

44

Proceedings of  
**2<sup>nd</sup> ICSIT 2010**

International Conference on  
Soft Computing, Intelligent System  
and Information Technology


1-2 July 2010, Bali, Indonesia



**ICSIT**

Supported by

**MERATUS** 

 **APTIKOM**

**IBM**

# Proceedings

# ICSIIT 2010

**International Conference on  
Soft Computing, Intelligent System and Information Technology**

1-2 July 2010

Bali, Indonesia

**Editors:**  
**Leo Willyanto Santoso**  
**Andreas Handojo**



**Informatics Engineering Department  
Petra Christian University**

**Center of Soft Computing and  
Intelligent System Studies**

# ICSIT 2010

## Table of Contents

Preface.....	xi
Organizing Committee.....	xii
Program Committee.....	xiii
<b>Human Language Technology: The Philippine Context .....</b>	<b>1</b>
<i>Rachel Edita Roxas, Allan Barra</i>	
Hybrid-Multidimensional Fuzzy Association Rules from a Normalized Database.....	10
<i>Rolly Intan</i>	
<b>Fuzzy Systems &amp; Neural Networks</b>	
A Context-Based Fuzzy Model for a Generator Bidding System.....	18
<i>Moeliono Widjaja</i>	
Neural Networks for Air-Conditioning Objects Recognition in Industrial Environments.....	24
<i>Enrique Dominguez, J.J. Carmona</i>	
Pattern Recognition Using Discrete Wavelet Transformation and Fuzzy Adaptive Resonance Theory.....	29
<i>Arnold Aribowo, Samuel Lukas, Joannes Franciscus</i>	
Resolving Occlusion in Multi-Object Tracking using Fuzzy Similarity Measure .....	33
<i>Rahmatri Mardiko, M. Rahmat Widyanto</i>	
Search Engine Application using Fuzzy Relation Method for e-Journal of Informatics Department Petra Christian University.....	39
<i>Leo Willyanto Santoso, Rolly Intan, Prayogo Probo Susanto</i>	
The Use of Gabor Filter and Back-Propogation Neural Network for the Automobile Types Recognition.....	45
<i>Gregorius Sotia Budhi, Rudy Adipranata, Fransisco Jimmy Hartono</i>	
<b>Genetic Algorithm &amp; Applications</b>	
A Linear Graph and Genetic Algorithm Approach for Evolving Manipulator Modelling.....	57
<i>Kok Kiong Tan</i>	

Comparing Genetic and Ant System Algorithms in Course Timetabling Problem..... <i>Djasli Djamarus</i>	56
Gas Distribution Network Optimization with Genetic Algorithm..... <i>K.A. Sidarto, L.S. Riza, C.K. Widita, F. Haryadi</i>	62
Hybrid Genetic Algorithm for Solving Strimko Puzzle..... <i>Samuel Lukas, Arnold Aribowo, James Nagajaya Dyalim</i>	68
Optimal Design of Hydrogen Based Stand-Alone Wind/Microhydro System Using Genetic Algorithm..... <i>Soedihyo, Heri Suryoatmojo, Imam Robandi, Mochamad Ashari, Takashi Hiyama</i>	71
Optimization of Steel Structure by Combining Evolutionary Algorithm and SAP2000..... <i>Mohammad Ghazi, Pujo Aji, Priyo Suprobo</i>	76
The Hydrophobic-Polar Model Approach to Protein Structure Prediction..... <i>Tigor Nauli</i>	82
University Course Scheduling Using the Evolutionary Algorithm..... <i>Ade Jamal</i>	86
<b>Artificial Intelligence &amp; Applications</b>	
Adaptive Appearance Learning Method using Simulated Annealing..... <i>Du Yang Kim, Elwan Yang, Moongsu Jeon, Vladimir Skin</i>	91
Bayesian Network and Minimax Algorithm in Big2 Card Game..... <i>Nur Ulfa Maulidevi, Hengky Budiman</i>	96
Cell Formation Using Particle Swarm Optimization (PSO) Considering Machine Capacity, Processing Time, and Demand Rate Constraints..... <i>Dedy Suryadi, Ferry Putra, Cynthia Juvono</i>	102
Computer Aided Learning for List Implementation in Data Structure..... <i>Ng Melissa Anega, Susana Limanto</i>	108
Development Weightless Neural Network on Programmable Chips to Intelligent Mobile Robot..... <i>Siti Nurmaini, Bambang Tutuko</i>	112
If-Statement Modification for Single Path Transformation: Case Study on Bubble Sort and Selection Sort Algorithms..... <i>Rahmadi Trimamanda</i>	116

Implementation of Particle Swarm Optimization Method in K-Harmonic Means Method for Data Clustering.....	120
<i>Ahmad Saikhin, Yoke Ohta</i>	
Implementation of Starfruit Maturity Classification Algorithm.....	177
<i>R. Amirulah, M.M. Mokji, Z. Ibrahim</i>	
Improving Choquet Integral Agent Network Performance by using Competitive Learning Algorithms.....	132
<i>Handri Santoso, Shusaku Nomura, Kazuo Nakamura</i>	
Improving Food Resilience with Effective Cropping Pattern Planning using Spatial Temporal-Based Updated Pranata Mangsa.....	138
<i>Kristika Dwi Hartomo, Sri Yulianto J.P., Krismiyati</i>	
Knowledge Based System in Defining Human Gender Based On Syllable Pattern Recognition.....	143
<i>Muhammad Fachrurozi</i>	
Maintaining Visibility of a Moving Target: The Case of an Adaptive Collision Risk Function.....	146
<i>Ashraf Elkogar, Ibrahim Al-Blawi</i>	
Measuring Interesting Rules in Characteristic Rule.....	152
<i>Spits Warnars</i>	
MIDI Composition Tools using JFugue Java API.....	157
<i>Kartika Gumadi, Liliana, Hendra Kurnia Wijaya</i>	
Mobile-based Interaction using Dijkstra's Algorithm for Decision Making in Traffic Jam System ...	150
<i>Puji Sularsih, Egi Wisnu Moyo, Fitria H. Siburian, Sigit Widiyanto, Dewi Agushinta R.</i>	
Model and Boarding Simulation for Reducing Seat and Aisle Interferences Between Passenger.....	164
<i>Bilqis Amaliah, Victor Hariadi, Antonius Malem Baris</i>	
Optimizing Rijndael Cipher using Selected Variants of GF Arithmetic Operators.....	170
<i>Petrus Mursanto</i>	
PCR Primer Design using Particle Swarm Optimization Combined with Piecewise Linear Chaotic Map.....	176
<i>Cheng-Hong Yang, Yu-Huei Cheng, Li-Yeh Chuang</i>	
Performance Analysis of Heterogeneous Computer Cluster.....	182
<i>Abdusy Syarif, Saiful Ikhsan, Muhammad Risky</i>	
Reduced Space Classification using Kernel Dimensionality Reduction for Question Classification in Public Health Question-Answering.....	187
<i>Hapnes Toba, Ito Wasito</i>	

<b>The Developing of Interactive Software for Supporting the Kinematics Study on Linear Motion and Swing Pendulum.....</b>	<b>193</b>
<i>Liliana, Kartika Gimadi, Yanathan Rindayanto Ongko</i>	
<b>University Timetabling Problems with Customizable Constraints using Particle Swarm Optimization Method.....</b>	<b>197</b>
<i>Pantus Mudjihartono, Wahyu Triadi Gumawan, The Jim Ai</i>	
<b>Knowledge &amp; Data Engineering</b>	
<b>A Design of Multidimensional Database for Content-based Television Video Commercial Mining.....</b>	<b>201</b>
<i>Yaya Heryadi, Yudha Giri Sucalya, Ariati Murni Arjumnurthy</i>	
<b>Applying Sound to Enhance the Comprehension of Sorting Algorithms.....</b>	<b>206</b>
<i>Lijana, Edwin Pramana</i>	
<b>Data Mining to Build a Pattern of Knowledge from Psychological Consultations.....</b>	<b>211</b>
<i>Sri Mulyana, Sri Hartati, Retantyo Wardova, Edi Winarko</i>	
<b>Data Warehouse Information Management System RSU Dr. Soetomo for Supporting Decision Making.....</b>	<b>215</b>
<i>Silvia Rostianingsih, Oviliani Yenti Yuliana, Gregorius Satia Budhi, Denny Irawan</i>	
<b>Development of an Electronic Medical Record (EMR) in Stayed Nursing Installation.....</b>	<b>220</b>
<i>Eko Handoyo, Aghus Sofwan, Mohammad Muttaqin</i>	
<b>Development of Supporting Sales Analysis Application using Frequent Closed Constraint Gradient Mining Algorithm (FCCGM).....</b>	<b>224</b>
<i>Susana Limanto, Dhiani Tresna Absari</i>	
<b>Implementation of KMS to Integrate Knowledge Management and Supply Chain Management Process.....</b>	<b>229</b>
<i>Vivine Nurcahyawati, Retno Aulia Vinarti, Mudjahidin</i>	
<b>Indonesian WordNet Sense Disambiguation using Cosine Similarity and Singular Value Decomposition.....</b>	<b>234</b>
<i>Sandra Sari, Ruli Mamuring, Mirna Adriani</i>	
<b>Influence of Electronic Media and External Reward Towards Knowledge Sharing Management to Learning Process in Higher Education Institution.....</b>	<b>240</b>
<i>Alexander Setiawan</i>	

193	Information and Technology Outsourcing Vendor Selection: An Integrative Literature Review.....	245
	<i>Jimmy</i>	
	Information Retrieval on MARC Metadata.....	251
	<i>Adi Wibowo, Rolly Intan, Irawan Arifin</i>	
197	Learning Management Systems' Integration.....	256
	<i>N.S. Linawati, Putra Sastra, P.K. Sudiarta</i>	
	Mining Sequential Pattern on Sequential Data of Paint Sales Transaction Flow.....	260
	<i>Agustinus Noertjahyana, Gregorius Saria Budhi, Henry Kusumawati Wibowo</i>	
201	Modeling School Bus for Needy Student Using Geographic Information System.....	265
	<i>Daniel Hary Prasetyo, Jamilah Muhammad, Rosmadi Fauzi</i>	
206	Optimization SQL Server 2005 Query using Cost Model and Statistic.....	272
	<i>Ihsan Gunawan</i>	
211	Spatial Autocorrelation Modelling for Determining High Risk Dengue Fever Transmission Area in Salatiga, Central Java, Indonesia.....	277
	<i>Sri Yulianto J.P., Kristoko Dwi Hartomo, Krismirati</i>	
215	Supply Chain Improvement with Design Structure Matrix Method and Clustering Analysis (A Case Study).....	281
	<i>Tanti Octavia, Siana Halim, Stefanus Anugraha Lubianto, Harvey Sutopo</i>	
220	The Comparison of Similarity Detection Method on Indonesian Language Document.....	285
	<i>Anna Kurniawati, Lily Walandari, I Wayan Simri Wicaksana</i>	
224	The Effects of Training Documents, Stemming, and Query Expansion in Automated Essay Scoring for Indonesian Language with VSM and LSA Methods.....	290
	<i>Heninggar Septiantri, Indra Budi</i>	
229	The Impact of Object Ordering in Memory on Java Application Performance.....	296
	<i>Amil A. Ilham, Kazsaki Murakami</i>	
234	Using Data Mining to Improve Prediction of 'No Show' Passenger on an Airline Reservation System.....	302
	<i>Johan Setiawan, Hobby Limantara</i>	
240	Using Frequent Max Substring Technique for Thai Keyword Extraction used in Thai Text Mining.....	309
	<i>Todsanai Chummatana, Kok Wai Wong, Hone Xie</i>	

Using the End-User Computing Satisfaction Instrument to Measure Satisfaction with Web-Based Information Systems .....	315
<i>Dedi Rianto Rahadi</i>	
<b>Imaging Technology</b>	
Batik Image Classification using Log-Gabor and Generalized Hough Transform Features .....	320
<i>Laksmi Rahadiani, Hadoiq R. Sanabila, Ruli Manurung, Anisati Murni</i>	
Burrows Wheeler Compression Algorithm (BWCA) in Lossless Image Compression .....	326
<i>Erijun Syahrul, Julien Dubois, Vincent Vajnovszki, Asep Juarna</i>	
Comparison of Random Gaussian and Partial Random Fourier Measurement in Compressive Sensing Using Iteratively Reweighted Least Squares Reconstruction .....	332
<i>Endru</i>	
Developing a Video Player Application for Phillips File Standard for Pictorial Data Format (NXPP): A Project View Approach .....	335
<i>Eko Handoyo, Restiono Djati Kusumo</i>	
Development Edge Detection Using Adhi Method, Case Study: Batik Sidomukti Motif .....	340
<i>Adhi Pranoto, Suyoto</i>	
Discriminating Cystic and Non Cystic Mass Using GLCM and GLRM-based Texture Features .....	346
<i>Hari Wibawanto, Adhi Susanto, Thomas Sri Widodo, S Maesadji Tjokronegoro</i>	
Fractal Terrain Generator .....	351
<i>Budi Hartanto, Monica Widiasri, Gunawan Widjaja</i>	
From Taiwan Puppet Show to Augmented Reality .....	356
<i>Yang Wang, Bo Rui Huang, Zih Huei Wang</i>	
Generating Iriscodes using Gabor Filter .....	362
<i>I Ketut Gede Darma Putra, Lie Jusa</i>	
Interpolation Technique to Improve Unsupervised Motion Vector Learning of Wyner-Ziv Video Coding .....	366
<i>I. M. Oka Widyanegara, N.P. Sastra, D.M. Wiharta, Wirawan, G. Hendratoro</i>	
Iris Segmentation and Normalization .....	371
<i>I Ketut Gede Darma Putra, I Nyoman Piarsa, Nazer Jawas</i>	
NEATS: A New Method for Edge Detection .....	377
<i>Maria Yunihe, Suyoto</i>	



15

20

26

32

38

44

50

56

62

68

74

80

86

92

Online Facial Caricature Generator ..... 383  
*Rudy Adipranata, Stephanus Surya Jaya, Kartika Gunadi*

Silny Approach to Edge Detection for Central Borneo Batik ..... 387  
*Silvia Siyanto*

**Internet, Web Services & Mobile Applications**

Cattle's Cost of Goods Sold System Information at CV Agriranch ..... 392  
*Lily Puspa Dewi, Yulia, Anita Nathania, Doddy Hartanto*

Compensation Method for Internet Grids using One-to-many Bargaining ..... 396  
*Andreas Kurniawan, Pujianto Yugopuspito, Johan Muliadi Kerta*

Mobile RSS Push Using Jabber Protocol ..... 406  
*Fajar Baskoro, Dwi Ardi Irawan*

Teacher's Community Building Website to Facilitate Networking and Life-Long Learning ..... 412  
*Ariituh Imam Kaharâjo, Yulia, Silvia Kostuaningsih*

Vision and Mission Educational Foundation (TYVM) Web-Based Project Management System ..... 417  
*Arlinoh Imam Rahardjo, Yulia, Edwin*

Web Based School Administration Information System on LOGOS School ..... 421  
*Djoni Haryadi Setiabudi, Ibnu Gunawan, Handoko Agung Puandty*

**Communication Systems & Networks**

Data Visualization of Modulated Laser Beam Communication System ..... 427  
*Zin May Ave*

Development of Steganography Software with Least Significant Bit and Substitution  
Monoalphabetic Cipher Methods for Security of Message Through Image ..... 432  
*Isnur Kumbara, Erwin*

Feasibility Analysis of Zigbee Protocol in Wireless Body Area Network ..... 436  
*Vera Suryani, Achmad Rizal*

Mobile TV with RTSP Streaming Protocol and Helix Mobile Producer ..... 439  
*Yunianto Purnomo, Andrew Jaya Efendy*

Quantitative Performance Mobile Ad-Hoc Network using Optimized Link State Routing  
Protocol (OLSR) and Ad-Hoc On-Demand Distance Vector (AODV) ..... 442

Spatial Rain Rate Measurement to Simulation Colour Noise Communication Channel Modeling for Millimeter Wave In Mataram.....	446
<i>Made Sutha Yadhya, Giamantyo Hendranta</i>	
The Effect of Maximum Allocation Model in Differentiated Service-Aware MPLS-TE.....	453
<i>Rayu Erfianto</i>	
User Accounting System of Centralized Computer Networks using RADIUS Protocol .....	457
<i>Heru Nurwarsito, Raden Arief Setyawan, Handoko D. Fatikno</i>	
Wireless Data Communication with Frequency Hoping Spread Spectrum (FHSS) Technique.....	463
<i>Khin Swe Myint, Zarli Cho</i>	
Wireless LAN User Positioning using Location Fingerprinting and Weighted Distance Inverse .....	465
<i>Justinus Andjarwirawan, Silvia Rostianingsih, Charlie Anthony</i>	
WLANXCHANGE: A New Approach in Data Transfer for Mobile Phone Environment .....	474
<i>Ary Mazharuddin Shiddiqi, Rogus Jati Santoso, Rio Indra Maulana</i>	
<b>Control &amp; Automation</b>	
Analysis Influence Internal Factors on Fuzzy Type 2 Performance of Swing Phase Exit Restoration .....	479
<i>Hendi Wicaksono</i>	
Design and Construction of Wind Speed Indicator Based on PIC Microcontroller System .....	484
<i>Khin Mar Aye, Khi Tar Oo</i>	
Fault Diagnosis in Batch Chemical Process Control System using Intelligent System .....	489
<i>Syahriil Ardi</i>	
Implementation of an Adaptive PID Controller using the SPSA Algorithm with Realistic Target Response.....	493
<i>Sofyan Tan</i>	
Induction Heating Efficiency Analysis Modeling Using COMSOL <sup>®</sup> Multiphysics Software.....	498
<i>Didi Istardi</i>	
Authors Index.....	504

# Reduced Space Classification using Kernel Dimensionality Reduction for Question Classification in Public Health Question-Answering

Hapnes Toba

Maranatha Christian University  
Bandung, Indonesia / University of  
Indonesia Depok, Indonesia

hapnes.toba@eng.ma ranatha.  
edu / hapnes.toba@ui.ac.id

Ito Wasito

Information Retrieval Laboratory  
University of Indonesia  
Depok, Indonesia

ito.wasito@cs.ui.ac.id

## ABSTRACT

One of the major problems in Question Answering System is how to classify a question into a particular class that further will be used to find exact answers within a large collection of documents. Kernel Dimensionality Reduction (KDR) is an alternative method that can be used for features reduction, and in the same time classify question type by using the most effective m-dimensional features in its vector space. In this experiment we used question-answer pairs data from public health domain and word (unigram) features construction. This research shows that KDR correct rate performance is better than SVM after a head-to-head comparison from 100 observations.

## Categories and Subject Descriptors

H.3.4 [Systems and Software]: Question-answering (fact retrieval) systems

## General Terms

Experimentation, Algorithms, Performance.

## Keywords

Kernel Dimensionality Reduction, Reproducing Kernel Hilbert Space, Supervised Machine Learning, Question Classification, Question Answering System

## 1. INTRODUCTION

Question answering system (QAS) is a form of information retrieval that used a natural language question as its input and returns explicit answers in the form of a single answer or snippets of text rather than a whole document or set of documents. One of the most challenges in QAS is how to classify a question into a particular class that further will be used to find exact answers within a large collection of documents. Two major approaches have been widely used in question classification, i.e. the pattern-based and machine learning approach [1]. While the pattern-based

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.  
ICSIT 2010, July 1–2, 2010, Bali, Indonesia.  
Copyright 2010 ICSIT ISBN ...

approach try to identify a question in its syntax form which can be resource intensive [2], on the other hand machine learning approach try to approximate in which class a question can be classified by using an already trained classifier [3], [4], [5].

The widely used algorithm in question classification using the machine learning approach are mainly based on supervised classification using the Support Vector Machines (SVM) [6], [7], [9], and Maximum Entropy [3], [4], [8] to analyze the semantic and syntactic structure. Due to the data sparsity and features selection problem in both algorithms, it is hard to choose the best features in a particular question class. In this paper we will introduce that Kernel Dimensionality Reduction (KDR) algorithm can be used to reduce word matrix features and in the same time classify question type using the most effective m-dimensional features in its vector space. The rest of this paper is organized as follow: section 2 will give an exploration of KDR and the feature selection method. Our research design will be described in section 3, followed by the experiments and their results in section 4. Some discussions, conclusions and future works will be presented in section 5.

## 2. METHODS EXPLORATIONS

### 2.1 KDR with Reproducing Kernel Hilbert Spaces

KDR is based on a particular class of operators on reproducing kernel Hilbert spaces (RKHS) [10]. A Hilbert space is an extension of a vector space. It requires the definition of an inner product on the vector space [15] which enables it to be called an inner product space. An example of an inner product on a finite vector space between any vector  $\underline{x}$  and  $\underline{y}$  is:

$$(x, y) = \sum_{i=1}^n x_i \cdot y_i$$

The KDR algorithm relates dimensionality reduction to conditional independence of variables, and use RKHS to provide characterizations of conditional independence and thereby design objective functions for optimization. The hypothesis is to find effective subspace that can be formulated in terms of conditional independence. In particular, it is assumed that there is an r-

dimensional subspace  $S \subset \mathbb{R}^m$  such that the following equality holds for all  $x$  and  $y$ :

$$pY|X(y|x) = pY|\prod_s X(y|\prod_s x) \dots (1)$$

Let  $(A, B)$  be an  $m$ -dimensional orthogonal matrix such that the column vectors of  $A$  span the subspace  $S$  (so  $A$  is  $m \times r$ ), and define  $U=A^T X$  and  $V=B^T X$ . Because  $(A, B)$  is an orthogonal matrix, we can derive that  $p(X(x))=p(U, V(u, v))$  and  $p(X, Y(x, Y))=p(U, V, Y(u, v, Y))$ .

Eq. (1) is thus equivalent to:

$$pY|U, V(y|u, v) = pY|U(y|u) \dots (2)$$

In this way, the effective subspace  $S$  is the one which makes  $Y$  and  $V$  conditionally independent given  $U$  [10] (see Figure 1).

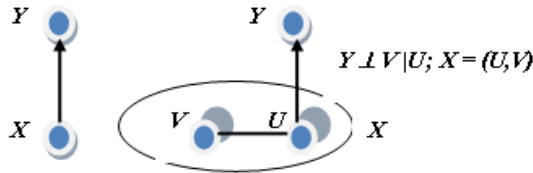


Figure 1 Representation of Dimensionality Reduction [10]

Another important viewpoint on the equivalence between conditional independence and the effective subspace is the mutual information condition that holds the dimensionality reduction. It is known that:

$$I(Y, X) = I(Y, U) + E_u[I(Y|U, V|U)] \dots (3)$$

where  $I(Y, X)$  is the mutual information between  $X$  and  $Y$ . Because Eq. (1) and (2) implies  $I(Y, X)=I(Y, U)$ , the effective subspace  $S$  is characterized as the subspace which retains the entire mutual information between  $X$  and  $Y$ , or equivalently, such that  $I(Y|U, V|U)=0$ . This produces again the conditional independence of  $Y$  and  $V$  given  $U$ .

KDR uses covariance operator on RKHS to produce an objective function for dimensional reduction. If there is a set  $\Omega$  consisting of feature vectors in its columns, RKHS is produced by using the kernel that has the following reducing property:  $\langle f, i(\cdot, x) \rangle_H = f(x)$  for all  $x$  the elements in the vector space and all  $f$  the functions (or features in this sense) in  $H$ , the reproduced space. Fukumizu et. al. in [10] uses the Gaussian kernel  $i(x_1, x_2) = \exp(-\|x_1 - x_2\|^2 / 2\sigma^2)$ . We will have thus  $(H, i)$  which is a reproducing kernel Hilbert space of functions on a set of random vectors in  $\Omega$  with a positive definite kernel  $i: \Omega \times \Omega \rightarrow \mathbb{R}$  and an inner product  $\langle \cdot, \cdot \rangle$  in  $H$ . The vector space that has been reproduced by the kernel function need to be further processed to guarantee the conditional probability and linear independency of the reduced kernel. This is achieved in KDR by using the cross covariance operator  $\Sigma_{YX}$  from  $H_1$  to  $H_2$  that defined by the relation:

$$\langle i, \Sigma_{YX} f \rangle_{H_2} = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] (= \text{Cov}[f(X), g(Y)]) \dots (4)$$

This relation implies that the covariance of  $f(X)$  and  $g(Y)$  is given by the action of the linear operator  $\Sigma_{YX}$  and the inner

product. Interested readers should refer to [10] for the complete mathematical proofs.

## 2.2 Feature Selection

The features in the  $m$ -dimensional space of documents are usually formed by its textual features. Information retrieval research suggests that word stems can be used effectively as representation units of a document. Such word stems are derived from the occurrence form of word by removing case and flections information [11]. This leads to an attribute-value representation of text. Each distinct word  $w_i$  (unigram feature) corresponds to a feature with term frequency  $TF(w_i, x)$ , the number of times word  $w_i$  occurs in the document  $x$ , as its value.

Refining this basic representation, it is better to scale down the dimension of feature vector with their inverse document frequency  $IDF(w_i)$  [12], which can be calculated from the document frequency  $DF(w_i)$  which is the number of documents the word  $w_i$  occurs in:

$$IDF(w_i) = \log\left(\frac{n}{DF(w_i)}\right) \dots (5)$$

Where  $n$  is the total number of documents. In this research we assume that a question is comparable as a document, and thus we called our feature as inverse question frequency of  $w_i$ ,  $IQF(w_i)$ .

Our word matrix representation will have  $i$ -rows, that equal the number of questions and  $j$  columns, which equal the number of features (see Figure 2). The problem with such representation is the sparsity of data in each feature ( $j$ -th column) that represent the occurrences of a term in the all  $i$ -questions. It is reasonable that not every word should be appeared in all questions. This kind of problem will be useful to evaluate the performance of KDR and compare it with other comparable supervised method, in this case the support vector machines.

	IQF-1st	...	IQF-jth
row 1-st			
.			
.			
row i-th			

Figure 2 Matrix Representation of Questions

## 2.3 Support Vector Machines

Support vector machines (SVM) are based on the Structural Risk Minimization principle from computational learning theory [13]. Joachims in [14] described the idea of SVM as structural risk minimization that try to find a hypothesis  $h$  for which we can guarantee the lowest true error. The true error of  $h$  is the probability that  $h$  will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis  $h$  with the error of  $h$  on the training set and the complexity of  $H$  (measured by VC-Dimension), the hypothesis space containing  $h$ . Support vector machines find the hypothesis  $h$  which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of  $H$ . The SVM will thus in particular define the criterion to be looking for a decision surface that is maximally far away from any data point [16]. This distance from the decision surface to the closest data point determines the margin of the

classifier. This method of construction necessarily means that the decision function for an SVM is fully specified by a subset of the data which defines the position of the separator. These points are referred to as the support vectors [16, 17] (see Figure 3).

Both KDR and SVM are promising to be compared because each method can handle the text classification properties, i.e.: high dimensional input space, few irrelevant features, document vectors are sparse, and most text categorization problems are linearly separable [14].

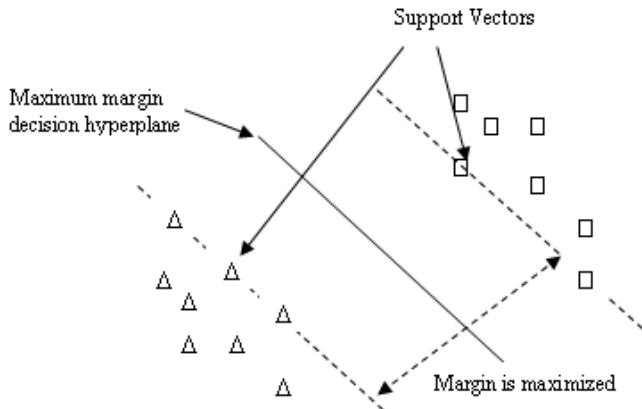


Figure 3 Maximization of margin of the support vectors [16]

### 3. EXPERIMENTAL SETTING

#### 3.1 Algorithms and Tools

During the experiments the following algorithms and tools are used:

1. KDR algorithm (© Fukumizu 2004);
2. SVM Linear and RBF Classification (© 2004-2007 The MathWorks, Inc);
3. KNN Classification (© 2004-2007 The MathWorks, Inc).

#### 3.2 Data

We used questions from the public health domain. We downloaded the question-answer pairs from the Singapore Ministry of Health FAQ pages<sup>1</sup> in the topics of swine flu (H1N1-2009) and gastric flu. In total there are 92 questions from those two topics (73 questions about H1N1 and 19 questions about gastric flu). The reason we have chosen those topics based on the assumption, that topics that share the same context (in this sense the “flu” context), will share the same features. This assumption is important to form an objective orientation when we try to classify a test question in classifier that was constructed from the same (randomize) dataset.

After we downloaded the FAQ, to obtain the version in Bahasa Indonesia, we used the Google translation tools<sup>2</sup>. The translation that we obtained was not directly used for the research. We reconstructed first some of the grammar and unmatched contextual terms that is used in daily Indonesian. After we have the final version of the translated FAQ in Bahasa Indonesia, we

use Perl programming language to convert the FAQ into the feature matrix as described in section 2.2. We have thus for the feature matrix 92 rows and 1137 columns as features.

#### 3.3 Performance Evaluation

In the benchmarking step, we trained first a classifier using the Linear and RBF SVM classification, and then tested it with random test data from a subset of the row data.

The complete procedures of the benchmarking steps are:

1. Using the whole matrix representation, we trained the data with SVM to separate two distinct classes using the [train, test] composition of [90, 10].
2. We tested the SVM classifier with the random generated test data from number 1, and find the correct rate for each composition in 100 runs.
3. We use the correct rate to evaluate the performance of each classifier, as follow:

$$CorrectRate = \frac{NumberOfCorrectClassifiedQuestions}{NumberOfTotalTestQuestions} \dots (6)$$

4. We saved the test-indexed question which will be used as the supervision vector in the KDR method.
5. We reduced the features matrix representation using KDR using the 2-dimensional reduction, 100 iterations and 0.1 learning rate.
6. Use the result of the already reduced KDR matrix as the input vector for the SVM training. In this step we use the concept of “build classifier in the reduced space”, as also described in [10].
7. We use the saved test-indexed question (number 4) as the same test data for KDR classification.
8. Compare the results of the original SVM and the enhanced KDR results.

In our research, besides comparing the whole matrix, we also compare the KDR with “manual”-reduced SVM. This “manual”-reduced SVM, is a reduced matrix that was formed by selecting the two most occurrence words that occur in all question classes after we applied the stemming and removed the stop words, and used them as the features. To compare the resulting KDR classification, we also took the KNN classifier [18], with K=2 and K=5, to see how close the distance among the classified question-answer pairs.

To compare the consistency of the KDR and the SVM classification, we used the head-to-head comparison over 100 evaluation runs on both methods, i.e. we run the experiment 100 times for each method and then count how many times a method outperforms the other. We also compute the mean and standard deviation for each method to see the performance in overall evaluation runs.

### 4. EXPERIMENTS AND ANALYSIS

We have run our experiments according to the setting which is described in Section 3.

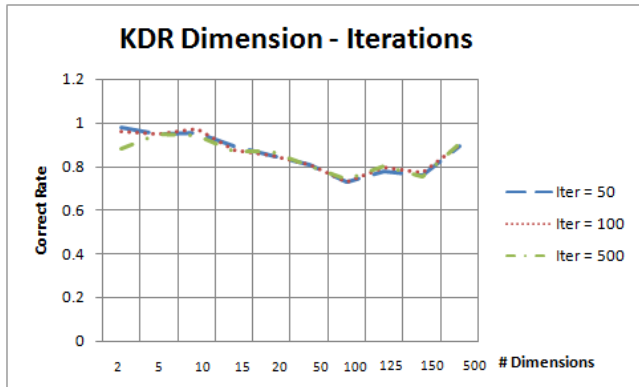
#### 4.1 KDR Iterations and Dimensions

The purpose of this experiment is to see the impact of KDR number of iterations and dimension construction. We run an experiment that used 50, 100 and 500 iterations to construct a

<sup>1</sup> [http://www.pqms.moh.gov.sg/apps/fcd\\_faqmain.aspx](http://www.pqms.moh.gov.sg/apps/fcd_faqmain.aspx) (menu: Illness and Diseases), accessed on February 2010

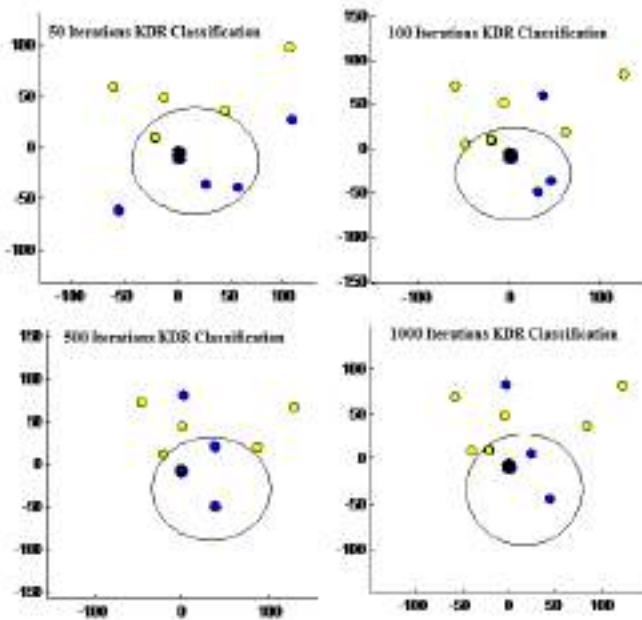
<sup>2</sup> [http://www.google.co.id/language\\_tools?hl=id](http://www.google.co.id/language_tools?hl=id)

dimension of 2, 5, 10, 15, 20, 50, 100, 125, 150 and 500. The result of this experiment can be seen in Figure 4.



**Figure 4 Performances of KDR Iterations and Dimensions**

From Figure 4, we can see that the performances of 2- and 10-dimension reduced matrix are the best for all iterations. This indicates that KDR can still perform it best in a small number of dimension (features) which is important in the benchmarking steps (cf. section 3.3). We can also see from Figure 4, that the correct rate patterns are almost identical for each iteration. This result indicates that the number of iterations has no direct impact to the number of dimensions. Figure 5 plots the impact of number of iterations (Blue = H1N1, Yellow = gastric).



**Figure 5 Vector Distribution of KDR Classification (50, 100, 500 and 1000)**

The number of iterations indicates how fast the learning rate closer to a convergence area in each training-classification session. Based on the results in this experiment, we choose the 2 dimensions and 100 iterations as our default setting in the benchmarking steps.

## 4.2 SVM and KDR Classification

We used the SVM classification with linear and RBF function (sigma = 1). The resulting “mean value” correct rate classification of each random generated [90, 10] composition for all 1137 features in four series of 100 training-classification runs can be seen in Table 1. For the “manual” reduced SVM, the most occurrence words that occur in both classes H1N1 and gastric flu are the word “flu” and “virus”.

**Table 1 SVM Results for All Features**

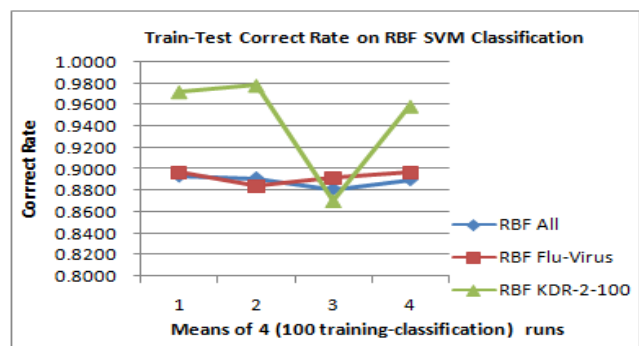
SVM all features		SVM Most 2-Words (flu-virus)	
RBF Sig=1	Linear	RBF Sig=1	Linear
0.8933	0.8867	0.8967	0.8822
0.8911	0.8744	0.8844	0.8933
0.8811	0.8978	0.8922	0.8689
0.8900	0.8878	0.8967	0.8856

Result in Table 1 shows that “all features” SVM Linear classification performed better than when we choose only “several selected features”. To compare the performance of Linear and RBF SVM from Table 1 against the KDR algorithm, we run KDR experiment with 2-dimensional features in 50, 100, and 500 iterations in four series of 100 training-classification runs, which further classified using the RBF and linear SVM classifier. The “mean value” of the correct rate in these experiments can be seen in Table 2. Results of these KDR experiments show again that the iterations number does not give any direct impact to the correct rate (cf. section 4.1).

**Table 2 KDR Performance**

KDR (2-dim, 50 iter)		KDR (2-dim, 100 iter)		KDR (2-dim, 500 iter)	
RBF Sig=1	Linear	RBF Sig=1	Linear	RBF Sig=1	Linear
0.9783	0.9783	0.9722	0.9783	0.8261	0.9783
0.9722	0.9783	0.9783	0.9722	0.8056	0.9722
0.8701	0.9565	0.8701	0.8837	0.9565	0.913
0.8837	0.8701	0.9588	0.9565	0.8701	0.8837

Interpretation plot from the result in Table 1 and 2 can be seen in Figure 6a and 6b.



**Figure 6a Comparison of Correct Rate RBF Linear SVM (right) on Different Train-Test Composition**

Those figures give us an insight that KDR reduction matrix which is trained using the RBF-SVM and linear-SVM has given an almost identical correct rate patterns during the 4 series of training-classification runs. Such result is also hold for the original “all features” and the “manual” constructed features.

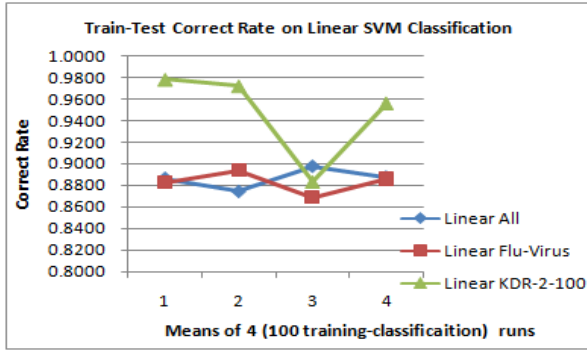


Figure 6b Linear SVM (right) on Different Train-Test Composition

### 4.3 Effectiveness of KDR

To see how effective the dimension reduction in KDR, we also plotted the “manual” constructed 2-dimension features and the KDR 2-dimension. The plot can be seen in Figure 7a and 7b.

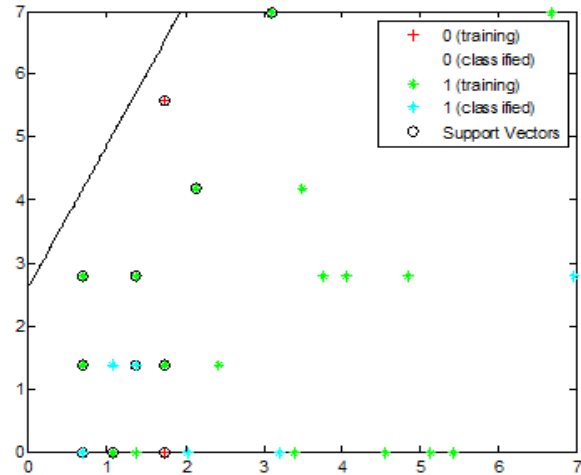


Figure 7a 2-“Manual” Selected Features Linear SVM

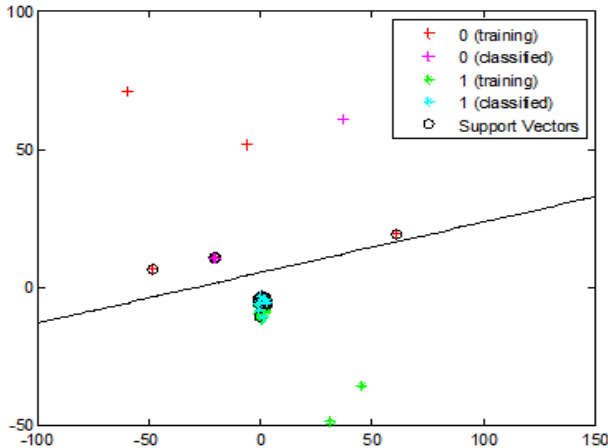


Figure 7b KDR 2-dim 100- iterations Classification

Figure 7a & 7b give an indication that KDR is very effective to classify a huge number of features into a much smaller dimensions. KDR (Figure 7b) has produced better classification than the “manual” selected features (flu-virus in Figure 7a).

### 4.4 KDR and KNN Classification

The comparison of KDR and KNN gives another view of the KDR classification. Besides the reduced dimension that has been achieved with KDR, it also produces the classification that comparable with KNN. We used the data from the “manual” reduced features for our 2-NN and 5-NN classification. Figure 8 shows the plot of our experiment with 5-NN Euclidean Distance Classification compare with KDR (100 iterations). The x and y-axis the vector value estimations of the classifications.

Figure 8 shows us that the distance between KDR classified vectors is much closer than the KNN classification. In other words, this means that KDR can classify the features in the right classification although the distances between the features are very close one to another.

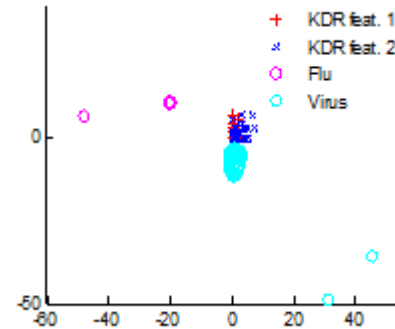


Figure 8 Comparison of Euclidean Distance (K=5) with “manual” selected features (flu-virus) and KDR (100 iterations)

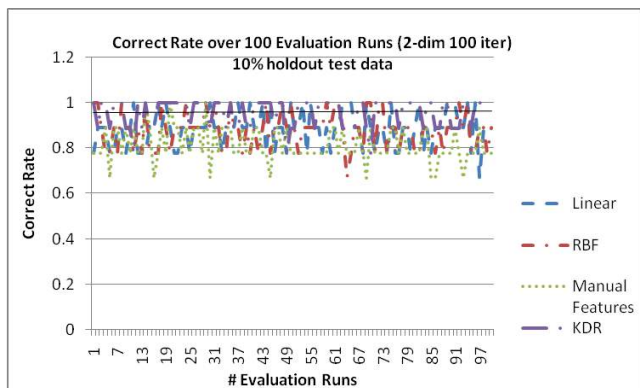
### 4.5 Overall Performance Evaluation

To observe the consistency of KDR, we evaluated the correct rate performance of SVM Linear and RBF with all features, SVM Linear with manual constructed features (virus-flu), against the KDR-2-dim-100 iterations, in 100 training-classification runs with 10% holdout test data. The overall performance of the evaluation runs can be seen in Table 3. The graphical interpretation of each method in this evaluation can be seen in Figure 9.

Table 3 Overall Performances in 100 Evaluation Runs

	SVM Lin	SVM RBF	SVM Man	KDR
Mean	0.882222	0.878889	0.813333	0.958889
Std Dev	0.080248	0.080674	0.073872	0.060457
Mean +	0.96247	0.959563	0.887205	1.019346
Mean -	0.801974	0.798215	0.739462	0.898432

Because the value of mean +/- the standard deviation of each method is overlapping (see also Figure 9), the result is not conclusive. We need thus to evaluate the head-to-head comparison. This comparison gives an insight about the performance of each method in each evaluation run. We will see how many times a method outperforms the other.



**Figure 9 Correct Rate over 100 Evaluation Runs**

The result of the head-to-head comparison of 100 observations can be seen in Table 4.

**Table 4 Head-to-Head Comparison**

Comparison	Lin	RBF	Man	KDR
Lin	0	66	86	37
RBF	34	0	90	39
Man	14	10	0	13
KDR	63	61	87	0

Each row in Table 4 gives the number of “winning” or equal correct rate of each method against the other. We can see from Table 4 that KDR classification outperforms the other methods. The “manual” constructed features perform the worst in each method; it indicates that such subjective selected features should be strengthened with some other features which will give better classification.

## 5. CONCLUSIONS & FUTURE WORKS

We found that KDR can be used as a promising alternative method to classify questions in Question Answering System. An important viewpoint is that KDR can effectively classify questions even with only very few features (words), i.e. 2-dimensions (cf. Section 4.1). KDR can also determine the best effective features in the vector space. The classifications of questions using the features reduction that KDR has determined in most of the time are better than the “manually” constructed features and the original “all feature” matrix (cf. Section 4.2 and 4.3). During the head-to-head comparison, we found that KDR outperforms significantly the SVM classification in many cases. This indicates that the features reduction that has been produced by KDR is very effective to be used in classification of questions (cf. Section 4.4 and 4.5).

As future works, we are going to apply KDR to strengthen the question classification and answer validation method in our ongoing research to build an Indonesian Question Answering System. In this sense, we are going to build a KDR classifier that can be used to anticipate a set of important features (words) from a question which could be classified into more than one question

class (multi-labeling). The classification that produced by KDR will be important to find the real context of the question.

## 6. REFERENCES

- [1] Purwarianti, et. al. 2006. Estimation of Question Types for Indonesian Question Sentence. Department of Information and Computer Sciences, Toyohashi University of Technology.
- [2] Toba, H & Adriani, M. 2009. Pattern Based Indonesian Question Answering System. Proceedings of the International Conference on Advanced Computer Systems and Information Systems (ICACSIS) University of Indonesia.
- [3] Ittycheriah, A. et. al. 2001. IBM’s Statistical Question Answering System. Proceedings of the 10th Text Retrieval Conference (TREC 2001).
- [4] Schlaefter, Nico. 2007. Deploying Semantic Resources for Open Domain Question Answering. Diploma Thesis. Language Technologies Institute School of Computer Science Carnegie Mellon University.
- [5] Li, X., Roth, D. 2002. Learning Question Classifiers. Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan.
- [6] Cruchet, S, et. al. 2008. Supervised Approach to Recognize Question Type in a QA System for Health. MIE.
- [7] Joachims, T. 1999. Transductive Inference for Text Classification using Support Vector Machines. International Conference on Machine Learning (ICML).
- [8] Berger, A. et. al. 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics 22(1).
- [9] Zhang, D & Lee, WS. 2003. Question Classification using Support Vector Machine. ACM SIGIR 2003.
- [10] Fukumizu, K., Bach, F.R., Jordan, M.I. 2004. Kernel Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. The Journal of Machine Learning Research. Volume 5. December 2004. Pages 73-99.
- [11] Porter, M. 1980. An Algorithm for Suffix Striping. Program (Automated Library and Information Systems). Volume 14. Number 3. Pages 130-137.
- [12] Salton, G. & Buckley, C. 1988. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management. Volume 24. Number 5. Pages 513-523.
- [13] Vapnik, V.N. 1995. The Nature of Statistical Learning Theory. Springer, New York.
- [14] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Springer.
- [15] Rijsbergen, van C.J. 2004. The Geometry of Information Retrieval. Cambridge University Press.
- [16] Manning, C.D., et. al. 2008. Introduction to Information Retrieval. Cambridge University Press.
- [17] Cristianini, N., and Shawe-Taylor, J. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, First Edition (Cambridge: Cambridge University Press). <http://www.support-vector.net/>.
- [18] Mitchell, T. 1997. Machine Learning. McGraw Hill.