Prediction Model on Web Mining

Pratik J. Bhise[1] Prajakta M. Ramteke[2] Mithun R. Funde[3]
[1,2,3]Department of Information Technology
[1,2,3]Tulsiramji Gaikwad Patil College of Engineering & Technology

Abstract— Because of the rapid growth of the word-wide-web, the problem of predicting a user's browsing behavior on a web-site has gained important and the need is personalize and influence a user's browsing experience. For solving this problem Markov models and its variations have been found. Personalization, building proper websites, promotion, getting marketing information, and forecasting market trends etc, the prediction result can be used. By using Markov model the users' browsing behaviors can be predicted at category level. Probability of present and infer users' browsing behaviors at webpage level by applying the Bayesian theorem. The system can effectively filter the possible category of the websites by Markov models and predict websites accuracy by the Bayesian theorem. In this paper we present different techniques for prediction of users' browsing behaviors and intelligently selecting parts of different order Markov model. The resulting model has improved prediction accuracy and reduced state complexity.

Key words: Web Usage Mining, Markov models, Bayesian theorem, Prediction

I. Introduction

To improve the web cache performance, recommend related pages, understand and influence buying patterns, improve search engines, and personalize the browsing experience. The problem of predicting a user's surfing behavior on a web-site has attracted a lot of research interest it is used. In recent years, various enterprises have changed the ways of doing business, which make the development of web usage mining skills important and enhance the rapid development of E-commerce directly. Data mining is the technique to seek extract knowledge from the web data.

Web Mining is the data mining techniques to automatically discover and extract information from web documents/services. There are main three types of web mining technique that is web structure mining, web content mining & web usage mining. To discovers knowledge from hyperlinks, which represent the structure of the web, the web structure mining is used. To extracts the useful information or knowledge from web page contents, the web content mining is used. Web usage mining mines customer access patterns from usage logs, which record clicks made by every user. In this paper we use the web usage mining technique.

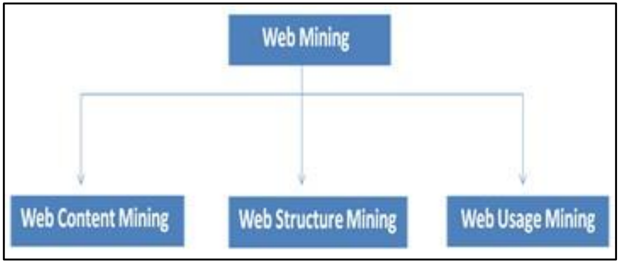


Fig. 1: Type of Data Mining

The technology of data mining that is web usage mining applies on the web services and Internet. It is used to find out, extract the pattern and the knowledge in the web usage.

It is mandatory to understand the user browsing behaviour and apply web usage mining technology, for achieve the purpose. To predict user's and costumer's browsing behaviour through mining the web data or web log files the web usage mining is used.

In this paper, costumer's browsing behaviour will be observed at two levels to meet the nature of the World Wide Web. First is the category level and second is web page level. On web site every category contains various web pages. The transaction probabilities of at web page level are less stable than category level. With the first order transaction probability, first level of prediction model, Markov model is used to predict the category of users' next step. It will help to reduce the operation scope.

II. Proposed System

Below figures clearly shows that the explanation overview of proposed system of our project.

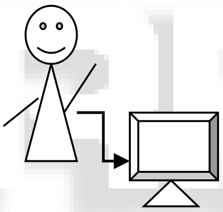


Fig. 2: First time visit

1) It shows that at first time user will be visiting first time.

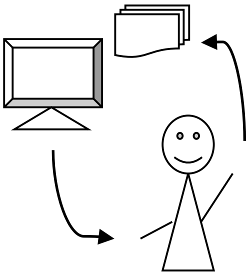


Fig. 3: Transaction

2) It shows that transaction between user and system will start

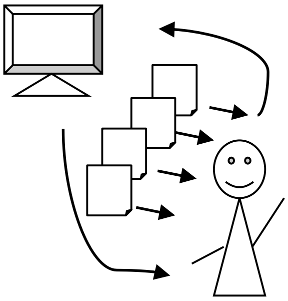


Fig. 4: Visit Various Pages

3) User will visit various pages, the transaction will be save in database for that user.

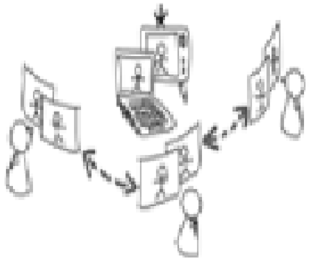


Fig. 5: Prediction

4) When that same user will visit same system, our system give prediction for user depend which depend upon the visiting history.

III. METHODOLOGY

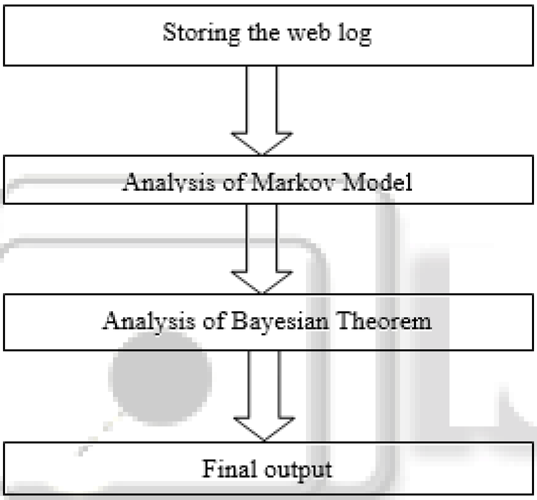


Fig. 5: PUCC Model

Above figure shows the PUCC model in which there are four stage using this we are predict the user future request. This four stage are as follows, first stage is used for cleaning the data in which we removed the unwanted log files. In the second stage, cookies were identified and removed. In third stage a graph partitioned clustering is used to navigate the user request. Final stage is the output stage in which we get the result.

IV. MARKOV MODEL ANALYSIS

Markov models have been mostly used to identify and analyze user or customer Web navigation data. Markov model are used to enable the identification of categories of is a excise way of representing a collection of sessions and they have been widely used to model a collection of customer sessions. In such context, each and every Web page in the web site corresponds to a state in the model, and each pair of web pages viewed in sequence corresponds to a state transition in the Markov model. A transition probability is calculated by the ratio of the number of times the transaction was traversed to the number of times the first state in the pair of web pages was visited. In many cases, the first-order Markov model predict less accuracy in achieving right predictions, which is why extensions to higher order models are compulsory. All higher order Markov model holds the promise of achieving upper prediction accuracies and improved coverage or its range than any single-order Markov model, at the expense of a dramatic rise in the state-space complexity. This led us to develop methods for intelligently combining different order of Markov models so that the resulting or final model has low state complexity, retains the coverage of the all higher order of above model and improved prediction accuracy.

A. *Similarity Matrix of Web Categories:*

Proposed Research framework of level one is to create the similarity matrix from web log file stored in database. At first, the situation of categories in each customer's session has to be count. The vector_i = <v_(l,i,..,) v_(h,i,…,) v_(m,i)> for each category i is gather the i^th element of session c from all m customer sessions, v_(l,i)= 1 means customer h visited web page of category i otherwise v_(l,i)= 0. Two categories can be calculated the Set of similarity and Euclidean distance simultaneously, by equation and Euclidean distance is further normalized equation. The final results are calculated by similarity and Euclidean distance. They are combined into a weight total similarity equation as equation.

$$\mathrm{SetSim}(\mathrm{A},\mathrm{B}) = \left[\frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{A} \cup \mathrm{B}}\right]$$

Euclidean distance

$$D(A|B) = [_{\mathrm{i}-1}^{\mathrm{m}}> (\mathrm{A_i} - \mathrm{B_i})^2]$$

Normalization

$$\mathrm{N}\big(\mathrm{D}(\mathrm{A}|\mathrm{B})\big) = \left[1 - |\frac{\sum_{\mathrm{i}-1}^{\mathrm{m}}(\mathrm{A_i} - \mathrm{B_i})^2}{\mathrm{m}}|\right]$$

Weight total similarity

$$S(A,B) = \big[\mathrm{SetSim}(\mathrm{A},\mathrm{B}).\mathrm{Wss} + \mathrm{N}\big(\mathrm{D}(\mathrm{A}|\mathrm{B})\big).\mathrm{W_D}\big]$$

Where

$$\mathrm{Wss} + \mathrm{W_D} = 1$$

$$\mathrm{W_D} = 1 - \mathrm{Wss}$$

After the similarity is computed, the similarity matrix S is a k x k matrix of similar category, where the similarity between $\mathrm{C_i}$ and $\mathrm{C_j}$ is $\mathrm{S_{ij}}$ .that is established by above equations.

$$\mathrm{S}= \begin{array}{cc} & [\mathrm{C_1} \quad \mathrm{C_2} \ldots \quad \mathrm{C_k}] \\ \begin{bmatrix} \mathrm{C_1} \\ \mathrm{C_2} \\ \vdots \\ \mathrm{C_K} \end{bmatrix} & \begin{bmatrix} \mathrm{S_{11}} & \mathrm{S_{12}} \ldots & \mathrm{S_{1k}} \\ \mathrm{S_{21}} & \mathrm{S_{22}} \ldots & \mathrm{S_{2k}} \\ \vdots & \vdots & \vdots \\ \mathrm{S_{k1}} & \mathrm{S_{k2}} \ldots & \mathrm{S_{kk}} \end{bmatrix} \end{array}$$

B. *Transition Matrix Of Markov Model:*

Proposed Research framework of level two is to create the transition matrix of Markov model P, which is based on web log file stored in database as well as similarity matrix. The P matrix is first-order transition matrix of Markov model and it is represented as follows:

$$\mathrm{S}= \begin{array}{cc} & [\mathrm{C_1} \quad \mathrm{C_2} \ldots \quad \mathrm{C_k}] \\ \begin{bmatrix} \mathrm{C_1} \\ \mathrm{C_2} \\ \vdots \\ \mathrm{C_K} \end{bmatrix} & \begin{bmatrix} \mathrm{P_{11}} & \mathrm{P_{12}} \ldots & \mathrm{P_{1k}} \\ \mathrm{P_{21}} & \mathrm{P_{22}} \ldots & \mathrm{P_{2k}} \\ \vdots & \vdots & \vdots \\ \mathrm{P_{k1}} & \mathrm{P_{k2}} \ldots & \mathrm{P_{kk}} \end{bmatrix} \end{array}$$

Each element in the P matrix presents a transition probability between two categories of web pages. P_{ij} represents a transition probability which is computed by equation between category i and category j. The numerator of set of equation that is the number of transition times between category i and category j, and the denominator is that of total number of transition times between category I, and every category k. The transition matrix of P_n can be computed by equation.

C. *Relevance Matrix:*

Proposed Research framework of level three is to creating the relevance matrix. The element Rij of relevance matrix is equal or same to the product of Sij and Pij. which are receiving from similarity matrix and transition matrix of Markov model simultaneously. In this paper, the relevance is an necessary factor of prediction between any two categories of product. The relevance can be used to infer the users' browsing behavior between two categories. The relevance matrix is represented as follows:

$$S= \begin{array}{cc} & [C_1 \quad C_2 \ldots \quad C_3] \\ \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_K \end{bmatrix} & \begin{bmatrix} R^n_{11} & R^n_{12} \ldots & R^n_{1k} \\ R^n_{21} & R^n_{22} \ldots & R^n_{2k} \\ \vdots & \vdots & \vdots \\ R^n_{k1} & R^n_{k2} \ldots & R^n_{kk} \end{bmatrix} \end{array}$$

$$\text{Where} \quad R^n_{ij} = [S_{ij}.P^n_{ij}.R^n_{ij}]$$

Represents a relevance, which is computed by equation, between category i at time t−n and category j at time t.

D. *Bayesian Theorem:*

Markov model as well as Bayesian theorem also can be used to predict the most possible customers' next request. It is assumed that the sample spaces S, M and N are two events of sample spaces S and P(N) > 0. The condition probability P(M| N) means the probability of event A while event
N occurs. It is represented as follows:

$$P(M|N) = \frac{P(M \cap N)}{P(N)}$$

In the Bayesian theorem, the some of information is used to revise the priority probability and obtain the posterior probability from that theorem. The above techniques is called the equation and Bayesian theorem as follows:

$$P(M|N) = \frac{P\left(\frac{N}{M}\right)P(M)}{P(N)}$$

If A is a partition then, A = {A1, A2, ..., An}.

In Bayesian theorem used to revise the prior probability and obtain the posterior probability as follows:

$$P\ (Mj/N) = \frac{P\left(\frac{N}{Mj}\right)P(Mj)}{P(N)}$$

$$= \frac{P\left(\frac{N}{Mj}\right)P(Mj)}{\sum_{r-1}^{n} P(Mr)P(N/Mr)}$$

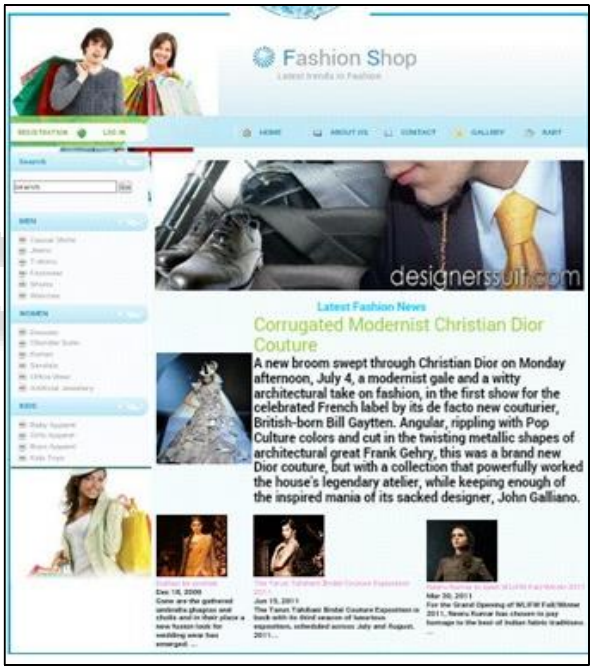
This research paper introduces an Efficient Two Level Prediction model to represent and analyze Web User behavior navigation data. Bayesian theorem and Markov model is used to enable the identification of customer navigation patterns and also used to the next link choice of a customer. It will help to minimize the operations in Level two in specific categories instead of all categories. After that, the web pages in specific categories are predicted by Bayesian theorem in the Level two of prediction model. It is expected that the two Levels of prediction model can minimize the scope of operation and increase the accuracy precision.

V. IMPLEMENTATION

Using above methodology we are designed one E-Commerce web site for which we are predicting the user's browsing behavior.

We have implemented this project in phases. First we have designed GUI of the project using PHP. Then for the database purpose we have used Sql Server 2005. For connecting GUI with database.

Following steps are used to actual execution and calculation.



Latest Fashion News

Corrugated Modernist Christian Dior Couture

A new broom swept through Christian Dior on Monday afternoon, July 4, a modernist gale and a witty architectural take on fashion, in the first show for the celebrated French label by its de facto new couturier, British-born Bill Gaytten. Angular, rippling with Pop Culture colors and cut in the twisting metallic shapes of architectural great Frank Gehry, this was a brand new Dior couture, but with a collection that powerfully worked the house's legendary atelier, while keeping enough of the inspired mania of its sacked designer, John Galliano.

Fig. 3: Home page

If user want to visit any web site that user required registered on that web site.

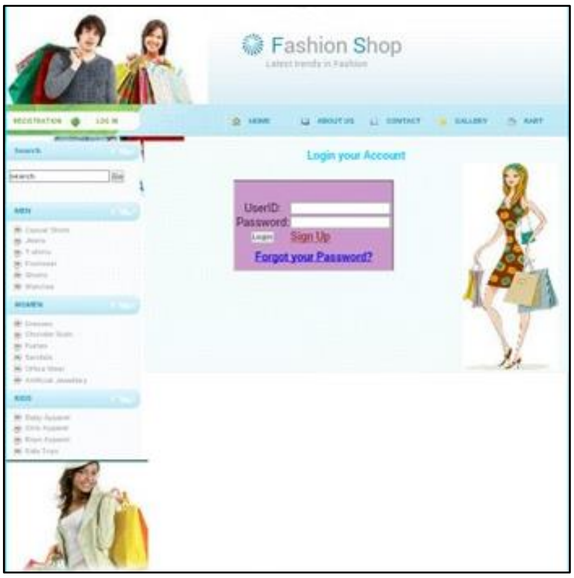


Fig. 4: log in page

After creating the account, user can registered or sign up the web side directly by entering the user name & password. If the user can not having their own account then he can create account by using "Registration" option & then account is conform.

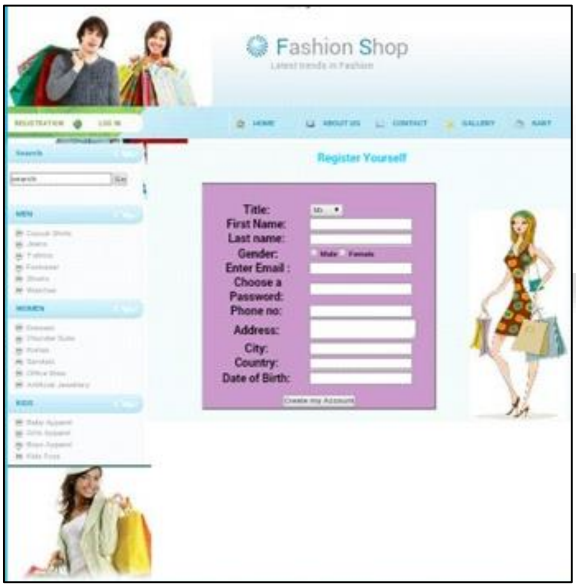


Fig. 5: Registration Page

First user will visit our website. He will registered on the home page, corresponding user data and password will save in database. User view or perches product from different category.

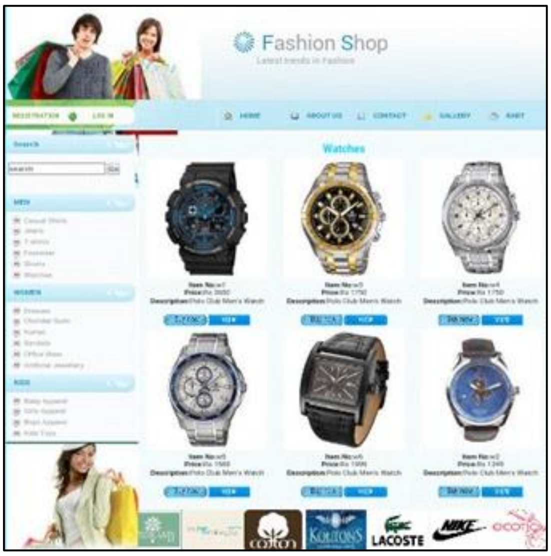


Fig. 6: Category (Watch)

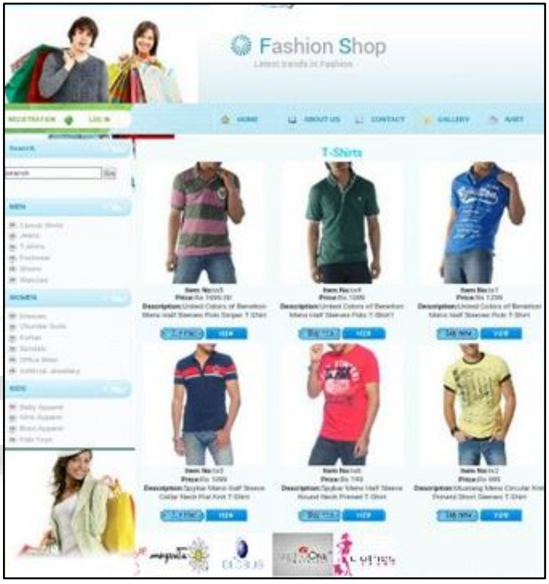


Fig. 7: Category (T-shirt)

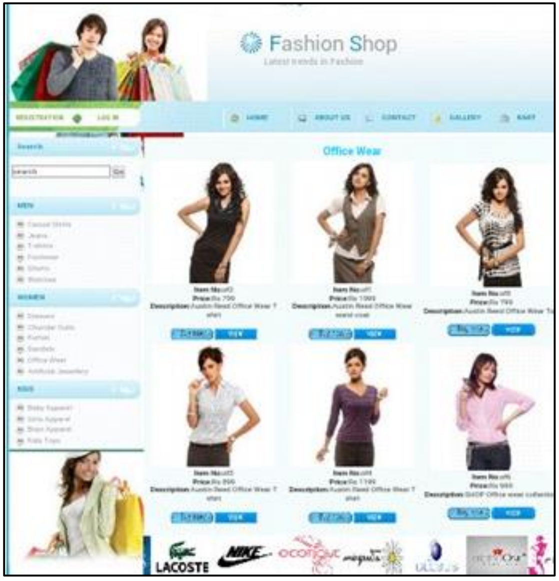


Fig. 8: Category (Formal ladies ware)

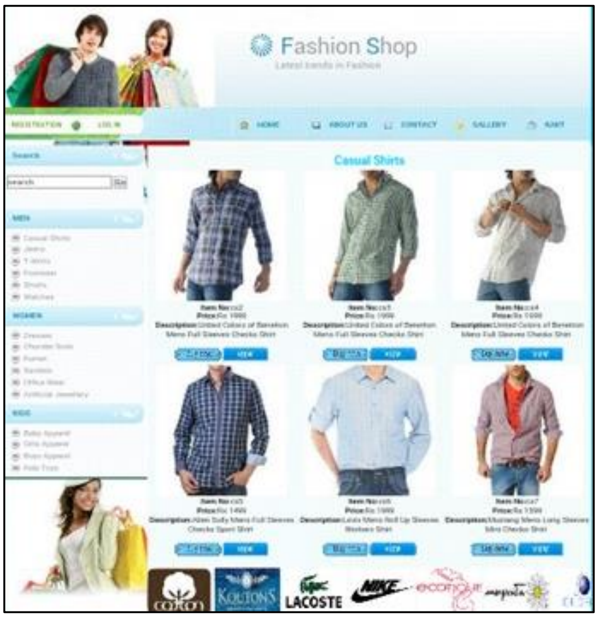


Fig. 9: Category (Formal shirt)

When different users views or purchases any product on site, the regarded information of particular product get stored in database.

The log files are generated to each user for each and every session and stored in database.

The prediction result is viewed when that user visit the web site next time again.

Access-pattern for each and every customer and maximum count for visited pages.

Using above business logic the prediction result gets calculated.

VI. Conclusion

The huge quantity of data of web pages on many portal sites, for convenience, are to assemble the web page based on category. In this paper, users browsing behavior will be observed at two levels to meet the nature of the portal. One is category stage and the other is web page stage. In stage one is to predict category. The unnecessary categories can be excluded. The scope of calculation is massively reduced. Next, using Bayesian theorem in the level two to predict the users' browsing page is more effective and accurate. The results of experiment prove the Hit Ratio is well in both stages.

References

[1] Chu-Hui Lee & Yu-Hsiang Fu,"The International Multi Conference of Engineers and Computer Scientists 2008". IMECS 2008, 19-21 March 2008, Hong Kong.(http://www.ieee.com)

[2] Mukund Deshpande & George Karypis,"Selective Markov Models for Predicting Web-Page Accesses", University of Minnesota, Department of Computer Science/Army HPC Research.

[3] http://www.webdatamining.net/usage.

[4] IEEE paper on Prediction Model for User's Browsing Behavior by Prof. Varsha Bhosale Information Technology, Vidyalankar Institute Of Technology.

[5] IEEE paper on Web Data Mining by Bing Liu Department of Computer Science University of Illinois at Chicago.

[6] Chu-Hui Lee & Yu-Hsiang Fu,"Web Usage Mining based on Clustering of Browsing Features" IEEE Transactions on Department of Information Management, Chaoyang University of Technology, 2008.

[7] Dilpreet kaur and Sukhpreet Kaur, "A Study on User Future Request Prediction Methods Using Web Usage Mining" IEEE Transactions on Computational Engineering, Volume 03, Issue 4, April 2013.

[8] V.V.R. Maheswara Rao & Dr. V. Valli Kumari," An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining " IEEE Transactions on Advances in Recent Technologies in Communication and Computing, 2010.

[9] http://kdd.ics.uci.edu/

[10] Mathias Gery, Hatem Haddad "Evaluation of Web Usage Mining Approaches for User‟s Next Request Prediction" WIDM‟03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8,2003.

[11] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", Procedia Engineering 30 ,2012.