K.E., Lyytinen, K. and De Gross, J. (eds.). The Impact of Computer Supported Technology on Information Systems Development. Amsterdam, Elsevier/North Holland.

Esther E. Klien, Thomas Tellefsen, Paul J. Herskovitch. (2007). in 'The use of group support systems in focus groups: Information technology meets qualitative research' *in Computers in Human Behaviour,* 2133-2132.

Hesse-Biber, S. N., & Levy, P. (2006). in the Practice of Qualitative Research. Thousand Oak: Sage

Hughes, J.A., Randall, D. and Shapiro, D. (1992). Faltering from Ethnography to Design, CSCW '92. ACM 1992 Conference on Computer-Supported Cooperative Work: Sharing Perspectives. New York, ACM Press, 115-123.

Lau, F. (1991). A Review on the Use of Action Research in Information Systems Studies. in Information Systems and Qualitative Research, A.S. Lee, J. Liebenau and J.I. DeGross (eds.), Chapman and Hall, London, 31-68.

Makoto Matsuo, Christina W.Y. Wong, Kee-hung Lai. (2008). Experience-based learning of Japanese IT professionals: A qualitative research. *Journal of strategic Information Systems* 17, 202 -213.

Mangan J. (2004). 'Combining quantitative and qualitative methodologies in logistic research', *International Journal of Physical Distribution & Logistics Management.* Vol 34 No. 7, 565-578.

Marzanah, J. A. (2007). A Framework for managing Knowledge and Competencies in A Group Project Implementation, PhD Thesis, Universiti Putra Malaysia.

Myers, M. D. (1997). in "Qualitative Research in Information Systems," *MIS Quarterly* (21:2), pp. 241-242.

Myers, M.D., & Michael Newman. (2006). in 'The qualitative interview in IS research:Examining the craft' Information and Organization., pp 2-26.

Northcutt N., & McCoy, D. (2004), Interactive qualitative analysis. Thousand Oaks. Sage.

Orlikowski W.J. & Baroudi J.J. (1991). 'Studying Information Technology in Organizations: Research Approaches and Assumptions' Info. Sys. Research 2, 1-28.

Orlikowski. (1993). W. "CASE tools are organizational change: Investigating Incremental and Radical Changes in Systems Development," *MIS Quarterly*, (17:3), 309-340.

Orlikowski, W. and C. Iacono. (2001). "Desperately Seeking the IT in IT Research – A Call to Theorizing the IT Artifact" Information Systems Research 12(2): 121-134.

Randall, D., Hughes, J., O'Brien, J., Rodden, T., Rouncefield, M., Sommerville, I. and Tolmie, P. (1999). "Banking on the Old Technology: Understanding the Organizational Context of 'Legacy' Issues," *Communications of the AIS*, (2:8), 1-27.

Rose, Anne, Plaisant, Catherine, and Shneiderman, Ben. (1995). "Using Ethnographic Methods In User Interface Re-engineering." Proceedings of the ACM conference for Designing Interactive Systems: processes, practices, methods, and techniques. New York, NY: ACM Press, 115-122. http://delivery.acm.org/10.1145/230000/225447/p115-rose.pdf

Straub, D., Gefen, D., and Boudreau, M.-C. (2004). "The ISWorld Quantitative, Positivist Research Methods Website", http://dstraub.cis.gsu.edu:88/quant/

Supriya Singh, Kylie Kassar Bartolo & Christine Satchell. (2005). in 'Grounded theory and user requirements: A challenge for qualitative research' *Australian Journal of Information System*, Vol 12(2).

Walsham G.. (1993). 'Interpreting Information Systems in Organizations', Wiley.

Yin, R. K. (2002). *Case Study Research, Design and Methods,* 3rd ed. Newbury Park, Sage Publications.

Table 1. Research Methodologies: Clarke (1997)

Non-Empirical	Scientific	Interpretive	Scientific/ Interpretive	Engineering
- Conceptual research. - Theorem proof. - Simulation. - Future research. - Scenario-building, and game- or role-playing. - Review of existing literature.	- Forecasting. - Field experimentation and quasi-experimental designs. - Laboratory experimentation	- Descriptive/ interpretive research. - Focus group research. - Action research. - Ethnographic research. - Grounded theory.	- Field study. - Questionnaire-based survey. - Interview-based survey. - Case study. - Secondary research.	- Conceptual - Design - Development

Table 2. Benefits and limitations of the qualitative research approach

Case Study	
Strength	Weakness
1. Excels in understanding complex issue or object and can extend experience or add strength to what is already known through previous research. 2. Captures the local situation in greater detail and with respect to more variables than is possible 3. Applicable to real-life, contemporary, human situations and its public accessibility through written reports.	1. Lack of control of variables. 2. Different interpretations by different people 3. Unintentional biases and omissions in the description due to intense exposure to the study 4. Study of a small number of cases can offer no grounds for establishing reliability or generality of findings. 5. Case study research as useful only as an exploratory tool.
Ethnography	
Strength	Weakness
1. Powerful assessment of users' needs: A crucial goal of an ethnographic study is to gain the capacity to view a system through the eyes of the user 2. It uncovers the true nature of the system user's job: A goal of an ethnographic study is to uncover all tasks and relationships that combine to form a user's job. It is often the case that a user performs tasks and communicates in ways that are outside of their official job description. 3. The ethnographer can play the role of the end-user: The high level of user understanding that an ethnographer can gain through his/her fieldwork can be a useful bonus. 4. The open-ended and unbiased nature of ethnography allows for discovery: Other HCI research methods, such as task analysis and controlled experimentation, must formalize, categorize, and/or theorize how users interact with a system in order to yield quantitative results.	1. Time consuming 2. (long time to do field work and analyzing) 3. In-depth knowledge only of particular context and situations (lack of generalization) 4. The highly qualitative nature of results can make them difficult to present in a manner that is usable by designers. 5. Use a small number of participants and a small-scale environment (Hughes et al., 1995). Increasing the scale can be extremely difficult as it imposes a much greater amount of cost, communication, and effort.
Action Research	
Strength	Weakness
1. Captures the local situation in greater detail and with respect to more variables than is possible. 2. It can be used in many research modes, both to generate new theory and to reinforce or contradict existing theory. 3. Participatory action research enriches the research community by drawing researcher-practitioners into the research process.	1. Lack of agreed criteria for evaluating action research 2. Action research "looks like" consulting. 3. Lack of control makes it difficult to apply action research as an instrument in an orchestrated research program

Grounded Theory	
Strength	Weakness
1. Ability to derive theory from within the context of data collected. 2. No worries about the formality in the method of usage. 3. Resulting theories are explicitly emergent, does not test a hypothesis. 4. Start collecting data as soon as there is a research phenomenon to study.	1. Sensitive to thoroughness and skills of individual researcher in interpreting data and the judgment to know when saturation is achieve 2. Does not favor the novice researcher as they are likely to find the approach more difficult than more conventional methodologies; and the more experienced researcher is likely to produce better theory.

Table 3. Research Framework

Philosophical / Conceptual Foundation	
Dimension & Criteria	
Aim/Question	An approach to manage prototype development of group knowledge and competencies
Assumptions	Knowledge can be extracted and competencies of group could manage
Perspective/Tradition	Interpretive
Stream	Participatory action research
Methodology Study Design	
Dimension & Criteria	
Background	The need and challenges in managing group knowledge and competencies
Intended Change	Introduction of an agent based KEPSNet framework to facilitate group tacit knowledge and competencies management
Site	Laboratory of learning and Multimedia Innovation, Multimedia and Software Institute, UPM
Participants	Technical Application Development Design Team
Data Sources	Participant observation, Interviews Electronic group discussion transcripts, Prototype evaluation and testing
Duration	18 months
Degree of Openness	System development design, Testing, Implementation
Access/Exit	Appointed as a researcher for the duration of three years. Enter as Quality Manager for the project.
Presentation	Case study with emphasis on method, results and implications on lessons learned.
Research Process	
Dimension & Criteria	
Problem Diagnosis	Problem in managing group knowledge and competencies. Need to improve business processes in the organizations.

Action Interventions	Introduced the KEPSNet framework and to manage the extraction of group knowledge and the competencies management in the group. Usage of the prototype framework during the development life cycle.
Reflective Learning	Best practices learned and reuse knowledge.
Iteration	Knowledge Capture Accumulation of knowledge and competencies Expertise Identification Creation of Directory of expertise Competencies matching. Project management and System Development Life Cycle
General Lessons	Group members' competencies profiling and knowledge network creation had competing effects, and organizational implications.
The Socialtechnical Design	
Dimension & Criteria	
Diagnosis and entry	From sociotechnical aspect, it is impossible to separate the organizational issues or social issues from the technical issues.
Management of the change process	During the design phase, the focus would be on how to capture group members' expertise and competencies as an informal process for the group.
System Design	In sociotechnical design analysis, analysis is based on the observation, and the understanding of the organizational context such as organizational structure, work and staff. As sociotechnical requirements are embedded in the work of activities of individuals, techniques such as participatory system design will provide a way of extracting the requirements.
Adjustment of coordinating mechanisms	Amendments to other subsystems. Changes in the system that might necessitate changes in other subsystem: Office Automation Application Human resource system Knowledge repositories
Implementation	Understand the outcome in the work practices and the social interaction in a group project implementation.
Role Expectation	
Dimension & Criteria	
Researcher	Facilitated group processes, collected and analyzed research data.
Participants	Groups took part in workshops, provided data through questionnaires, interviews, etc. Team helped develop, define process and deliverables.
Competency	Staff can use the proposed knowledge extraction and competencies profiling framework. Increased group productivity, perceived outcome quality with the proposed framework.
Ethics	A contract with deliverables defined; team involved in decision making.

Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

Unsupervised Coreference Resolution with HyperGraph Partitioning

Jun Lang (Corresponding author)
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: bill_lang@ir.hit.edu.cn

Bing Qin
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: qinb@ir.hit.edu.cn

Ting Liu
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: tliu@ir.hit.edu.cn

Sheng Li
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: lis@ir.hit.edu.cn

The research is supported by National Natural Science Foundation of China (60675034, 60803093), National High Technology Research and Development Program of China (863 Program) (2008AA01Z144). (Sponsoring information)

Abstract

Unsupervised-learning based coreference resolution obviates the need for annotation of training data. However, unsupervised approaches have traditionally been relying on the use of mention-pair models, which only consider information pertaining to a pair of mentions at a time. In this paper, it is proposed the use of hypergraph partitioning to overcome this limitation. The mentions are modeled as vertices. By allowing a hyperedge to cover multiple mentions that share a common property, the additional information beyond a mention pair can be captured. This paper introduces a hypergraph partitioning algorithm that divides mentions directly into equivalence classes representing individual entities. Evaluation on the ACE dataset shows that our unsupervised hypergraph based approach outperforms previous unsupervised methods.

Keywords: Coreference resolution, HyperGraph partitioning, Unsupervised learning

1. Introduction

Coreference resolution is the process of partitioning mentions into different real world entities. It is a key component of many Natural language processing (NLP) applications. Especially, due to its important role in Information extraction

(IE), coreference resolution was defined as an IE subtask and officially evaluated in the Message Understanding Conference (MUC) and Automatic Content Extraction (ACE) programs. So far, supervised-learning-based approaches have been widely applied to coreference resolution, which requires a set of training data to build a classifier for coreference judgment (Soon et al., 2001; Ng and Cardie, 2002). However, coreference annotation is a difficult task, which involves not only deep linguistic knowledge, but also background knowledge related to the domain. For this reason, the size of existing annotated coreference corpora is quite small (e.g., 599 documents in the ACE2005 corpus) compared with other NLP tasks, and is limited only in some specific domains.

To deal with the lack of the training data, several unsupervised approaches were proposed which require no training data for coreference resolution and are adaptive to different domains. For example, Cardie and Wagstaff (1999) suggested recasting coreference resolution to a clustering problem, which tries to group noun phrases into different coreference clusters. They defined a distance function to measure the incompatibility of two mentions. Given a document, mentions are processed backwards one by one. Two mentions are placed into the same cluster if their distance is below a threshold, and no mentions from their respective clusters are incompatible. Wagstaff (2002) further enhanced this method by adding more linguistic constraints (must-link and cannot-link) during clustering.

However, there are several problems with the previous clustering based unsupervised methods.

(1) As with many other learning based approaches to coreference resolution (e.g., Soon et al. (2001), and Ng and Cardie (2002)), these methods adopt a mention-pair model. The distance function is only based on the information of two given mentions. However, as individual mentions lack adequate information about the entities they refer to, the distance may be not accurate to represent the (in)compatibility of two mentions. For example, the compatibility between mentions "*Powell*" and "*She*" may be different, depending on the gender information of "*Powell*" which cannot be determined from the mention alone.

(2) As the clustering is agglomerative, the wrong linking decision could not be undone and would lead to cascading errors. Suppose we have three mentions "*Mr. Powell*", "*Powell*", "*She*". If "*She*" is wrongly linked to "*Powell*", the cluster cannot be broken and will prevent the subsequent linking of "*Powell*" with "*Mr. Powell*".

To overcome the above problems, this paper proposes an unsupervised coreference resolution approach with hypergraph partitioning. Hypergraph is a special graph in which an edge connects more than two vertices (Berge, 1989). To model coreference resolution, mentions could be viewed as vertices. A set of mentions is covered by a hyperedge if they show a specific common property. As a hyperedge can describe information shared by two or more mentions, it has a more powerful representation capability for knowledge than a traditional mention-pair feature. By using a partitioning algorithm, mentions are divided into equivalence partitions representing individual entities. The partitioning process can avoid the cascading errors in the clustering-driven unsupervised approaches. In our experiments, we evaluated our approach on the ACE data and our experimental results show that our approach is effective for coreference resolution.

The following sections are organized as following: Section 2 describes some related works for coreference resolution and hypergraph with its applications. Section 3 introduces the basic concepts of hypergraph and the partitioning algorithm. Section 4 describes the hypergraph-based model for coreference resolution. Section 5 gives the experimental results with some discussions. Finally, section 6 summarizes the conclusion and presents future works.

2. Related work

Supervised-learning-based approaches are widely adopted in coreference resolution. It was first proposed by using decision tree approach (McCarthy and Lehnert, 1995), and later many other systems follow. A typical one of them is presented by (Soon et al., 2001). In it, coreference resolution is deemed as a classification problem. A training or testing instance is formed by two mentions, with a feature vector describing their properties and relationships, including the information of gender, number, person, semantic, string match, appositive, name alias, and so on. When testing, a mention to be resolved is checked against its preceding mentions, and is linked with the closest one that is classified as positive. The work is further enhanced by expanding the feature set and adopting "best-first" linking strategy (Ng and Cardie, 2002).

Such a mention-pair-based model only considers information related to two mentions in question, and would cause triangular contradiction errors at a testing time. Suppose we have three mentions "Mr. Powell", "Powell", and "she" in a document. The model tends to link "*she*" with "*Powell*" because of their proximity, and link "*Mr. Powell*" with "*Powell*" since head string matching. Merging the two pairs together, nevertheless, would lead to gender disagreement between "*she*" and "*Mr. Powell*".

Several researchers proposed to use graph theory to deal with the triangular contradiction errors in coreference resolution. They converted a document to a graph in which mentions in the document are mapped to vertices in the graph. An edge connecting two vertices represents the coreference relationship between the two corresponding mentions. The weight of an edge accounts for the confidence of the coreference relationship and is derived from coreference classification. Then, some graph partitioning algorithms can be used for global optimization, such as BESTCUT

(Nicolae and Nicolae, 2006). Similarly, Bell-Tree global searching (Luo et al., 2004) and triangular contradiction constraint learning with Conditional Random Field (McCallum and Wellner, 2003) are proposed for such problem. However, they all are supervised learning methods.

Hypergraph has shown many advantages in clustering and classification problems (Zhou et al., 2006). In recent years, it is also employed in NLP applications like sentence parsing (Klein and Manning, 2001; Huang, 2008), word sense disambiguation (Klapaftis and Manandhar, 2007) and document clustering (Shinnou and Sasaki, 2007). However, to our knowledge, our work is the first effort to adopt this technique to the coreference resolution task.

3. Basic concepts of hypergraph

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set, $H = \{E_1, E_2, \ldots, E_m\}$ be a family of subsets of X. The family H is said to be a hypergraph on X if

$$E_i \neq \phi (i = 1,2,\ldots,m) \tag{1}$$

$$\bigcup_{i=1}^{m} E_i = X \tag{2}$$

$H = (X: E_1, E_2, \ldots, E_m)$ is called a *hypergraph*. $|X| = n$ is the order of the hypergraph. The elements $x_1, x_2, \ldots, x_n$ are *vertices* and the sets $E_1, E_2, \ldots, E_m$ are called *hyperedges*.

An example hypergraph is shown in Figure 1. An edge E_i with $|E_i| > 2$ is drawn as a curve encircling all of its vertices. An edge E_i with $|E_i| = 2$ is drawn as a line connecting its two vertices. An edge E_i with $|E_i| = 1$ is drawn as a loop as in a graph. If $|E_i| \leq 2$ for all i, a hypergraph is reduced to a common graph. In a hypergraph, two vertices are said to be adjacent if there is a hyperedge E_i that contains both of these vertices. Two hyperedges are said to be adjacent if their intersection is not empty.

The incidence matrix of hypergraph $H = (X: E_1, E_2, \ldots, E_m)$ is a matrix $A = (a_{ij})$ with m rows that represent the hyperedges of H and n columns that represent the vertices of H, that is,

$$a_{ij} = \begin{cases} 1, if\ x_j \in E_i \\ 0, if\ x_j \notin E_i \end{cases} \tag{3}$$

Each (0, 1)-matrix is an incidence matrix of a hypergraph if no row or column contains only zeros. For illustration, the hypergraph of Figure 1 can be converted to the following one.

$$A = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \end{array} \begin{array}{c} \begin{array}{cccccccc} x_1 & x_1 & x_1 & x_1 & x_1 & x_1 & x_1 & x_1 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \tag{4}$$

There exist quite a few hypergraph partitioning algorithms that have been proved effective in different practical problems, such as partitioning circuit netlists, clustering categorical data, and segmenting images. In our study, we chose hMETIS (2.0pre1) (Note 1) which is capable of providing high quality partitions with a high speed. The algorithm in hMETIS is based on multilevel hypergraph partitioning algorithm (Selvakkumaran and Karypis, 2006). In our study, we used the direct k-way partitioning scheme of hMETIS. The overall quality of the obtained partitioning can be computed using the following quality measures (Note 2):

3.1 Scaled Cost: defined for *k-way* partitioning as

$$\frac{1}{n(k-1)} \sum_{i=1}^{k} \frac{|E(P_i)|}{|P_i|} \tag{5}$$

where $|E(P_i)|$ is the number of hyperedges that are incident but not fully contained inside the partition P_i.

3.2 Absorption: This is defined as

$$\sum_{i=1}^{k} \sum_{e \in E | e \cap P_i \neq \phi} \frac{|e \cap P_i| - 1}{|e| - 1} \tag{6}$$

where E is the set of hyperedges, $|e \cap P_i|$ is the number of vertices of a hyperedge e in partition P_i, and $|e|$ is the number of vertices of e.

The hMETIS program performs partitioning by minimizing the *Scaled Cost*, while maximizing *Absorption* at the same time.

4. HyperGraph modeling for coreference resolution

To recast coreference resolution to a hypergraph partitioning problem, we view mentions as vertices, and use various kinds of knowledge for coreference resolution to create hyperedges. In this section, we will focus on several important aspects of the hypergraph model: designing hyperedges, choosing proper weights of hyperedges, performing partitioning, and setting the stopping criterion.

4.1 Hyperedges

Traditional learning based coreference resolution systems represent knowledge in terms of features. For coreference resolution with hypergraph partitioning, however, we represent knowledge with hyperedges. As introduced in Section 3, a hyperedge is a special edge that covers more than one vertex. We can convert mentions in a document to vertices in a hypergraph. Several mentions are thought of being covered by a hyperedge if they share a specific common property. In this way, we can capture the information of multiple mentions at the same time, instead of a mention-pair as in tradition learning-based approaches to coreference resolution. Hence, the hypergraph would provide us a more powerful representation capability for knowledge.

In our study, we define the following types of hyperedges. For illustration, we use the text in Table 1 as an example.

1) *FullString*: This type of hyperedge covers the mention vertices that have the same string (excluding the determiners). For example, in Table 1, mentions m_2 and m_7 have the same string, and so do mentions m_6 and m_{10}. Then we will have two *FullString* hyperedges that cover $\{m_2, m_7\}$ and $\{m_6, m_{10}\}$, respectively.

2) *Head*: This type of hyperedge covers mentions with the same head string.

3) *Gender*: This type of hyperedge covers the mentions that have the same gender type. To considering only effective partitioning of mentions, there are only two types of gender (Note 3), *Male* and *Female*. A hypergraph has at most two *Gender* hyperedges. A mention with a neuter gender (such as "it", "the president") is not covered by a hyperedge of type *Male* for *Female*. In Table 1, mentions m_4, m_5 and m_9 are male and thus will be covered by a hyperedge {4, 5, 9}.

4) *Number*: This type of hyperedge covers the mentions that have the same number type. A hypergraph may have two *Number* hyperedges for singular or plural mentions.

5) *Person*: This type of hyperedge covers the mentions that have the same person type. There are only two *Person* hyperedges for mentions that are persons or non-persons.

6) *Semantic*: This type of hyperedge covers the mentions that have the same semantic type. A hypergraph may contain five hyperedges of this type for the semantic types *Organization*, *GPE*, *Person*, *Location*, and *Facility* defined in the ACE annotation scheme. Features for semantic types were obtained from the gold annotations. In Table 1, we can get three *Semantic* hyperedges {1, 6, 8, 10, 11}, {2, 7} and {3, 4, 5, 9}, for *Organization*, *GPE* and *Person*, respectively.

7) *ThreeSentences*: This type of hyperedge covers a pronoun (Note 4) and the preceding mentions in the current sentence and previous two sentences. In Table 1, for pronoun its_2, its_7 and He_9, the mentions in the 3-sentence window are $\{m_1 - m_2\}$, $\{m_1 - m_7\}$, $\{m_1 - m_9\}$, respectively. Thus, we have two hyperedges: {1, 2, 3, 4, 5, 6, 7} and {1, 2, 3, 4, 5, 6, 7, 8, 9}. Note that we do not generate a hyperedge {1, 2} as its vertices are fully contained in the hyperedge {1, 2, 3, 4, 5, 6, 7}.

8) *TwoSentences*: This type of hyperedge is similar to *ThreeSentences*, but it just considers mentions within a two-sentence window.

9) *OneSentence*: This type of hyperedge is similar to *ThreeSentences*, but it just considers mentions within the same sentence.

10) *NameAlias*: This type of hyperedge covers the mentions that are name alias of one another in a document. Consider Table 1, $\{m_4, m_5\}$, $\{m_1, m_6, m_{10}\}$ are name-alias groups and thus we can have two *NameAlias* hyperedges {4, 5}, {1, 6, 10}.

11) *Appositive*: This type of hyperedge covers the mentions that are in the same appositive structure.

12) *CannotLink*: As suggested by Wagstaff (2002), we enforce cannot-link constraints during partitioning. For this purpose, we create hyperedges to cover a pair of mentions $<m_i, m_j>$ that are not likely to corefer. In our study, we consider the following constraints:

a. m_j is an indefinite noun phrase.

b. m_i and m_j are three sentences apart and do not have the same head word.

c. m_i and m_j are pronouns and m_i and m_j do not agree in number or gender.

d. m_j is a pronoun and m_i and m_j are three sentences apart.

In Table 1, mention pairs m_2 - m_9 and m_7 - m_9 violate the third constraint and thus are covered by two *CannotLink* hyperedges {2, 9} and {7, 9}, respectively.

The hyperedges generated for Table 1 are in Table 2.

4.2 Weights for hyperedges

We classify the hyperedges into six categories based on their confidence level for a positive coreference determination, as shown in Table 3.

The hypergraph partitioning algorithm tends to divide vertices covered by a hyperedge with a low-weight, and retain in the same partition vertices covered by a high-weight hyperedge. Thus, we manually assign a higher weight to a hyperedge that covers mentions that are likely to corefer, while a lower weight to a hyperedge that covers mentions that are unlikely to corefer. Table 3 shows the different weights for different levels of hyperedges. The hyperedge *CannotLink* was assigned the lowest weight of zero (Note 5). The hyperedges *FullString*, *Appositive,* and *NameAlias*, which are strong indicators of coreference relationship (Soon et al., 2001), were given the highest weight.

4.3 Coreference resolution with mention partitioning

Given a document, all the mentions are placed into a large cluster initially. As described, we map each mention to a vertex in a hypergraph, and find out all the possible hypergraphs for the vertices. Then we can invoke the hMETIS program to perform mention partitioning. The process is done in an unsupervised way. After the partitioning stops, a generated partition could be deemed as a coreferential cluster for a single entity, with all the mentions in the same cluster being coreferential with each other.

4.4 Stopping Criterion

One problem with mention partitioning is when the process should stop. In other partitioning tasks, the number of target clusters is predefined. However, for our task, it is not possible to give a predefined cluster number as the number of entities in a document is unknown before resolution. Therefore, we need to design a stopping criterion for partitioning.

As described in Section 3, to certain partitioning clusters number k, the hMETIS program performs partitioning by minimizing the *Scaled Cost* (5), while maximizing *Absorption* (6) at the same time. After inner optimization, hMETIS would find the best partitioned clusters with the final Scaled Cost and Absorption values, named as $ScaledCost_{final}(k)$ and $Absorption_{final}(k)$ respectively.

Enlarging the target entity number k would make the former value increase while the latter decrease. Without any prior knowledge, we try to find a k which compromise on the two costs varying trends at the same time. For this consideration, we define a stopping criterion based on the product of final generated $ScaledCost_{final}(k)$ and $Absorption_{final}(k)$ values after optimization:

$$P(k) = ScaledCost_{final}(k) * Absorption_{final}(k) \quad (7)$$

We prefer a partition with high product of *ScaledCost* and *Absorption*. For a given document, we put all the mentions in a cluster (k=1) and perform partitioning repeatedly. The process stops at a round when the value of P reaches the peak, or when each cluster contains only one mention. The generated clusters are output as the coreference resolution result. Actually, in our study such stop criterion achieved good result.

5. Experiments and results

5.1 Experimental Setup

In our study, we did evaluation on the ACE-2 V1.0 corpus (NIST, 2003) that is divided into three domains: broadcast news (BNews), newspaper (NPaper), and newswire (NWire). As we conducted unsupervised learning, we did not use the training data and just ran the system on the test data. However, in the comparing supervised systems, the training and testing data were used together. The number of entities with more than one mention, as well as the number of the contained mentions, is summarized in Table 4.

For both training and resolution, an input raw document was processed by a pipeline of NLP modules of OpenNLP (Note 6), including sentence boundary detector, tokenizer, and part-of-speech tagger. The boundaries of a mention are directly from annotation in the corpus. Our experiment setup just follows the official "diagnose" evaluation of ACE in which coreference resolution is done and evaluated on the perfect mentions, which allows the validation of the utility of the hypergraph method under an environment of accurate mention features. We used the mention's head, boundaries, and the semantic type information from "gold" annotation. Other features, like string-matching, apposition, name-alias, distance and so on, were all computed at a running time. Following tradition, all results are reported using recall, precision and F-measure based on the MUC-6 scoring algorithm (Vilain et al., 1995).

5.2 Results and discussion

Table 5 lists the performance of different coreference resolution systems.

For comparison, we first duplicated the traditional unsupervised learning based system by Cardie and Wagstaff (1999) as baseline. The first line of Table 5 shows the results of such a system, which adopted the same clustering radius threshold (i.e., r = 4) as in Cardie and Wagstaff (1999)'s system. Our duplicated system (denoted by Cardie99r4) achieves a recall of 66.30% and a precision of 50.99%, obtaining an F-measure of 57.64%. The F-measure (57.64%) is higher than their results (52.8%) reported on the MUC-6 data.

Cardie and Wagstaff (1999)'s radius value was fined-tuned for the MUC-6 data, and is not necessarily optimal for the ACE data. In our experiments, we examined the performance of the duplicated system under different radius value from 1 to 10. We found that the system achieves the best result when r = 6 (Cardie99r6), with 67.34% recall, 51.73% precision and 58.5% F-measure. The recall, precision, F-measure results for each domain are consistently higher than those of Cardie99r4. It indicates that the performance of their system is significantly affected by the threshold value.

In the experiments, we were also interested in comparing performance difference between the system with unsupervised learning and the system with supervised learning. For this purpose, we implemented the classical decision trees-based coreference resolution system by Soon et al. (2001) (denoted by Soon01), and the results are shown in the third line of Table 5. Compared with Cardie99r6, the system has a drop in recall (up to 11.83%), but achieves a large improvement in precision (up to 21.23%). Overall, it produces an average 55.51% recall, 72.96% precision, and 63.05% F-measure. The F-measure is 4.54% higher than Cardie99r6. This is in line with Wagstaff (2002)'s report that Cardie and Wagstaff (1999)'s unsupervised approach got an F-measure about 9% lower than the supervised system.

The fourth line of Table 5 summarizes the performance of our system with hypergraph partitioning. From the table, the system produces a higher recall of 79.37%, 12.03% than Cardie99r6, with just only 2.87% loss in precision. Overall, the F-measure is about 2% higher than Cardie99r6. The difference against the supervised based system (Soon01) is reduced to 2.57%, and the results are encouraging considering that our approach did not use any training data.

One interesting finding of the table is that unsupervised approaches tend to produce a lower precision but a higher recall than supervised approaches. This should be the case because our hypergraph method is based on top-down partitioning. Mentions tend to be retained in the same cluster unless they have some inconsistency. By contrast, a supervised approach is based on bottom-up merging, mentions are only merged together if some coreference indicators, like string matching, name alias or appositive can be satisfied. The merge is comparatively conservative and thus leads to a higher precision but a lower recall.

We were also concerned how much each type of hyperedge affected the resolution performance. Table 6 summarizes the performance contribution of each kind of hyperedges to our system of HyperGraph. The last three columns show the gain or loss in recall, precision and F-measure, respectively, because of subtracting a particular hyperedge while keeping the rest in the HyperGraph system.

As our approach is partitioning-driven, the low-weight hyperedges play an important role in dividing mentions. We were also concerned how much the hyperedge *CannotLink* affects the resolution performance. The last line of Table 6 shows the loss of performance by removing the *CannotLink* from the system. From the table, the removal of *CannotLink* results in a drop of by 37.56% in recall and 6.14% in precision. Overall, the F-measure decreases by 18.22%. Similarly, the hyperedge contribution to whole system F-measure decreased like *Semantic*(1.27%), *NameAlias*(0.54%), *Number*(0.35%), *ThreeSentences*(0.24%), *Gender*(0.14%), *HeadString*(0.08%), *Appositive*(0.05%).

Interestingly, when only subtracting *FullString*, *Person*, *TwoSentences*, and *OneSentence*, the final F-measure increased 0.25%, 0.11%, 0.30%, 0.21%, respectively. In other words, the four kinds of hyperedges decreased the whole system performance using all features. After deep analysis, we found that the *FullString* with high weight is little repeated by *HeadString* with middle weight. Moreover, *Person* is just a kind of semantic. The *Person=True* hyperedges are replaced by *Semantic=Person* hyperedges. When replaced, the hyperedges are redundant, and hence decrease the final resolution result. Meanwhile *TwoSentences* is repeated to some extent by *ThreeSentences* and *OneSentence*. Similarly, so does *OneSentence* by *ThreeSentences* and *TwoSentences*.

Our results show that reducing feature redundancy is a practical problem for unsupervised coreference resolution hypergraph partitioning. Actually, we experimented on subtracting any two, three or all of the above four kinds of hyperedges while keeping the rest in the HyperGraph system. The results were all worse than using all features. It was because all the features were intersecting in the hypergraph.

6. Conclusion

This paper presented an unsupervised learning approach for coreference resolution based on hypergraph partitioning. It converts a document to a hypergraph where a vertex corresponds to a mention in the document. It uses a hyperedge to cover mentions that share a specific common property, which can capture information about multiple mentions, instead

of only two mentions as in the traditional approaches based on the mention-pair model. Our approach adopts a hypergraph partitioning algorithm to divide mentions into clusters each representing a single entity. The partitioning process can avoid the cascading errors in the previous clustering-based unsupervised approaches.

In the paper, we described the resolution framework, the definition of hyperedges, and the stopping criteria of partitioning. The evaluation on the ACE data set shows that the hypergraph partitioning approach performs better than the previous clustering-based unsupervised approach (with up to 1.97% in F-measure), and the gap between the supervised approach is only 2.57% in F-measure.

Our current work focuses on the framework of coreference resolution with hypergraph partitioning. There are several directions for future work:

(1) For simplicity, we currently just used some common knowledge, represented as hyperedges, for coreference resolution. We would like to explore more effective knowledge, such as grammar roles, context template information, and others proposed in Ng and Cardie (2002).

(2) In the current system, the weights for hyperedges were all heuristically designed. We intend to try some weights learning mechanisms, e.g., the genetic algorithm.

(3) The stopping criterion has a big influence on the final resolution performance. However, our current stop criterion was defined in a heuristic way. We would like to incorporate more prior knowledge related to coreference resolution.

(4) Feature redundancy is another problem for hypergraph partitioning. We will try to process it as a learning problem.

References

Andrew McCallum and BenWellner. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. *Proceedings of the IJCAI-03 Workshop on IIWeb*, pp.79-84.

C. Berge. (1989). *Hypergraphs*. North-Holland, Amsterdam

C. Cardie and K. Wagstaff. (1999). Noun phrase coreference as clustering. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.82-89.

C. Nicolae and G. Nicolae. (2006). Best cut: A graph algorithm for coreference resolution. *Proceedings of the EMNLP*, pp.275-283.

D. Klein and C.D. Manning. (2001). Parsing and hypergraphs. *Proceedings of the IWPT*.

D. Zhou, J. Huang, and B. Schölkopf. (2006). Learning with hypergraphs: Clustering, classification, and embedding. *Proceedings of the NIPS*, pp.1601-1608.

Hiroyuki Shinnou and Minoru Sasaki. (2007). Ensemble document clustering using weighted hypergraph generated by nmf. *Proceedings of the ACL*, pp.77-80.

I.P. Klapaftis and S. Manandhar. (2007). UOY: A Hypergraph Model For Word Sense Induction & Disambiguation. *Proceedings of the SemEval*, pp.414-417.

Joseph F. McCarthy and Wendy G. Lehnert. (1995). Using decision trees for coreference resolution. *Proceedings of the IJCAI*, pp.1050-1055.

Kiri Lou Wagstaff. (2002). *Intelligent Clustering with Instance-Level Constraints*. Ph.d. thesis, Cornell University,

Liang Huang. (2008). Forest reranking: Discriminative parsing with non-local features. *Proceedings of the ACL-08:HLT*, pp.586-594.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. (1995). A model-theoretic coreference scoring scheme. *Proceedings of the 6th MUC*, pp.45-52.

N. Selvakkumaran and G. Karypis. (2006). Multiobjective Hypergraph-Partitioning Algorithms for Cut and Maximum Subdomain-Degree Minimization. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 25(3):504-517.

V. Ng and C. Cardie. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the ACL*, pp.104-111.

W.M. Soon, H.T. Ng, and D.C.Y. Lim. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521-544.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. *Proceedings of the ACL*, pp.135-142.

Notes

Note 1. http://glaros.dtc.umn.edu/gkhome/fetch/sw/hmetis/hmetis-2.0pre1.tar.gz

Note 2. These definitions can be extended in a straightforward manner for hypergraphs with weighted hyperedges, as described in http://glaros.dtc.umn.edu/gkhome/fetch/sw/hmetis/manual.pdf

Note 3. The gender type of a person name was obtained from a name-gender list provided by the corpora from NLTK package, while the gender of a common noun (e.g., mother, son, president) was got from WordNet (if the gender of a mention, such as "the president", is not available in WordNet, we set the gender type as neuter).

Note 4. These sentence hyperedges only aim for pronoun resolution. They do not cover non-pronoun anaphora, as the sentence factor plays little influence on the coreference determination for non-pronoun resolution.

Note 5. Intuitively, negative weight for *CannotLink* is better. However, the hypergraph toolkit hMETIS could not accept negative weights. So here, zero is chosen.

Note 6. http://opennlp.sourceforge.net/

Table 1. This is an example about tables

[Microsoft Corp.] $_1$ *announced* ***[[its]*** $_2$ ***new CEO]*** $_3$ ***[Steve Ballmer]*** $_4$ *yesterday.* ***[Mr. Ballmer]*** $_5$ *said* ***[Microsoft]*** $_6$ *would try* ***[its]*** $_7$ *best to compete with* ***[Google]*** $_8$*.* ***[He]*** $_9$ *also mentioned that* ***[Microsoft]*** $_{10}$ *had been challenged by* ***[some companies]*** $_{11}$ *in internet area during the late years.*

Table 2. An example of generated hyperedges

Type	Hyperedges
FullString	{2,7}, {6,10}
HeadString	{2,7}, {6,10}, {4,5}
Gender	{4,9}
Number	{1,2,3,4,5,6,7,8,9,10}
Person	{3,4,5,9}, {1,2,6,7,8,10,11}
Semantic	{1,6,8,10,11}, {2,7}, {3,4,5,9}
ThreeSentences	{1,2,3,4,5,6,7}, {1,2,3,4,5,6,7,8,9}
TwoSentences	{1,2,3,4,5,6,7}, {5,6,7,8,9}
OneSentence	{1,2}, {5,6,7}
NameAlias	{4,5}, {1,6,10}
Appositive	{3,4}
CannotLink	{2,9}, {7,9}

Table 3. Weights for four kinds of hyperedges

Level	Hyperedges	Weight
Cannot	*CannotLink*	0
Low-1	*ThreeSentences*	5
Low-2	*TwoSentences*	10
Low-3	*OneSentence*	15
Middle	*HeadString*, *Gender*, *Number*,*Person*, *Semantic*	20
High	*FullString*, *Appositive*, *NameAlias*	30

Table 4. Statistics of entities (length > 1) and contained mentions for the test data set in ACE

Domain	#entity	#mention
BNews	468	2493
NPaper	365	2290
NWire	411	2304

Table 5. Results of different systems for coreference resolution

System	BNews			NPaper			NWire			Total		
	R	P	F	R	P	F	R	P	F	R	P	F
Cardie99r4	72.06	61.28	66.23	60.58	44.89	51.57	66.26	48.04	55.70	66.30	50.99	57.64
Cardie99r6	73.22	61.89	67.08	62.18	46.26	53.05	66.60	48.19	55.92	67.34	51.73	58.51
Soon01	60.63	81.03	69.36	51.92	65.53	57.94	53.91	72.77	61.94	55.51	**72.96**	**63.05**
HyperGraph	86.83	55.12	67.43	72.44	45.75	56.08	78.81	45.84	57.97	**79.37**	48.86	60.48

Table 6. Results of different systems for features contribution comparison

System	Total			Gain(+)/Loss(-)		
	R	P	F	R	P	F
All Features	79.37	48.86	60.48			
-*ExtentString*	79.87	48.99	60.73	**0.50**	**0.13**	**0.25**
-*HeadString*	79.34	48.76	60.40	-0.03	-0.10	-0.08
-*Gender*	79.06	48.79	60.34	-0.31	-0.07	-0.14
-*Number*	78.67	48.66	60.13	-0.70	-0.20	-0.35
-*Person*	79.41	48.98	60.59	**0.04**	**0.12**	**0.11**
-*Semantic*	77.34	47.96	59.21	-2.03	-0.90	-1.27
-*ThreeSentences*	79.02	48.67	60.24	-0.35	-0.19	-0.24
-*TwoSentences*	79.91	49.05	60.78	**0.54**	**0.19**	**0.30**
-*OneSentence*	79.58	49.05	60.69	**0.21**	**0.19**	**0.21**
-*Appositive*	79.30	48.82	60.43	-0.07	-0.04	-0.05
-*NameAlias*	78.73	48.39	59.94	-0.64	-0.47	-0.54
-*CannotLink*	41.81	42.72	42.26	-37.56	-6.14	-18.22

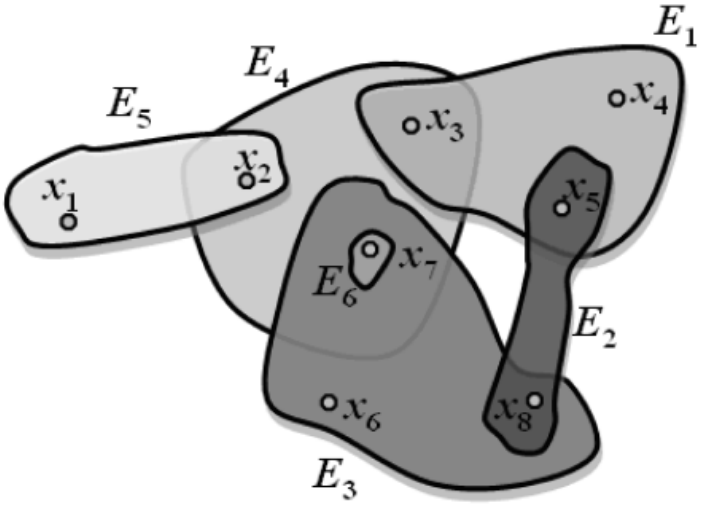


Figure 1. An example of hypergraph

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Reader Perspective Emotion Analysis in Text through Ensemble based Multi-Label Classification Framework

Plaban Kumar Bhowmick (Corresponding author)

Indian Institute of Technology Kharagpur

Kharagpur, India - 721302

E-mail: plaban@gmail.com

Anupam Basu

Indian Institute of Technology Kharagpur

Kharagpur, India - 721302

E-mail: anupam@iitkgp.ac.in

Pabitra Mitra

Indian Institute of Technology Kharagpur

Kharagpur, India - 721302

E-mail: pabitra@cse.iitkgp.ernet.in

Abstract

Multiple emotions are often triggered in readers in response to text stimuli like news article. In this paper, we present a novel method for classifying news sentences into multiple emotion categories using an ensemble based multi-label classification technique called RAKEL. The emotion data consists of 1305 news sentences and the emotion classes considered are disgust, fear, happiness and sadness. Words are the most obvious choice as feature for emotion recognition. In addition to that we have introduced two novel feature sets: polarity of subject, verb and object of the sentences and semantic frames. Experiments concerning the comparison of features revealed that semantic frame feature combined with polarity based feature performs best in emotion classification. Experiments on feature selection over word and semantic frame features have been performed in order to handle feature sparseness problem. In both word and semantic frame feature, improvements in the overall performance have been observed after optimal feature selection.

Keywords: Emotion classification, Multi-label classification, Ensemble classifier, Feature selection

1. Introduction

In the area of Natural Language Processing (NLP), syntactic and semantic level processing of text has been the focus of attention for decades. The related tasks like parts of speech tagging, parsing, machine translation, semantic role labeling have been solved to an acceptable accuracy for different languages. With syntactic and semantic tools in the disposal, the NLP researchers are looking forward to solve the challenges that deal with social and humanistic dimensions of text like emotion, sentiment, attitude, belief etc.

Analysis of the views of the users towards a particular entity is the focus of study in Opinion mining or sentiment analysis (B. Pang & L. Lee., 2004). This task judges an entity in the dimension of positivity or negativity, i.e., whether a particular product is liked by the users or not. On the other hand, emotion analysis of text goes beyond positive-negative dimension to discrete emotion categories like happiness, sadness etc.

Facial or audio expressions are the most notable and prominent clues and have widely been used in analyzing emotion. Though emotion is not a linguistic entity (Z. Kovecses., 2003), in many situations, emotion is expressed through language in day-to-day speech communications or published communications.

Emotion can be analyzed from two different perspectives: From the writer/speaker perspective, where we need to understand the emotional intent of the writer/speaker and from the reader's perspective, where we try to identify the emotion that is evoked in a reader in response to a language stimulus. In the current study, we aim at performing sentence level emotion analysis from a reader's perspective which includes the following challenges.

- *Triggering of multiple emotions*: Given a sentence, a mix of multiple emotions can be triggered in a reader. For example, the following sentence may evoke *fear* and *sad* emotion in readers mind.

A 23-year-old pregnant woman succumbed to swine flu at a city-based hospital.

- *Study of suitable features*: Emotion analysis of text being in its infancy, appropriate feature set required for emotion analysis has not been investigated properly.

- *Feature sparseness*: While emotion analysis in discourse or paragraph level may provide larger number of cues as features, in a single sentence, the number of features is less indicating a feature sparseness problem.

Selection of data source is an important issue. We have considered an age old popular concept in news media for writing emotionally charged news articles called *Emotional framing* (P. E. Corcoran., 2006). According to this theory, each news item is shaped into a form of story with layered dramatic frames, e.g., fear caused by danger; sorrow and grief arising from violence, crime and death; exhilaration and joy resulting from good luck or victory. As a result, amount of news articles capable of evoking emotions in readers is huge. Accordingly, we have rested our study on a set of sentences collected from news articles and headlines.

The contributions of this work are as follows:

- *Multi-Label model of emotion*: The problem reader perspective emotion analysis has been modeled in a multi-label classification framework where the sentence has belongingness in multiple emotion categories. Consequently, the problem of reader emotion classification in text data can be mapped to a multi-label text categorization problem. In this work, we use an ensemble based method, RAKEL (Tsoumakas, 2007), for emotion classification.

- *Feature space exploration*: A thorough exploration of features for reader emotion analysis has been performed in the work. Word feature and word co-occurrence statistics have been used in the earlier works towards reader perspective emotion analysis. In addition to word feature, we have introduced two new features, namely the polarity feature (subject, object and verb) and the semantic frame feature. In the baseline study, word occurrence feature based model is considered. The description and extraction methods for these features have been provided. Semantic frames are generalization of terms or words in the lexicon. Use of semantic frames as feature provides the facility of dimensionality reduction and feature generalization. Thus, semantic frame based emotion recognition model outperforms other feature group based models by a considerable margin. Semantic frame feature, coupled with polarity feature performs best in the selected multi-label classification framework.

- *Feature selection study*: Selection of appropriate features is important as it will help in filtering out the redundant and noisy features from the feature space. Feature selection experiments (χ^2 feature selection) on word and semantic frame features have been performed to train the classifier with optimal feature set.

The rest of the paper is organized as follows: In section 2, we review some of the previous works in writer and as well as reader perspective emotion analysis. In section 3, we point out the limitations in the previous works. A formal representation of the multi-label emotion classification problem and a brief description of the multi-label classification framework used in this study have been provided in section 4. We description and statistics of the emotion data set has been presented in section 5. The features used in this study have been provided in section 6. In section 7, we discuss the experimental set up and present the outcomes of different experiments.

2. Related Works

As stated earlier, emotion analysis can be performed in two different perspectives. So, we provide overview of previous works on emotion analysis from both the perspectives.

2.1 Emotion Analysis in Reader Perspective

Affective text analysis was the task set in *SemEval-2007 Task 14* (C. Strapparava & R. Mihalcea., 2007). A corpus of news headlines collected from Google news and CNN was considered in this task. Two types of tasks were considered: To classify headlines into positive/negative emotion category and as well as distinct emotion categories like anger, disgust, fear, happiness, sadness and surprise.

The system UA-ZBSA (Z. Kozareva, B. Navarro, S. Vazquez & A. Montoyo., 2007) computes statistics from three different search engines (MyWay, AllWeb and Yahoo) to label the news headlines with emotion classes. The work derives the PMI score of each content word of a headline with respect to each emotion by querying the search engines

with the headline and the emotion. The accuracy, precision and recall of the system are reported to be 85.72%, 17.83% and 11.27% respectively.

UPAR7 (F. R. Chaumartin., 2007) adopt a rule-based approach towards emotion classification. The system performs emotion analysis on news headline data provided in SemEval-2007 Task 14. The common words are decapitalized with the help of parts of speech tagger and Wordnet (C. Fellbaum., 1998) in the preprocessing step. Each word is first rated with respect to emotion classes. The main theme word,which is detected by parsing a headline, is given a higher weight than the other words. The emotion score boosting to the nouns are performed based on their belongingness to some general categories in Wordnet. The word scoring also considers some other factors like human will, negation and modals, high-tech names, celebrities etc. The average accuracy, precision and recall of the system are 89.43%, 27.56% and 5.69% respectively.

A supervised approach has been adopted by the system SWAT (P. Katz, M. Singleton & R. Wicentowski., 2007) towards emotion classification in news headlines. The system develops a word-emotion map by querying the Roget's New Millennium Thesaurus. The score each word in the headline is assigned with the help of the created map. The average score of the headline words are considered while labeling it with a particular emotion. The reported classification accuracy, precision and recall are 88.58%, 19.46% and 8.62% respectively.

The work by Lin and Chen (K. H. Y. Lin, C. Yang & H. H. Chen., 2008a; K. H. Y. Lin, & H. H. Chen. 2008b) deals with the method for ranking reader's emotions in Chinese news articles from Yahoo! Kimo News. Eight emotional classes are considered in this work. Support vector machine has been used as the classifier. Chinese character bigram, Chinese words, news metadata, affix similarity and word emotion have been used as features. The best reported system accuracy is 76.88%.

2.2 Emotion Analysis in Writer Perspective

Subasic and Huettner (P. Subasic & A. Huettner. 2001) proposed Fuzzy Semantic Typing approach where manually developed fuzzy lexicon was used. In this lexicon, one word may belong to multiple emotion categories with varying intensity and membership values. Many other works (Y. H. Cho & K. J. Lee., 2006) make use of emotion lexical resources for writer perspective emotion analysis. Wordnet-Affect (A. Valitutti, C. Strapparava & O. Stock., 2004) is used in emotion detection task (C. Ma, H. Prendinger & M. Ishizuka., 2005).

Mishne (G. Mishne., 2005) performs emotion analysis on blogpost corpus. The feature set considered in this task consists of frequency counts, parts of speech (POS) and lemma of words; length related features, PMI-IR features, emphasized words and special punctuation symbols. Classification has been performed using support vector machine. Positive-negative emotion classification accuracy is reported to be 60% and that reported for distinct mood labels is 55%.

Emotion analysis on text corpus consisting of 22 children's fairy tales has been performed in (C. O. Alm, D. Roth & R. Sproat., 2005). The classification was performed with three classes, namely, positive emotion, negative emotion, and neutral. Sparse Network of Winnows learning architecture has been used for the classification task with features like first sentence of the story, quote, thematic role type, Wordnet emotion words etc. The F-score reported for neutral, positive emotion and negative emotion classes are 69%, 32% and 13% respectively.

Leshed et al. (G. Leshed & J. J. Kaye., 2006) considered LiveJournal blog posts as the emotion text corpus and perform emotion classification on 50 topmost emotions appearing in the blog posts. The 'bag of word' model of information retrieval combined with tf-idf feature has been used in SVM classifier to assign emotion labels to the blogposts. The average accuracy of the system is reported to be 78%.

Mihalcea et al. (R. Mihalcea & H. Liu., 2006) consider a corpus based approach to classify blog posts from LiveJournal in 'happy' and 'sad' category. Naive Bayes classifier has been used with unigram features for classification task. The accuracy of the system is reported to be 79.13%.

The *a priori* algorithm for association rule mining and Separable Mixture Model (SMM) techniques have been used for text emotion detection in (C. H. Wu, Z. J. Chuang & Y. C. Lin., 2006). The emotion recognition model was evaluated with a dialog corpus consisting of the students' daily expressions. The model was trained with the dialog corpus and achieved 75.14% precision under a recall rate of 65.67%. The model was then tested with another corpus from different domain (broadcast drama). In this case, the precision was 61.18% at 45.16% recall.

Jung et al. (Y. Jung, H. Park & S. H. Myaeng., 2006) takes a hybrid approach in mood classification in blogposts considering four mood classes: happy, sad, angry and fear. SVM classifier has been used to assign mood labels to documents using the features like term-frequency, n-gram and PMI. The hybrid approach achieves accuracy of 81.80%.

Abasi et al. (A. Abbasi, H. Chen, S. Thoms & T. Fu., 2008) provide an extensive comparison of different features and techniques used for emotion analysis on different corpus and finally propose a support vector regression correlation ensemble (SVRCE) method for text emotion recognition.

The articles taken from periodicals were considered for emotion analysis study in (J. Wang & L Zhang, 2009). Three different models have been constructed in this study: term frequency, semantic characteristics and cognition appraisal theory based models. The micro-average accuracy in cognition appraisal theory based model is reported to be 45%.

3. Limitations of the Previous Works

Based on the study of above mentioned works, following observations can be made.

- Most of the previous works perform emotion analysis in the perspective of the writer, where one text segment portrays only one emotion. On the other hand, one particular text segment can evoke multiple emotions in reader perspective emotion analysis. The multi-labelness of emotion text has not been explored in the previous studies.
- Most of previous studies perform emotion recognition task in document level which may be coarse grained in many applications. Thus finer level analysis like sentence may be explored.
- The performance evaluation measures for multi-label classification are different from that of multi-class or single label classification. So, performance of emotion analyzer should be measured with the metrics defined for multi-label classification.

The above mentioned limitations of the previous works provide the basic motivation behind our work.

4. Multi-Label Emotion Classification Problem and RAKEL

The problem of multi-label emotion classification is defined as follows: Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of emotional sentences and $\xi = \{e_i | i = 1, 2, \ldots, |\xi|\}$ be the set of emotion classes (e.g., happy, sad etc.). The task is to find a function $h : S \rightarrow 2^{\xi}$, where 2^{ξ} is the powerset of ξ.

The problem of reader emotion classification from text data can be mapped to the multi-label text categorization problem. Multi-label classification algorithms have been categorized into two classes: algorithm adaptation methods and problem transformation methods. In algorithm adaptation methods, existing single label classification algorithms are adapted to handle multi-label data whereas, the multi-label data instances are transformed into single label by applying some transformation techniques in problem transformation based methods.

Binary relevance classifier and the Label Powerset classifiers (G. Tsoumakas & I. Katakis., 2007) are examples of problem transformation method. One common problem transformation method is to consider each different subset of ξ as single label. Label Powerset (LP) classifier learns one single classifier $h : S \rightarrow 2^{\xi}$. Random k-Label sets classifier (RAKEL) (G. Tsoumakas & I. Katakis., 2007) builds an ensemble of a number of LP classifiers trained using a different small random subset of ξ. RAKEL has been selected as the representative of problem transformation method in our study as it is reported to outperform the other problem transformation methods and algorithm adaptation methods. Below we provide a brief description of the RAKEL algorithm.

A *k-labelset* is defined as follows:

Definition 1. *A* k-labelset *is defined to be the set* $E \subseteq \xi$ *with* $k = |E|$. *The term* ξ^k *denotes the set of all distinct k-labelsets on* ξ, *where* $|\xi^k| = {}^{|\xi|}C_k$.

The algorithm works in two distinct phases:

- *Ensemble Production:* In this phase, an ensemble of *n* LP classifiers is constructed through *n* iterations. At each iteration, $i = 1, 2, \ldots, n$, a distinct *k*-labelset, E_i, is selected from ξ^k and an LP classifier $h_i : S \rightarrow 2^{\xi i}$ is learnt. The parameter *n*, the number of models, is a user specified one and can assume values ranging from 1 to $|\xi^k|$. The permissible range for another user specified parameter, *k*, is 2 to $|\xi| - 1$.
- *Ensemble Combination:* The multi-label classification is performed by combining votes from individual LP classifiers constructed in the first phase. For a test item *t*, each model h_i provides binary decisions $h_i(t, l_j)$ for each label l_j in the corresponding *k*-labelset *Ei*. Finally the average decision for each label $l_j \in \xi$ is computed. The test instance is labeled with l_j if the average vote is greater than a user specified threshold τ.

5. Emotion Data

The emotion text corpus consists of 1305 sentences extracted from *Times of India* news paper archive. The emotion label set consists of four emotions: disgust, fear, happiness and sadness. A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence to more than one emotion categories. An example annotation is presented in Table 1.

The statistics related to the gold standard data is provided in term of

- Number of sentences in emotion category
- Label Density (LD): It is defined as the average density of labels and is given by

$$LD = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{|E_i|}{|\xi|} \quad (1)$$

where E_i, the proper subset of ξ, is the label set associated with i[th] sentence.

- Label Cardinality (LC): It is defined as the average number of labels associated per sentence and is given by

$$LC = \frac{1}{|S|}\sum_{i=1}^{|S|}|E_i| \quad (2)$$

The statistics of the gold standard data set used in this work is presented in Table 2.

6. Features for Emotion Classification

Three types of features have been considered in our work as given below:

- *Word Feature (**W**)*: Words sometimes are indicative of the emotion class of a text segment. For example, the word 'bomb' may be highly co-associated with fear emotion. Thus words present in the sentences are considered as features. Before creating the word feature vectors, stop words and named entities are removed and the words are stemmed using Porter's stemmer.
- *Polarity Feature (**P**)*: Polarity of the subject, object and verb of a sentence may be good indicators of the emotions evoked. The subject, object and verb of a sentence is extracted from its parse tree and the polarity for each phrase is extracted from manual word level polarity tagging with a set of simple rules. The extraction of polarity related features involves several steps which are presented below with the following polarity tagged (manual) example sentence (Ne → Negative polarity, P → Positive polarity and no tag implies neutral word).

The [shameful]/Ne [scam]/Ne [stains]/Ne the [clean]/P image of the country.

- STEP 1: In this step, the head words of the verb, subject and object phrases are extracted with the help of the dependency relations obtained by parsing the sentence with Stanford Parser (D. Klein & C. D. Manning., 2003). The output in this step is as follows.

Subject head word: *scam*
Verb head word: *stains*
Object head word: *image*

- STEP 2: The modifier words for verb, subject and object head words are determined by consulting the dependency relations. This step yields the following phrases (M → modifier word, H → head word).

Subject phrase: *shameful/M scam/H*
Verb phrase: *stains/H*
Object phrase: *clean/M image/H*

- STEP 3: The polarity of the verb, subject and object phrases are determined with the help of some rules defined over the modifier word, head word and the dependency relation connecting them. The yield of this step is as follows.

Subject phrase: *Ne*
Verb phrase: *Ne*
Object phrase: *P*

- *Semantic Frame Feature (**SF**)*: The Berkeley FrameNet project (C. J. Fillmore, 2003) is a well known resource of frame-semantic lexicon for English. Apart from storing the predicate-argument structure, the frames group the lexical units. For example, the terms 'kill', 'assassin' and 'murder' are grouped into a single semantic frame 'Killing'. In this

work, we shall be exploring the effectiveness of the semantic frames feature in emotion classification. The semantic frame assignment was performed by SHALMANESER (K. Erk, & S. Padó., 2006). Semantic parsing of an example sentence is presented in Figure 1.

7. Experimental Setup and Results

In this section, we present results of experiments of emotion classification with RAKEL which is an ensemble of a number of base classifiers. The base classifier used in our experiment is C4.5. RAKEL deals with three pre-specified parameters that need to be estimated prior training: i) the number of models, ii) the subset size and iii) the threshold for multi-label output generation. The optimal parameters were estimated via 3-fold cross-validation by varying the number of models from 1 to 500, the subset size from 2 to 3 and the threshold from 0.1 to 0.9 with a step of 0.1. For evaluation 5-fold cross-validation were performed for each experiment.

7.1 Experimental Design

In the multi-label classification framework, we aim at exploring the set of suitable features for emotion recognition. We intend to perform different explorations as follows:

- *Exploration of feature sets:* Experiments on finding proper feature combination out of word, polarity and semantic frame features have been performed.
- *Feature selection:* All the word or semantic frame features are not relevant for emotion classification. Feature selection experiment has been conducted in order to find optimal word and semantic frame feature sets.

7.2 Evaluation Measures

We evaluate our emotion classification task with respect to different sets of multi-label evaluation measures:

- Example based measures (G. Tsoumakas & I. Katakis., 2007): Hamming Loss (HL), Partial match accuracy (P-Acc), Subset accuracy (S-Acc) and F1.
- Ranking based measures (M. L. Zhang & Z. H. Zhou., 2007): One Error (OE), Coverage (COV), Ranking Loss (RL), Average Precision (AVP).

7.3 Comparison of Features

Experiments have been performed with different feature combinations. Table 3 summarizes the results of emotion classification with different features and their combinations with best results presented in bold face.

When the assessment of individual features is concerned, the performance of the emotion classifier with polarity feature (P) deteriorated as compared to the baseline classifier (using word feature (W)) for all the evaluation metrics. This explains how important the terms present in the text are for emotion classification.

On the other hand, use of semantic frames (SF) as features improves the performance of emotion classification significantly. The improvements on partial match accuracy (P-Acc), subset accuracy (S-Acc) and F1 are 6.8%, 9.5% and 5.4% respectively. This significant improvement may be attributed to two different transformations over the word feature set.

- *Dimensionality Reduction:* A significant reduction in the dimension of semantic frame feature set as compared to word feature set has been observed(semantic frame feature dimension = 279 and word feature dimension = 2345).
- *Feature Generalization:* Semantic frame assignment to the terms in the sentences is one generalization technique where conceptually similar terms are grouped into a semantic frame. For example, the terms ‘kill’, ‘assassin’ and ‘murder’ are grouped into a single semantic frame ‘Killing’. In semantic frame feature set, unification of these features is performed resulting in less skewedness in feature distribution.

General observations over the feature comparison experiment are as follows.

- The P+SF feature combination performs best in emotion classification with RAKEL. The SF feature performs closer to P+SF as compared to other feature combinations. In case of ranking based measures, the P+SF feature combination outperforms SF by a better margin.
- The polarity feature (P) is inefficient than other combinations but whenever combined with other feature combinations (i.e., W vs. W+P, SF vs. SF+P and W+SF vs. W+SF+P), improvement in performance has been observed. This improvement can be explained with the fact that the polarity feature may help the word or semantic frame based models by classifying the data set into positive and negative category.
- Whenever W feature is coupled with SF, degradation in performance has been noted (i.e., SF vs. W+SF, P+SF vs. W+P+SF). This degradation in performance is due that fact that SF is a generalization over word feature and introduction of word feature only adds noise to the system.

7.4 Feature Selection

All the words and semantic frame features are not important for emotion classification. So, it is better to filter out the words or semantic frames that are not informative enough for discrimination process. To achieve this we have computed χ^2 statistics (Y. Liu, H. T. Loh & A. Sun., 2009) of the word and semantic frame features. Chi-square measures the lack of independence between a word w and a class e_i and is given by

$$\chi^2(w, e_i) = \frac{|S| [P(w, e_i)P(\overline{w}, \overline{e_i}) - P(w, \overline{e_i})P(\overline{w}, e_i)]^2}{P(w)P(\overline{w})P(e_i)P(\overline{e_i})} \quad (3)$$

The χ^2 value for a word w is given by

$$\chi^2(w) = \sum_{i=1}^{|\xi|} \chi^2(w, e_i) \quad (4)$$

The plot word feature χ^2 value vs. rank (see Figure 2) follows the Zipfian distribution (power law fit with equation $y = \alpha x^{-\beta}$ where $\alpha = 236.43$, $\beta = 0.82$ and goodness of fit $R^2 = 0.89$) having a long tail which is strong indication of feature sparseness problem.

We performed experiment on selecting optimal W feature set size based on their χ^2 values. Top 40% of the total W feature set is found to be optimal feature set. The relative performance after feature selection for W is shown in Figure 3. Similar experiment was performed to select important SF features. Top 80% out of the total set was selected as optimal feature set for SF feature. The relative performance with the selected SF feature set is presented in Figure 4.

It is evident from results that the there is a slight improvement in performance after adopting feature selection strategy for both the feature sets. With P+SF feature combination being the close competitor, best performance is achieved with P+80%SF (HL = 0.110, P-Acc = 0.769, F1 = 0.821, S-Acc = 0.670).

7.5 Comparison with Other Systems

The existing methods model the emotion recognition as a single label classification problem. Whereas multi-label classification based approach has been adopted in our approach. As the performance measures for single label classification tasks are different from that of multi-label one, direct comparison of the existing system with ours is not possible. Comparisons may only be performed based on the micro-averaged label based measures like accuracy, precision, recall, and F1 (Tsoumakas, 2007). The comparison of our system with the existing approaches based on these measures is provided in Table 4.

8. Conclusions

In this paper, we have presented a multi-label classification based emotion analysis model. The emotion corpus considered in this study consists of 1305 sentences collected from news archive. An ensemble based multi-label classification technique called *random k-label set (RAKEL)* has been used in our study.

Apart from traditional word feature, we have introduced two other feature groups, namely, polarity based features and semantic frame based features. Experiments with different feature combinations reveal that semantic frame feature combined with polarity based feature performs best in RAKEL framework.

The spurious word or semantic frame features that may not be important in emotion classification task should be removed. To achieve this, we have adopted χ^2 statistics based feature selection strategy. Improvements in performance have been observed after feature selection in both word and semantic frame feature.

References

A. Abbasi, H. Chen, S. Thoms & T. Fu. (2008). Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168–1180.

A. Valitutti, C. Strapparava & O. Stock. (2004). Developing affective lexical resources. *PsychNology Journal*, vol. 2, no. 1, pp. 61–83, 2004.

B. Pang, & L. Lee. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL 2004,* Barcelona, Spain, Main Volume, pp. 271–278.

C. Fellbaum. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press.

C. J. Fillmore (2003) Background to framenet. *International Journal of Lexicography*, vol. 16, no. 3, pp. 235–250.

C. H. Wu, Z. J. Chuang & Y. C. Lin. (2006). Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, no. 2, pp. 165–183.

C. Ma, H. Prendinger & M. Ishizuka. (2005). Emotion estimation and reasoning based on affective textual interaction. In *ACII* pp. 622–628.

C. O. Alm, D. Roth & R. Sproat. (2005) Emotions from text: machine learning for text-based emotion prediction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 579–586.

C. Strapparava & R. Mihalcea. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic.

D. Klein & C. D. Manning. (2003). Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 423–430.

F. R. Chaumartin. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 422–425.

G. Leshed & J. J. Kaye. (2006). Understanding how bloggers feel: recognizing affect in blog posts. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, pp. 1019–1024.

G. Mishne. (2005). Experiments with mood classification in blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access*.

G. Tsoumakas & I. Katakis. (2007). Multi label classification: An overview. *International Journal of Data Warehouse and Mining*, vol. 3, no. 3, pp. 1–13.

K. Erk, & S. Padó. (2006). Shalmaneser - a toolchain for shallow semantic parsing. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 527–532.

K. H. Y. Lin, C. Yang & H. H. Chen. (2008a). Emotion classification of online news articles from the reader's perspective. In *Web Intelligence*, pp. 220–226.

K. H. Y. Lin, & H. H. Chen. (2008b). Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 136–144.

M. L. Zhang & Z. H. Zhou. (2007). ML-kNN: A lazy learning approach to multilabel learning. *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048.

P. E. Corcoran. (2006). Emotional framing in Australian journalism. In *Australian & New Zealand Communication Association International Conference*. Adelaide, Australia: ANZCA.

P. Katz, M. Singleton & R. Wicentowski. (2007). SWAT-MP: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 308–313.

P. Subasic & A. Huettner. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transaction on Fuzzy Systems*, vol. 9, no. 4, pp. 483–496.

R. Mihalcea & H. Liu. (2006). A corpus-based approach to finding happiness. In *AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*. AAAI Press, pp. 139–144.

Y. H. Cho & K. J. Lee. (2006). Automatic affect recognition using natural language processing techniques and manually built affect lexicon. *IEICE - Transactions on Information and Systems*, vol. E89-D, no. 12, pp. 2964–2971.

Y. Jung, H. Park & S. H. Myaeng. (2006). A hybrid mood classification approach for blog text. In *PRICAI*, pp. 1099–1103.

Y. Liu, H. T. Loh & A. Sun (2009). Imbalanced text classification: A term weighting approach. *Expert systems with Applications,* vol. 36, no. 1, pp. 690-701.

Z. Kovecses. (2003). Language and emotion concepts. In *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge: Cambridge University Press.

Z. Kozareva, B. Navarro, S. Vazquez & A. Montoyo. (2007). UA-ZBSA: A headline emotion classification through web information. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 334–337.

Table 1. An example annotation (1 → Emotion evoked in reader, 0 → Emotion not evoked)

Sentence	**Disgust**	**Fear**	**Happiness**	**Sadness**
The four terrorists in the Taj Mahal hotel have killed virtually anyone and everyone they saw.	0	1	0	1

Table 2. Statistics for emotion data

Sentence Distribution		**LC**	**LD**
Class	No of Sentences		
Disgust	307	1.30	0.26
Fear	371		
Happiness	282		
Sadness	735		

Table 3. Comparison of features (W → Word feature, P → Polarity feature and SF → Semantic frame feature)

Measure	**W**	**P**	**SF**	**W+P**	**P+SF**	**W+SF**	**W+P+SF**
Example based measures							
HL	0.166	0.237	**0.112**	0.155	**0.112**	0.123	0.129
P-Acc	0.685	0.531	0.753	0.710	**0.764**	0.752	0.748
F1	0.744	0.589	0.798	0.766	**0.812**	0.809	0.806
S-Acc	0.566	0.414	0.661	0.597	**0.666**	0.635	0.627
Ranking based measures							
OE	0.343	0.342	0.156	0.215	**0.146**	0.171	0.169
COV	0.910	1.119	0.672	0.785	**0.663**	0.684	0.684
RL	0.235	0.347	0.203	0.241	**0.194**	0.198	0.199
AVP	0.796	0.770	0.893	0.857	**0.898**	0.887	0.887

Table 4. Comparison of proposed system with other emotion classification systems

System	**Accuracy**	**Precision**	**Recall**	**F1**
UPAR7	89.43	27.56	5.69	9.43
UA-ZBSA	85.72	17.83	11.27	13.81
SWAT	88.58	19.46	8.62	11.95
Li and Chen	76.88	--	--	--
Our System	**88.2**	**84.42**	**79.93**	**82.1**

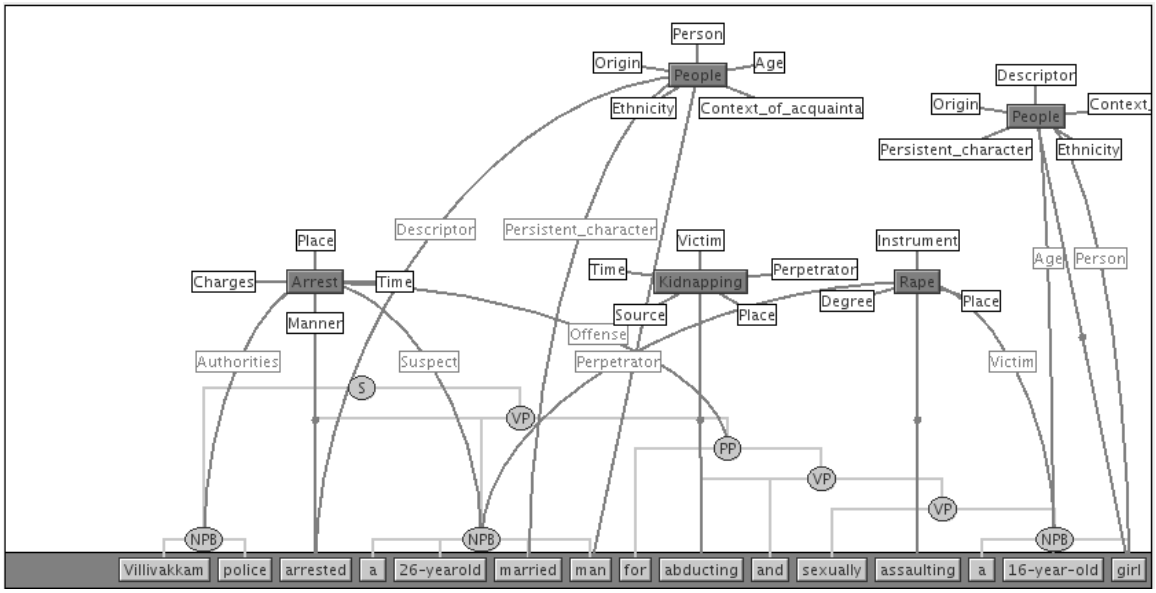


Figure 1. Semantic parsing of an example sentence

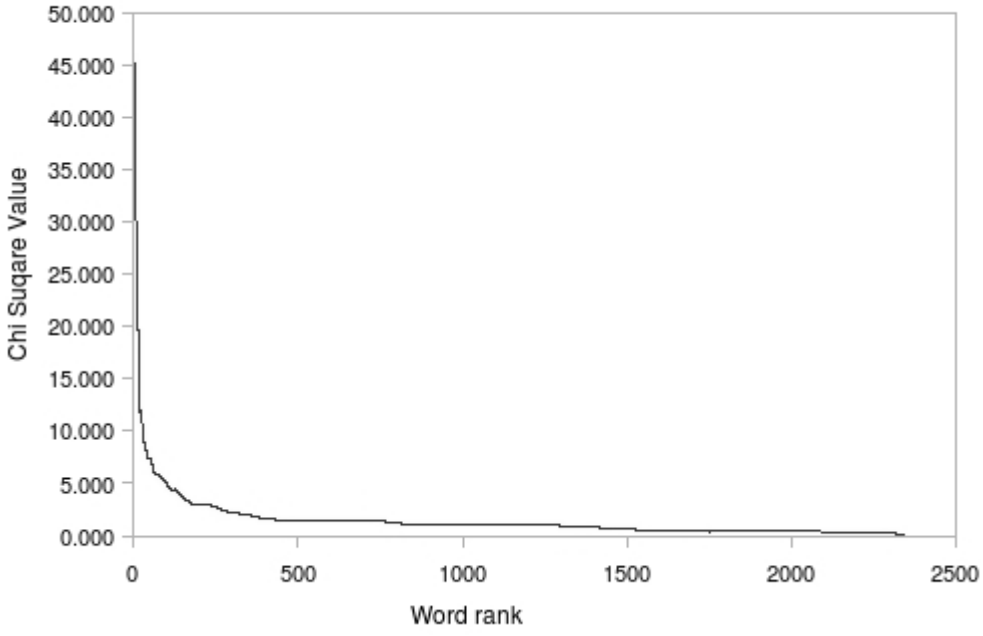


Figure 2. χ^2 vs. rank plot for word feature

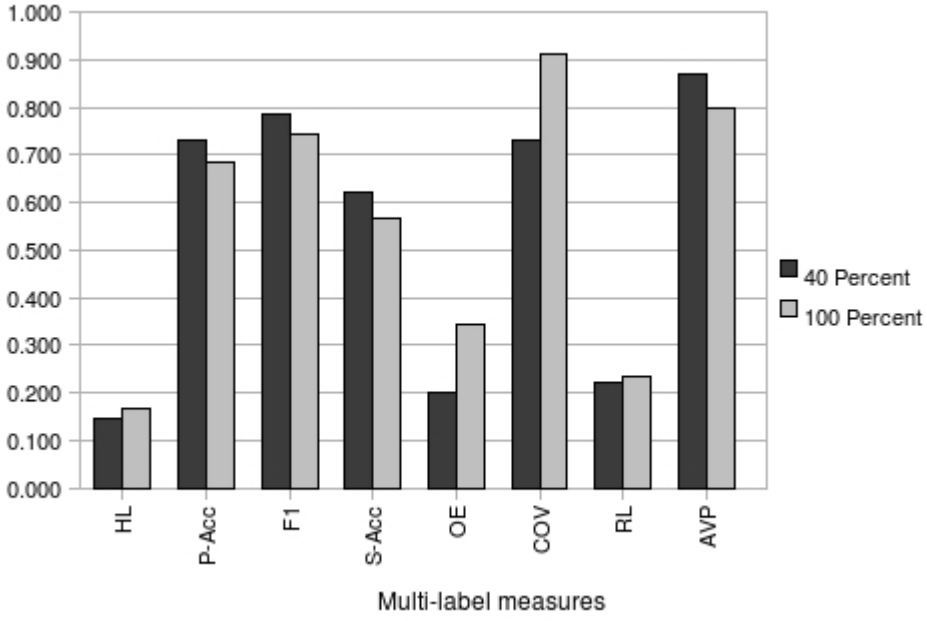


Figure 3. Relative performance after χ^2 word feature selection

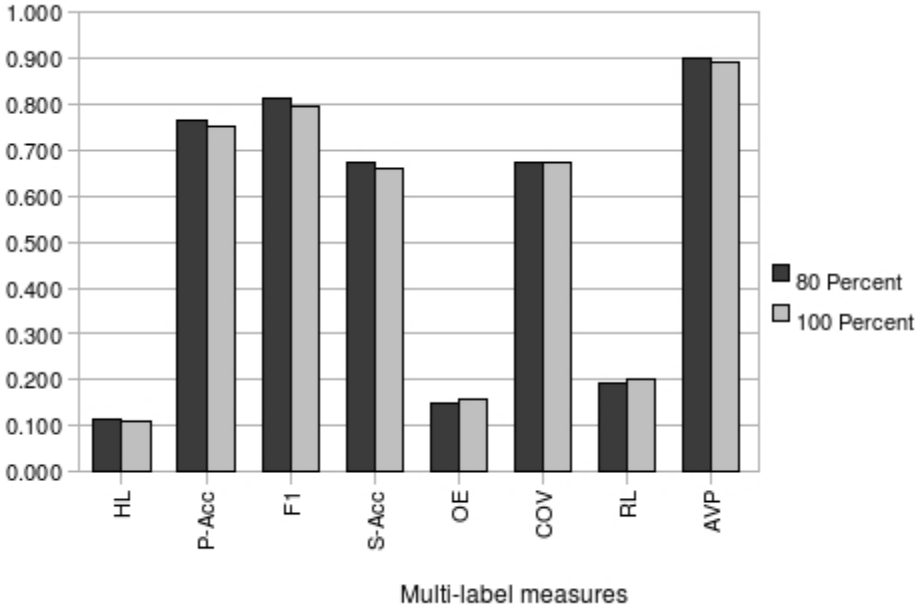


Figure 4. Relative performance after χ^2 semantic frame feature selection

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Artificial Fish Swarm Algorithm-Assisted and Receive-Diversity Aided Multi-user Detection for MC-CDMA Systems

Zhicheng Dong, Wei Xiao & Xiping Zhang

Electronic information engineering department

School of Engineering, Tibet University, Lhasa 850000, China

Tel: 86-817-288-4617 E-mail: dongzc666@163.com

Abstract

Artificial fish swarm algorithm (AFSA) assisted multi-user detection (MUD) is proposed for the receive-antenna-diversity-aided multi-carrier code-division multiple-access (MC-CDMA) systems in frequency selective fading channel. Due to the receive-diversity, the signals received at the different antennas are faded independently, resulting in an independent objective function for each antenna. To resolve the multi-objective dilemma when choosing one signal estimation for multiple receive antenna-branches, the individuals associated with the AFSA are selected based on the concept of Pareto optimality, which uses the information from the antennas independently. Simulation results showed that: with the same computation complexity, the strategy has much better bit error rate (BER) performance than the convention one. Comparisons with the conventional multiuser detector and the decorrelator verified the effectiveness of the proposed scheme.

Keywords: Antenna-diversity, MC-CDMA, Artificial fish swarm algorithm, MUD, Pareto optimal

1. Introduction

Recently, the multi-carrier code-division multiple access (MC-CDMA) based on the combination of OFDM (Orthogonal Frequency Division Multiplexing) technique and conventional CDMA has attracted great interests in both practical and theoretical studies of modern mobile communications because of its high spectral efficiency and high data rate transmission(Abrao, T.; Ciriaco, F.; de Oliveira, L.D.; Jeszensky, P.J.E. 2006)(B. Steiner. 1997). Due to multiple users face serious multi-access interference (MAI) in the uplink and limit the capacity, the optimum multi-user detector (OMD) proposed by Verdu(C. Ergun, K. Hacioglu. 2000) in 1986. But the computation complexity of this detector increases exponentially with the number of users and make it impractical for an actual system. Consequently, the major portion of the scholars focus on suboptimal detectors, which are near-far resistant, have reasonable computational complexity, and approach the performance of OMD.

The mathematical model of MUD can be considered as a combinatorial optimization problem, which is also a NP-complete problem. Genetic Algorithm(GA) and Particle Swarm Optimization(PSO) can solve these problems effectively, so many scholars proposed the GA-based multi-user detection (GAMUD) and the PSO-based multi-user detection (PSOMUD)(D.E. Goldberg. 1989)(E. Zitzler, L. Thiele. 1999)(G. Syswerda. 1989)(Jianmei Xiao; Xiaoming Zheng; Xihuai Wang; Youfang Huang, 2006).

Artificial fish Swarm Algorithm (AFSA) is a new kind of intelligence optimization algorithm, which has some advantages that GA and PSO do not have. Multi-user detection based on AFSA has been proposed (K. Yen, L. Hanzo. 2003) and show that have better performance than GAMUD and PSOMUD.

In the receive diversity aided systems, the antennas are assumed to be sufficiently separated such that the received signals at the antennas are faded independently, resulting in an independent log-likelihood function (LLF) for each antenna. This poses a problem of multi-objective optimization due to the fact that while a specific signal estimation may be deemed optimum on the basis of the LLF of one antenna, the same estimation may not necessarily be deemed optimum in terms of the LLF of the other antenna. To independently use the fading information of different antennas, Yen and Hanzo (S. Verdu. 1986) proposed a genetic algorithm (GA) based approach which takes the Pareto optimality (U. Fawer, B. Aazhang. 1995) into the consideration. However, many researches show that AFSA approach not only has better performance than the GA approach but also has lower computational complexity (Yang Yu, Ya-fei Tian, Zhi-feng Yin. 2005).In order to resolve this multi-objective dilemma in a more effective and less complex way, this letter

proposes an AFSA based MUD with Pareto optimal.

This paper is organized as follows. Section II describes the synchronous MC-CDMA system communicating over frequency-selective fading channel using some antennas. Section III describes the AFSA used to implement our proposed detector in conjunction with diversity reception, and analyses complexity issues. Our simulation results are presented in Section IV, while Section V concludes the letter.

2. System Model

We consider the uplink of the synchronous MC-CDMA systems (B. Steiner. 1997), where the base-station receiver is equipped with M far-spaced receive antennas ,shown in Fig .1,the sub-carriers and the length of the frequency domain spreading code is N and K synchronous users are accessed simultaneously .When considering the kth user, the information symbol $b_k(t)$ emitted by the kth user is multiplied with the frequency domain spreading code $C_k = 1/\sqrt{N}[c_k(1),..c_k(N)]^T$,where $c_k(n) = \pm 1$,with $(\square)^T$ denoting transpose; the path gain of user k from it's transmit antenna to the mth receive antenna is denoted as $\mathrm{h}_{(k,m)}$, $\mathbf{H}_{k,m} = \mathbf{F}\mathbf{h}_{(k,m)}$ where $\mathbf{F}$ denotes the discrete Fourier transform (DFT). $\sqrt{A_k}$ denoted the received energy for the kth user.

The signal vector received at the mth receive antenna can be expressed in the following form

$$\mathbf{y}_m(t) = \sum_{k=1}^{K} \mathbf{H}_{k,m} C_k \sqrt{A_k} b_k(t) + \mathbf{n}_m(t) = \mathbf{H}_m \mathbf{A}\mathbf{b}(t) + \mathbf{n}_m(t) \tag{1}$$

where $\mathbf{b}(t) = [b_1^T(t),...,b_K^T(t)]^T$, $\mathbf{A} = diag(\sqrt{A_1},...,\sqrt{A_K})$ is the diagonal matrix of the received bit energy for the K users. $\mathbf{H}_m = [\mathbf{H}_{1,m} \square C_1,...,\mathbf{H}_{K,m} \square C_K]$, where $\square$ denoted multiplying element by element. where $\mathbf{n}_m(t)$ is the noise part received at the mth antenna, is the zero-mean complex additive white Gaussian noise (AWGN) with independent real and imaginary components, each having a double-sided power spectral density of $N_0/2$.

At the receiver-side, the MLD detects all users' data to jointly minimize the effects of MAI (Y. Du, K.T. Chan, 2003). For the mth antenna, the log-likelihood function (LLF) is defined as

$$\Lambda_m(b) = 2\mathrm{Re}\left\{b^H H_m^H y_m\right\} - b^H H_m^H H_m b \tag{2}$$

The decision rule for the MLD associated with the mth antenna is:

$$\hat{b}_m = \arg\left\{\max_b [\Lambda_m(b)]\right\} \tag{3}$$

Since the channel fading of different receive antennas are independent, normally there exists $\Lambda_m(b) \neq \Lambda_{m'}(b)$ where, $(m \neq m')$,When there exists deep fading in some antennas, the estimations of the transmitted signal corresponding to different antennas are not coincident, ie.

$$\arg\left\{\max_b [\Lambda_m(b)]\right\} = \hat{b} \neq \arg\left\{\max_b [\Lambda_{m'}(b)]\right\} \tag{4}$$

Nevertheless, for the MLD, the LLF corresponding to different antennas are combined according to (Ying Zhao; Junli Zheng, 2004)

$$\Lambda(b) = \sum_{m=1}^{M} \Lambda_m(b) \tag{5}$$

The decision rule is then to find the signal estimation that maximizes (5).

3. AFSA-Based Multi-user Detection with Diversity Reception

AFSA performs a parallel search in the solution space to find the optimum solution by simulating the behavior of fish including searching food, congregate and follow (Yang Yu, Ya-fei Tian, Zhi-feng Yin. 2005). In other words, this algorithm searches the optimum solution based on the cooperation and competition of the fish individuals .In this letter, we will employ AFSA in order to detect the estimated transmitted bit vector, shown in Fig .2.

AFSA commence their search for the optimum K bit state at the so-called y=0th generation by randomly creating P legitimate K bit state, where the pth artificial fish (AF) is expressed here as $\hat{\mathbf{b}}_p^y = [\hat{b}_{p,1}^y, \hat{b}_{p,2}^y, ..., \hat{b}_{p,K}^y]$.

3.1 The behavior of selection

Generally, in convention AFSA-MUD, the new AF swarm will been all selected for exploitation in the subsequent generation. In this letter, T AFs will been selected from the school, where $2 \le T < P$,then produced a new population of offspring using the so-called uniform crossover(Z. Li, M. J. Juntti, M. Latva-aho, 2005) process.

In the convention strategy, each K bit AF is associated with a fitness value denoted as $f(b_p) = \Lambda(b_p)$, which is a function of the LLF of (5)(S. Verdu. 1986). AFs having the T highest fitness values in the population, where $2 \le T < P$, are then selected and placed in the mating pool.

The second AF-selection strategy of the AFSA –assisted multi-user detection is based on the concept of the so-called Pareto optimality(U. Fawer, B. Aazhang. 1995). This strategy favors the so-called non-dominated AF and ignores the so-called dominated AF. Here, the pth K bit AF is associated with M+1 fitness values denoted as $f(b_p)=[\Lambda_1(b_p),..\Lambda_M(b_p),\Lambda(b_p)]$, where the first M fitness values are functions of the LLF of (2), while the last fitness value

is a function of the LLF of (5). Then the ith K bit AF is considered to be dominated by the jth AF iff (U. Fawer, B. Aazhang. 1995)

$$\begin{aligned} &\forall m \in \{1,\ldots,M\} : \Lambda_m(b_j) \geq \Lambda_m(b_i) \\ &\wedge \exists m' \in \{1,\ldots,M\} : \Lambda_{m'}(b_j) > \Lambda_{m'}(b_i) \end{aligned} \tag{6}$$

If an AF is not dominated in the sense of (6) by any other K bit AFs in the population, then by definition it is considered to be non-dominated. According to our second AF-selection strategy, all the non-dominated K bit AFs are selected and placed in the mating pool. Hence, the value of T in this case is not fixed, since it depends on the number of non-dominated AFs. If there is only one non-dominated AF in the current population, the next non-dominated AFs will be selected, so that there will be more than one AF in the mating pool.

3.2 The behavior of crossover

Two K bit AFs in the mating pool are then selected as parents based on their corresponding figure of merit in (5) .The antipodal bits of the K bit parent vectors are then exchanged using uniform crossover(Z. Li, M. J. Juntti, M. Latva-aho, 2005) process, in order to produce two K bit offspring. The selection of K bit parents from the mating pool is repeated, until a new population of P offspring is produced(Z. Li, M. J. Juntti, M. Latva-aho, 2005).

3.3 The behavior of congregation

The new population of P offspring updates their state by the behavior of searching food, congregate and follow .According to the character of the problem, the AF evaluates the environment at present, and then selects an appropriate behavior. In the letter, the AF executes the behavior of congregate first, if its state cannot be improved , then executes the behavior of searching food(Yang Yu, Ya-fei Tian, Zhi-feng Yin, 2005).

Let us assume that b_p is the pth AF state at present. b_p explores the center position b_c and the number n_f of its fellow within the visual distance. If $1 \leq n_f < P$, and $\Lambda(b_c) > \Lambda(b_p)$, which means state of b_c is better than b_p and the surroundings is not very crowded, then the AF moves from b_p to b_c .otherwise executes the behavior of searching food.

3.4 The behavior of searching food

Let us assume that b_p is the pth AF state at present. The AF selects a state b_j randomly within the visual distance(Yang Yu, Ya-fei Tian, Zhi-feng Yin, 2005). If $\Lambda(b_c) > \Lambda(b_p)$, which means the state of b_j is better than b_p ,the AF move from b_p to b_j ;otherwise select a state b_j randomly again and justify if it satisfy the forward requirement. If it can not meet the requirement after TN times, select a state randomly.

Evaluate the scalar fitness value of each AF by computing the fitness of the equation (5), and compare with the fitness value of AF_g . If its fitness is higher than the fitness value of AF_g , instead the AF_g by its state; Otherwise the AF_g remains unchanged. We identify the lowest merit K bit AF in the population and replace it with AF_g under elitism. This will ensure that the highest merit AF is propagated throughout the evolution process.

The AFSA terminates after Y-1 number of generations. The AF corresponding to the highest scalar fitness value in (5) is the detected K number of users' bit vector.

3.5 Complexity Issues

The number of computations multiplications and additions required to detect K bits for the conventional multi-user detector is $2KN$; The number of computations for the decorrelator is $2K(K+N)+INV$ (S. Verdu. 1986),where INV denotes computation for the inverse correlation matrix, is on the order of K^3 ; The number of computations for the proposed AFSA based detector is $[2K(K+N)+5K+1+TN]PY$,where TN denotes the time of selection a state randomly. Hence, the values of P,Y and TN can be adaptively selected, in order to find a tradeoff between the computational complexity and the performance.

4. Simulations Result

In system we have adopted the following parameters: the spread sequence are selected as pseudo-noise(PN) with processing gain $N = 32$; the receiver, consists of two antennas separated spatially the number of active synchronous users is $K = 16$;two-paths slow Rayleigh channels with the second ray delay T_c from the first . Perfect power control and CIR estimation was assumed. The strategy based on the sum of the figures of merit from both antennas, will be denoted as S1, while the strategy based on the Pareto optimality will be denoted as S2.

Fig. 3 shows the BER performance against the average signal-to-noise ratio (SNR) for the ASFA-based multi-user detector employing AF-selection strategy S1 and S2.For the sake of comparison, the BER performance of a decorrelator and conventional multi-user detector is also shown. An error floor is observed for the results shown in the figure. This is due to the limitations of the ASFA associated with the particular set of P and Y values, not due to the multiple access interference (MAI). It is seen in Fig. 3 that the BER performance for improved when number of generations size was

increased from 10 to 20.However, this also increased the computational complexity. Hence, the value of Y can be selected, in order to find a tradeoff between computational complexity and performance. More importantly, we also see from Fig. 3 that the ASFA employing S2 performs better, exhibiting a lower error floor than S1.

Fig. 4 shows the BER performance against the average signal-to-noise ratio (SNR) for the ASFA-based multi-user detector. It is seen in Fig. 4 that the BER performance for improved when population size was increased from 10 to 20. However, this also increased the computational complexity. Hence, the value of Y can be selected, in order to find a tradeoff between computational complexity and performance. We also see from Fig. 4 that the ASFA employing S2 performs better, exhibiting a lower error floor than S1.

Fig. 5 shows the BER performance against the average signal-to-noise ratio (SNR) for the ASFA-based multi-user detector. It is seen in Fig. 5 that the BER performance for improved when the times of searching food TN was increased from 2 to 6. However, this also increased the computational complexity. Hence, the value of TN can be selected, in order to find a tradeoff between computational complexity and performance. We also see from Fig. 5 that the ASFA employing S2 performs better, exhibiting a lower error floor than S1.

5. Conclusions

AFSA is a new kind of intelligence optimization algorithm. Considering the MUD from a combinatorial optimization viewpoint, many scholars have employed the intelligence optimization algorithm such as GA and PSO to solve this problem. We developed a suboptimal multi-user detector based on AFSA. To mitigate the effects of fading, dual-antenna diversity techniques were used. We have shown that AFSA employing the strategy based on the Pareto optimality always exhibit a lower BER compared to those employing the convention strategy. We have also shown that the BER performance can be improved by increasing, number of generations size, the population size and the times of searching food. Hence, the value of Y, P and TN can be selected, in order to find a tradeoff between computational complexity and performance.

References

Abrao, T. Ciriaco, F. de Oliveira, L.D. Jeszensky, P.J.E. (2006). Particle Swarm and Quantum Particle Swarm Optimization Applied to DS/CDMA Multiuser Detection in Flat Rayleigh Channels. Page(s):133 – 137 Aug. 2006.

B. Steiner. (1997). Uplink performance of a multicarrier-CDMA mobile radio system concept[C]. *IEEE Proc of VTC97*, 3:1902-1906.

C. Ergun, K. Hacioglu. (2000). "Multiuser detection usiing a genetic algorithm in CDMA communications systems," IEEE Trans. Commun., vol. 48, pp. 1374-1383, Aug. 2000.

D.E. Goldberg. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning[M]. *MA: Addison-Wesley*.

E. Zitzler, L. Thiele. (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Trans Evol Compute,* 3(4): 257-271.

G. Syswerda. (1989). "Uniform crossover in genetic algorithms," in Proc. 3rd Int. Conf. Genetic Algorithms, J. D. Schaffer, Ed.. San Mateo, CA.

Jianmei Xiao; Xiaoming Zheng; Xihuai Wang; Youfang Huang. (2006). A Modified Artificial Fish-Swarm Algorithm. Volume 1, Page(s):3456 – 3460.

K. Yen, L. Hanzo. (2003). Antenna-diversity-assisted genetic-algorithm-based multiuser detection schemes for synchronous CDMA systems. *IEEE Trans Commun*.

S. Verdu. (1986). Minimum probability of error for asynchronuous Gaussian multiple-access channels. *IEEE Trans Inform Theory*, 32(1):85-96.

U. Fawer, B. Aazhang. (1995). A multiuser receiver for code-division multiple-access communications over multipath channels[J]. *IEEE Trans Commun*, 1995, 43(2): 1556-1565. Feb.-Apr. 1995.

Y. Du, K.T. Chan. (2003). Feasibility of applying genetic algorithms in space-time block coding multiuser detection systems[C]. Proc of ICWOC.

Yang Yu, Ya-fei Tian, Zhi-feng Yin. (2005). Multiuser Detector Based on Adaptive Artificial Fish School Algorithm. China.

Ying Zhao; Junli Zheng. (2004). Particle swarm optimization algorithm in signal detection and blind extraction. 10-12, Page(s):37 – 41, May 2004.

Z. Li, M. J. Juntti, M. Latva-aho. (2005). "Genetic algorithm based frequency domain multiuser detection for MC-CDMA systems," *IEEE Proc of VTC2005*, vol. 2, pp. 983-987, June 2005.

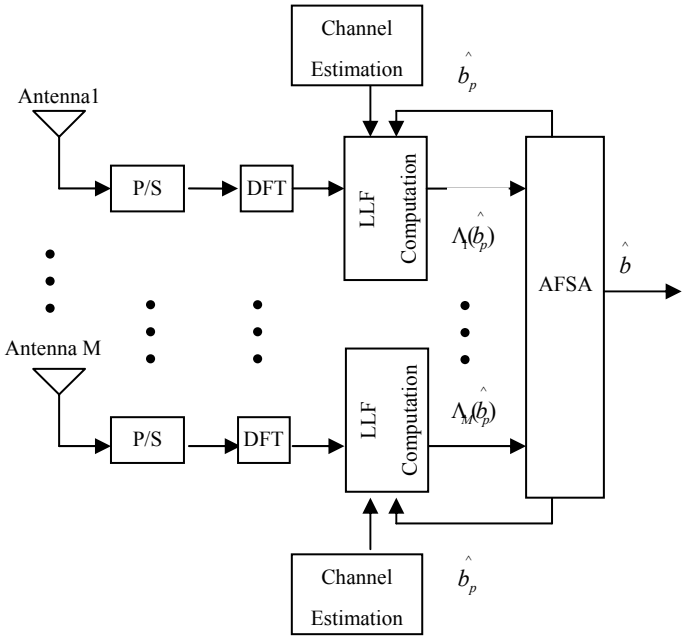


Figure 1. Block diagram of the receiver model

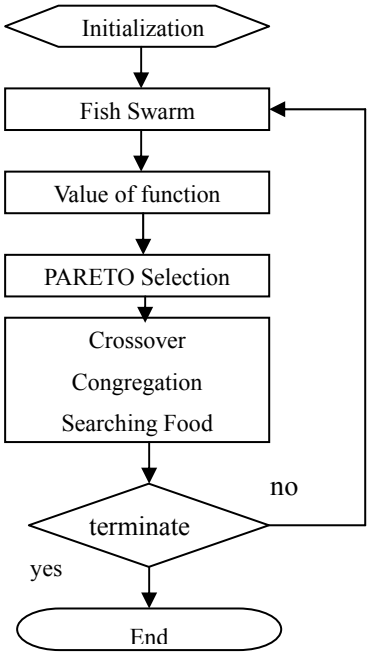


Figure 2. AFSA Program

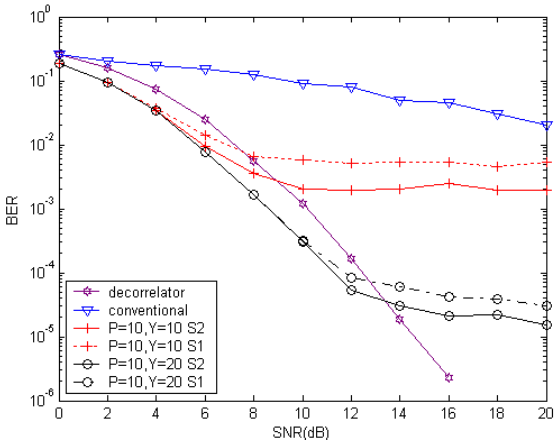


Figure 3. BER performance of the AFSA-based multi-user detector employing the AF-selection strategies of S1 and S2 with population sizes of P = 10, Y=10; 20. $TN = 2$, using signature sequences of length 32and supporting K = 16.

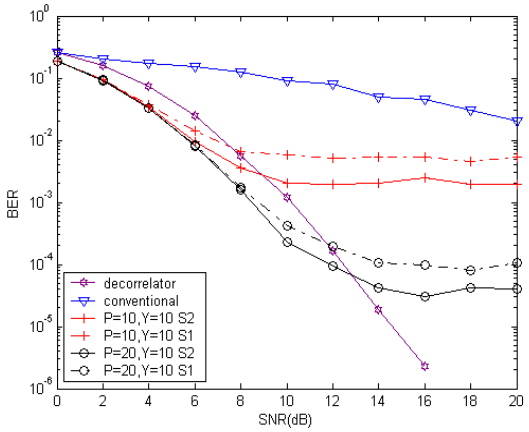


Figure 4. BER performance of the AFSA-based multi-user detector employing the AF-selection strategies of S1 and S2 with population sizes of Y = 10, P=10; 20. $TN = 2$, using signature sequences of length 32and supporting K = 16.

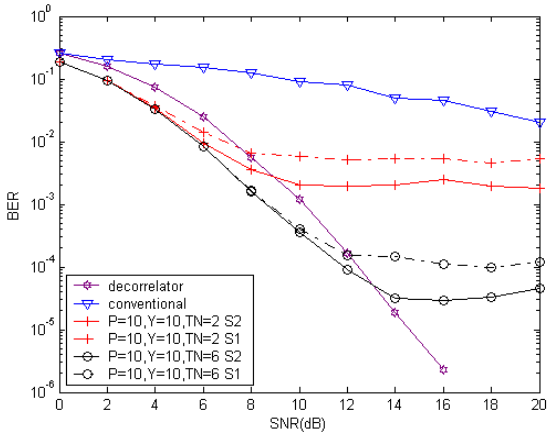


Figure 5. BER performance of the AFSA-based multi-user detector employing the AF-selection strategies of S1 and S2 with population sizes of Y =10, P=10, TN=2; 6, using signature sequences of length 32and supporting K = 16.

Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 4

November 2009

Strict versus Negligence Software Product Liability

Farhah Abdullah

Department of Law

Universiti Teknologi MARA (UiTM)

Dungun Campus, Dungun 23000, Terengganu, Malaysia

Tel: 60-1360-48211 E-mail: farha523@tganu.uitm.edu.my

Kamaruzaman Jusoff (Coresponding author)

TropAIR, Faculty of Forestry, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Tel: 60-3-8946-7176 E-mail: kjusoff@yahoo.com

Hasiah Mohamed

Universiti Teknologi MARA (UiTM)

Faculty of Computer and Mathematical Sciences

Dungun Campus, Dungun 23000, Terengganu, Malaysia

Tel: 60-19989-7963 E-mail: hasia980@tganu.uitm.edu.my

Roszainora Setia

Universiti Teknologi MARA (UiTM)

Language Academy

Dungun Campus, Dungun 23000, Terengganu, Malaysia

Tel: 60-1-9919-9909 E-mail: roszainora@tganu.uitm.edu.my

Abstract

The law of products liability in tort is designed to maintain a reasonable balance between the inevitable social costs and the benefits of innovative product technologies. Technological development must be supported not only for the best interests of the public but also the side effect namely product defect into one of the following: (1) manufacturing defect; failures to correctly implement safety measures from the design; and (2) design defects: failures of the design itself to exhibit socially acceptable levels of safety. Software has been described as an artifact with fundamentally different properties than other engineered artifacts. This article will discuss several issues such as identifying the features of high technology products which lead to difficulties in applying traditional tort notions to them.

Keywords: Software, Strict liability, Negligence, Damages, Risk

1. Introduction

Software is a relatively new technological artifact to reach the consumer market. It has offered many technological possibilities and also, the potential for personal injury (Stephen, 1998). Such products have already been involved in cases of personal injury and death (Leveson & Turner, 1993). When a consumer is injured and the common law of products liability is invoked, two different legal standards are available: negligence and strict liability. Software is the new tool; its nature is totally different from the application of traditional legal and engineering tools and defect classifications used to determine the standard of liability.

The modern law of products liability developed from roots in negligence and warranty causes of action. In the 1960s strict products liability in tort developed in response to the perceived inadequacies of negligence and warranty causes of action when applied to products of modern complexity involved in personal injury (S. L. Birnbaum, 1980). It was to be based on proof of product defect rather than proof of fault. Once a product defect was proven to have caused injury to the person,

damages could be awarded. But, for a negligence case, unreasonable conduct must be proved and the injury may be merely economic in nature. Fault (unreasonable conduct) must be proved, the injury may be merely economic in nature, and a product need not be the instrumentality of the injury (it could be a service) (Ravi Belani, n.d.).

Here are some selected examples of the following list of software errors that resulted in recalls of medical equipment (Armour & Humphrey, 1993a), namely incorrect match of patient and data, incorrect readings in an auto analyzer, faulty programming provided false cardiac FVC values, faulty programming caused pacer telemetry errors, incorrect software design caused lockup of cardiac monitor, incorrect calculations, failure of central alarm in arrhythmia monitor, table top moved without a command, detector head could hit the patient, algorithm error caused low blood pressure readings, and over infusion due to programming error. The reasonableness of human conduct presents many arguable issues, and this raises expected costs for the plaintiff in prosecuting a successful case (Keeton & Dobbs, 1984).

The common "limitations of liability" and bold statements negating the manufacturer's responsibility for the behavior of the product have no bearing on a products liability case (Ravi Belani, n.d.). In order to apply, the strict products liability standard, personal injury must be involved. Mere dissatisfaction with the product or economic damages will not support a case; neither will the provision of services. Thus, if a software defect threatens the person or property of a customer or a third party, the injured party is entitled to bring a strict liability claim against the supplier notwithstanding contractual disclaimers, limitations of remedies, and limited warranties. The virtue of a strict liability action from the claimant's perspective is that there is no need to prove that the supplier was negligent. If the product was defective and it caused the claimant's loss, strict liability applies (Armour & Humphrey, 1993b).

There is some debate about whether software is a product or a service. There is growing evidence, however, that courts will consider software a product. The instrumentality causing the injury must be considered a “product”. Thus, the first critical legal inquiry may be whether a product is involved in the injury. Next, the defect must be classified. The term “defect” has evolved to encompass two general categories: defects in “manufacture” and defects in design. Both categories have risen to the same legal label: strict products liability. However, the defect in design has been a controversial category respecting its proper inclusion under the strict products liability rubric which should be subject to a negligence standard. This section will continue to discuss on the problem of vendor liability in respect of a critical issue of whether software is a “Product” or a “Service”.

As computer technology evolves, more powerful computer systems and software are available to more individual users and businesses (Gemignani, 1980). As a result, software vendors are likely to face increasing exposure to lawsuits alleging that software did not perform as expected. The consequences of such lawsuits to software vendors could be catastrophic.

Several examples illustrate the extent of the potential liability faced by software vendors. For instance, in the Therac 25-Accidents (Nancy Leveson & Turner) an error ("bug") in a computerized therapeutic radiation machine caused it to administer incorrect dosages. Two people were killed and several others were seriously injured (Zammit & Savio, 1987). In another case, a construction company alleged that a bug in a spreadsheet program caused the company to underbid a $3 million contract. The company sued the manufacturer of the program for $245,000, claiming it had lost that amount as a result of the incorrect bid (Blodgett, 1986). Finally, in Scott v. White Trucks (1983), the defendant's truck was equipped with computer- controlled anti-lock brakes. After the brakes failed and the truck crashed, the driver of the truck brought a product liability action alleging defects in the software.

Few software product liability cases have been litigated, so little directly applicable judicial guidance is available to pinpoint steps a software vendor should take to minimize liability. However, by understanding the legal theories upon which a hypothetical plaintiff may rely in a suit against a software vendor, it is possible to identify and understand where the risk of liability lies.

2. Threshold issue: Is software a “product" or a "service"?

Some authors like Chris and Angel believed that the classification of software into products and services is illogical (Reed & Angel, 2000). Their late 1990’s argument was based on s.61 of the Sale of Products Act (SOGA) 1979 which defines "products" includes all personal chattels other than things in action and money, and in Scotland all corporeal moveables except money; and in particular "products" includes emblements, industrial growing crops, and things attached to or forming part of the land which are agreed to be severed before sale or under the contract of sale;. However, software producers might argue that software recorded in the form of tangible medium supplied does not covered under the said SOGA 1979 (Reed & Angel, 2000).

A threshold issue is whether computer software is a "product" or a "service." One reason this is important is that sales of products, but not of services, are subject to the damages and warranty provisions of the Uniform Commercial Code (UCC) unless the context otherwise requires, the Article applies to transactions in products. Another reason is that manufacturers of products, but not providers of services, may be subject to suit under a strict liability theory of tort (Bimbaun, 1988). A review of relevant cases shows that courts have used various analytical approaches to decide this issue. Software is

generally licensed and not sold, and the UCC by its terms applies to sales. However, when courts have found that software is a "product," they have not allowed the fact that software is licensed to insulate vendors from liability under the UCC.

3. Contractual liability

It is vital to distinguish between the two different elements of software supply i.e. the license of intellectual property and the development and/or supply of a copy of the software. According to Reed & Angel (2000), the former has one real contractual risk is that a third party may possess intellectual property rights which are superior to those of the licensee. The nature and extent of this risk and the drafting of suitable provisions to control is not the concern. Liability will arise either from the express terms of the contract or from those implied by law (Reed & Angel, 2000).

Breach of contract can be subdivided into (Reed & Angel, 2000) namely,

(a) Contracts of sale or supply; contracts to provide services; and license contracts.

(b) Product Liability, for physical injury or property damage caused by a defective product.

(c) Negligence claims for physical injury or property damage,

(d) Negligence claims for financial loss, divided into two, namely the consequential losses because the software is unusable; and losses caused by reliance on information, produced by the software and addressed to the human mind.

4. Strict liability vs. negligence

Under a strict liability interpretation, a person who is harmed in some way by a software failure would have the right to obtain damages either from the manufacturer of the software or the institution operating the software when the error occurred (eg: a patient suing a hospital because of an x-ray machine software malfunction). Under the negligence interpretation of liability, the victim would need to prove that the manufacturer of the software failed to develop and test its product well enough to the point where it was reasonably confident that the product was safe to operate, or that the operator of the software failed to use the software correctly or grossly failed to interpret the software's findings correctly ("Strict Liability vs. Negligence,").

Under current law, strict liability principles are not applicable to doctors and hospitals, although strict liability is being applied more frequently these days to manufacturers of medical software. It is often difficult to prove negligence, as it can be very difficult to prove where in the chain of production the defect occurred. Potential sources of a defect include ("Strict Liability vs. Negligence,") software manufactures, equipment manufacturers, program distributers, programmers' consultants, companies using the software, and software operators.

4.1 Applying strict liability

Courts have not addressed the issue of whether software is a product for the purposes of applying strict liability in the same way that they have addressed the issue of whether software is a product for the purposes of applying the UCC. Certain cases, however, may be relevant. In one series of cases, courts have held that information is a product and that strict liability therefore applies to it. First, in Saloomey v. Jeppesen & Co. (1983) navigational charts were found to be a product, not a service. In this case, inaccuracies in the charts caused a fatal airplane crash, and the decedents' administrators brought a wrongful death suit against the publisher of the charts (Saloomey v. Jeppesen & Co. 1983). The court said that since the charts were mass- produced and it was likely that purchasers substantially relied on them without making any alteration to them, the publisher had a duty to insure that consumers would not be injured by the use of the charts (Saloomey v. Jeppesen & Co. 1983).

Second, in Brocklesby v. United States (1985) the court held a publisher of an instrument approach procedure for aircraft strictly liable for injuries incurred due to the faulty information contained in the procedure. Strict liability applied because the product was defective, even though the publisher had obtained the information from the government (Lawrence B. Levy & Bell). In contrast, other cases have not applied strict liability to information contained in books. For example, in Cardozo v. True (1977) a woman was poisoned when she ate an exotic ingredient called for by a recipe in a cookbook. The court in this case held that the information contained in the book was not a product for the purposes of applying strict liability, and, therefore, the seller of the book was not liable on that theory for failure to warn of the nature of the ingredients used in the recipe. These cases may be relevant because software usually either contains information or assists in the generation or manipulation of information.

There are strong policy reasons to apply the UCC to sales of software. For example, software sales would then be subject to a concise and statutory, predictable to both vendors and users (Rodau, 1986). It would also be consistent with the parties' expectations; although they typically are licensees, especially when the same program is made available to multiple users, software users frequently regard themselves as purchasers of products (Beckerman-Rodau, 1986). Finally, since the UCC provides for such measures as warranties, disclaimers, and limitations of liability (L. N. Birnbaum, 1988), the UCC provides a mechanism to allow the parties to allocate the risk of product performance (Lawrence B. Levy & Bell). However, if software is considered a "product" in order to apply the UCC, it increases the likelihood that it will be

characterized as a product for other purposes, such as the application of strict product liability law. Software vendors may be held liable for damages caused by their products under various contract and tort theories.

4.2 Contract theories of liability

Often a contract (such as a development or license agreement) exists between the injured user and the software vendor. If so, it will usually be the first place to which the user will turn for a theory of liability under which it may recover against the vendor for damages resulting from defects in the software. For example, the user may claim that the defect breached the terms of a software warranty contained in the contract. The elements and defenses of such a case would follow the normal rules of local contract law. In addition, if the transaction is determined to be a sale of products (Rodau, 1986), the provisions of UCC Article 2 will apply to the interpretation of the contract and the recoverable damages.

Moreover, a court may hold a vendor liable under a warranty theory for statements about the software made outside the formal agreement. "Unless the context otherwise requires, this Article applies to transactions in products (“Uniform Commercial Code”, 1989). In addition, if a court determines that the transaction was a sale of products subject to the UCC, in the absence of a warranty disclaimer it may hold the vendor liable for breach of warranties implied by that statute. One such warranty is the implied warranty of merchantability, which requires that the software be reasonably fit for the general purpose for which it is sold (Levy, 1988). Another warranty is the implied warranty of fitness for a particular purpose, which applies when (1) the vendor knew of a particular purpose for which the software was required; and (2) the vendor knew that the user relied on the vendor's skill and judgment to furnish a suitable program. "Unless excluded or modified ... , a warranty that the products shall be merchantable is implied in a contract for their sale if the seller is a merchant with respect to products of that kind." and "Products to be merchantable must be at least such as ... are fit for the ordinary purposes for which such products are used ...." (“Uniform Commercial Code”, 1989).

Under a contract theory, a plaintiff may recover the difference between the market value of the software as delivered and the contract price of the software (“Uniform Commercial Code”, 1989). Plaintiffs may also be able to claim additional damages such as consequential damages, incidental damages, lost profits, or other losses that the sellers should have reasonably anticipated, such as the loss to the construction company in completing the contract mentioned above (“Uniform Commercial Code”, 1989). The possibility of large damages from these additional categories poses a great threat to software vendors faced with a breach of contract claim.

Since these are contractual claims, the vendor frequently can draft its contracts to substantially limit its exposure (Lawrence B. Levy & Bell). However, in Pennsylvania, privity of contract is not required to recover for breach of UCC warranties of merchantability and fitness for a particular purpose under Sections 2-314 and 2-318 of the Pennsylvania Uniform Commercial Code. Holding that Section 2-318 should be co-extensive with the state's doctrine of strict product liability, a Pennsylvania court allowed a purchaser to recover damages from a computer manufacturer for breach of implied warranty even though the purchaser had actually purchased the computer system from a reseller who had purchased it from the computer manufacturer. The disclaimers and limitations in the manufacturer's contract with the reseller did not apply to the reseller's customer (Spagnol Enter., Inc. v. Digital Equipment Corp. 1989) there may be limits to this ability to minimize liability.

4.3 Tort theories of liability

Since it may be possible for vendors to draft agreements that limit their exposure to contract damages for defective software (Lawrence B. Levy & Bell), injured users also look to a variety of tort theories for recovery. Contractual liability disclaimers may not be effective in limiting vendor liability to suits brought under tort theories. In addition, injured third parties may rely on tort theories, whose applicability is not predicated upon the existence of a contractual relationship with the vendor.

4.4 Misrepresentation

One tort theory involves claims that the vendor fraudulently misrepresented the capabilities of the software (Lawrence B. Levy & Bell). In order to prevail under this theory, the plaintiff must show that it was damaged because (1) the vendor misrepresented a material fact concerning the software, and (2) the plaintiff justifiably relied on this misrepresentation. Plaintiffs apparently have had some success under this theory (Reece, 1987). For example, in Laurie Financial Corp. v. Burroughs Corp.(1985), a vendor claimed that its computer system, including software, would be suitable for a prospective customer's business needs, when in fact the system was not (Reece, 1987). The court held the vendor liable to the customer on a fraudulent misrepresentation theory. It found that the customer justifiably relied on the vendor's misrepresentations because the customer had no previous experience with the particular system involved and had not made an independent investigation of the system's capabilities (Reece, 1987).

The defendant's misrepresentation usually must be intentional for the plaintiff to recover (Reece, 1987), but some jurisdictions recognize a cause of action for negligent misrepresentation whereby (Harper Tax Servs Inc. v. Quick Tax Ltd 1987). Courts may allow some "puffing" in marketing efforts, but seem to be more inclined to find misrepresentation when the vendor makes misstatements concerning facts that are exclusively within the vendor's knowledge (Conley,

1987). Under New York law a claim of negligent misrepresentation is stated only in special circumstances and upon certain allegations including the existence of a nexus of intimacy between the parties approaching that of privity and closer than that of ordinary buyer and seller.

A fraudulent misrepresentation claim is especially threatening to software vendors because under this theory, a plaintiff may sue when it suffers damages solely to its intangible economic interests (such as business reputation), rather than personal injuries or damage to tangible personal property. In the case of Computer Sys. Eng'g, Inc. v. Quantel Corp (1984), held that fraud claim upheld on appeal where purchaser neither knew nor could have known of defects in software.

4.5 Negligence

A second tort theory is that the vendor was negligent in developing the software. The plaintiff must show that the vendor had a duty to use a specific standard of care, usually "reasonableness," and that the vendor breached that duty. There are many actions or omissions that might constitute such a breach. These might include, for example, a failure to (1) write or test the program properly, (2) correct significant bugs in the program, (3) warn of limitations in the program, (4) instruct users how to operate the program (Computer Sys. Eng'g, Inc. v. Quantel Corp. 1984), or (5) provide adequate security for the system. The standard of care in each instance will depend on the specific circumstances. In addition, as technology evolves, there is the possibility that courts will hold vendors liable for less obvious breaches of duty such as the failure to insure that the software does not contain any hidden destructive programs (known as viruses) (Midgely v. S.S. Kresge Co. 1976). Finally, the plaintiff must show that the vendor's breach caused the plaintiff's injury, and that the plaintiff suffered damages from its injury (Branscomb, 1990). Recoverable damages usually are restricted to bodily injury or property loss (Zammit & Savio, 1987).

For a variety of reasons, plaintiffs have had less success claiming damages from software vendors under this theory than under other theories (Invacare Corp. v. Sperry Corp. 1984) Demonstrating that a vendor's conduct is unreasonable is difficult and expensive. Moreover, the defenses of assumption of risk and contributory negligence are available to vendors. For example, vendors have successfully defended on the grounds that the buyer was negligent in using defective data or in hiring incompetent operators. As a result, it may be difficult to prove that the plaintiff's injury was caused by the vendor's breach of duty, such as supplying defective hardware or software, rather than by the plaintiff's own error, such as use of incorrect data. Finally, some courts hold that a negligence action for economic loss is barred if a contract exists between the parties (Reece, 1987).

In the American case of Scott v District of Columbia (1985), where the claimant alleged false arrest and wrongful imprisonment against the police, who had relied on a warrant erroneously issued by a computer system, her suit failed because the officers who arrested her had not been negligent in relying on the computer. It would be legally irrelevant that a computer was involved. But first we must identify the type of damage suffered by the claimant. For instance, the noise from computer printer constituted a nuisance or an information retrieval system contained defamatory material as the same noise from a typewriter or defamation in a book.

4.6 Strict Liability

"Product" is defined in section 1(2) as "any products or electricity" including components. "Products" are defined in section 45 as including "Substances:, and it also defines "substance" as "any natural or artificial substance" (Reed & Angel, 2000). The Directive also defines "product" as any moveable. What about software which is installed by copying it onto the purchaser's system from a medium (e.g. a tape or disk) ownership of which is transferred to the purchaser (Reed & Angel, 2000). Section 5(1) in the Act also limits liability to death or personal injury or damage to property which is ordinarily intended for private use or consumption and was so intended to be used by the claimant by referring to Section 5(1) and (3). Thus, if software is purchased by a business and used solely within the business, claims can only be expected from outsiders who are injured by the business's activities where a cause was a defect in the software. Section 2 (1) requires that the damage be caused by a 'defect' in the product, and this is defined in section 3.A product is defective if it does not provide the level of safety that persons generally are entitled to expect. Once a claimant has established that he has suffered damage, he has a choice of defendants. In addition, Reed and Angel criticized the Consumer Protection Act 1987 as limited whereas scope and the majority of claims in respect of software will be for financial loss.

Another tort theory that might be applicable is strict liability. It is the theory often used in cases involving defective consumer products, and it would only apply if software is characterized as a product (Lawrence B. Levy & Bell). For strict liability to apply to the manufacturer of software, the user must have used the product in a reasonable fashion and the product must have reached the user without substantial change.Thus far efforts to find strict liability as to the services themselves have entirely failed (Bimbaum, 1988). If the user is injured while using the product, the user need show only that the product caused the injury, as stated by the court in *Scott v. White Trucks,* a plaintiff "without fault on his part" may recover under this theory if he "proves ... that the product was defective when it left the hands of the manufacturer." (Scott v. White Trucks 1983) and that the product was sold in a defective or unreasonably dangerous condition. The alleged defect could be a defect in the design or manufacturing of the software, or it could simply be a failure to warn of hazards.

An important feature of the strict liability theory is that it renders legally irrelevant the issue of whether the vendor acted reasonably (Zammit & Savio, 1987). By preventing the vendor from presenting exculpatory arguments, this theory in effect forces software manufacturers to guarantee the safety of their products (Zammit & Savio, 1987). The strict liability theory also has an effect on potential defendants and on recoverable damages. If it is applied, everyone in the chain of distribution of the product may be liable for the plaintiff's damages. However, users are not generally compensated for economic loss under a strict liability theory, but only for personal injury or property damage.

To date, courts have been reluctant to apply the theory of strict product liability to computer software. A Pennsylvania court, however, held that a purchaser may recover for a breach of statutory warranties against a remote manufacturer for purely economic loss (Frank, 1987; Spagnol Enters., Inc. v. Digital Equipment Corp.) Proponents have offered several arguments in favor of doing so. They argue that the software manufacturer should be held liable because it is in the best position to prevent defects, and can spread the risk of liability through insurance or by increasing the cost of the product; that strict liability assures compensation of the victim; that negligence is too difficult to prove; and that the application of strict liability will deter the manufacturer from producing defective software (Lawrence B. Levy & Bell). It is partially these types of considerations that have led courts to apply strict product liability to certain types of information (James, 1988).

On the other hand, the application of strict liability theory might hinder the development of software. Consider again the medical program that regulates the amount of radiation exposure for cancer patients (Lawrence B. Levy & Bell). It is likely that clinical judgments and heuristic rules for making decisions, rather than fixed engineering and mathematical principles, would be used to develop the software (Lawrence B. Levy & Bell). The software developer is therefore not necessarily able to eliminate defects, any more than a medical practitioner can guarantee positive results from radiation treatments. Nevertheless, under the strict liability theory, the software manufacturer might be held liable for large personal injury damages if the radiation treatments do not help, or injure, the patient. Moreover, everyone in the chain of distribution may be liable, including the hospitals and medical practitioners that used the program (Lawrence B. Levy, & Bell, S. Y).

In the coming years, we can expect to see each of these theories tested in the courts as the number of damage claims alleging defective software increases. The prudent software vendor should be cognizant of these risks and take actions to limit its exposure. We now turn to these preventive actions.

5. Malaysian approach

Technology in IT runs so quickly before an ex ante law upon software engineers could be drafted to minimize the number of accidents. The Government agency such as Multimedia Department will impose regulations on the software industry when the software involved is safety critical by raising the cost of software (AR. Rizal, 2004). If government regulation is to be imposed, it should be limited to software whose failure might endanger lives or cause serious damage.

5.1 The Sale of Goods Act 1957

Under the Malaysia's the Sale of Goods Act 1957, section 3 defines products in a very limited and narrow definition. It only applies to every kind of moveable property other than actionable claims and money; and includes stock and shares, growing crops, grass and things attached to or forming part of the land which are agreed to be severed before sale or under the contract of sale; a person is said to be "insolvent" who has ceased to pay his debts in the ordinary course of business, or cannot pay his debts as they become due, whether he has committed an act of bankruptcy or not.

5.2 Consumer Protection Act 1999

Whereas under the Consumer Protection Act 1999. "products" means products which are primarily purchased, used or consumed for personal, domestic or household purposes and includes products attached to or incorporated in, any real or personal property, animals, including fish, vessels and vehicles, utilities and trees, plants and crop whether on, under or attached to land or not, but does not include chooses in action, including negotiable instruments, shares, debentures and money.

While "services" in section 3 in Consumer Protection Act 1999 includes any rights, benefits, privileges or facilities that are or are to be provided, granted or conferred under any contract but does not include rights, benefits or privileges in the form of the supply of products or the performance of work under a contract of service. If we compare the two (2) provisions under Consumer Protection Act, section 71 : Prohibition on exclusion from liability means The liability of a person under this Part to a person who has suffered damage caused wholly or partly by a defect in a product, or to a dependent of such a person, shall not be limited or excluded by any contract term, notice or other provision.

Moreover, section 62 of The Sale of Goods Act 1957: Exclusion of implied terms and condition as to where any right, duty or liability would arise under a contract of sale by implication of law, it may be negatived or varied by express agreement or by the course of dealing between the parties, or by usage, if the usage is such as to bind both parties to the contract, it gives two conflicting views on the part of the liability of the software programmer. Thus, an exact and

appropriate policy recommendations for software liability laws should be made so that a distinction can be made between safety critical and normal software applications as different liability rules will induce different levels of care, any software liability statute or doctrine should differentiate between regular and safety-critical applications such as exacting levels of care should be demanded from programmers whose failures could result in the injury or lose of a human being.

For regular (non-safety-critical) applications, a negligence standard should be imposed. Guidelines for the standard should be drafted by computer professional organizations. In these cases, the burden of proof would be the consumer to show that the company did not meet the guidelines that would discourage lawsuits (AR. Rizal, 2004). Problem regarding the software industry is its high level of innovation. So, to impose strict liability on software engineers would expose individual programmers as well as companies to a floodgate of litigation. The fundamental nature of software engineering makes bug-free software unattainable and the unique characteristics of software that make error-free software should not be expected. Negligence standard should be tailored to the concept of software engineers as professionals. Now, software engineers' skill demands professional responsibility. As their product can be accessible by anyone who rely on their skill, they should be held to high standards of duty such as an auditor whose degree of care and skill are expected of their level of expertise of that profession. Nevertheless, the government should draft regulations on software projects that deal with safety – critical-software. Strict liability should be used in litigation dealing as well as regulations on the software projects by having authorized certificate of programmers and the government performed testing before release to the public.

The authors agree that strict liability is more appropriate to be used as opposed to negligence as software programmers and engineers are not professional having own professional body although their contribution is enormous and the possibility of failure results will emerge.

6. Conclusion

As society increasingly relies on software to perform critical functions in everything from manufacturing to life support systems, the risk that an error in a software program will lead to economic loss, property damage, or personal injury increases. Prudent software developers will be cognizant of these risks and will take steps to minimize their exposure to this type of liability. One of the solutions is to suggest that the syllabus of the Information Technology Law and Cyber Law subjects are not merely discussed ways to apply the new software technology but also adequate methods to make it work safely. Determining how far a software developer should go in this effort requires balancing the degree of exposure to software product liability against the adverse impact as product liability cases in tort are not subject to contract or license disclaimers of liability (Henningsen v. Bloomfield Motors,Inc., 1960), if any, of the steps required to limit this exposure on the software vendor's ability to market and sell its software. This liability issue is still evolving. For the time being, applying strict liability principle is better than negligence as the latter is harder for the consumers to prove.

References

Abdul Rahman,Rizal (2004). *Liability of Software Producers and Users (Handout)*, Malaysia : National University of Malaysia : Masters in Law.

Armour, J., & Humphrey, W. S. (1993a). Software product liability. *Software Engineering Institute, TR CMU/SEI-93-TR-13, ESC-TR-93-190*, 3.

Armour, J., & Humphrey, W. S. (1993b). Software product liability. *Software Engineering Institute, TR CMU/SEI-93-TR-13, ESC-TR-93-190*, 7.

Beckerman-Rodau, A. (1986). Computer Software: Does Article 2 of the Uniform Commercial Code Apply? *Emory Law Journal, 35*, 853.

Birnbaum, L. N. (1988). Strict products liability and computer software. *Computer/Law Journal, 8*, 135-156.

Birnbaum, S. L. (1980). Unmasking the Test for Design Defect: From Negligence [to Warranty] to Strict Liability to Negligence. *Vand. L. Rev., 33*, 593.

Blodgett. (1986). *Suit Alleges Software Error,* A.B.A. J., Dec. 1, at 22.

Branscomb, A. W. (1990). Rogue computer programs and computer rogues: Tailoring the punishment to fit the crime. *Rutgers Computer and Technology Law Journal, 16*, 1.

Brocklesby v. United States 767 F.2d 1288 (9th Cir. 1985).

Computer Sys. Eng'g, Inc. v. Quantel Corp., 740 F.2d 59, 65-66 (1st Cir. 1984).

Consumer Protection Act 1999.

Contra Invacare Corp. v. Sperry Corp., 612 F. Supp 448, 454 (N.D. Ohio 1984), at 396.

Frank (1987), *Tort Adjudication and the Emergence of Artificial Intelligence Software,* 21 SUFFOLK U. L. REV. 623, 647.

Gemignani, M. C. (1980). Product liability and software. *Rutgers Computer and Technology Law Journal, 8*, 173.

Harper Tax Servs., Inc. v. Quick Tax Ltd., 686 F. Supp. 109, 113 (D. Md. 1988).

Henningsen v. Bloomfield Motors,Inc., 161 A.2d 69 (NJ 1960).

Keeton, W. P., & Dobbs, D. B. (1984). Prosser and Keeton on torts. *St. Paul, MN: West Publishing*.

Lawrence B. Levy, & Bell, S. Y. Software Product Liability: Understanding and Minimizing The Risks. *Journal*. Retrieved from http://www.law.berkeley.edu/journals/btlj/articles/vol5/Levy/html/text.html.

Leveson, N. G., & Turner, C. S. (1993). An investigation of the Therac-25 accidents. *IEEE computer, 26*(7), 18-41.

Midgely v. S.S. Kresge Co. (1976), 55 Cal. App. 3d 67, 127 Cal. Rptr. 217.

Nancy Leveson, & Turner, C. S. An Investigation of the Therac-25 Accidents. *Journal*. Retrieved from http://courses.cs.vt.edu/~cs3604/lib/Therac_25/Therac_1.html.

Ravi Belani, Charles Donovan, Howard Loo, & Jessen Yu,(n.d.) Strict Liability vs. Negligence. *Journal*. Retrieved February 25, 2005 from http://cse.stanford.edu/class/cs201/projects-95-96/liability-law/Liability&Neg.html.

Rodau (1986), *Computer Software: Does Article 2 of the Uniform Commercial Code Apply?,* 35 EMORY L.J. 853, 857-60.

Reece (1987), *Liability for Defective Computer Software,* COMPUTER L. REP. 853, 855.

Reed, C., & Angel, J. (2000). *Computer law* (Fourth ed.): Blackstone London.

Saloomey v. Jeppesen & Co.,707 F.2d at 671-77 (2d Cir. 1983).

Scott v District of Columbia (1985) 493 A 2d 319.

Scott v. White Trucks 699 F.2d 714 (5th Cir. 1983).

Spagnol Enter., Inc. v. Digital Equipment Corp., 568 A.2d 948, 390 Pa. Super. 372 (1989).

Stephen, R. S. (1998). *Classical and Object-Oriented Software Engineering W/ Uml and C++*: McGraw-Hill, Inc.

Strict Liability vs. Negligence. *Journal*. Retrieved from http://cse.stanford.edu/class/cs201/projects-95-96/liability-law/Liability&Neg.html.

The Sale of Goods Act 1957.

Uniform Commercial Code (1952). Article 2.

Zammit & Savio (1987). *Tort Liability For High Risk Computer Software,* 23 PLI/PAT 373, at 391.