LINEAR REGRESSION MODEL FOR KNOWLEDGE DISCOVERY IN ENGINEERING MATERIALS

Doreswamy[1], Hemanth K S[2] and Manohar M G[3]

[#1,2,3]Department of Computer Science,
Mangalore University, Mangalagangotri-574 199,
Karnataka, INDIA.
doreswamyh@yahoo.com, reachhemanthmca@gmail.com,
manoharg.phd@gmail.com

ABSTRACT

Nowadays numerous interestingness measures have been proposed to disclose the relationships of attributes in engineering materials database. However, it is still not clear when a measure is truly elective in large data sets. So there is a need for a logically simple, systematic and scientific method or mathematical tool to guide designers in selecting proper materials while designing the new materials. In this paper, linear regression model is being proposed for measuring correlated data and predicating the continues attribute values from the large materials database. This method helps to find the relationships between two sub properties of mechanical property of different types of materials and helps to predict the properties of unknown materials. The method presenting here effectively satisfies for engineering materials database, and shows the knowledge discovery from large volume of materials database. Studying on regression analysis suggests that data mining techniques can contribute to the investigation on materials informatics, and for discovering the knowledge in the materials database, which make the manufacturing industries to hoard the waste of sampling the newly materials.

KEYWORDS

Mechanical property, Materials database, Knowledge discovery , Regression ,Correlation .

1. INTRODUCTION

The rapid development of novel technologies in designing new engineering materials for different applications has derived numerous new materials. Subsequently it is very difficult to manage and make decision from such kind of engineering materials[16]. Efficient techniques for data storing and effective data analysis models for knowledge discovery from storage data are required for the effective decision making purpose during engineering materials design phase and manufacturing process. Hence, data mining/Knowledge Discovery from databases is rising in a broad sense in engineering materials design and their development[11]. Though the manufacturing technology is improving day-by-day, still, trial and error methods have been executing in composite materials design applications in manufacturing industry[18]. Creation of dynamic data repository for vast amount of newly emerging materials data and their management is really a changeling task for engineers. Concerning such vast materials database-making activity, it is required to construct effective and efficient databases[12]. Creation of centralized database/Data warehouse has potential research scope for data mining. Therefore, Application of Data Mining on Engineering Materials suits for the extraction of nontrivial ,implicit, previously

David Bracewell, et al. (Eds): AIAA 2011,CS & IT 03, pp. 147–156 , 2011.
 DOI : 10.5121/csit.2011.1313

unknown and potentially useful information(such as knowledge rules, constraints, regularities) from data in databases[1][2][3].

Processes and technologies have been accepted for many years, still there is indeed of great potential for mining knowledge to integrate manufacturing, product characteristics, and the engineering design processes. Engineering design is a multidisciplinary, multidimensional, and non-linear decision-making process where parameters, actions, and components are selected. The selection of engineering materials is often done based on historical data, information, and knowledge. It is therefore a prime area for data mining applications and although as yet only a few papers have reported applications of data mining in engineering design [12], and this has been an area of increased research interests in recent years. There are lots of related works have been done by researchers on fuzzy based systems for knowledge discovery from engineering materials[13][14][15][17], and an "if -then" rule based approach for the selection of materials' relevant information for decision making[16].

In this paper, a statistical model is being proposed for predicting the mechanical property of engineering materials using the existing data and information associated to other materials. The concepts and procedures adopted for predicting such properties are organized in the rest of this paper. A brief introduction to material database is given in section 2. Determination of materials properties and their significance are discussed in section 3. Section 4 describes the Methodology used for the proposed work. Experimental results are discussed in the section 5 and conclusions and future scopes are made in the section 6.

2 Material database

Materials database is an organized collection of materials data sets. Each data set characterizes an engineering material with their properties. It is being frequently accessible during materials design applications in concurrent engineering design process , as shown in figure 1 .

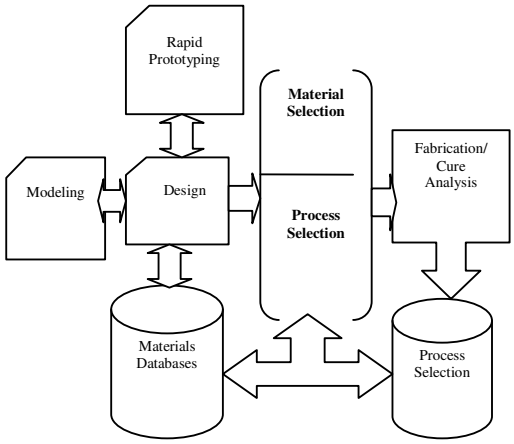


Figure 1. Concurrent Engineering Materials Design Process

However, there are a few material property databases that are available on demand with cost constraints[4][5][10]. Therefore, materials database with about 5600 data sets is designed by referring many textbooks[7][8], handbooks and website[9] of engineering materials. The objective of the engineering material database is to offer information for the materials selection during the concept development of a new composite materials. While designing a new material, researchers have to be gone though many data resources such as materials property database and materials handbook in ordered to select right materials that satisfy the design requirements. It is a time consumption process and expensive for referring the relevant bibliographic materials.

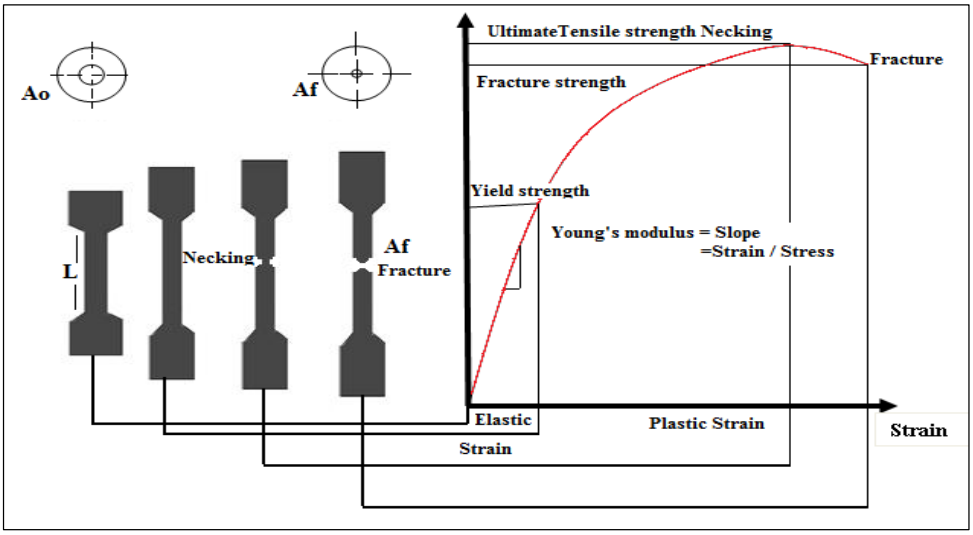
There is an enormous range of possible new materials, and it is often difficult to physically model the relationships between constituents, and processing, and final properties. Therefore, data mining is becoming one of the increasingly valuable tools in the general area of materials design and their development. Materials properties and their sub property ranges identified from different data sources and trivial importance for predicting the correlated properties of materials are listed in the table 1.

Table 1 : Materials sub properties and their range values indentified from different sources of data.

Attribute	Sub Property	S Name	Polymers	Ceramics	Metals
#1	Tensile Strength	TS	1 – 30	31 – 100	101 – 300
#2	Yield Strength	YS	1 –26	0	26 – 250

3. Determination of Material Property

When studying materials and especially for selecting materials for a project/design, it is important to understand key properties of materials. The most important properties are tensile strength, elasticity, plasticity, ductility and tensile strength. Here in this paper, tensile strength, which is a sub property of the mechanical property, of materials[6] is considered for prediction. The mechanical properties of a material describe how it will react to physically applied forces. Mechanical properties occur as a result of the physical properties inherent to each material, and are being determined through a series of standardized mechanical tests[7][8]. One of the mechanical test is tensile test where yield and tensile strength values can be observed in the test[14]. A tensile test can be performed by incrementing the sample with an electrical device to measure strain and then stretching the sample until it fails. The stretch, both elastic and plastic, is called strain. Using the original cross-sectional area of the sample, the load is converted into stress, and a stress-strain diagram is obtained as shown below in figure 2.



Figurer 2: Shape of ductile specimen at various stages of testing

There are three very important mechanical properties, which can be determined from this diagram. They are yield strength, tensile strength and modulus of elasticity. From the diagram, yield strength can be determined first and tensile strength can be determined later. Using these two properties , modulus of elasticity can be determined. The yield strength is defined as the stress at which a predetermined amount of permanent deformation occurs. The graphical portion of the early stages of a tension test is used to evaluate yield Strength. Yield strength is not a characteristic that can be calculated. It must be derived through experiment and then it can be calculated by using the following formula:

$$Stress \ , \sigma = \frac{L}{A}$$

Where L is applied and A is the area of cross section. Using this formula, yield strength can be computed. So tensile strength and yield strength values can be computed and stored in the materials database through tensile test. As there are two attribute values(tensile strength and yield strength attribute values) coming up from a single experimental test, where these two properties showing the correlation between each other, predictive attribute values can be computed. Therefore, a regression analysis is proposed on the correlated property values to predict the other property values without going for further experimental tests. This leads to manufacturer to hoard the time and work for finding the tensile strength. The methodology, which is being proposed for correlated data sets is described in the following section.

4.METHODOLOGY

Linear Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences. It well known as knowledge discovery method in data mining field[2][3] in Computer science. Its main objective is to explore the relationship between a dependent variable tensile strength(Y) and independent variable, yield strength(X) of an engineering material. **Co-efficient of Determination** is used to measure the accuracy of the linear regression model and **Least-Squares Estimation method is proposed** for estimating parameters by minimizing the squared discrepancies between tensile strength(Y) and yield strength(X) of an engineering material.

4.1 Linear Regression Model

Giving the measures of tensile strength and yield strength attributes, linear regression analysis can be measured how strongly one attribute implies the other, based on the available data. For numerical attributes say X and Y, we can evaluate the correlation between these two attributes by computing the Correlation Coefficient, r as follows:

$$r = \frac{\sum_{i=1}^{N} (x_i . y_i) - N \overline{X}\,\overline{Y}}{N \sigma_x \sigma_y} \qquad (1)$$

Where N is the number of tuples, x_i and $\boldsymbol{y_i}$ are the respective attribute values of X and Y in a tuple i, $\overline{X}$ and $\overline{Y}$ are the respective mean values of X and Y, σ_x and σ_y are the respective standard deviations of X and Y and $\sum x_i y_i$ is the sum of the cross-product, XY (that is, for each tuple, the value for x is multiplied by the value for y in that tuple). Note that $-1 \le r \le +1$.

$$\sigma_x = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2 \qquad (2)$$

Where $\bar{x}$ is the mean value of the observations.

$$\sigma_y = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 \quad (3)$$

Where $\bar{y}$ is the mean value of the observations.

In simple linear regression analysis, the data are modelled to fit a straight line and to approximate a set of data. For example, a random variable, y (called a response variable), can be modelled as a linear function of another random variable, x (called a predictor variable), with the equation given below

$$y = ax + b \quad (4)$$

where *b* is a constant, the point at which the line crosses the y-axis when x = 0, *a* is a coefficient representing the "slope" of the line, x, is the observed value of the independent variable. For the i^{th} case, the slope of the line is given by the formula:

$$a = \frac{\sum\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sum\left(x_i - \bar{x}\right)^2} \quad (5)$$

Where $\sum\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]$ is the sum of cross product of deviations of x and y from their means $\overline{X}$ and $\bar{Y}$ respectively. $\sum\left(x_i - \bar{x}\right)^2$ is the sum of the squared deviation of x from x- mean, $\overline{X}$.

$$b = \hat{Y} - a\hat{X} \quad (6)$$

Where $\hat{Y}$ is the mean of y values and $\hat{X}$ is the mean of x values .

Regression line goes through the point whose co-ordinates are the mean values of the variables X and Y.

4.2 Least-Squares Estimates

Least-Squares Estimation method is proposed for estimating parameters by minimizing the squared discrepancies between tensile strength(y) and yield strength(x) of engineering materials. This method is studied in the context of a regression problem, where the variation in the response variable y , can be partly explained by the variation in the co variable x.

$$y = ax + b + \varepsilon \quad (7)$$

representing the true linear relationship between yield strength and tensile strength for all materials, the error term, ε is needed to account for the indeterminacy in the model, the residuals $\left(y_i - \Pr edcator\right)$ are estimates of the error terms, $\varepsilon_i, i = 1,2,\ldots\ldots,n.$= equation (7) is called regression equation.

4.3 Co-efficient of Determination

Co-efficient of determination measure, r^2 is used to assess how well a linear regression model explains and predicts future outcomes. The coefficient of determination , r^2, is computed by

$$r^2 = \frac{SSR}{SST} \tag{8}$$

where, SSR is sum of square regression and SST is sum of square total. Therefore, it can be derived from the following:

$$\text{SSR} = \frac{\left[\sum xy - (\sum x)(\sum y)/n\right]^2}{\sum x^2 - (\sum x)^2/n} \tag{9}$$

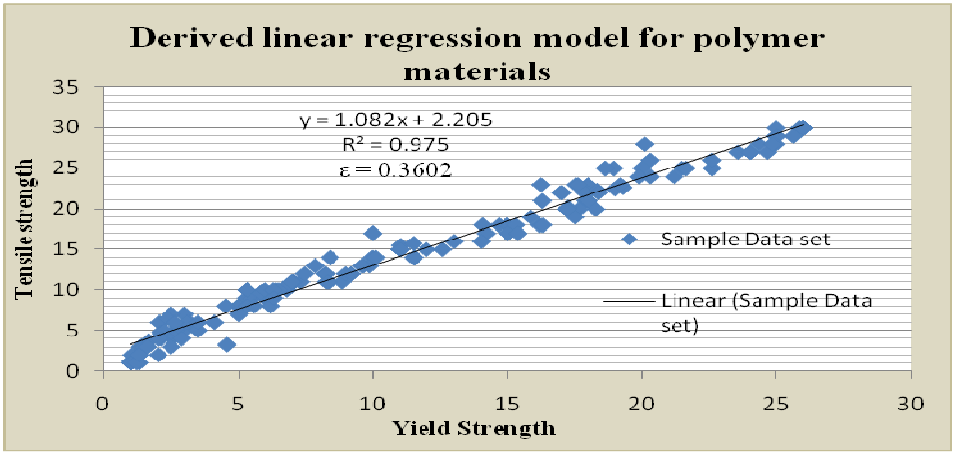
$$\text{SST} = \sum y^2 - \left(\sum y\right)^2/n \tag{10}$$

5. Experimental Results

Experiments on different materials data sets, created by referring various websites and materials hand books related to engineering materials, are done.

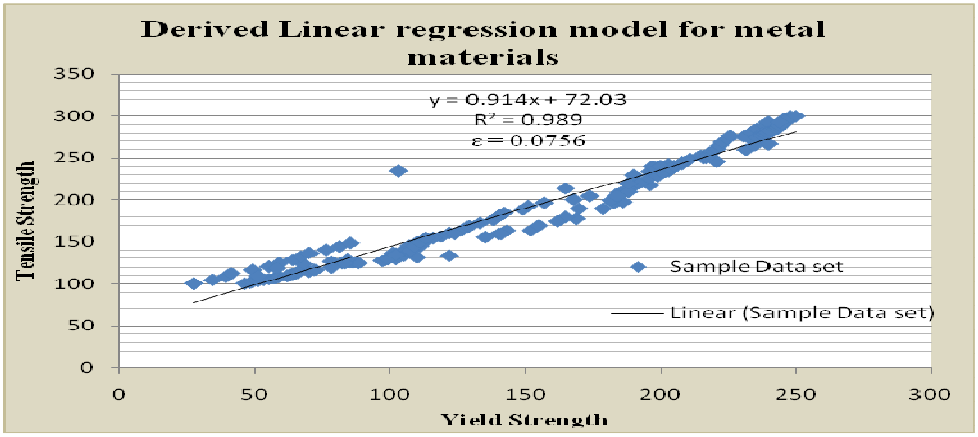
5.1. Data sets used for deriving Linear Regression Models

For deriving a linear regression model that approximates the relationship between the yield strength and tensile strength values of polymer data sets, 150 positively correlated polymer data sets, whose yield strength and tensile strength values lies respectively in the ranges (1MPa - 30MPa) and (1MPa - 26MPa) are considered for deriving the linear regression model. The polymer data sets, which are positively correlated between yield strength and tensile strength, considered for deriving the regression model is shown in the figure 3.



Figurer 3 . Positively correlated polymer materials and their yield strength and tensile strength.

Correlation coefficient model is computed on the positively correlated 150 metal data sets whose, yield strength and tensile strength values lies respectively in the ranges (101MPa – 300MPa) and (26MPa – 250MPa). The Metal data sets, which best fits the linear regression line, are shown in the figure 4.



Figurer 4 .Positively correlated Metal materials and their yield strength and tensile strength.

By using positively correlated polymer and Metal data sets, the parameters such as coefficients, a and b of regression model, and accuracy measuring parameters such as coefficient of determination, r^2 and least square estimate ε, are computed for the polymer and metal data sets.

Linear regression model, Y = 1.082(x) + 2.205 is derived from the positively correlated polymer data sets in the database and its coefficient of determination represents prediction accuracy, which is about $r^2 = 0.975$ with minimized least square estimates , $\varepsilon = 0.3602$. Similarly the linear regression model, Y = 0.914(x) + 72.03 for positively correlated metal data sets determines the coefficient of determination, $r^2 = 0.9890$ of prediction accuracy with minimized least square estimates, $\varepsilon = 0.0756$. Experimentally determine parameters are tabulated in the table 2.

Table 2. Derived regression models and accuracy measuring parameters for Polymer and Metal materials.

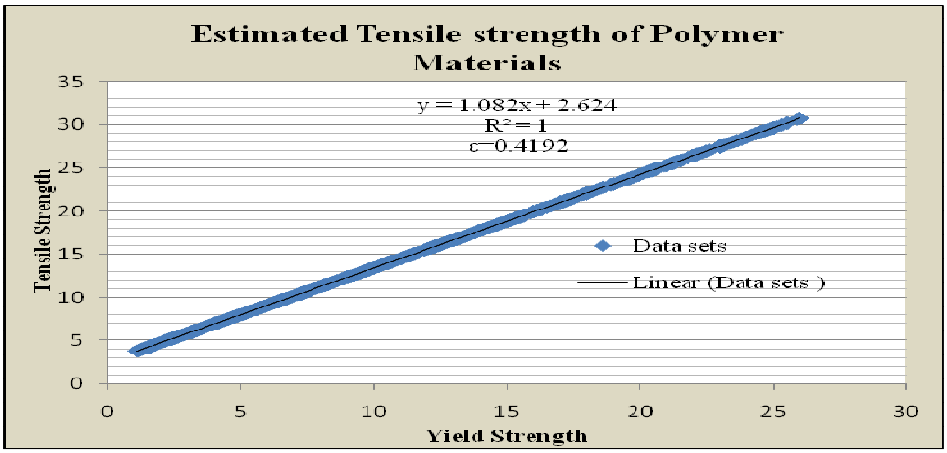
Materials Types	Regression Models	Co-efficient of Determination(r^2)	Least square estimates ε
Polymer materials	Y=1.082(x)+ 2.205	0.975	0.3602
Metal materials	Y=0.914(x)+72.03	0.989	0.0756

5.2. Data Sets Used For Validating Linear Regression Models

The above models are validated on randomly selected data sets of both Polymer and Metal types.

I. Polymer Data sets:

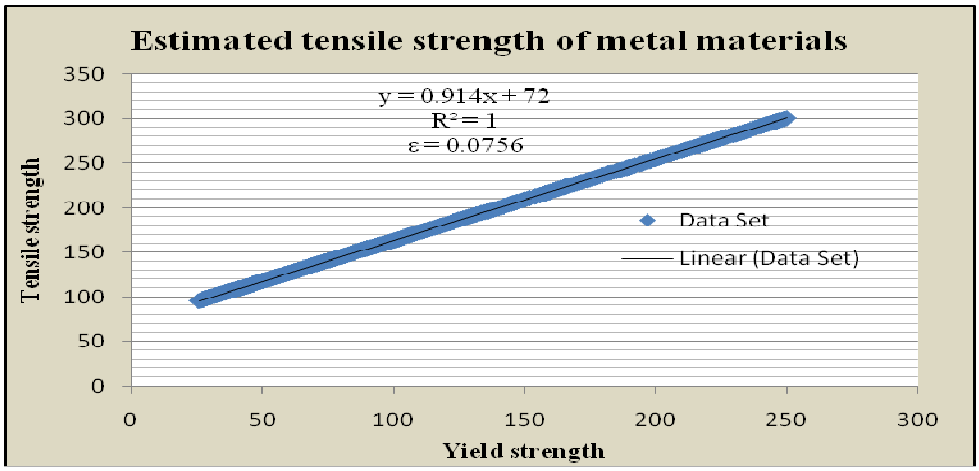
In order to validate the linear regression model for polymer data sets, 1500 polymer data sets are randomly selected from material database. For any polymer materials, whose yield strength (x) lies in the range (1.00MPa- 26.00MPa), tensile strength(Y) is being predicted by the liner regression model and best fits the line Y= 1.082(x) +2.624 with $r^2 = 0.9976$ and $\varepsilon = 0.4192$, also lies in the range((1Mpa- 30MPa). Linear regression line that best fits all the 1500 polymer data sets that have linear relationship between their yield strength and tensile strength is shown in figure 5.



Figurer 5. Show the Polymer materials straight line with predication

Metal Data Sets:

Similarly 1500 Metal data sets are randomly selected from material database. For any metal materials, whose yield strength (x) lies in the range (26MPa- 250MPa), tensile strength(Y) is being predicted by the liner regression model and best fits the line Y= 0.914(x) + 72.01 with r^2 = 0.9994 and ε = 0.0.0757., also lies in the range (101MPa- 300MPa) Regression model that best fits all the 1500 Metal datasets that have linear relationship between their yield strength and tensile strength is shown in figure 6.



Figurer 6. Show the Metals materials straight line with predication

From the data sets used for validating the derived models shown in figure 5 and 6, regression models and accuracy measuring parameters for both polymer and metal materials are tabulated in the table 3.

Table 3. Validated regression models and accuracy measuring parameters for Polymer and Metal materials.

Materials Types	Regression Models	Co-efficient of Determination(r^2)	Least square estimates ε
Polymer materials	Y=1.082(x)+ 2.624	0.9976	0.4192
Metal materials	Y=0.914(x)+72.01	0.9994	0.0757

By comparing the table 2 and 3, it s found that, coefficient of determination, performance of the predictive model increases as the number of data sets increase and best fits al the data sets with linear relationship.

6.Conclusion and future scope

In this paper, linear regression models is implemented and proposed on engineering materials data sets for predicting the linear relationship between the yield strength and tensile strength of engineering materials. As we absorbed in the materials database two materials properties values are correlated with each other and found to have a linear relationship between yield strength and tensile strength. So by knowing yield strength of materials, which are experimentally determined in nature, we can predict tensile strength of those materials. This helps design engineers to reduce their work in finding tensile strength through physical experiment. The predicated results analyzed in this research depict that linear regression technique can be used for predicting the mechanical properties of engineering materials. This prediction analysis can be more useful for mechanical design engineers in the manufacturing industry to overcome by the wear and tear of the cost and time of the design and manufacturing industries.

Further, multiple linear regression model is proposed for predicting more than one mechanical properties by knowing the existing properties.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support from the University Grant Commission (UGC), INDIA for the Major Research Project “Scientific Knowledge Discovery Systems (SKDS) For Advanced Engineering Materials Design Applications” vide reference F.No. 34-99\2008 (SR), 30th December 2008, and also gratefully acknowledge the unanimous reviewers for their kind suggestions and comments for improving this paper.

REFERENCES

[1]. Lukasz A. Kurgan and Petr Musilek,. " A survey of Knowledge Discovery and Data Mining process models" The Knowledge Engineering Review, Vol. 21:1,pp 1–24 (2006).
[2]. Dan Braha. "Data Mining for Design and Manufacturing " Springer, pp.544.(2002).
[3]. Osmar R. Zaiane,.:(1999)"Introduction to Data Mining" Principles of Knowledge Discovery in Databases.
[4]. SONG Qinggong "A preliminary investigation on materials informatics " Chinese Science Bulletin Vol. 49 No.2 pp.210-214(2004).
[5]. WEI Qunyi, PENG Xiaodong, LIU Xiangguo & XIE Weidong "Materials informatics and study on its further development " Chinese Science Bulletin Vol. 51 No. 4.pp. 498—504, (2006).
[6]. E.J.Pavlina and C.J.Van Tyne. "Correlation of yield strength and tensile strength with Hardness for steels " Journal of materials Engineering and performance Vol.17(6),pp.888-893(2008).
[7]. Kenneth G.Budinski (2000)"Engineering Materials Properties and selection" 5th edition June .
[8]. W.D. Callister Jr.,(2000) Materials Science and Engineering. 5th ed., John Wiley and Sons, New York, NY, USA.
[9]. Engineering materials data sets available at http://www.matweb.com.
[10]. Y Kaji, H Tsuji, M Fujita, Y Xu, K Yoshida, S Mashiko, K Shimura,S Miyakawa4 and T Ashino "Development of a Knowledge Based System Linked to a Materials Data Base" Data Science Journal , Vol 3, 21 pp,88-95. (2004).
[11]. Gang Yu1,2, Jingzhong Chen1, Li Zhu "Data mining techniques for materials informatics: datasets preparing and applications 2009 Second International Symposium on Knowledge 2009 Second International Symposium on Knowledge Acquisition and Modeling published be IEEE pp 189-192.(2009).

[12]. Doreswamy, Hemanth K S, Nagaraju S." Naive Bayesian Classifier For Knowledge Discovery From Materials Informatics" International Conference on Information and Communication Technology ICICT – 2010.pp 228-233.
[13]. Markus J.Buehler , Jef Dodson, Adri C.T. van Duin, Peter Meulbroek, William A. Goddard "The computational Materials Design Facility(CMDF): A powerful framework for mutliparadigm multi-scale simulations" Mat. Res. Soc. Proceedings (Combinatorial Methods and Informatics in Materials Science), Vol. 894, LL3.8, 2006.
[14]. Park SungHo, Park NoSeok, Kim JaeHoon,."A Statistical study on tensile characteristics of stainless steel at elevated temperatures" 15th International Conferences on the stength of Materials (ICSMA-15)Conferences Series 240(2010)012083.
[15]. Md.Noor-E-Alam, Tahmina Ferdousi Lipi , M. Ahsan Akhtar Hasin, A.M.M.S. Ullah" Algorithms for fuzzy multi criteria decision making (ME-MCDM) " Knowledge -Based systems Vol. 24.pp.367-377(2011).
[16]. A.M.M Shair Ullah, Khlifa H. Harib" An intelligent method for selecting optimal materials and its application " Advanced Engineering Informatics Vol.22. pp.473-483(2008).
[17]. LUKASZ A. KURGAN and PETR MUSILEK "A survey of Knowledge Discovery and Data Mining process models"The Knowledge Engineering Review, Vol. 21:1,pp.1-24.
[18]. A. KUSIAK " Data mining: manufacturing and service applications" International Journal of Production Research, Vol. 44, No.18–19,pp.4175-4191(2006).

Authors

Doreswamy received B.Sc degree in Computer Science and M.Sc Degree in Computer Science from University of Mysore in 1993 and 1995 respectively. Ph.D degree in Computer Science from Mangalore University in the year 2007. After completion of his Post-Graduation Degree, he subsequently joined and served as Lecturer in Computer Science at St.Joseph's College, Bangalore from 1996-1999 and at Yuvaraja's College, a constituent college of University of Mysore from 1999-2002. Then he has elevated to the position Reader in Computer Science at Mangalore University in year 2003. He was the Chairman of the Department of Post-Graduate Studies and Research in Computer Science from 2003-2005 and from 2008-2009 and served at varies capacities in Mangalore University and at present he is the Chairman of Board of Studies in Computer Science of Mangalore University. His areas of research interests include Data Mining and Knowledge Discovery, Artificial Intelligence and Expert Systems, Bioinformatics, Molecular Modeling and Simulation, Computational Intelligence, Nanotechnology, Image Processing and Pattern Recognition. He has been granted a Major Research project entitled **"Scientific Knowledge Discovery Systems(SKDS) for advanced Engineering Materials Design Applications"** from the funding Agency University Grant Commission, New Delhi, INDIA. He has published about 30 contributed peer reviewed papers at National/International Journals and Conferences. He received **SHIKSHA RATTAN PURSKAR** for his outstanding achievements in the year 2009 and **RASTRIYA VIDYA SARAWATHI AWARD** for outstanding achievement in chosen field of activity in the year 2010.

Hemanth K S received B.Sc degree and MCA degree in the years 2006 and 2009 respectively. Currently working as Project Fellow under UGC Major Research Project and is working towards his Ph.D degree in Computer Science under the guidance of Dr. Doreswamy in the Department of Post-Graduate Studies and Research in Computer Science, Mangalore University.

Manohar M G received B.Sc and M.C.A Degree from Karnataka University Dharwad in 1999 and 2002 respectively. After completion of his Post-Graduation Degree, he subsequently joined Honeywell Software technology Solutions Pvt. Ltd. Bangalore India and served for an around a year as software engineer. Then in the year 2003 he joined Mangalore University and currently working as System Programmer/Analyst. Meanwhile he received the MPhil Degree in Computer Science in the year 2011 and continuing the research work for his Ph.D degree in Computer Science under the guidance of Dr. Doreswamy in the Department of Post-Graduate Studies and Research in Computer Science, Mangalore University