ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

# Adaptive Relevance Feature Discovery for Text Mining with Simulated Annealing Approximation

C.Kanakalakshmi[1], Dr.R.Manicka chezian[2]

Research Scholar, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[1]

Associate Professor, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[2]

**ABSTRACT:** The field of text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The discovery of relevant features in real-world data for describing user information needs or preferences is a new challenge in text mining. Relevance of a feature indicates that the features is always necessary for an optimal subset, it cannot be removed without affecting the original conditional class distribution. In this paper, an adaptive method for relevance feature discovery is discussed, to find useful features available in a feedback set, including both positive and negative documents, for describing what users need. Thus, this paper discusses the methods for relevance feature discovery using the Simulated annealing approximation and genetic algorithm, a population of candidate solutions to an optimization problem toward better solutions.

**KEYWORDS:** Feature Discovery, Mutual Information, Simulated Annealing, Genetic Search.

## I. INTRODUCTION

Dimensionality reduction [1] [2] is one of the most popular techniques to remove irrelevant and redundant features. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively and increase the performance of the classifiers. Text mining applications need to deal with large and complex datasets of textual documents that contain much relevant and noisy information. Feature selection [3] [4] aims to remove the irrelevant and noisy information by focusing only on relevant and informative data for use in text mining. The relevance of features [14] is assessed as the capability of distinguishing different classes. For example, a feature $f_i$ is said to be relevant to a class $c_j$ if $f_i$ and $c_j$ are highly correlated. By focusing on the selected subset of features, simple and fast models can build by using only the subset and gain better understanding of the processes described by the data. Many techniques [10] are developed for selecting an optimal subset of features from a larger set of possible features. To, effectively utilize and update discovered features is still an open research issue, especially in the domain of text mining. The objective of relevance feature discovery is to find useful features available in text documents, for describing the text mining results.

## II.RELATED WORK

C. Kanakalakshmi and Dr.R.Manicka chezian[1] discuss the various feature selection techniques and compared the performance of Feature Selection methods with various classifiers. Vasantha et al [2] presents a comparative evaluation of several attribute selection methods based on the performance accuracy of different tree based supervised classification for mammogram images of MIAS database. Ramaswami et al [3] present the work on Feature Selection methods used in the educational field for the purpose of extracting useful information on the behaviors of students in the learning process. Isabelle Guyon et al [4] discuss the variable and feature selection has become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. Nikita Katariya et al [6] present different steps involved in text preprocessing. The important pre-processing techniques namely stop word removal; stemming and indexing are discussed in detail. Shivani Patel et al [8] present about genetic algorithm in detail and discussed how GA can be beneficial in text mining. Sayantani Ghosh et al [9] gives the detail about the text mining process and the algorithms used for text classification such as Naïve Bayes, SVM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 10, October 2015**

and Genetic algorithm. Ning Zhong et al[13] presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Lei Yu et al [14] discusses a new framework that decouples relevance analysis and redundancy analysis and develop a correlation-based method for relevance and redundancy analysis. Yuefeng Li [15] presents an innovative model that discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features. It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns.

## III.THE PROPOSED METHOD

The method comprises of series of steps that effectively discovers the relevant features in text dataset and efficient revision and updating weight of extracted features in the vector space. An Adaptive information filtering system called Adaptive Relevance Feature Discovery (ARFD) is built upon Relevance Feature Discovery (RFD) [15] method. The new ARFD model evaluates the new feedback documents using the knowledge (called base knowledge) that the system has. The documents that are correctly classified by the system will be discarded because they are redundant documents. Then, new knowledge can be extracted from the selected documents in the new training dataset. A merging function has been developed to merge both base knowledge and the new knowledge. The goal of the adaptive model is to update the system efficiently with new knowledge to improve the effectiveness. The overview of proposed method is given in Figure 1.

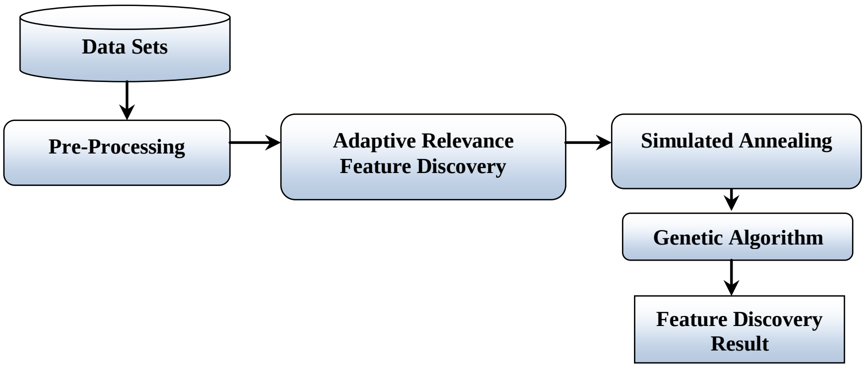

**Figure 1:** Overview of the Proposed Method

### *A. PRE-PROCESSING*

Preprocessing [5] method plays a very important role in text mining techniques and applications. For the relevance feature discovery, we use the two steps of preprocessing namely, Training set extraction and Feature Attribute selection.

*Training Set Extraction*

The training set feature set, compute the cross validation classification error for a large number of features and find a relatively stable range of small error. This range is called **Ω**. The optimal number of features (denoted as *n**) of the training set is determined with in **Ω**. The whole process includes three steps:

- The Training feature selection is to select *n* (a preset large number) sequential features from the input *X*. This leads to *n* sequential feature sets $F_1 \subset F_2 \subset \ldots \subset F_{n-1} \subset F_n$.
- The *n* sequential feature sets $F_{1,}$ …, $F_k$, …, $F_n$,($1 \leq k \leq n$) to find the range of *k*, called **Ω**, within which the respective (cross-validation classification) error $e_k$ is consistently small (i.e., has both small mean and small variance).
- Within **Ω**, find the smallest classification error $e_k$ = min $e_k$. The optimal size of the candidate feature set, *n**, is chosen as the smallest *k* that corresponds to *e**.

*Feature Attribute Selection*

The Feature attribute selection [6] is a statistical technique that can reduce the dimensionality of data as a by-product of transforming the original attribute space. Transformed attributes are formed by first computing the covariance matrix of the original features, and then extracting its sorting. The attribute selection defines a linear transformation from the original attribute space to a new space in which attributes are uncorrelated.

## IV.ADAPTIVE RELEVANCE FEATURE DISCOVERY

The adaptive relevance feature discovery process considers the mutual-information-based feature selection for both supervised and unsupervised data. For discrete feature variables, the integral operation in (1) reduces to summation. In this case, computing mutual information is straightforward, because both joint and marginal probability tables can be estimated by tallying the samples of categorical variables in the data.

Given two random variables $x$ and $y$, their mutual information [12] is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$:

$$MI(x,y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (1)$$

Given $N$ samples of a variable $x$, the approximate similarity function *Simm* ($x$) has the following form:

$$Simm(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x^{i}, h) \quad (2)$$

where $\delta(.)$ is the sampling window function as explained below, $x^{(i)}$ is the *ith* sample, and $h$ is the window width.

## V. SIMULATED ANNEALING APPROXIMATION

Simulated annealing (SA) [7] is a probabilistic Meta heuristic approach for optimization problems like locating a good feature approximation to the global optimum of a given function in a large dimensional search space. Another approach for feature selection is the simulated annealing wrapper method which conducts a search for a good subset using an induction algorithm. The algorithm runs on the large array data, usually partitioned into internal learning and external test sets. The feature subset with highest evaluation is chosen as the final set on which to build a classifier. Maximal classification accuracy on a separate test can be obtained using this approach as the feature subset selection is able to couple tightly with the decision mechanism of the classifier. The simulated annealing algorithm consists the following steps.

```
Input: $ARFD$, Length $l$
Initialize;
   M:=0;
repeat Process (config.$arfd$ →config.$arfd$.$idx$, $\Delta arfd_{ij}$);
if $\Delta arfd_{ij}$ <=0 then accept else
            if exp(-$\Delta arfd_{ij}$ /$l$) >random [0,1) then accept;
if accept then UPDATE(configuration $j$);
until search is approached sufficiently closely;
$arfd_{M+1}$:=f($arfd_M$);
$M$:=$M$+1;
until stop criterion = true (System is frozen);
end.
```

## VI.GENETIC FEATURE REDUCTION

Genetic Algorithms (GA) [8] [9] are generally quite effective for rapid search of large, nonlinear and poorly understood spaces. Unlike classical feature selection strategies where one solution is optimized, a population of solutions can be modified at the same time. This can result in several optimal (or close-to-optimal) feature subsets as output. A feature subset is typically represented by a binary string with length equal to the number of features present in the dataset. A zero or one in the j$^{th}$ position in the chromosome denotes the absence or presence of the j$^{th}$ feature in this particular subset. An initial population of chromosomes is created; the size of the population and how they are created are important issues. From this pool of feature subsets, the typical genetic operators (crossover and mutation) are applied.

The initial population consists of 100 randomly generated feature subsets, the probability of mutation and crossover set to 0.4 and 0.6 respectively, and the number of generations is set to 100. The genetic feature reduction algorithm has the following steps.

```
Step 1: Take the Simulated annealing feature input into the information system along with feature values of
        each data objects in every class.
Step 2: Initialize $\gamma_{prev}$ = 0.0, $\gamma_{best}$ = 0.0, flag = 0, count-of-generation = 0.
Step 3: do repeat until ( flag == 1 ) :
            Count-of-generation ++1.
            Select randomly number of features to be taken.
                number-features ←− rand()
                number-features = number-features +1.
            Generate the combination set containing all combinations of number − features number of
                attributes.
            Select a combination from the combination set.
                comb-num ←− rand() % total-number-of-elements in combination set.
            Take the reduced information system
                number−features number of attributes for $comb-num^{th}$ combination of the combination set.
            Find out the crossing over probability.
            Modify the roughest information system (x) as required after mutation.
            Call function variation ($x$, no-of-objects, no-of attributes) for generating different combination of
                attributes.
```

## VII.RESULT AND DISCUSSIONS

For the evaluation of the proposed model which is a supervised approach that needs a training set including both relevant documents and irrelevant documents. A two-stage feature selection algorithm is presented. In the first stage, a candidate feature set using the ARFD incremental selection method is found. In the second stage, other more sophisticated schemes are used to genetic search a compact feature subset from the candidate feature set.

This work used two popular data sets to test the ARFD proposed model: Reuters Corpus Volume 1, a very large data collection; and Reuters-21578, a small one.

The effectiveness of a model is usually measured by the following means [10] [11]: the average precision of the top-20 documents, F1 measure, and mean average precision (MAP), the break-even point (b/p), and interpolated average precision (IAP). The F-beta (Fβ) measure is a function to describe both Recall (R) and Precision (P), together with a parameter beta β. The parameter β = 1 was used in this paper, which denotes that precision and recall were weighed equally. Therefore, Fβ is denoted by

$$F_1 = \frac{2PR}{(P+R)} \qquad (3)$$

**Table 1**: Comparison of All Models on Reuters-21578(R8)

| Model | Top-20 | *b/p* | MAP | $F_\beta = 1$ | *IAP* |
|---|---|---|---|---|---|
| ***ARFD*** | **0.796** | **0.682** | **0.792** | **0.668** | **0.763** |
| *RFD*$_2$ | 0.794 | 0.669 | 0.745 | 0.600 | 0.746 |
| *Rocchio* | 0.706 | 0.594 | 0.633 | 0.527 | 0.632 |
| *BM25* | 0.675 | 0.556 | 0.582 | 0.508 | 0.590 |
| *SVM* | 0.794 | 0.693 | 0.729 | 0.557 | 0.709 |
| $X^2$ | 0.263 | 0.245 | 0.211 | 0.260 | 0.242 |
| *Lasso* | 0.719 | 0.627 | 0.657 | 0.536 | 0.651 |

In the above Table 1, *ARFD* is compared with the state-of-the-art term-based methods[15] under pinned by *Rocchio*, *BM25*, *SVM*, *x2* and *Lasso* for each variable top - 20, B/P, MAP, IAP and $F_\beta = 1$ on Reuters-21578 dataset. From the above table , it is observed that the ARFD method has the higher value when compared to other methods.

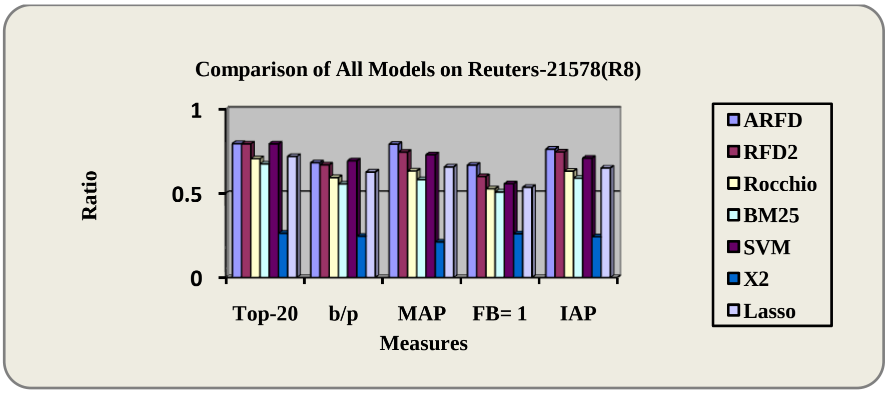

**Figure 2:** Comparison Chart of all models on Reuters-21578 dataset

From the above Figure 2, the comparison is made to proposed method ARFD with other feature selection models and it is shown that the ARFD has the highest value than the other existing methods .

The proposed method using ARFD was also compared with popular feature selection models [14] including *Rocchio*, *BM25*, *SVM*, $x^2$ and *Lasso*. The experimental results on RCV1 dataset is reported in Table 2.

**Table 2:** Comparison Results of All Models on RCV1 dataset

| Model | Top-20 | *b/p* | MAP | $F_\beta = 1$ | *IAP* |
|---|---|---|---|---|---|
| ***ARFD*** | **0.566** | **0.483** | **0.504** | **0.475** | **0.593** |
| *RFD*$_2$ | 0.561 | 0.473 | 0.493 | 0.470 | 0.513 |
| *Rocchio* | 0.501 | 0.424 | 0.440 | 0.433 | 0.459 |
| *BM25* | 0.445 | 0.407 | 0.407 | 0.414 | 0.428 |
| *SVM* | 0.453 | 0.408 | 0.409 | 0.421 | 0.435 |
| $X^2$ | 0.322 | 0.326 | 0.319 | 0.355 | 0.345 |
| *Lasso* | 0.506 | 0.434 | 0.460 | 0.445 | 0.480 |

From the above Table 2, it is observed that the proposed ARFD method has the high value when compared with other state-of-art methods.

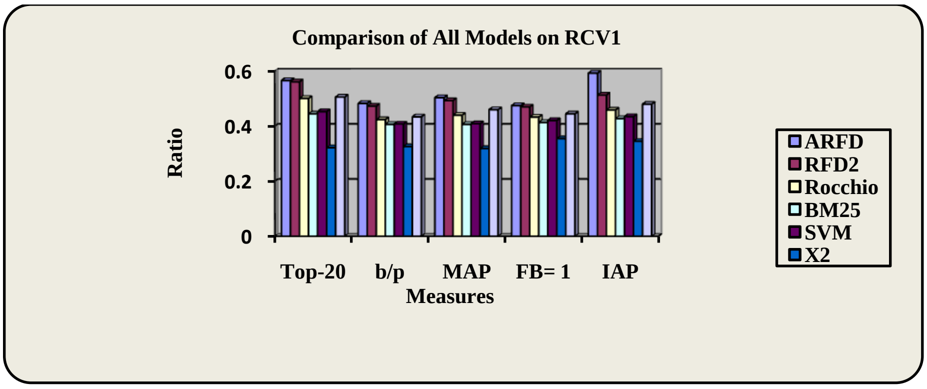

**Figure 3 :** Comparison Chart of All models on RCV1 dataset.

From the Figure 3, in which comparison is made with ARFD method and other methods like SVM, X2, it is shown that the proposed method has the high value when compared to other methods.

The proposed method ARFD is compared with the pattern based model RFD$_2$ and PTM [13] which uses positive patterns only in Reuters-21578. The results are given in Table 2.

**Table 2:** Comparison of the Proposed Model with the Pattern Based Model on Reuters-21578 dataset

| Model | Top-20 | *b/p* | MAP | $F_\beta = 1$ | *IAP* |
|---|---|---|---|---|---|
| ***ARFD*** | **0.797** | **0.727** | **0.763** | **0.661** | **0.753** |
| *RFD$_2$* | 0.794 | 0.704 | 0.747 | 0.601 | 0.748 |
| *PTM* | 0.731 | 0.633 | 0.661 | 0.564 | 0.664 |

From the Table 2, it is observed that the proposed method has the high value when compared with the pattern based model on Reuters-21578 dataset.

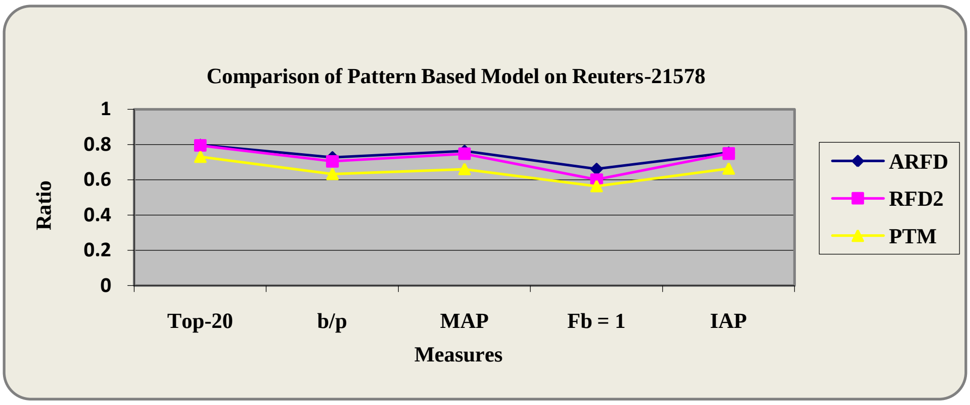

**Figure 4:** Comparison chart of proposed method with the Pattern based Model.

From the above Figure 4, the comparison chart shows that the proposed method ARFD has the highest values than other pattern based methods like RFD and PTM.

## II. CONCLUSION

This paper presents the method on the concept of developing an effective pattern-based Adaptive Relevance Feature Discovery model (ARFD) that uses both positive and negative feedback. In the ARFD model, some of the new training documents are selected using the knowledge currently held by the system. Then, specific features will be extracted from selected training documents. Different methods have been used to merge and revise the weights of features in a vector space. Compared with the baseline models that use batch training documents, the experiments on Reuters-21578 and RCV1 topics demonstrate that the efficiency of updating the system using the proposed model is significantly improved and maintains almost the same level of effectiveness.

## REFERENCES

1. C.Kanakalakshmi, Dr.R.Manickachezian "Feature Selection Approaches with Text mining for Categorical Variable Selection" in International Journal for Research in Science Engineering and Technology, Volume 2, Issue 6, August 2015 , pp: 1-8.
2. M.Vasantha, V.Subbiah Bharathy "Evaluation of Attribute Selection Methods with Tree based Supervised Classification-A Case Study with Mammogram Images" International Journal of Computer Applications,Volume 8, Issue 12, October 2010, pp :35-38
3. M. Ramaswami , R. Bhaskaran "A Study on Feature Selection Techniques in Educational Data Mining" Journal Of Computing, Volume 1, Issue 1, December 2009, pp: 7-11.
4. Isabelle Guyon, Andr´e Elisseeff "An Introduction to Variable and Feature Selection" Journal of Machine Learning Research, Volume 3, 2003, pp: 1157-1182.
5. Vijayarani, Ilamathi, Nithya "Preprocessing Techniques for Text Mining - An Overview" International Journal of Computer Science & Communication Networks, Volume 5, Issue 1, pp: 7-16.
6. Nikita Katariya, Chaudhari "Text Preprocessing for Text Mining using side information" International Journal of Computer Science and Mobile Applications, Volume3, Issue 1, January 2015, pp: 01-05.
7. Dimitris Bertsimas, John Tsitsklis, "Simulated Annealing", Statistical Science, Volume 8, Issue 1, 1993, pp: 10-15.
8. Shivani Patel, Prof. Purnima Gandhi " A Detailed Study on Text Mining using Genetic Algorithm" International Journal of Engineering Development and Research, Volume 2, Issue 22, 2013, pp: 101-105.
9. Sayantani Ghosh, Mr. Sudipta Roy, Samir Bandyopadhyay "A tutorial review on Text Mining Algorithms" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012 pp: 223-233.
10. Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proceedings of ACM SIGKDD Knowledge Discovery Data Mining, 2010, pp. 753–762.
11. S.T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proceedings of IEEE Conference on Data Mining, 2006, pp. 1157–1161.
12. Alper Unler , Alper Murat, Ratna Babu Chinnam, "mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification", Information Sciences , Volume 181,2011, pp: 4625–4641.
13. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Volume 24,Issue 1, January 2012, pp: 30-44.
14. Lei Yu, Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learning Research, Volume 5 ,2004, pp: 1205–1224.
15. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, "Relevance Feature Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 6, June 2015, pp: 1656-1669.

## BIOGRAPHY

**C.Kanakalakshmi** is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi. She received her Master of Computer Applications (M.C.A) in 2011 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has presented papers in International/National conferences and attended Workshop, Seminars and published paper in international journal. Her research focuses on Data Mining.

**Dr. R.Manickachezian** received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published more than One Hundred and Twenty papers in international/national journal and conferences. He is a recipient of many awards like Desha Mithra Award, Life Time Achievement Award in Computer Science and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.