International Journal of Research and Analytical Reviews

UGC Approved Research Journal

Periodicity - Quarterly

Atman Publishing Academy

International Journal of Research and Analytical Reviews

An open Access, peer reviewed, refereed, online and print research journal

Atman Publishing Academy

2061-C/2/B, Nr. Adhyatma Vidya Mandir, Sanskar Mandal, Bhavnagar-364002.
Contact : 9427903033 E mail : editorsijrar@gmail.com, ijrar1@gmail.com
Thoughts, language vision and example in published research paper are entirely of author of research paper. It is not necessary that both editor and editorial board are satisfied by the research paper. The responsibility of the matter of research paper/article is entirely of author.

2. Editing of the IJRAR is processed without any remittance. The selection and publication is done after recommendations of at least two subject expert referees.

3. In any condition if any National/International University denies accepting the research paper/article published in IJRAR, than it is not the responsibility of Editor, Publisher and Management.

4. Only the first author is entitle to receive the copies of all co-author.

5. Before re-use of published research paper in any manner, it is compulsory to take written permission from the Editor – IJRAR, unless it will be assumed as disobedience of copyright rules.

6. All the legal undertakings related to IJRAR is subject to Bhavnagar Jurisdiction.

Editor
International Journal of Research and Analytical Reviews
Atman Publishing Academy
2061-C/2/B, Nr. Adhyatma Vidya Mandir, Sanskar Mandal, Bhavnagar-364002.
Contact : 9427903033 E mail : editorsijrar@gmail.com, ijrar1@gmail.com
An open Access, peer reviewed, refereed, online and print research journal

**Editor in chief**

Dr. R. B. Joshi

**Senior Advisory Board**

<table>
<thead>
<tr>
<th>Name</th>
<th>Position and Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. H. O. Joshi</td>
<td>Retd. Prof. &amp; Head, Department of Education, Saurashtra University, Rajkot, Gujarat.</td>
</tr>
<tr>
<td>Dr. Bhaves Joshi</td>
<td>Associate Professor, College of Food Processing Technology &amp; Bioenergy, Agricultural University, Anand – 388110, Gujarat.</td>
</tr>
<tr>
<td>Vasantkumar Pathak</td>
<td>Director, Pathak Group of Schools &amp; College, Rajkot.</td>
</tr>
</tbody>
</table>

**Editorial Board**

<table>
<thead>
<tr>
<th>Name</th>
<th>Position and Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof. (Dr.) Ami Upadhyay</td>
<td>Director, Department of Humanities And Social Sciences, Dr. Babasaheb Ambedkar Uni. A’Bad.</td>
</tr>
<tr>
<td>Dr. A. Heidari</td>
<td>Faculty of Chemistry California South University (CSU) Irvine, California, U. S. A.</td>
</tr>
<tr>
<td>Dr. A. K. Lodi</td>
<td>H.O.D. Faculty of Education, Integral University, Lucknow(UP)</td>
</tr>
<tr>
<td>Dr. Anil Ambasana</td>
<td>Retd. Prof. &amp; Head, Department of Education, Saurashtra University, Rajkot.</td>
</tr>
<tr>
<td>Dr. Bhavesh Joshi</td>
<td>Associate Professor, College of Food Processing Technology &amp; Bioenergy, Agricultural University, Anand – 388110, Gujarat.</td>
</tr>
<tr>
<td>Dr. Bharat Ramanuj</td>
<td>Professor &amp; Head, Department of Education, Saurashtra University, Rajkot.</td>
</tr>
<tr>
<td>Dr. Dushyant Nimavat</td>
<td>Associate Professor Department of English, Gujarat University, Gujarat, India</td>
</tr>
<tr>
<td>Dr. Awa Shukla</td>
<td>Asst. Professor &amp; Director, Social Sciences Dept. Babasaheb Ambedkar Open University, Ahmedabad.</td>
</tr>
<tr>
<td>Dr. Manahar Thaker</td>
<td>Principal G. H. Sanghavi college of Education, Bhavnagar, Gujarat.</td>
</tr>
<tr>
<td>Dr. K. S. Meenakshisundaram</td>
<td>Director, C. A. A., Great Lakes Institute of Management, Chennai</td>
</tr>
<tr>
<td>Dr. J. D. Dave</td>
<td>I/c Principal P.D. Malviya Graduate Teachers' College, Rajkot, Gujarat.</td>
</tr>
<tr>
<td>Dr. Nahla Mohammed Abd El-Aziz</td>
<td>Assistant professor - Entomology Department, Faculty of Science Cairo University, Egypt.</td>
</tr>
<tr>
<td>Dr. Trupti Pathak</td>
<td>Assistant Vice President(Tech.) Claris life Sciences, Ahmedabad. Gujarat.</td>
</tr>
<tr>
<td>Dr. M. B. Gaijan</td>
<td>Associate Professor, Shamaldas Arts College, Bhavnagar.</td>
</tr>
<tr>
<td>Dr. Trupti Pathak</td>
<td>Assistant Vice President(Tech.) Claris life Sciences, Ahmedabad. Gujarat.</td>
</tr>
<tr>
<td>Dr. K. Ramadevi</td>
<td>Associate Professor Department of Civil Engineering Kumarakuru College of Technology, Coimbatore, Tamilnadu.</td>
</tr>
<tr>
<td>Dr. Jayant Vyas</td>
<td>Professor &amp; Head, Department of Education, M. K. Bhavnagar University, Bhavnagar</td>
</tr>
<tr>
<td>Dr. Sandeep R. Sirsat</td>
<td>Associate Professor &amp; Head, Department of Computer Science, Shri Shivaji Science &amp; Arts College, Chikhli, Dist: Buldana (M.S.-India)</td>
</tr>
<tr>
<td>Dr. Manahar Thaker</td>
<td>Principal G. H. Sanghavi college of Education, Bhavnagar, Gujarat.</td>
</tr>
<tr>
<td>Dr. A. K. Lodi</td>
<td>H.O.D. Faculty of Education, Integral University, Lucknow(UP)</td>
</tr>
<tr>
<td>Dr. Anil Ambasana</td>
<td>Retd. Prof. &amp; Head, Department of Education, Saurashtra University, Rajkot.</td>
</tr>
<tr>
<td>Dr. Awa Shukla</td>
<td>Asst. Professor &amp; Director, Social Sciences Dept. Babasaheb Ambedkar Open University, Ahmedabad.</td>
</tr>
<tr>
<td>Dr. Bhavesh Joshi</td>
<td>Associate Professor, College of Food Processing Technology &amp; Bioenergy, Agricultural University, Anand – 388110, Gujarat.</td>
</tr>
<tr>
<td>Dr. Bharat Ramanuj</td>
<td>Professor &amp; Head, Department of Education, Saurashtra University, Rajkot.</td>
</tr>
<tr>
<td>Dr. Dushyant Nimavat</td>
<td>Associate Professor Department of English, Gujarat University, Gujarat, India</td>
</tr>
<tr>
<td>Dr. Awa Shukla</td>
<td>Asst. Professor &amp; Director, Social Sciences Dept. Babasaheb Ambedkar Open University, Ahmedabad.</td>
</tr>
<tr>
<td>Dr. Bhavesh Joshi</td>
<td>Associate Professor, College of Food Processing Technology &amp; Bioenergy, Agricultural University, Anand – 388110, Gujarat.</td>
</tr>
<tr>
<td>Dr. Bharat Ramanuj</td>
<td>Professor &amp; Head, Department of Education, Saurashtra University, Rajkot.</td>
</tr>
<tr>
<td>Dr. Dushyant Nimavat</td>
<td>Associate Professor Department of English, Gujarat University, Gujarat, India</td>
</tr>
</tbody>
</table>

URL: http://ijrar.com

ISSN 2349-5138

E-ISSN 2348-1269
# Review Committee

**Editor & Head of Review Committee**

<table>
<thead>
<tr>
<th>Name</th>
<th>Position/Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dr. S. Chelliah</strong></td>
<td>Professor &amp; Head, Dept. of English and Comparative Literature, Madurai Kamraj University, Madurai-21, India.</td>
</tr>
<tr>
<td><strong>Mr. Zeeshan Shah</strong></td>
<td>Senior Lecturer, Department of Multimedia and Communication, University College of Bahrain, Kingdom of Bahrain.</td>
</tr>
<tr>
<td><strong>Harish Mahuvakar</strong></td>
<td>Associate Professor &amp; Head, Dept. of English, Sir P. P. Institute of Science, Bhavnagar, Gujarat, India.</td>
</tr>
<tr>
<td><strong>Dr. A. Kusuma</strong></td>
<td>Assistant Professor, Department of Social Work, Vikramasimhapuri University, Nellore.(AP)</td>
</tr>
<tr>
<td><strong>Prof. Rajeshkumar N. Joshi</strong></td>
<td>I/C Dean, Faculty of Arts &amp; Humanities, C. U. Shah University, Gujarat, India.</td>
</tr>
<tr>
<td><strong>Nahla Mohammed Abdelazez</strong></td>
<td>Assistant Professor, Faculty of Science, Cairo University, Giza Governorate, Egypt.</td>
</tr>
<tr>
<td><strong>Dr. Samira Shahbazi</strong></td>
<td>Plant Protection &amp; Biotechnology Research Group, Nuclear Agricultural Research School, Nuclear Science &amp; Technology Research Institute (NSTRI), Iran</td>
</tr>
<tr>
<td><strong>Dr. Mainu Devi</strong></td>
<td>Assistant Professor (Sr. Grade) in Zoology, Diphu Govt. college, Karbi Anglong – Assam India.</td>
</tr>
<tr>
<td><strong>Asim Gokhan YETGIN</strong></td>
<td>Assistant Professor, Faculty of Engineering, Dumlupinar University, Kutahya, Turkey.</td>
</tr>
<tr>
<td><strong>Sunita. B. Nimavat</strong></td>
<td>Assistant Professor of English, N.P.College of Computer &amp; Mgt., Kadi (North Gujarat).</td>
</tr>
<tr>
<td><strong>Dr. Riyad Awad</strong></td>
<td>Associate professor, Structural Engineering, An-Najah National University, Nablus, Palestine.</td>
</tr>
<tr>
<td><strong>Dr. Amer A. Taqa</strong></td>
<td>Professor Dept. of Dental Basic Science, College of Dentistry, Mosul University, Masul, Iraq.</td>
</tr>
<tr>
<td>Article Number</td>
<td>Title</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>APPROACH OF MULTIFACTOR AUTHENTICATION WITH DIFFERENT COMBINATION USING 3D PASSWORD</td>
</tr>
<tr>
<td>2</td>
<td>DEVELOPING A PROTOTYPE MODEL FOR ATTAINING AUTHENTICATION IN VIRTUAL CURRENCY TRANSACTIONS USING ASYMMETRIC CRYPTOSYSTEM – ELLIPTIC CURVE CRYPTOGRAPHY</td>
</tr>
<tr>
<td>3</td>
<td>ANALYSIS ON DISTRIBUTED QUERY PROCESSING AND OPTIMIZATION TECHNIQUES</td>
</tr>
<tr>
<td>4</td>
<td>A SURVEY: PERFORMANCE EVALUATION METRICS FOR DEHAZING ALGORITHMS</td>
</tr>
<tr>
<td>5</td>
<td>A SURVEY ON IMAGE SEGMENTATION TECHNIQUES</td>
</tr>
<tr>
<td>6</td>
<td>ROLE OF WIENER FILTERING IN IMAGE ENHANCEMENT FOR COMPRESSION AND SEGMENTATION</td>
</tr>
<tr>
<td>7</td>
<td>IMPROVING STUDENT’S PERFORMANCE USING K-MEANS CLUSTERING</td>
</tr>
<tr>
<td>8</td>
<td>A SURVEY ON INTELLIGENT TRANSPORT SYSTEM: TRANSPORT MANAGEMENT THROUGH MODERN TECHNOLOGY</td>
</tr>
<tr>
<td>9</td>
<td>SECURE SYMMETRIC KEY ALGORITHM FOR INFORMATION EXCHANGE USING MODULO 69</td>
</tr>
<tr>
<td>10</td>
<td>SOFTWARE TESTING TECHNIQUES</td>
</tr>
<tr>
<td>11</td>
<td>COMPARATIVE STUDIES FOR TEXT MINING</td>
</tr>
<tr>
<td>12</td>
<td>PARSING TECHNIQUE IN NATURAL LANGUAGE PROCESSING - AN OVERVIEW</td>
</tr>
<tr>
<td>13</td>
<td>SURVEY PAPER FOR SOFTWARE PROJECT ARRANGEMENT PROBLEM</td>
</tr>
<tr>
<td>14</td>
<td>PERSONALIZED RECOMMENDATION IN E-LEARNING SYSTEM - A SEMANTIC APPROACH</td>
</tr>
<tr>
<td>15</td>
<td>A PROBE ON TEXT MINING TOOLS, ALGORITHMS AND METRICS</td>
</tr>
<tr>
<td>16</td>
<td>CONVEX ANALYSIS OF FLOOD AND REGIONAL FREQUENCY IN HADOOP ANALYSIS</td>
</tr>
</tbody>
</table>
International Conference
on
Advanced Computing Technologies
Organized by
Department of Computer science
Madurai Sivakasi Nadars Pioneer Meenakshi Women’s College
Accredited with ‘B’ Grade by NAAC
(Affiliated to Alagappa University, Karaikudi)
Poovanthi
Sivaganga District.

Organizing Committee:

Convener:
Ms. K.Sudharani, Head of the Department

Co-Convener:
Ms.P.Priya, Assistant Professor
Ms.C.Sulochana, Assistant Professor

Committee Members:
Ms.T.Ramaporkalai, Assistant Professor
Ms.V.Jeyalakshmi, Assistant Professor
Ms.K.Rajeswari, Assistant Professor
Ms.S.Priyadharshini, Assistant Professor
Ms.B.Gomathi, Assistant Professor
Ms.S.Ashika Parvin, Assistant Professor
EDITORS OF SPECIAL ISSUE

Dr. S. NagaNandhini Sujatha
Assistant Professor
Department of Computer Applications
Sri Meenakshi Government Arts College for Women
Madurai

Ms. K. Sudharani
Head of the Department
Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College
Poovanthi

Ms. T. Ramaporkalai
Assistant Professor
Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College
Poovanthi

Ms. V. Jeyalakshmi
Assistant Professor
Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College
Poovanthi
INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING TECHNOLOGIES
Organized by Dept. of Computer science, Madurai Sivakasi Nadars Pioneer Meenakshi Women’s College

APPROACH OF MULTIFACTOR AUTHENTICATION WITH DIFFERENT COMBINATION USING 3D PASSWORD

Maheswari K.P & S.Nirmala Devi
1,2 Department of CS & IT, Subbalakshmi Lakshmipathy College of Science, Madurai

ABSTRACT: Authentication plays a vital role in all system for the purpose of security. Password protection is one of the simplest and the traditional way to avoid unauthorized users to access the credential data. Now a day, there are many different kinds of authentication techniques like textual, graphical, biometric, etc are available to strengthen the security. But at the other end, password theft is constantly evolving as hackers employ methods like key logging, phishing, pharming, etc. The reason can be some insufficiency in approaching the password or may be the loopholes in password generation. In this proposed system, the above mentioned problem is focused in different dimension. The system is developed based on the concept of 3D password and the multifactor is applied. Different combinations are worked out with different paradigm to enrich the security aspect and prevent hacker’s breach. This method also avoids the complexity in generating passwords.

Keywords: Authentication, textual passwords, graphical passwords, biometrics passwords, multifactor, 3D password.

Introduction
Overview
The user’s identity is verified through the process of authentication. This process is highly needed in all sectors for security. The main aim of doing authentication is to avoid unknown persons to access the confidential data. Mainly passwords are used to authenticate the persons. There are many types of authentication methods [1], Such as

1. Textual Authentication
2. Graphical Authentication
3. Biometric Authentication

Each has its own pros and cons. For any case one of the ways to eliminate or minimize the weakness is trying out the combinations. Thus, the concept of 3D password evolved. It provides advanced security and much stronger than other methods.

Textual Authentication
These passwords are created with the combinations of characters, digits and special symbols. These are weak passwords easy to guess and crack. It is vulnerable to many kinds of attacks [3].

Graphical Authentication
Graphical password schemes are better than the textual passwords. It works by having the user select from images, in a specific order, presented in a graphical user interface (GUI). For this reason, the graphical-password approach is sometimes called graphical user authentication (GUA). The possibility of risk is less comparatively to textual type.

Biometric Authentication
Biometric identification can provide extremely accurate, secured access to information. Fingerprints, Retinal and Iris scans produce absolutely unique data sets when done properly.

3D Password
A 3D password is a multifactor authentication scheme that combines Recognition+Recall+Tokens+Biometrics in one authentication system.

Caesar Cipher
It is one of the substitution ciphers. In this type the cipher text is generated by shifting the plain text based on the key. More complex encryption schemes employ the Caesar cipher as one element of the encryption process. The strength of this algorithm is depends on the key. More complex key value can avoid vulnerabilities.

Proposed System
In this paper, a method to strengthen the authentication has been proposed by multifactor logic with the concept of 3D password. The implementation is done by encrypting the factors such as textual password, graphical password and finger print as a biometric password in different combination. It is
worked out with three algorithms by having Caesar cipher as an encryption algorithm. The work was carried out by creating application using PHP & MYSQL. The system is developed in three phases.

**Phase I:** The factors needed for authentication are collected from the user through registration.

**Phase II:** The preprocess based on the combinations such as plaintext and key generation are performed.

**Phase III:** Execution of algorithms for encryption and results are stored in database to authenticate.

### Process Flow

![Process Flow Diagram](image)

### Algorithms

The proposed combinations are generated with the following algorithms.

#### Algorithm I

**[Encryption using Caesar Cipher with Text Password as Plain Text and Coordinates of Pixel as key]**

1. **Step 1:** Read Text password from the user (Plain Text - P)
2. **Step 2:** Read the coordinates (x, y) of pixel points from the image selected by the user (Key)
3. **Step 3:** Key Generation: Add the values of x and y until single digit as a key value (K).
4. **Step 4:** Apply Caesar cipher with the obtained plain text and a key.
   \[ E_n(P) = (P + K) \mod 26 \]
5. **Step 5:** Save the cipher text for verification.

#### Algorithm II

**[Encryption using Caesar Cipher with Text Password as Plain Text and pixel coordinates from image as key]**

1. **Step 1:** Read Text password from the user (Plain Text - P)
2. **Step 2:** Read Real End Point (p) & Real Branch Point (q) from the finger print of the user (Key)
3. **Step 3:** Key Generation: Add the values of p and q until single digit for a key value (R).
4. **Step 4:** Apply Caesar cipher with the obtained plain text and a key.
   \[ E_n(P) = (P + R) \mod 26 \]
Step 5: Save the cipher text for verification.

Algorithm III

[Encryption using Caesar Cipher with coordinates of the Pixel Points as Plain Text and Finger Print points as key]

Step 1: Read pixel points \((x, y)\) from the image selected by the user.
Step 2: Add the values of \(x\) and \(y\).
Step 3: Convert the resultant value in to text & let it be plain text \((P)\).
Step 4: Read Real End Point \((p)\) & Real Branch Point \((q)\) from the finger print of the user
(Key)
Step 5: Key Generation: Add the values of \(p\) and \(q\) until single digit for a key value\((S)\).
Step 6: Apply Caesar cipher with the obtained plain text and a key.
\[E_c(P) = (P + S) \mod 26\]
Step 7: Save the cipher text for verification.

Using the above algorithms authentication is carried out and it is strengthened by key generation. The first combination deals with textual password and graphical image. In this case, the selected coordinates \((x,y)\) from the image were summed up into a single digit for a key value. For second combination textual password and the Real End Point \((p)\) and Real Branch Point \((q)\) from the finger print were taken [4] and the key is evaluated as like in the first case. The coordinates\((x,y)\) of pixels from the image and the point \((p,q)\) from the finger print were used for third combination. This is a special case in which the coordinates \((x,y)\) are converted into text and used as a plain text for encryption. Like other cases here key is also important which is generated from finger print points.

Each case has different approach in authentication which can trim down Password Guessing, Password Cracking, Social Engineering attacks, Shoulder Surfing, Brute Force attack, Dictionary attacks etc at the maximum level. It is considered to be a novel approach of authentication which over comes the traditional way of using single factor authentication.

Analysis and Discussion

The proposed combination of multifactor authentication provides choices for the users to perform authentication based on their needs (i.e) passwords can be generated with any one of the above said methods. Through this option the various levels of authentication can be done. The system is strongly secure and easy for the users, at the same time very difficult for the third persons [5]. Hence the system suit both security and usability metrics.

By using the combination of the factors, the security is reinforced by overcoming the problems in the individual parameters like text based, graphical based and biometric based. This is achieved with the help of algorithms, especially for the graphical and biometric passwords. While considering the authentication through image, Pass Point Algorithm is focused [6]. Instead of working with multiple points in an image, single point coordinates are taken to avoid more time to complete the authentication process [7].

For biometric password, users’ finger print is taken. The two end points, real end point and real branch point in the finger prints acts as a unique factor for the identification of the user. It is processed as like an image, and stored in the database [4]. The preprocess works were done to encrypt the password for better security.

Caesar cipher, one of the substitution techniques is used to avoid the complexity in running algorithm with the better outcome. This is simple in structure and minimum security provider, because of its simple key nature. This drawback can be eliminated, when it is implemented with the keys generated using the images and finger print points. By doing so, it provides highest level of security by making the key difficult to predict.

Conclusion

The authentication with the combination of multifactor such as Textual, Graphical and Biometric passwords are tried out as an approach to strengthen the authentication. The complexity of verification of the factors for authentication is minimized by storing the combinations in form of cipher text in the database. Since the processed details in different combinations are stored, verification process is easy with less time consumption. The security is also enhanced with the help of algorithms.
References

DEVELOPING A PROTOTYPE MODEL FOR ATTAINING AUTHENTICATION IN VIRTUAL CURRENCY TRANSACTIONS USING ASYMMETRIC CRYPTOSYSTEM – ELLIPTIC CURVE CRYPTOGRAPHY

Anuradha.N¹ & Andal Priyadharsini.R²
¹,²Department of CS & IT, Subbalakshmi Lakshmipathy College of Science, Madurai

ABSTRACT: Virtual Currency plays a vital role in today’s digital era. This is also known as Crypto currency or digital currency and is in a form of electronic money. They do not physically exist as coins or currencies notes. In future this kind of technology will be used in every money transaction. Hence security should be ensured over these transactions to avoid intrusion. The authentication service can be implemented in Virtual currency through cryptosystems. Private key cryptosystems are not suitable for providing the authentication problem, because key management and key distribution is tricky here. Hence, public key cryptosystems are appropriate to surmount this difficulty. Public key cryptosystems not only provide authentication service but also provide key distribution and management, data integrity and non-repudiation services. It is sensible to build cryptographic schemes which support both secure and inexpensive. As ECC needs less memory requirement, low computational cost and low power consumption than other asymmetric systems, this prototype model is developed using ECC. Here Python programming language is used to develop a prototype model for attaining authentication in virtual currency transactions.

Keywords: Virtual Currency, Authentication, Private key, Public key, ECC, Python

Introduction

The money transactions through internet mode achieve a level of privacy by using the trusted third party. The public can see that someone is transferring an amount to someone else [1]. As it leads to tampering the security over money transaction, virtual currency is introduced without the intervention of third party. Cryptosystem plays an essential role in providing secrecy of virtual money transaction [2-3]. Generally, the cryptosystem can be constructed in two ways such as Private Key cryptosystems and Public key cryptosystems based on the number of keys supported by them.

The widely accepted asymmetric cryptosystems to augment security are RSA, Diffie-Hellman, Digital Signature Standard, ElGamol and Elliptic Curve Cryptography (ECC)[3]. As ECC is a contemporary and well-organized type of public key cryptography, used to generate digital signature and encrypt the data.

This kind of security is based on the difficulty to resolve discrete logarithms on the field defined by particular equations work out over a curve. The main advantage of using this algorithm is that the size of a key is significantly smaller than other public key cryptosystem algorithms like RSA or DSA[4][7]. As Python supports built-in libraries for internet protocols and widely accepted open source software, the prototype model is developed by this language.

Elliptic Curve

The elliptic curve is just representation of cubic equation ie \( y^2 = x^3 + ax + b \). As \( y = \sqrt{x^3 + ax + b} \), each curve is symmetric [1][3]. It is represented as \( E_p(a, b) \), where \( p \) is a prime number. Elliptic curve may be of singular or non-singular elliptic curve. If \( 4a^3 + 27b^2 = 0 \), then the elliptic curve is singular elliptic curve. It may not have three distinct roots. This kind of curve is not suitable for cryptography purpose. If \( 4a^3 + 27b^2 \neq 0 \), then the elliptic curve is known as non-singular elliptic curve. It has three distinct roots. This kind of curve is suitable for cryptography.

Elliptic curve can be defined as EC over \( Z_p \) (prime curve) and EC over \( GF(2^m) \) (Binary curve). Prime curves are used in software implementation and binary curves are in hardware implementation. The prime curve is represented as \( y^2 \mod p = (x^3 + ax + b) \mod p \). The variables and coefficients are restricted to elements of a finite field. The values should be limited from 0 through \( p-1 \). If the values exceed this range, \( \mod p \) operation is to be performed. The curve is to be focused in only one quadrant from \((0,0)\) through \((p-1,p-1)\) containing non-negative numbers. The logics over functions that work on curve point has been developed by python.
In an elliptic curve, adding two points \( P(x_p, y_p) \), \( Q(x_q, y_q) \) in the curve will give the third point \( R(x_r, y_r) \). Formula to find the slope
\[
\lambda = \frac{y_q - y_p}{x_q - x_p} \text{ if } P \neq Q
\]
\[
\lambda = \frac{3x_p^2 + a}{2y_p} \text{ if } P = Q \text{ where is taken from } E_p(a, b).
\]
To find the point \( R(x_r, y_r) = P + Q \)
\[
x_r = \lambda^2 - x_p - x_q \quad ; \quad y_r = \lambda(x_p - x_r) - y_p
\]

Fig. 1 Sample Elliptic curve to produce R point by adding P&Q

**Elliptic Curve Cryptography**

Elliptic Curves are used in public key cryptography to create short encryption keys.

<table>
<thead>
<tr>
<th>Symmetric key Size</th>
<th>Public key size</th>
<th>Keysize ratio ECC RSA/DSA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSA</td>
<td>RSA</td>
</tr>
<tr>
<td>80</td>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>112</td>
<td>2048</td>
<td>2048</td>
</tr>
<tr>
<td>128</td>
<td>3072</td>
<td>3072</td>
</tr>
<tr>
<td>192</td>
<td>7680</td>
<td>7680</td>
</tr>
<tr>
<td>256</td>
<td>15360</td>
<td>15360</td>
</tr>
</tbody>
</table>
They are in the form of $y^2=x^3+ax+b$. ECC reduces the processing overhead, because of the reduced key size. The most popular curve Secp256k1 is used in key generation process of bitcoins and is defined with $a=0$ and $b=7$. Hence the elliptic curve equation may be $y^2 = x^3 + 7$ [2]. A plot for elliptic curve $y^2 = x^3 + 7$ is

![Elliptic curve with a=0 & b=7 ie Secp256k1 curve](image)

An elliptic curve with $a=486662$ and $b=1$ is known as curve25519 and is used in the applications of SSH.

**Key generation and Exchange in ECC**

**Key Generation**

In Elliptic Curve Cryptography, the sender selects a private key ($P_r$) as a random number with 256 bit value. Any one point on the elliptic curve is considered and this point is multiplied with the private key chosen by the sender. This product is known as the public key ($P_u$) of the sender [2].

$$P_r = \text{random}(2^{256})$$

$$P_u = P_r \times P$$ where $P_r$ is the private key (random number) and $P$ is the point in the elliptic curve. The advantage of ECC is even if the intruder knows the values of $P$ and $P_u$, it is very challenging to find the value of $P_r$.

**Key Exchange**

The Flow for key exchange model is be like

**Sender 'Alice' Receiver 'Bob'**

Private key of Alice private key of $A\_P_r < P_r$.

Bob $B\_P_r < P_r$ Calculates public key of Alice $A\_P_u = A\_P_r \times P$.

Calculates public key of Bob $B\_P_u = B\_P_r \times P$.

Both sender and receiver exchange their public keys to each other and calculate shared key.

Shared Key of Alice $K = A\_P_r \times B\_P_u$ Shared Key of Bob $K = B\_P_r \times A\_P_u$ $K=A\_P_r \times B\_P_u \times P$

To encrypt a message $M$,

$$C = \{K \times A\_P_r, M + K \times B\_P_u\}$$

To decrypt the cipher text

$$M + K \times B\_P_u - B\_P_r \times (K \times P) = M + K \times (B\_P_r \times P) - B\_P_r \times (K \times P) = M$$

The logics over functions that work on keypair generation and exchange of those keys have been developed by python3.0 as follows.
### Keypair generation and ECDSA ###

```python
def make_keypair():
    private_key = random.randrange(1, curve.n)
    public_key = scalar_mult(private_key, curve.g)
    return private_key, public_key

print "Basepoint:\t", curve.g
aliceSecretKey, alicePublicKey = make_keypair()
bobSecretKey, bobPublicKey = make_keypair()

print "Alice's secret key:\t", aliceSecretKey
print "Alice's public key:\t", alicePublicKey
print "Bob's secret key:\t", bobSecretKey
print "Bob's public key:\t", bobPublicKey
print "================"יים
sharedSecret1 = scalar_mult(bobSecretKey,alicePublicKey)
sharedSecret2 = scalar_mult(aliceSecretKey,bobPublicKey)
print "================"
print "Alice's shared key:\t",sharedSecret1
print "Bob's shared key:\t",sharedSecret2
```

During the time of the above code execution, the key pair (private & public key) is generated and shared key is produced after exchanging the public keys as in the following output.

![Shared Key Output](http://ijrar.com/)

**Applications of ECC**

ECC is applied in key exchange applications and digital signature generation applications. These both applications are very much used in virtual currency transaction with very secured manner.

**Virtual Currency**

Virtual Currency is nothing but an electronic currency with collection of digital signatures. The best example for this technology is BitCoin, Etherium, LiteCoin, etc. The core of the technology is block chain...
cryptosystem which is spoken for how the currency has been transmitted from sender to receiver. The coin holder passes on the coin to the next by digitally signing with the hash value of the previous transaction and the public key of the next owner and adding these to the end of the coin. Each person can check the signatures to verify the chain of ownership.

Relationship of ECC & Virtual Currency Addressing

The magic of ECC is applied in Virtual Currency (bitcoin) addressing to sign transactions. Using ECC, the sender generates a random number with 256-bit value, which is his private key. The key is converted into a Base-58 form (which includes 1-9, A-Z and a-z and excludes 0, 'O', 'I' & 'l'). This is in Wallet Interchange Format (WIF). This key should not be known to anyone. Then, the sender has to produce his 512-bit public key (as discussed above)[6]. After this, it is then hashed with SHA-256 and RIPEM-160 to produce a public key hash value. This is then converted, using Base-58, into sender’s Bitcoin ID.

Conversion of Base 58 form:

The magic of ECC is applied in Virtual Currency (bitcoin) addressing to sign transactions. Using ECC, the sender generates a random number with 256-bit value, which is his private key. The key is converted into a Base-58 form (which includes 1-9, A-Z and a-z and excludes 0, 'O', 'I' & 'l'). This is in Wallet Interchange Format (WIF). This key should not be known to anyone. Then, the sender has to produce his 512-bit public key (as discussed above)[6]. After this, it is then hashed with SHA-256 and RIPEM-160 to produce a public key hash value. This is then converted, using Base-58, into sender’s Bitcoin ID.

Conversion of Base 58 form:

Ignore‘0’,'o','I','l','+

The flow for converting key into Base58 value:

1. Summation of ASCII (n)
2. Find the equivalent character for the values in R[i] to get Base58 value.

The python code for converting key into Base58 value:

```python
from hashlib import sha256
print(sha256)
# 58 character alphabet used
alphabet = '123456789ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz'
if bytes == str:
    iseq = lambda s: map(ord, s)
    bseq = lambda s: ''.join(map(chr, s))
    buffer = lambda s: s
```

Another application of ECC is signing. The sender Alice may generate a key pair, and find the hash of the message. Then he encrypts that hash code using the private key. She then sends the message and the signed hash to receiver Bob, who takes his own hash of the message, and decrypts Alice’s hashed version with her public key. If the hashes match, he has proven that Alice sent the message and that the message has not changed [2][7].

Prototype model

A prototype model is defined for verifying and signing the virtual currency transactions as in figure 3. Private key of Bob is given as an input for Base58 format. The output is in WiF (Wallet Interchange Format) Private key. The Private key of Bob is multiplied with a point in an elliptic curve to produce public key. The hash value of the public key is found and the output is passes through Base58 format. If this output is matched with WiF private key, the verifying and signing process is correct.
Conclusion

In this Paper, a Prototype model for ensuring using public key crypto system Elliptic Curve Cryptography (ECC). As ECC supports shorter key size than other public key cryptosystem, it provides high efficiency and allows for a high security. Because of smaller key size it reduces the processing over head during the signing process done in virtual currency transaction. As Python supports pre-defined libraries for cryptography implementation, this prototype model is illustrated using Python in an efficient manner authentication in virtual currency is developed.

References

3. Wei Bi*, Xiaoyun Jia, Maolin Zheng “ A Secure Multiple Elliptic Curves Digital Signature Algorithm for Blockchain” Seele Tech Corporation, San Francisco, USA
ANALYSIS ON DISTRIBUTED QUERY PROCESSING AND OPTIMIZATION TECHNIQUES

S. Jeya
Department of Computer Science, The Madura College

ABSTRACT: Query processing remains one of the important challenges of object oriented database management systems. Query optimization is the procedure of improving the evaluation strategy of the query to make the evaluation faster and more effective. Distributed query optimization is more difficult than relation query optimization since it involves more complex issues, such as fragmentation or replication of relations, additional communication costs and parallel execution. This presentation proposes techniques to efficiently handle distributed queries by replacing hash join operation by a series of semijoin operations, and to use bushy processing trees instead of linear query trees.

Keywords: Query Processing, Query Optimization, Bushy processing trees, Semijoin.

Introduction
The database query optimizer is a very important and complex module in database management systems. It receives a query optimization request with a query tree as a parameter and returns an optimized execution plan. There are several data models (e.g. distributed, object-oriented, relational, and semi-structured/XML) suitable to store information for different kinds of applications.

Query processing is classically done in two steps. The query optimizer first generates an "optimal" Query Evaluation Plan for a query using a query optimization technique. The Query Evaluation Plan is then executed within the distributed architecture using appropriate query execution techniques. Obviously, the query optimization technique and the query execution techniques are highly correlated since the former assumes the existence of the latter.

The query-processing module consists of query optimizer and query execution engine. The query optimizer maps the logical operations with the physical operation implemented with some algorithms in the query execution engine. A single logical operation is typically implemented by several alternative physical operators, for example, a query execution engine typically supports several join algorithms.

The result of the query optimizing is query execution plan, which is a sequence of physical operators such as various join and access methods implemented in the query execution engine. During the translation process, the query optimizer tries to find a plan that minimizes the chosen optimization cost, usually disk access or CPU cost.

Bushy Processing Trees
Instrumental to understanding how to optimize complex queries understands how query plans are generated. A query is compiled into a tree of operators and several different formats exist for structuring this tree of operators. The operator tree results from the macro-expansion of the query tree. Nodes represent atomic operators that implement relational algebra and the edges represent data flow.

A query plan is usually compiled into a tree of operators called a join sequence tree, where a leaf node represents an input relation and an internal node represents the resulting relation from joining two relations associated with its two child nodes.

The two major categories of processing trees are:
1. Linear trees (Left-deep and Right-deep trees)
2. Bushy trees.

If all the internal nodes of a tree have at least one leaf as child then they are called deep trees otherwise they are called bushy trees. The different formats that exist for query tree construction range from simple to complex. A simple query tree format i.e linear format is one in which the format of the tree is restricted in some manner. But, there is the danger that a restricted query plan will not be capable of representing the optimal query plan.

Left-deep trees and right-deep trees represent the two extreme options of restricted format query trees. Bushy trees, on the other hand, have no restrictions placed on their construction. Since they comprise the design space between left-deep and right-deep query trees, they have combinatorial benefits of both the strategies.
The following figures show different forms of processing trees:

![Query Tree Representations](image1)

**Fig. 1 Query Tree Representations**

![Bushy Query Tree](image2)

**Fig. 2 Bushy Query Tree**

A significant amount of research efforts has been elaborated upon developing linear execution trees, to improve the query execution time. The bushy tree join sequences, on the other hand, did not attract as much attention as sequential ones in the last decade since it was generally deemed sufficient, by many researchers, to explore only sequential trees for desired performance. This can be in part explained by the reasons that in the past time power/size of the multi processor system was limited and that the query structure used to be too simple to require further parallelizing as a bushy tree.

**Semijoin operation**

The execution of a query is represented syntactically as an annotated join tree where the internal node is a joint operation and each leaf node is a base relation. The annotations provide the details such as selection predicates, the choice of access paths, join algorithms and projection attributes of the result relation. The set of all annotated join trees for a query that is considered by the optimizer will be called the execution space of the query.

In the case of Distribute databases, data is stored in several sites (nodes), geographically or administratively across multiple systems and each site is running an independent DBMS.

![Distributed Query Processing](image3)

**Fig. 3 Distributed Query Processing**
Semijoin queries are the core of any integration service that integrates data from multiple data sources and possibly high volumes of data. The evaluation of multi-join queries faces increasing scalability concerns. Parallel processing has been applied to tackle the problem.

To handle distributed queries efficiently, Parallel Semijoin processing has been explored. With the obvious trend towards multiprocessors, attention has focused on efficiently parallelizing the join operation. Semijoins can be used to efficiently implement joins.

The semijoin acts as a size reducer such that smaller relations need to be transferred. The semijoin is beneficial if the cost to produce and send it to the other site is less than the cost of sending the whole operand relation and of doing the actual join.

Among various join methods, the semijoin has been the focus of much research effort and reported to have performance superior to that of others, particularly because it presents an opportunity for different kinds of parallelism.

Conclusion

In this paper the Bushy processing trees are used with Semijoin technique and the problem of distributed query optimization is explored. To increase the efficiency of getting optimal plan, Bushy processing trees are considered instead of Linear trees (left-deep or right-deep) and independent parallelism is combined together with Bushy trees.

To efficiently optimize the multi-join problem, Semijoin technique is adopted, which saves the memory space and execution time than any other join methodologies. Not only it improves the optimization but also it provides the better way for parallelism. Thus the above methodology ensures the increasing efficiency by surely obtaining optimal plan and applicable for distributed environment.

References
A SURVEY: PERFORMANCE EVALUATION METRICS FOR DEHAZING ALGORITHMS

Ramesh kanthan.M1 & Dr.Naganandini Sujatha.S2

1Research and Development Centre, Bharathiar University, Coimbatore
2Dept. of Computer Application, Sri Meenakshi Govt. Arts College for Women, Madurai

ABSTRACT: The paper presents a comprehensive study and evaluation of existing single image dehazing algorithms, using a large scale Benchmark consisting of both synthetic and real-world hazy images. Dehazing has emerged as a promising technology to recover the clear image from an input hazy image, such that the image quality can be significantly enhanced. However, it is still an open issue to quantitatively evaluate the performance of existing dehazing algorithms. To fill this gap, this paper provides a rich variety of criteria for dehazing algorithm evaluation, ranging from full-reference metrics, to no-reference metrics, to subjective evaluation and the novel task-driven evaluation. A hazy image is based on the physical model which is strongly related to the depth information, an effective depth estimation method that combines the geometry and edge information should be the good evaluator of the algorithm. This paper proposes evaluation methods with the comparisons and limitations of state-of-the-art dehazing algorithms, and suggests promising future directions.

Keywords: Dehazing, Detection, Dataset, Evaluations. Estimated depth maps, image quality evaluation

Introduction

The image quality is obtained through the accuracy and the visual perception. However, outdoor images captured in a bad weather and lighting condition can be degraded. To address this issue, single image dehazing algorithms pertaining extensive research efforts [1][12]. They can be classified into two main categories. They are enhancement-based approach and restoration based approach. More precisely, the enhancement-based approach attempts to enhance the visibility of hazy images without considering the reasons for image degradation, such as methods [1][3]. The restoration-based approach aims to study the physical process of the image degradation and establish the model of the hazy image.

The lack of a perfect quality haze-free image is the great challenge for this issue. Ancuti O C and Ancuti C. [6] propose an evaluation approach based on synthetic hazy images, which refers to two methods of obtaining synthetic hazy image by using professional computer software. Subsequently, the full reference (FR) evaluation indexes are used to get an evaluation result. To sum up, there is no a commonly acceptable technique for evaluating dehazing performance. In this paper, we analyze the existing approach to evaluate the performance of the dehazing algorithms based on a synthetic outdoor hazy images dataset, which is different from the existing methods.

Image Quality assessment

Images captured in bad weather suffer from low contrast and faint color. Recently, plenty of dehazing algorithms have been proposed to enhance visibility and restore color. However, there is a lack of evaluation metrics to assess the performance of these algorithms or rate them.

Among them, mean squared error (MSE), signal to noise ratio (SNR), and peak signal to noise ratio (PSNR) are the simplest and most widely used methods, but these methods are not adaptable. The weighted PSNR (WPSNR) is defined as an extension of the traditional PSNR, which takes into account the local human visual system (HVS) sensitivity.

Efforts have been made to create objective measures of quality. For many applications, a valuable quality metric correlates well with the subjective perception of quality by a human observer. Quality metrics can also track unperceived errors as they propagate through an image processing pipeline, and can be used to compare image processing algorithms.

If an image without distortion is available, you can use it as a reference to measure the quality of other images. For example, when evaluating the quality of compressed images, an uncompressed version of the image provides a useful reference. In these cases, you can use full-reference quality metrics to directly compare the target image and the reference image. If a reference image without distortion is not available. Then it can use a no-reference image quality metric instead. These metrics compute quality scores based on expected image statistics.
Full-Reference Quality Metrics

Full-reference algorithms compare the input image against a pristine reference image with no embedded reference algorithms:

- Mean-squared error (MSE). MSE measures the average squared difference between actual and ideal pixel values. This metric is simple to calculate but might not align well with the human perception of quality.
- Peak signal-to-noise ratio (PSNR). PSNR is derived from the mean square error, and indicates the ratio of the maximum pixel intensity to the power of the distortion. Like MSE, the PSNR metric is simple to calculate but might not align well with perceived quality.
- Structural Similarity (SSIM) Index. The SSIM metric combines local image structure, luminance, and contrast into a single local quality score. In this metric, *structures* are patterns of pixel intensities, especially among neighboring pixels, after normalizing for luminance and contrast. Because the human visual system is good at perceiving structure, the SSIM quality metric agrees more closely with the subjective quality score. Because structural similarity is computed locally, ssim can generate a map of quality over the image.

No-Reference Quality Metrics

No-reference algorithms use statistical features of the input image to evaluate the image quality. These no-reference algorithms include:

- Blind/Reference less Image Spatial Quality Evaluator (BRISQUE). A BRISQUE model is trained on a database of images with known distortions, and BRISQUE is limited to evaluating the quality of images with the same type of distortion. BRISQUE is *opinion-aware*, which means subjective quality scores accompany the training images.
- Natural Image Quality Evaluator (NIQE). Although a NIQE model is trained on a database of pristine images, NIQE can measure the quality of images with arbitrary distortion. NIQE is *opinion-unaware*, and does not use subjective quality scores. The tradeoff is that the NIQE score of an image might not correlate as well as the BRISQUE score with human perception of quality.
- Perception based Image Quality Evaluator (PIQE). The PIQE algorithm is opinion-unaware and *unsupervised*, which means it does not require a trained model. PIQE can measure the quality of images with arbitrary distortion and in most cases performs similar to NIQE. PIQE estimates block-wise distortion and measures the local variance of perceptibly distorted blocks to compute the quality score.

Evaluating Dehazing Methods

Since there is no specialized evaluation procedure of the dehazing techniques we searched the recent literature for an appropriate method for this task. Tarel and Hautiere [11] evaluate the quality of dehazing techniques based on a visibility restoration procedure [12]. The intermediate results of the dehazing methods are shown in figure 1. Because this procedure only applies to grayscale images and is mainly focused on finding the most visible edges, we searched for a more general method that is able to perform a pixel-wise evaluation of the dehazing process.

Image Quality Assessment (IQA)

Quality measure introduced recently by Aydin et al. [9], which compares images with radically different dynamic ranges. This metric, carefully calibrated and validated through perceptual experiments, evaluates both the contrast and the structural changes yielded by tone mapping operators. The IQA metric is sensitive to three types of structural changes: *loss of visible contrast* (green) – contrast that was visible in the reference image is lost in the transformed version, *amplification of invisible contrast* (blue) - contrast that was invisible in the reference image becomes visible in the transformed image, and *reversal of visible contrast* (red) - contrast visible in both images, but with different polarity.
As a general interpretation, contrast loss (green) has been related with image blurring, while contrast amplification (blue) and reversal (red) have been connected to image sharpening.

**Fog aware density evaluator**

Fog Aware Density Evaluator [9] (FADE) is a newly proposed contrast descriptor which predicts the visibility of a foggy scene by measuring deviations from statistical regularities in natural scene foggy and fog-free images. A lower value of FADE implies better performance of bility enhancement. The method extracts 12 fog-related features from the test image, and fits these features to a Multivariate Gaussian (MVG) model. Then, it computes the deviation using a Mahalanobis-like distance between the MVG fit of the test image and the MVG fit of a corpus of 500 fog-free images and another corpus of 500 foggy images, respectively. Each corresponding distance is defined as a foggy level $D_f$ and a fog-free level $D_{ff}$. Finally, the perceptual fog density $D$ of the test image is expressed as the ratio of the foggy level to the fog-free level.

**CNC measurement system**

The CNC measurement system (Contrast-Naturalness-Colorfulness), proposed by Guo et al, valuates the dehazing effect from three aspects: contrast $e'$, Color Naturalness Index (CNI) and Colorfulness Index (CCI). The CCI index reflects the degree of color vividness, which is defined as the summation of average saturation and standard deviation of saturation of the test image. The CNI index reflects the degree of color naturalness. It classifies pixels of test image into three classes ('skin' 'grass' and 'sky') according to their hue, then measuring the differences between the average saturation and the prototypical saturation of each class through a Gaussian function. The CNC index claims that it correlates well with human visual perception of the overall dehazing effect. However, colors of natural scene are rich and colorful, and will vary with time and light condition. It is impossible to use just three prototypical colors to represent the whole natural colors.

**The proposed Color Contrast estimation method**

The underlying principle of haze-line theory is that a haze-line in the hazy image respects a color cluster in the corresponding haze-free image, and pixels belong to the same color cluster have similar colors. Our proposed method based on the observation that deviations of different colors in a local area can reflect the local contrast of color image. The larger the color deviation of different colors in the haze free image, the higher the contrast. The proposed Color Contrast estimation method is composed of four essential steps:

1. Estimating the air light,
2. Clustering the pixels into haze-lines,
3. Computing inter cluster deviations in each local patch
4. Computing the final Color Contrast of the whole image.
Some overall remarks and empirical hypotheses made by the authors are Deep learning methods especially with the end-to-end optimization towards reconstruction loss are advantageous under traditional PSNR and SSIM metrics. The image quality metrics based on color is shown in figure 2 or few sample image results. However, the two metrics do not necessarily reflect human perceptual quality, and those models may not always generalize well on real-world hazy images. Classical prior-based methods seem to generate results favored more by human perception. That is probably because of their priors explicitly emphasized illumination, contrasts, or edge sharpness, to which human eyes are particularly sensitive. On the other hand, the typical MSE loss used in deep learning methods tend to over-smooth visual details in results, which are thus less preferred by human viewers.

Conclusion

Images captured in dull weather condition may suffer from low contrast and faint color. Recently, the research proposes plenty of dehazing algorithms to enhance visibility and restore color. However, there is a lack of evaluation metrics to assess the performance of these algorithms or rate them. In this paper, an indicator of color contrast enhancement factors are proposed basing on the existing evaluation theory. The theory assumes that colors of a haze-free image are well approximated by a few hundred distinct colors, which form tight clusters in RGB space. The presence of haze makes each color cluster forms a line, which is named haze-line. By using these haze-lines, we assess performance of dehazing algorithms designed to enhance the contrast by measuring the inter-cluster deviations between different colors of dehazed image.

References

ABSTRACT: In recent times segmentation of images is a growing and a research based field, as image segmentation is a very important part in the field of image processing and computer vision. Image segmentation is to classify or cluster an image into several parts or regions according to the feature of image, for example the pixel value. Now lots of image segmentation algorithms exist and be extensively applied in science and daily life. According to their segmentation method, we can approximately categorize them into region based, edge based segmentation and data clustering. Each one of them has its own advantages and disadvantages.

Introduction

Image segmentation is useful in many applications. It can identify the region of interest in a scene. The goal of image segmentation is to divide the image into multiple segments i.e., set of pixels, pixels in a region are similar to each other in some criteria such as color intensity or texture, so as to locate and identify objects and boundaries in an image. We categorize the existing segmentation algorithm into region based segmentation, data clustering and edge based segmentation includes the seeded and unseeded region growing algorithm. For data clustering the concept of them is based on the whole image and considers the distance between each data. The basis method of data clustering can be divided into hierarchical and partition clustering. Finally the edge based segmentation generally applies edge detection.

Region Based segmentation Methods

Region based segmentation methods are powerful tools for object detection and recognition. In this method we examine neighboring pixels and detect whether the neighbor pixel should be added or not. The common procedure is to compare one pixel with its neighbor. If a similarity criterion is satisfied, the pixel can be set belong to the cluster as one or more of its neighbor. We discuss the four algorithm in region based segmentation methods such as the Seeded region growing, the Unseeded region growing, the region splitting and merging and Fast scanning algorithm.

Seeded Region Growing (SRG)

It performs a segmentation of an image with examine the neighboring pixels of a set of points known as seed points[3]. The procedure of this method, to compute the difference of pixel value of the initial seed point and its neighboring points. If the difference is similar than the threshold then we can define the neighboring point could be classified into cluster. And recomputed the boundary of cluster and set those boundary points as new seed point [6].The following figure 1 is the result of segmentation using by the Seeded Region Growing(SRG) [6]

Unseeded Region Growing (URG)

In this segmentation procedure the seeds could be generated automatically. So this method can perform fully automatic segmentation[2]. The process initializes with cluster it containing a single image pixel. We define the set of all unsigned pixels which borders at least one of those cluster. To choose a point
and the point is less than the predefined threshold then the pixel is clustered. After that the pixel has been allocated to the cluster.

Region splitting and merging
The main goal of region splitting and merging is to distinguish the homogeneity of the image. Its concept is based on quadtrees[4]. In this method’s procedure initially a threshold is required as an input. The merging and splitting of the various region depends on this threshold value, it is done on the basis of the difference in the minimum and maximum intensities of each region. If the difference between the regions is within the threshold then those regions are merged into one single region[6]. Otherwise difference exceeds the threshold the region is split into half. This mechanism is based on a quadtree structure.

Fast Scanning Algorithm
Fast scanning algorithm which has been employed on food, sport and medical images. It scans all pixels is the image and cluster each pixel according to the upper and left neighbor pixels. This function used the gray value in the image’s pixels and the level of the image that is more the threshold is converted into intensity value between 0 and 1.

Edge based Segmentation
Edge based segmentation generally indicates the segmentation method based on the edge in an image. The simple methods apply some edge detection methods before segmentation. Edge based segmentation represents a large group of methods based on information about edges in the image. Edge based segmentations rely on edges found in an image by edge detecting operators. These edges mark image locations of discontinuities in gray level, color, texture, etc, Segmentation methods based on discontinuities find for abrupt changes in the intensity value. There are many ways to perform edge detection. The following figure.2 is one of the result of edge based segmentation [12].

Gray Histogram Technique
This technique is based on a threshold value. It is defined for the division of the image into foreground and background. The complexity of the method is in the proper selection of the threshold as the range of gray histogram is not uniform due to the presence of noise.

Gradient Based Method
Gradient refers to a generalized version of a derivative for image. Whenever there is an abrupt change in the intensity of the image near its edge[7]. There are many edge detection operators that can be used in gradient based method such as Sobel operator, canny method, Laplacian of Gaussian operator[13] and etc.,

Sobel Edge Detection
The sobel edge detection method is introduced by sobel in 1970. It proceeds the edges at those points where the gradient is highest. The sobel techniques performs a 2D spatial gradient quantity on an image and highlights region of that correspond to edges. It is used to find the absolute gradient magnitude at each point in input grayscale image[12].
Prewitt Edge Detection

The Prewitt edge detection is proposed by Prewitt in 1970. To estimate the magnitude and orientation of an edge Prewitt is a correct way. Even though different gradient edge detection wants a quiet time consuming calculation to estimate the direction from the magnitudes in the X and Y directions[12].

Kirsch Edge Detection

Kirsch edge detection in introduced by Kirsch in 1971. This technique defined by considering a single mask and rotating it to eight main directions such as North, Northwest, West, Southwest, South, Southeast, East and Northeast[8].

Rabinson Edge Detection

The Rabinson method (1977) is similar to kirsch masks but is easier to implement because they consider only on coefficients of 0, 1 and 2. The masks are symmetrical about their directional axis, the axis with the zeros[8].

Marr-Hildreth Edge Detection

The Marr-Hildreth (1980) technique is a method of detecting edges in digital images that is continuous curves whenever there are well-built and fast variations in image brightness.

LoG Edge Detection

The Laplacian of Gaussian was proposed by Marr (1982). A Laplacian of Gaussian uses a Gaussian filter to blur the images and a Laplacian to enhance edges. Edge location is done by finding Zero crossing. The Laplacian is generally used to found whether a pixel is on the dark or light side of an image.

Canny Edge Detection

Canny Edge detection technique is one of the standard edge detection techniques. It was created by John Canny in 1986. But it still outperforms many of newer algorithms that have been developed. Canny edge detector is an edge detection operator that uses a multistage algorithm to detect a wide range of edge in images. Canny method is a better method without disturbing the features of the edges in the image.

Threshold Based Segmentation

Thresholding is the simplest and most common method of image segmentation. From a gray scale image thresholding can be used to create binary images. Binary images produced from color images by segmentations. In image processing, thresholding is used to split an image into smaller segments using at least one color or gray scale value to define their boundary. The common way to convert a gray level image to a binary image is to select a single threshold value. The input to a thresholding operation is typically a gray scale or color image. In the simplest implementation, the output is a binary image representing the segmentation. Black pixels correspond to background and white pixels correspond to foreground and vice versa. This method of segmentation applies a single fixed criterion to all pixels in the image simultaneously. “Image segmentation=divide image into regions or sets of pixels.” Here the pixels are partitioned on their intensity value. There are three types of thresholding algorithms such as Global thresholding, Local thresholding and Adaptive thresholding.
Global Thresholding algorithm

The threshold has been a popular technique in many years. When the pixel values of the components and that of background are fairly consistent in their respective values over the entire image. So the global thresholding could be used, “Global Thresholding=Choose threshold T that separates object from background.

\[ G(x,y) = \begin{cases} 1 & \text{if } f(x,y) \geq T \\ 0 & \text{otherwise} \end{cases} \]

Global thresholding methods can fail when the background illumination is uneven[9].

Threshold Selection

Segmentation using thresholding technique is the choice of selecting threshold value. Automatically selected threshold value for each image by the system without human intervention is called an automatic threshold scheme[9].

Histogram based Threshold selection

The histogram based technique is dependent on the success of the estimating the threshold value that separate the two homogeneous region of the object and background of an image. So the histogram based thresholding is applied in the image[9].

EMT Technique

The threshold image by using Edge Maximization Technique (EMT) is used when there is more than one homogeneous region in image. To this reason any of the automatic threshold selection techniques performance becomes much better in images with large homogeneous and well separated regions[9]. This techniques segmentation about the maximum edge threshold in the image to start segmentation that image with the help the edge detection techniques operators.

Cluster Based Segmentation

A Cluster is a collection of pixels which are similar between them and are dissimilar to the pixels belonging to the different clusters. The clusters are formed on the basis of a certain criteria such as color, texture, size, etc.[12] The quality of the results of the clustering methods is dependent on the similarity measure and its implement.
Table 1.Comparison of image segmentation techniques and its algorithms

<table>
<thead>
<tr>
<th>Technique</th>
<th>Methods</th>
<th>Algorithms</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Based Segmentation</td>
<td>Segmentation methods based on</td>
<td>Gray Histogram Technique</td>
<td>It works well for images having good</td>
</tr>
<tr>
<td></td>
<td>discontinuities find for abrupt</td>
<td>Gradient Based method</td>
<td>contrast between regions.</td>
</tr>
<tr>
<td></td>
<td>changes in the intensity value.</td>
<td>Sobel Edge Detection</td>
<td>It is not a trivial job to produce a closed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prewitt Edge Detection</td>
<td>curve or boundary.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kirsch Edge Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Robinson Edge Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Marr-Hildreth Edge Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LoG Edge Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Canny Edge Detection</td>
<td></td>
</tr>
<tr>
<td>Threshold based Segmentation</td>
<td>The common way to convert a</td>
<td>Global thresholding</td>
<td>No need prior information of the image.</td>
</tr>
<tr>
<td></td>
<td>gray level image to a binary</td>
<td>Local thresholding</td>
<td>Easy to computation.</td>
</tr>
<tr>
<td></td>
<td>image is to select a single</td>
<td>Adaptive thresholding</td>
<td>It may be difficult to identify significant</td>
</tr>
<tr>
<td></td>
<td>threshold value.</td>
<td></td>
<td>peaks and valleys in the image.</td>
</tr>
<tr>
<td>Cluster Based Segmentation</td>
<td>A Cluster is a collection of</td>
<td>K-means clustering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pixels which are similar</td>
<td>Fuzzy clustering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>between them and are</td>
<td>Easy to Implementation</td>
<td>Result is sensitive to the selection to the</td>
</tr>
<tr>
<td></td>
<td>dissimilar to the pixels</td>
<td>Faster</td>
<td>initial random centroids.</td>
</tr>
<tr>
<td></td>
<td>belonging to the different</td>
<td></td>
<td>We cannot show the clustering details.</td>
</tr>
<tr>
<td></td>
<td>clusters.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

K-Means Clustering

K-means algorithm is an unsupervised clustering algorithm that classifies the input data points into distance from each other. In K-means algorithm data vectors are grouped into predefined number of clusters. After all the pixels are clustered, the mean of each cluster is recalculated [10].

Fuzzy Clustering

Fuzzy clustering techniques have been proposed for image segmentation to extract regular geometric regions. So they are not used for segmenting arbitrary-shaped objects. It is the most well-known algorithm as it preserves the maximum information. Fuzzy clustering method can be considered to be superior to those of their hard counter parts since they can represent the relationship between the input pattern data and cluster more naturally [11].

Conclusion

For various applications there are suitable segmentation methods that can be applied. The various segmentation methodologies applied for digital image processing is explained briefly. Though many techniques are developed, not all types are useful for all types of images. If the requirement is that the pixels
of each cluster should be linked, then region-based algorithms, especially fast scanning are preferred. Typically all techniques are used for image segmentation in various purposes. And the segmentation techniques and its various algorithms are compared in table 1. The abrupt result of image segmentation is depends on many factors, that is, pixel color, texture, intensity, image content and similarity of images. So all the segmentation methods can perform well for a particular type of image.

References

6. Yu-Hsiang Wang (王昱翔) “ Tutorial:image Segmentation” E-mail: r98942059@ntu.edu.tw Graduate Institute of Communication Engineering National Taiwan University, Taipei, Taiwan, ROC
ROLE OF WIENER FILTERING IN IMAGE ENHANCEMENT FOR COMPRESSION AND SEGMENTATION

Lydia Packiam Mettilda.M
Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Womens College, Poovanthi

ABSTRACT: In the present scenario IT plays an vital role in every fields of human survival. Due to its gigantic development in the information processing system, and the huge data base become a challenging tasks. Due to the various issues in the text processing, image processing has been emerged to provide a solution to such issues using various stages viz., image acquisition, image enhancement and image retrieval. This paper describes type of noise that can affect an image. It also evaluates the cause of noise in an image. Ever filter has its own property and it must be effective for a specific type of noise. so here we have taken wiener filter. Wiener filters, particularly appear to be very powerful for images since it is rather simple to obtain a good estimate of the noise spectrum.

Keywords: PSNR, MSE, Compression ratio (CR), Gaussian filters.

Introduction
Digital image processing is of emerging field in the area of Information Technology. It consists of various stages such as Image acquisition, Enhancement, Segmentation and recognition etc. Among them image segmentation is an important and the most difficult task. In this paper two methods are used and compared with the existing methods [1]. This work proposes a method of image segmentation and noise removal techniques with wiener filter, also known as edge preserving linear filter.

Preprocessing is the first stage and a gray scale image is given as input for noise removal [2]. Preprocessing of images commonly involves removal of noises, normalizing the intensity of the individual contents of images and enhancing the images prior to computational processing. Generally, noises will occur due to malfunctioned pixels in camera sensors, faulty memory location in hardware or error in data transmission. Hence, the wiener filter is used to preprocess the image.

The next step is to compress the filtered image using wavelets. The wavelet based image compression used a threshold technique to eliminate all the insignificant values, and transmit the significant coefficients. The haar wavelet transform method is also used for image compression. The Wavelet compression suppresses the redundant attribute of an image.

Finally, the marker controlled watershed algorithm is applied for edge detection and control over segmentation in gradient image. Using wiener filter provides the accuracy of segmentation high. The method is experimented with Berkley Segmentation Dataset and provides better results. The results show that the proposed method effectively reduce the over segmentation effect and achieve more accurate segmentation results than the existing method.

The implementation activities take part in this paper “An Adaptive Image Enhancement Using Wiener Filter With Compression and Segmentation”

Preprocessing
Preprocessing of images commonly involves in removing low frequency background noise, normalizing the intensity of the individual particles images, removing of enhancing data images prior to computational processing. Generally noise will occur due to malfunctioning pixels in camera sensors, faulty memory location in hardware or error in data transmission.

Types of noises
- Salt and pepper noise
  This kind of noise occurs due to sharp, sudden disturbances in the image signal, it is randomly scattered white or black (or both) pixels. It can be modeled by random values added to an image [3].
Gaussian Noise
Gaussian noise is an idealized form of white noise, which occurs by random fluctuations in the signal [2].

Wiener filter
The wiener function is derived from the Wiener filter techniques which is also been a type of linear filter. Applying the wiener filters in an image adaptively, tailoring itself to the local image variance. It smoothen the image at low variance. Similarly, it also smoothen the image more when the variance high. This filter provides better results compared to the linear filter. It performs well when the noise is constant-power "white" additive noise, such as Gaussian noise [4].

Mean Square Error (MSE)
MSE refers to a sort of average or sum of squares of the error between two images.

\[ MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - y(i,j))^2 \]  

Where, 
- **x** - Original image
- **y** - Segmented image
- **i** - \(i^{th}\) row pixel
- **j** - \(j^{th}\) column pixel
- **M** - Number of pixels in row
- **N** - Number of pixels in column

Peak Signal to Noise Ratio (PSNR)
Peak signal to Noise Ratio is the ratio between signal variance and reconstruction error variance. PSNR is usually expressed in Decibel scale. PSNR is used as an approximation to human perception of segmentation quality. PSNR has been accepted as a widely used quality measurement in the field of image segmentation.

\[ PSNR = 10 \log_{10} \frac{R^2}{\sqrt{MSE}} \]  

Where, **R** - Maximum fluctuation in the input image data type.

The table 1.1 shows the performance measure of PSNR and MSE value to filtering process for the experimented images.

Filtered image compression
Storage of image data occupies huge storage space and also the data transformation requires a wide channel capacity. The method of reducing the amount of data required to represent an image is called image compression. The pre processed image is compressed using wavelet. The basic idea of wavelet based image compression is to use a threshold to eliminate all the insignificant values, and only transmit the significant coefficients. There are two steps involved in image compression are listed below,

1. Compression
2. Decompression (Reconstruction)

Compressing an image is the process by which a large image size is converted to a smaller image size, such as a 10 KB image converted to a 1 KB image. Decompressing an image is the process by which a compressed image is converted back to its actual size, such as 1 KB image were converted back to a 10 KB image. Image compression can be classified as follows.

- Lossy Compression
- Lossless Compression

Lossy Compression
Lossy compression results in losing of data and the quality of an image from its original representation. It is typically associated with image files, such as JPEG,MP3 files etc. The “lossyness” of an image file may show up as jagged edges or pixilated areas. Lossy compression removes data from the original file; the resulting file occupies less disk space than the original.
Lossless Compression

Lossless compression [5] reduces a file’s size without loss of the image quality. This seems to be an efficient technique of reducing file sizes and it can be applied to both image and audio files. Lossless compression basically rewrites the data of the original file in a more efficient way. In this method, the resulting files are typically much larger than image files compressed with lossy compression. Also with lossless compression, a file may be compressed one tenth the size of the original.

Fig 1. Process of Compression & Decompression

Wavelet Compression Technique

It is one of the lossy compression techniques. A ‘wavelet’ is a small wave which has its energy concentrated in time.

Fig 2. Representation of a wave

Fig 3. Representation a wavelet

The wavelet packet compression procedure involves four steps:

Step 1: Decomposition, compute the wavelet packet decomposition of signal x at level N.
Step 2: Computation of the best tree.
   For given entropy, compute the optimal wavelet packet tree. This step is optional.
Step 3: Thresholding.
   The thresholding of wavelet packet coefficients for each packet (except for the approximation), select a threshold and apply thresholding to coefficients. The initial threshold based on balancing the amount of compression and retained energy. This threshold is a reasonable first approximation for most cases. However, in general the thresholds have to refine by trial and error so as to optimize the results to fit your particular analysis and design criteria. The experimentation with different thresholds makes it easy to alter the tradeoff between amount of compression and retained signal energy.
Step 4: Reconstruction.
   Compute wavelet packet reconstruction based on the original approximation coefficients at level N and the modified coefficients.
Wavelet-Based Image Compression is an effective Algorithm. Because of overlapping basis functions and better energy compaction property of wavelet transforms [6], Wavelet-Based Image Compression yields high quality result at low bit rate. Digital image compression techniques are normally analyzed with objective fidelity measuring metrics like Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE), Compression Ratio (CR).

**Compression Ratio (CR)**
Compression ratio is defined as the ratio between the original image size and compressed image size.

\[
\text{Compression Ratio (CR)} = \frac{\text{Original Image Size}}{\text{Compressed Image size}}
\]

**Image Segmentation**
Image segmentation [7] is a process that partitions an image into its constituent regions or objects. A digital image is composed of a finite number of elements, each of which has a particular location and value. In digital image processing, image segmentation is widely used in every field such as medical imaging, radar imaging, remote sensing imaging, traffic imaging etc. The level to which the subdivision should be stopped based on the objects of interest in an application has been isolated [8].

**Edge Detection**
Edge detectors are local image processing methods designed to detect edge pixels in an image[9]. Edge detection uses the difference in color between the background color and the foreground color. The end result is an outline of the borders. The types of edge detection techniques are Gradient Based Edge Detection. The morphological gradient of the compressed image is computed to overcome over-segmentation problem. When the morphological transition is applied to the gray scale image, it returns to high values when sudden transitions in gray level values (along the object edges) are detected, and returns to low values if neighborhood pixels are similar. So that boundaries of the catchment basin could be located on high gradient points.

**Computing Internal and External Markers**
Next step is computing markers. Direct application of watershed segmentation algorithm leads to over segmentation due to noise and other local irregularities of the gradient. An approach used to reduce over segmentation is based on the concept of markers. Marker is nothing but a connected component. There two types of markers, internal and external markers. Internal markers are associated with foreground pixels of an image. External markers are associated with background pixels of an image [10].

There is a procedure for foreground markers, which must be connected blobs of pixels inside each of the foreground objects. Here morphological techniques called opening by reconstruction and closing by reconstruction to clean up the image. These operations will create flat maxima inside each object that can be located using image opening [11].

Next the openings by reconstruction operations are performed. Following the opening with a closing can remove the dark spots and stem marks. Compare a regular morphological closing with a closing-by-reconstruction. Calculate regional maxima of reconstructed image to obtain good foreground markers.

In the cleaned up image, the dark pixels belong to the background, it starts with a thresholding operation. The background pixels are in black, but the background markers are too close to the edges of the objects try to segment.

Then thin the background by computing the skeleton by influence zones. This can be done by computing the watershed transform of the distance transform and then looking for the watershed ridge lines of the result.

**Experimental Result**
After the extraction of foreground and background markers, the image is ready for segmentation. Here watershed based segmentation [12] is applied on marker images. Dilation operation is used to superimpose the segmented object boundaries on the original image.
The table 1.1 shows the performance measure of watershed segmentation using with compression and the performance measure of watershed segmentation using without compression. The PSNR value of watershed segmentation using without compression is increased and MSE is decreased than the watershed segmentation using without compression.

The three images considered are represented as image number they are, 3096 of 481x321 dimensions, 45096 of 481x321 dimensions and 143090 of 481x321 dimensions.

Table 1.1 - Result

<table>
<thead>
<tr>
<th>S. No</th>
<th>Image No.</th>
<th>Wiener Filter</th>
<th>Watershed Segmentation with Compression</th>
<th>Watershed Segmentation without Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3096</td>
<td>0.39</td>
<td>52.2581</td>
<td>1274.68</td>
</tr>
<tr>
<td>2</td>
<td>45096</td>
<td>0.64</td>
<td>50.084</td>
<td>2586.86</td>
</tr>
<tr>
<td>3</td>
<td>143090</td>
<td>0.84</td>
<td>48.9167</td>
<td>8159.48</td>
</tr>
</tbody>
</table>

Fig 4. Resultant Images

Conclusion

In this section we have to present the results and discuss it with published theoretical values. In the results and discussion section, we present our results and discuss them by: commenting on the results obtained, interpreting what the results mean and explaining any results which are unexpected. We also need to identify any sources of error in our measurements and if possible, suggest how our experiment could have been performed more accurately.

An image is equal to thousand words. So here am going to show some of the screen shots of my two proposed methods. Segmentation subdivides an image into its constituent regions or objects. The proposed marker controlled watershed segmentation algorithm decreases the over segmentation in gradient image. The proposed methods include noise removal techniques with the wiener filter. The results show that the proposed method achieves more accurate segmentation results than the existing method.

Future Enhancement

In future this technique will be applied in medical images for diagnosis purpose and in satellite images for segmenting the objects.

- To propose a new filter for noise removal.
- To propose a novel compression model.
- To develop a new segmentation algorithm.
- To develop a common framework for an Image enhancement and segmentation.
References

ABSTRACT: The students' performance plays a crucial role in the success of a higher educational institution. With the significant increase in the number of higher educational institutions, institutions are becoming performance oriented and accordingly the goals and strategies are to be set for their achievement. Clustering in data mining helps us to analyze the voluminous data. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to the similarities. In this article we used k-means clustering to evaluate students' performance on the basis of internal tests. As we get cluster on the basis of students' mark, will help evaluate their performance and help the teacher to develop strategies to improve the students' performance for the summative examination. The students can also be prepared for the summative examination according to their learning needs.

Keywords: k-means clustering, data mining, clustering, SPMF

Introduction
Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace... etc, to increase their business efficiency. Therefore, the need for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background.

The growing part played by information technology in the expansion of society calls for a vigorous response to the challenges of the information society. The information society challenges the teaching system. In the period of technology, IT helps ample of resources to enhance the teaching skills and learning ability.

Clustering in higher education means it classifies the student by their academic performance. Lack of deep and enough knowledge in higher educational system may prevent system management to achieve quality objectives, data clustering methodology can help bridging this knowledge gaps in higher education system.

Clustering Algorithms
It can be classified into two main categories; Linear clustering algorithms and Non-linear clustering algorithms.

- Linear clustering algorithm
  - k-means clustering algorithm
  - Fuzzy c-means clustering algorithm
  - Hierarchical clustering algorithm
  - Gaussian(EM) clustering algorithm
  - Quality threshold clustering algorithm

- Non-linear clustering algorithm
  - MST based clustering algorithm
  - kernel k-means clustering algorithm
  - Density-based clustering algorithm

K-means clustering algorithm
Clustering techniques consider data tuples as objects. They partition the objects into groups, or clusters, so that objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters. Similarity is commonly defined in terms of how “close” the objects are in space, based on a distance function. The “quality” of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid.
The best number of k-clusters leading to the greatest separation (distance) is not known as a prior and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \| x_i - c_j \|^2$$

### Algorithm
1. Cluster the data into \( k \) groups where \( k \) is predefined.
2. Select \( k \) points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. A practical approach is to compare the outcomes of multiple runs with different \( k \) and choose the best one based on a predefined criterion. In general, a large \( k \) probably decreases the error but increases the risk of over fitting.

### Implementation of K-Means Algorithm
For implementing this method, we are using the marks secured by the students in the Database management system subject of I & II internal tests in the SPMF v2.35 tool.

- Step 1: Accept the number of clusters to group data and the dataset to cluster as input values
- Step 2: Initialize the first \( K \) clusters – Take first \( k \) instances or - Take Random sampling of \( k \) elements
- Step 3: Compute the arithmetic means of each cluster created in the dataset.
- Step 4: K-means assigns each record in the dataset to only one of the initial clusters – Each record is allocated to the nearest cluster using a measure of distance (e.g. Euclidean distance).
- Step 5: K-means re-assigns each record in the dataset to the most similar cluster and recalculates the arithmetic mean of all the clusters in the dataset.

### Use of K-means in Evaluation
We grouped the students according to their score in 3 different ways:
Students who have scored 11 – 15 marks are classified as “Advanced Learners”, those who have scored 6 – 10 marks are categorized as “Mediocre Learners” and 0 - 9 are categorized as “Slow Learners”.

<table>
<thead>
<tr>
<th>Category</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Learners</td>
<td>11 – 15</td>
</tr>
<tr>
<td>Mediocre Learners</td>
<td>6 – 10</td>
</tr>
<tr>
<td>Slow Learners</td>
<td>0 – 5</td>
</tr>
</tbody>
</table>

**Table 1. Category of Learners**

Clustering on the basis of the students score in the internal exam will help the teacher to evaluate and prepare them for the summative examination. The advanced learners are trained according to their level so that they will be able to score university ranks. The mediocre learners are prepared such that they will be stimulated to the next level. The Slow learners are trained according to their learning level and the students are able to stimulate to the next level i.e mediocre learners.
The Fig.1 shows the clustering output of the implementation in the SPMF v.2.35 tool. Here Cluster1 shows slow learners, Cluster2 shows mediocre learners and Cluster3 shows the advanced learners.

In the above figure, the students are clustered based on their I and II Internal tests. X-axis represents the students' score in the I Internal test and Y-axis represents the score in the II Internal test. The students in the Cluster I (Slow Learners) need more concentration and the teacher may train them with the repeated questions to boost them up to the level of pass percentage. The students in the Cluster II (Mediocre Learners) may be trained according to their level so that they will be able to score above 50%. The students in the Cluster III (Advanced Learners) are trained in such a way that they are trained to score university ranks. Thus the clustering method will help the teacher to devise strategies for improving the performance of the students according to their level.

**Results and Discussion**

After applying the above mentioned algorithm, we derived the following observations and conclusions:

The cluster I grouped the students who have scored between 0 - 5 (slow learners) and the percentage of slow learners is 10%. In Cluster II (Mediocre Learners) the students' percentage is 45 and Cluster III (Advanced Learners) percentage is also 45%.

<table>
<thead>
<tr>
<th>Category</th>
<th>Marks</th>
<th>No. of Students</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow Learners</td>
<td>0-5</td>
<td>5</td>
<td>10.20</td>
</tr>
<tr>
<td>Mediocre Learners</td>
<td>6-10</td>
<td>22</td>
<td>44.90</td>
</tr>
<tr>
<td>Advanced Learners</td>
<td>11-15</td>
<td>22</td>
<td>44.90</td>
</tr>
</tbody>
</table>

Table 2. Statistics of the data used
The results will help the teacher to prepare the students for the summative examination according to their level. Thus the results obtained will help the student as well as the teacher to improve the performance of the student.

The following are the advantages of using K-Means Clustering

1. Easy to implement
2. With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
3. K-Means may produce higher clusters than hierarchical clustering
4. An instance can change cluster (move to another cluster) when the centroids are recomputed

Conclusion

The key toward successful robust learning lies in finding the several strategies of personalization that will help both the teacher and student to improve the performance. In this paper, a simple methodology is being used to monitor and evaluate the performance of the students. This will help the teachers to plan accordingly while preparing the students for the summative examination. This paper proposes applications of k-means clustering algorithm for robust learning and personalized teaching by the teachers. The proposed models will rationally, broadly and meaningfully analyze the data and make the learning skill further flexible to the learners.

References

2. Data Mining: Introductory and Advanced topics, Margaret H. Dunham and S. Sridhar, Pearson Education Inc., 2003
3. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber and Jian Pei, Elsevier, 2012.
A SURVEY ON INTELLIGENT TRANSPORT SYSTEM: TRANSPORT MANAGEMENT THROUGH MODERN TECHNOLOGY

Ramaporkalai.T
Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women’s college, Poovanthi

ABSTRACT: With the rapid increase of modern economical and technical development, the Intelligent Transportation System becomes more and more important and essential for a country. Intelligent Transportation System applies advanced technologies of electronics, communications, computers, control and sensing and detecting in all kinds of transportation system in order to improve safety, efficiency and service. Road traffic congestion is a very big problem all over the world. So every country is actively exploring ITS technology to solve traffic problems. ITS bring significant improvement in transportation system performance, including reduced congestion and increased safety and traveler convenience. In this paper I have presented Intelligent Transport Management system through Modern Technologies.

Keywords: Intelligent Transport system (ITS), Modern Technologies.

Introduction

ITS[1] is acronym from Intelligent Transportation Systems. Under this name hides a collection of different technologies and management techniques which are used in public transport to improve efficiency of its. In addition, these systems help to protect natural environment and improve the safety of road users. The name ITS was accepted on first world congress of this subject in 1994, Paris. The system was created in response to the growing environmental awareness, protesting against the negative effects of the automotive industry and also in response to the lack of expected effects of the constantly innovating investments that did not solve the problem of traffic systems bandwidth. As it turned, ITS is system without which is hard to imagine a smooth and safe movement 2/9 of vehicles in urban areas. Years of research conducted in the American and Canadian agglomerations shows that the use of these systems reduces costs of transport infrastructure 20% to 30%. In this paper I explained Survey on Intelligent Transport Management system through Modern Technologies and comparisons between foreign countries, highlighting the advancements in this field of each one of them was conducted.

ITS Architecture

The system architecture of the ITS [2] as shown in Figure: 1 explains the data acquisition and evaluation technology, communication Network, digital mapping, video monitoring etc. This information helps in developing a system of traffic organization that enables information sharing the managers and users of traffic.
Benefits of ITS

ITS is viewed as an effort to channel technology applications to improve the performance of surface transportation systems in many ways. To this end it is expected to improve mobility (i.e., reduce congestion), reduce transportation-generated environmental impacts and energy consumption, enable improved safety, enhance quality of life including improved economic viability of communities, and increase the productivity of existing infrastructure.

Related Work

Congestion on the road is the major issue now these days. In India there are several reasons which attribute to high level of congestions [3] on the roads. The single road is shared by different types of vehicles i.e. 2-wheelers, 3-wheelers or 4 wheelers motorized or non-motorized vehicles. There by creating congestion. In developed countries the specific lane is meant for specific transport vehicle following lane discipline. So there is less chance of congestion on the roads. In the developed countries like in Japan, South Korea, Singapore, United States, Australia and United Kingdom ITS [4] had played an important role, and aided in reduction of congestion on the roads and improve traveler’s experience. While moving on the road if there is congestion with say accident, then commuters need be informed that there is congestion ahead.

Potholes are the major concern and can be seen on any road. It grows in size day by day if it is not properly maintained. P. Mohan, Eriksson et.al. [3, 5] suggested to get the position of pothole and send the information to the central server, sensors like Accelerometer, GPS and GPRS are used of the Smartphone. When the vehicle goes into the pothole, variation in the accelerometer is noted and GPS value will be recorded on that particular location. The whole information is sent to the central server via GPRS where the data is analyzed and stored in the database.

Too much honking on the road means there is congestion ahead. R. Sen, S. Roy et al. [6, 7] suggested to find the level of congestion on the road. Two microphones are implemented on the road side to detect the honks.

The road chaotic conditions are due to congestion which can be happened at any time. The passengers on the bus stop have to wait for long time. J.Biagioni et al. [8] solution predicted the location and arrival time of the bus. The experiment is done by setting an application on bus so that tracking of the bus is easy.

The regular use of the GPS over a long time may drain the battery at the faster rate. A.Thiagarajan et al. [9] suggested that GPS signals are weak and will not be able to work when it inside pocket, inside house or inside tunnel. There is an alternative to GPS is to use Wi-Fi. Using the Wi-Fi there are some benefits like the battery will not drain quickly as GPS and another reason is, it will give us the approximate location up to 40 meters of radius.

Using GPS we can find out the location but we cannot find the history of the place or navigate the vehicle on the road with real time traffic. P. Angin et al. [10] suggested using camera sensor present in the Smartphone to get navigated and assisted in the real time.

Live feeds will be sent to the user for which he is subscribed to. J. Chen et al. [11] suggested for getting the live feeds from the central server the user must the first subscribed to the service. Like using the services of Face book or Twitter, the user should be subscribed to the services first, after subscription he can easily publish or subscribe the post. The user can obtain information without regard of their location or source.

V. Jacobson et al. [12] suggested for publishing and subscribing services the model NDN is concerned. The model NDN (Named Data Networking) is based on Content-Centric Network (CCN) [17] where the data is stored centrally.

The data flow to each and every hardware that is router or the switches, communicate using XML language. W. Fenner, M.Ott et.al. [13, 14] suggested that, to query for accessing the data from the database of the subscription should be done using XML.

The central server will be overloaded when more and more queries from the user arrives. With this the response time is slow and thus the user has to wait. T. Hunter et al. [15] suggested that instead to use single server, the whole architecture should be shifted to cloud infrastructure.

The traffic on the road changes dramatically every minute thereby increases the incidents. Wang et al [16] suggested a hybrid approach for automatic incident detection.
Bret Hull et al. [17], proposed a computing system which is designed to collect, process and deliver and visualize data from the sensors located in smart phones. The data from the sensors sent to the central server wirelessly.

Technologies for ITS

ITS enabling many technologies to improve transportation conditions, safety and services. Some of them are listed.

Global Positioning System

The Global Positioning System (GPS) is the most common and accessible technique for vehicle localization. However, conventional localization techniques which mostly rely on GPS technology are not able to provide reliable positioning accuracy in all situations. This paper presents an integrated localization algorithm that exploits all possible data from different resources including GPS, radio-frequency identification, vehicle-to-vehicle and vehicle-to-infrastructure communications, and dead reckoning. A localization algorithm is also introduced which only utilizes those resources that are most useful when several resources are available. A close-to-real-world scenario has been developed to evaluate the performance of the proposed algorithms under different situations [18]. Several countries, notably Holland and Germany, are using satellite-based GPS devices to record miles traveled by automobiles [19].

Dedicated Short Range Communication (DSRC)

DSRC is dramatically increased for enhancing the road safety applications. The main task of DSRC is to protect the vehicles by communicating the warning message regarding the vehicle changing conditions, traffic occurrence and dangers over the road in the network. So, it is necessary to maintain the accurate communication timely with high reliability by implementing the appropriate protocol [9]. At present, DSRC systems in Europe, Japan, and the United States are generally not compatible.

Wireless Networks

Similar to technology commonly used for wireless Internet access, wireless networks allow rapid communications between vehicles and the roadside, but have a range of only a few hundred meters. However, this range can be extended by each successive vehicle or roadside node passing information onto the next vehicle or node. South Korea is increasingly using WiBro, based on WiMAX technology, as the wireless communications infrastructure to transmit traffic and public transit information throughout its transportation network.

Mobile Telephony

ITS applications can transmit information over standard third or fourth generation (3G or 4G) mobile telephone networks. Advantages of mobile networks include wide availability in towns and along major roads. However additional network capacity may be required if vehicles are fitted with this technology, and network operators might need to cover these costs. Mobile telephony may not be suitable for some safety-critical ITS applications since it may be too slow.

Radio wave or Infrared Beacons

Japan's Vehicle Information Communications System (VICS) uses radio wave beacons on expressways and infrared beacons on trunk and arterial roadways to communicate real-time traffic information. VICS uses 5.8GHz DSRC wireless technology.

Roadside Camera Recognition

Camera- or tag-based schemes can be used for zone-based congestion charging systems (as in London), or for charging on specific roads. Such systems use cameras placed on roadways where drivers enter and exit congestion zones. The cameras use Automatic License Plate Recognition (ALPR), based on Optical Character Recognition (OCR) technology, to identify vehicle license plates; this information is passed digitally to back office servers, which assess and post charges to drivers for their use of roadways within the congestion zone.

Probe Vehicles or Devices

Several countries deploy so-called “probe vehicles” that report their speed and location to a central traffic operations management center, where probe data is aggregated to generate an area-wide picture of
Sensing Technologies

Sensing Technology in transportation system is an unprecedented technology. In a technology the embedded sensors work, wireless sensor nodes are normally low-cost and low power, data processing and wireless communication capabilities examples are microchip, RFID etc. The WSNs (Wireless sensor networks) have a large number of sensor nodes represent a significant efficiency over traditional sensors. Sensor technologies designs for dissimilar image scenarios in intelligent transportation systems applications for instant car communication with each other traffic condition monitoring. These applications run on real time; therefore, end-to-end delay and synchronization is critical for such systems. In WSN the sensors or nodes, which are deployed for data gathering and one or more sink nodes connected through different long range connections i.e. satellite, Wi-Fi, WiMAX, etc.

Conclusion

In the growing need of the transportation systems the use of technology plays very crucial role to make it more affordable and efficient. The developments in the ITS modules using advanced tools made it more practical and adaptable in this environments.

The new developments in sensor technology and communication technology boosted the growth in ITS implementation. The use of these ITS tools according to the need of the environment and regional factors is a crucial decision making factor. Particularly in developing countries where ITS has different challenges to be faced in executing it.

There will be future scope of study to explore the most appropriate ITS modules which may lead to effective tackling for traffic related issues in developing countries.

References


15. Timothy Hunter, Teodor Moldovan, Matei Zaharia, Samy Merzgui, Justin Ma, Michael J. Franklin, Pieter Abbeel, Alexandre M. Bayen in Scaling the Mobile Millennium System in the Cloud. SOCC’11, 2011.


ABSTRACT: Exchange of data over the internet is increasing day by day. Information Security has become an important issue in modern world as the popularity and infiltration of internet commerce and communication technologies has emerged, making them a prospective medium to the security threats. To surmount these security threats modern data communications uses cryptography an effective, efficient and essential component for secure transmission of information by implementing security parameter counting Confidentiality, Authentication, accountability, and accuracy. In this paper, Secure Symmetric Key algorithm using modulo 69 is proposed. Here not only alphabets and numbers are used, but special characters have also been included. This study's main goal is to reflect the importance of security in network and provide the better encryption technique for currently implemented encryption techniques in simple and powerful method. Three keys are used in which one is a natural number which is relatively prime to 69 and finding the inverse modulo 69 of it and the other key is a random number generated by the proposed key generation method. The proposed a new algorithm is used for Encryption and Decryption.

Keywords: Cryptography, Symmetric Key, Encryption, Decryption, Key Generation

Introduction
Cryptography [1] is the science of devising methods that allow information to be sent in a secure form in such a way that the only person able to retrieve this information is the intended recipient. The highly use of networking leads to the data exchange over the network while communicating to one and another system. While communication it is very important to encrypt the message so that intruder cannot read the message. Network security is highly based on cryptography. Cryptography is an art of hiding information by encrypting the message using algorithms. The cryptography system is a system which performs encryption and decryption process. The encryption process takes plain text as input and produce an output called cipher text using keys. The decryption process performs same as encryption but in reverse order. Cryptography algorithm mainly falls under two categories i.e. Asymmetric and Symmetric encryption techniques. A plain text is encrypted using an algorithm called “encryption algorithm” and it is decrypted using an algorithm called “decryption algorithm”. A key is used at the time of encryption and decryption process. The security level of cryptography is determined by the key space (size of key).

In secret key cryptography, a single key is used for both encryption and decryption. The sender uses the key to encrypt the plaintext and sends the cipher text to the receiver. The receiver applies the same key to decrypt the message and recover the plaintext. Because a single key is used for both functions, secret key cryptography is also called symmetric encryption. With this form of cryptography, it is obvious that the key must be known to both the sender and the receiver; that, in fact, is the secret.

Related work
Encryption/Decryption has become a key component in any business competitive strategy. Organizations are gaining opportunities and benefits such as global presence and improved competitiveness from web-based security. But now a day’s hacking has become a common practice in society which made such cryptographic algorithms no longer safe. In this paper we have studied number of such symmetric key algorithms and selected one of them for reference in the proposed algorithm.

SKED – Secure Key Encryption Decryption Algorithm Using Modulo 69 BY Dr.E.George Dharma Prakash Raj, G.Muthu Kumar(July-2014),IJsrcsams.[2]

In this paper, a new Symmetric Key algorithm called as SKED (Secure Key Encryption Decryption) using modulo 69 is proposed. Here not only alphabets and numbers are used, but special characters have also been included. This study’s main goal is to reflect the importance of security in network and provide the better encryption technique for currently implemented encryption techniques in simple and powerful method. Three keys are used in which one is a natural number which is relatively prime to 69 and finding
the inverse modulo69 of it and the other key is a random number generated by the proposed key generation method. The proposed a new algorithm is used for Encryption and Decryption.

“Implementation Of Security Through Simple Symmetric Key Algorithm Based On Modulo 37” by Dr.Saeed Q Y Al-Khalidi, Prakash Kuppuswamy, (Oct-2012).[3]

It uses two keys: k1 and K2. K1 for positive number and k2 for negative number, find the inverse of selective number using modulo 37. The value is assigned for message A=1,B=2,...,Z=26, 0=27,...,9=36, Space=37. For Encryption: Calculate the assigned value with modulo37 CT=(integer value*k1)mod37, CT1=(CT*k2)mod37=Cipher Text. For Decryption: (CT1*k1’*k2’) mod37, Alphabets and numbers have been used in this algorithm.


It uses two keys: K2 is generated through user entered key length calculation using Modulo 69. K1 is natural number, inverse of natural number is calculated using modulo 69. Assigning synthetic value for message including special characters. A=1,B=2,...,Z=26, 0=27,...,9=36, Space=3 7,!=38,...,~=69. For Encryption: Calculate with modulo69. C1=V1*K1, C2=C1+K2, C2mod69=Cipher Text. For Decryption: P1=V2-K2, P2=P1*N1, P2mod69=Plain text. In this algorithm alphabets and special characters used.

“Secure Symmetric Key Algorithm [SSKA] For Information Exchange Using Modulo 67” BY Mr.H.Mohamed Ashik, Mrs.S.Vydehi (Feb 2014).[5]

It uses three keys: K1, K2and K3. K1 is generated by user entered key is converted into Ascii value and then ascii value is converted into binary value then this original binary value is shifted 1 bit and produced new binary value then original binary value and new binary value is XOR ed, so we can get new shifted binary value then the value is calculated with modulo67. K2 is selecting a natural number say, n1. Find the Inverse of the number using modulo67. say K1.K3 is selecting a natural number say, n2. Find the Inverse of the number using modulo67. say K2.A=1,B=2,...,Z=26, 0=27,...,9=36, Space=3 7,!=38,...,~=69. For Encryption: Calculate the assigned value with modulo67. P1=(integer value) , P2=PT1*N1, P3=PT2*N2, PT4=PT3*K1, CT=PT4mod67. For Decryption: CT1=integer value, CT2=K1-CT1, CT3=K3*-1, CT4mod67=Plain text. In this algorithm alphabets and special characters used.

“RDA Algorithms: Symmetric Key Algorithm” By Dinesh Goyal, Vishal Srivastava, 2012,[6]

Proposed a new model by combining the Vignere Cipher Model and ECB (Electronic Code Book) VignereCipherModel: -Encryption: Ci=P+i*Ki (mod 26), Decryption: Pi=Ci+Ki (mod 26) RDA Algorithm: Encryption: Ci=Pi+Ki (mod 256), Decryption: Pi= Ci+Ki (mod 256). Use a dynamic key pattern by shifting the key matrix of times for m ‘rounds’ during the process of encryption and decryption so that every time new key is used. Key length is 1024

“Multiphase Encryption: A New Concept in Modern Cryptography” By Himanshu Gupta, Vinod Kumar Sharma (Aug-2013).[7]


“Cryptography: a Mathematical Approach” by P. M. Durai Raj Vincent, Syed Amber Iqbal (Dec-2013).[8]

Keys used for encryption are (x, y, z) where x = ai % 26, y = bi % 26, z = ci. Keys used for decryption are (p, q, r) where p = ak % 26, q = bk% 26, r=ck. For each letter in the original message, obtain the equivalent decimal number Ni and perform the following operation: Ci = a* (Ni)2 + b* (Ni) + c ; where a, b ,c are the keys used for encryption. Let C be the summation of {C1, C2 , C3 , ... , Cn} decimal numbers and E be the summation of {E1, E2, E3, ..., En} decimal numbers of the cipher text. Then, Hash Code = Log (C + E)/Log (d) where d = a + b + c.

“Encryption and Decryption of Data Using QR Authentication System” By Atul Hole, Mangesh Jadhav (Apr-2014).[9]

The information which is encrypted are entered inside the QR code and QR code will be also printed with the original data of document.We use MASS encryption algorithm, which was combination of three different cryptographic modules: generalized modified Vernam cipher, KJS and DMASS, for the encryption purpose of data in the QR Code. After encrypting the data, we embed the data in the QR Code using a set of different protocols and ultimately generate the encrypted QR Code.

“Network Security Using A Symmetric Key Cryptography” by Mrinmoy Saha, Mr. Tarak Nandy (Feb-2014).[10]

A transposition cipher is a method of encryption by which the positions held by units of plaintext are shifted. Mathematically an objective function is used on the characters positions to encrypt and an
To encrypt a message using this technique we have to write a line in zigzag way across the page, and then read of each row. The decryption process for the Rail Fence Cipher involves reconstructing the diagonal grid used to encrypt the message.

Proposed Work
In this paper, propose a new algorithm which follows the symmetric key mechanism. Symmetric key cryptography is an encryption system in which the sender and receiver of a message share a single, common key that is used to encrypt and decrypt the message. To propose a new symmetric key algorithm and a new key generation method. The proposed algorithm is used for encryption and decryption process, using modulo69 and inverse modulo69. This algorithm is used for encryption and decryption, in which the same key is used both for encryption as well as decryption.

A message may consist of alphabets from A-Z, numbers from 0-9 and special characters such as +,-,%,< and so on. In this algorithm, firstly assign integer values for each letters, digits and special characters. Alphabet ‘A’ is assign with integer value ‘1’, B=2, C=3...so on till Z=26. Next we assign the integer value 0=27, 1=28,...so on till 9=36. Space=37,! =38, "=39,...so on as shown below. The second part is the key generation process. By using the proposed key generation method we generate the second key i.e., K2. The third part is the key generation process. By using the proposed key generation method we generate the second key i.e., K3. The first key i.e., K1 is an integer value taken from the user. The inverse modulo69 of K1 is generated and stored in ‘n1’. Hence using the keys K1, K2, K3 and ‘n1’, the encryption and decryption process is carried out.

Key Generation
Here three keys will be used k1, k2 and k3. The first key i.e., k1, can be a natural number and k2 will be derived from the key entered by the user which can be a combination of characters, numbers and special characters. And m=69.

a. Generating ‘k2’, firstly user enters a key. The length of the key is stored in 'kl'. Hence k2 is generated as follows: K2=(\sum_{i=0}^{kl-1} 2^i * kl * val) mod 69
b. Generating ‘k3’, firstly user enters a key. The length of the key is stored in 'kl'. Hence k3 is generated as follows: K3=(\sum_{i=0}^{kl-1} 3^i * kl * val) mod 69 Where, 'i' is the position of each character in the key. 'kl' is the length of the key. 'val' is the integer value that has been assign to each character as shown in table 1 above.
c. For ‘k1’, select any natural number say ‘k1’ where k1≠0 and must be relatively prime to d. 'm'(i.e., 'k1' should not have factors in common with 'm').
e. Find inverse modulo69 of ‘k1’ and store it in ‘n1’

Encryption Algorithm
for(i=0;i<message.length;i++)
{
    num1=hm.get(str.charAt(i));
    n1=(num1*key1);
    n2=(n1+k2);
    n3=(n2-k3)%69;
    for(Map.Entry < Character, String > entry : hm1.entrySet())
    {
        str1=""+n3;
    if(entry.getValue().equals(str1))
    {
        ch=entry.getKey();
        encryptedmsg=encryptedmsg+ch;
    }
    }
}

Steps
1. Firstly substitute or assign integer value for plain text.(v1)
2. Multiply Synthetic value with first key k1. (c1=v1*k1)
3. Now add the result from step2 above with second key k2. (c2=c1+k2)
4. Subtract ‘k3’ from the result. (c3 = c2 - k3)
5. Then calculate with modulo 69. (c3 mod 69)
6. Get Cipher text values.

**Decryption Algorithm**

```
// Algorithm for Decryption
for(int m=0;m<encmsglen;m++)
{
    num1=hm.get(str3.charAt(m));
    num2=num1+k3;
    num3=((num2-k2)*v)%69;
    for (Map.Entry<Character, String> entry : hm1.entrySet())
    {
        str4=""+num3;
        if(entry.getValue().equals(str4))
        {
            ch1=entry.getKey();
            decryptedmsg= decryptedmsg +ch1;
        }
    }
}
```

**Steps**
1. Assign integer value for cipher text (v2)
2. Now add the k3 with integer value. (p1 = v2 + k3)
3. Subtract ‘k2’ from step2 above the result. (p2 = p1 - k2)
4. Multiply above result with inverse modulo 69 of ‘k1’ i.e., ‘n1’. (p3 = p2 * n1)
5. Finally calculate with modulo 69. (p3 mod 69)
6. Get Plain text value.

**Implementation**

In this paper, Symmetric key encryption method is used. The symmetric encryption approach is divide in two type one is block cipher symmetric cryptography technique and another is stream cipher symmetric cryptography but here we are choosing stream cipher type because its efficiency and security. Basically private key concept is the symmetric key concepts where plain text is converting into encrypted text known as cipher text using private key where cipher text decrypted by same private key into plain text. The encryption key is trivially related to the decryption key, in that they may be identical or there is a simple transform to go between the three keys.

**Key Generation Process**
1. Let us suppose key enter by user is as follows: Key = HAI%
   With positions, i = 0 1 2 3, Key Length, kl = 4
2. Hence, K2 = \(\sum_{i=0}^{k-1} 2^i \ast kl \ast val\) mod 69
   = \{(2^0 \ast 4 \ast 8) + (2^1 \ast 4 \ast 1) + (2^2 \ast 4 \ast 9) + (2^3 \ast 4 \ast 42)\}
   = 24 + 8 + 144 + 1344
   = 1520 mod 69
   = 2.
3. Hence,
   K3 = \(\sum_{i=0}^{k-1} 3^i \ast kl \ast val\) mod 69
   = \{(3^0 \ast 4 \ast 8) + (3^1 \ast 4 \ast 1) + (3^2 \ast 4 \ast 9) + (3^3 \ast 4 \ast 42)\}
   = 32 + 12 + 324 + 4536
   = 4904 mod 69
   = 5.
4. Now select a natural number say, K1 = 4 which is relatively prime to 69.
5. Find the inverse of k1 denoted by n1 = 14 (Verification: 4*14 modulo 69 = 1)
Encryption Process
The process is converting plain text into cipher text. Encryption techniques are used to send secret message through an insecure channel. It requires an encryption algorithm and a key.
Let, Plain Text = BHARATHIDASAN
Each character in the plain text is assign with integer value as discussed above. The encryption process is using keys K1, K2 and K3. Hence for the Plain Text = BHARATHIDASAN
Cipher Text= 9{4*4> {B) 4/40

Decryption Process
It is reverse process of encryption where it converts text into plain text. Decryption takes place at receiver side to obtain the original message from non-readable message. It requires decryption algorithm and a key. Now using K3, K2 and n1 (which is inverse mod69 of k1), the cipher text is decrypted.

Conclusion
Secure Symmetric Key Algorithm is used to protect data from attackers. This algorithm used inverse modulo69 function and generated a key using the proposed key generation method. The important thing of proposed method is that it is almost impossible to break the encryption algorithm without knowing the exact key value. Cryptography is used to achieve few goals like Confidentiality, Data integrity, Authentication etc., in this algorithm three keys are used to encrypt and decrypt the data. So, the data is more secure than other existing algorithm. New symmetric algorithm is very authoritative and straightforward and more secure. For large amount of data transaction and commercial communication purpose this algorithm will work very smoothly.

References
2. SKED – Secure Key Encryption Decryption Algorithm Using Modulo 69 BY Dr.E.George Dharma Prakash Raj, G.Muthu Kumar(July-2014), IJsrcsams.
4. “KED - A Symmetric Key Algorithm for Secured Information Exchange Using Modulo 69 ”By Dr.E.George Dharma Prakash Raj, JanailinWarjri (Aug-2013), IJCNIS.
5. “Secure Symmetric Key Algorithm [SSKA] For Information Exchange Using Modulo 67” BY Mr.H.Mohamed Ashik, Mrs.S.Vydehi (Feb 2014), IJESRT
7. “Multiphase Encryption: A New Concept in Modern Cryptography” By Himanshu Gupta, Vinod Kumar Sharma (Aug-2013), IJCTE.
10. “Network Security Using A Symmetric Key Cryptography” by Mrinmoy Saha, Mr. Tarak Nandy (Feb-2014), IJETTCS.
11. “A Comparative Analysis on Symmetric Key Encryption Algorithms” By G. Muthukumar, Dr. E. George Dharma Prakash Raj (Feb-2014). IJARCET.
SOFTWARE TESTING TECHNIQUES

Sulochana .C
Department of Computer Science, Madurai Sivakasi Nadars Pioneer Meenakshi Women’s college, Poovanthi

ABSTRACT: The importance of software testing and its impact on software cannot be underestimated. Software testing is a fundamental component of software quality assurance and represents a review of specification, design and coding. The greater visibility of software systems and the cost associated with software failure are motivating factors for planning, through testing. It is not uncommon for a software organization to spent 40% of its effort on testing. During testing the software engineering produces a series of test cases that are used to “rip apart” the software they have produced. Testing is the one step in the software process that can be seen by the developer as destructive instead of constructive. Software engineers are typically constructive people and testing requires them to overcome preconceived concepts of correctness and deal with conflicts when errors are identified.

Keywords: Quality assurance, preconceived, rip apart

Introduction
A strategy for software testing integrates software test case design techniques into a well-planned set of steps that cause the production of software. A software test strategy provides a road map for the software developer, the quality assurance organization, and the customer. Any testing strategy needs to include test planning, test case design, test execution, and the resultant data collection evaluation. A software test strategy should be flexible enough to promote the creativity and customization that are required to adequately test all large software-based systems.

Strategic Approach To Software Testing
Testing is a group of activities that can be planned in advance and performed systematically. For this reason a set of stages that we can place particular tests case design techniques and test approaches should be developed for the software engineering procedure. A number of testing strategies have been identified, which provide a template for testing and all have the following features:

- Testing starts at the modular level and works outward towards the integration of the complete system.
- Diverse testing techniques are appropriate at diverse points in time.
- Testing is performed by the developer of the software and an independent test group.
- Testing and debugging are diverse activities, but debugging must be included in any testing strategy.

A strategy for testing must include low-level tests that are required to verify that a small source code segment has been implemented correctly as well as high-level tests that that validate major system functions based on customer requirements.

A software testing strategy
The software engineering procedure can be seen as a spiral. Initially the systems engineering states the role of the software and lead the software requirement analysis, where the information domain, function, behaviour, performance and validation criteria for the software are identified. Moving inwards along the spiral, we come to design and finally coding.

A strategy for software testing may be to move upward along the spiral. Unit testing happens at the vortex of the spiral and concentrates on each unit of the software as implemented by the source code. Testing happens upwards along the spiral to integration testing, where the focus is on design and the production of the software architecture. Finally we perform system testing, where software and other system elements are tested together.

Criteria for Completion Testing
A fundamental question in software testing is how do we know when testing is complete. Software engineers need to have rigorous criteria for establishing when testing is complete. Musa and Ackerman put forward an approach based on statistical response that states that we can predict how long a program will...
go before failing with a stated probability using a certain model. Using statistical modeling and software reliability theory, models of software failure as a test of execution time can be produced. A version of failure model, known as logarithmic Poisson execution-time model, takes the form \( f(t) = \frac{1}{p} \ln[l_0pt + 1] \) where

\[ f(t) = \text{cumulative number of failures that are anticipated to happen once the software has been tested for a particular amount of execution time t} \]

\[ l_0 = \text{the initial failure intensity at the start of testing} \]

\[ p = \text{the exponential reduction in failure intensity as errors are discovered and repairs produced.} \]

The instantaneous failure intensity, \( l(t) \) can be derived by taking the derivative of \( f(t) \):

\[ (a) \quad l(t) = \frac{l_0}{l_0pt + 1} \]

Using the relationship noted in equation (a), testers can estimate the drop off of errors as testing progresses. The actual error intensity can be plotted against the estimated curve. If the actual data gained during testing and the Logarithmic Poisson execution-time model are reasonably close to another over a number of data points, the model can be used to estimate the total test time required to produce acceptably low failure intensity.

**Methodology**

![Fig.1 Block Diagram](image)

**TRT** - Test Case Reduction Technique

**CTG** - Common Test Case Generation

**IPTI** - Individual Path Case Identifier

**PTE** - Parallel Test Case Executor

**Dynamic Domain Reduction (DDR)**

DDR is the technique that creates a set of values that executes a specific path. It transforms source code to a Control Flow Graph (CFG). A CFG is a directed graph that represents the control structure of the program. Each node in the graph is a basic block, a junction, or a decision node.

**Test Case Reduction Technique**

DDR uses the Get Split algorithm to find a split point to divide the domain. In the dynamic domain reduction procedure, loops are handled dynamically instead of finding all possible paths. The procedure exits the loop and continues traversing the path on the node after the loop. This eliminates the need for loop unrolling, which allows more realistic programs to be handled.

For a larger the input domain, the more exhaustive the testing would be. To avoid this problem, a minimum set of test cases needs to be created using an algorithm to select a subset that represents the entire input domain and gives maximum coverage. In addition, particularly for regression testing where every change in the program demands repeat testing, the testing itself would take longer to run, when test cases are larger. Therefore, reducing number of the test cases does have advantage for the regression testers.
Get split algorithm: Getsplit (LeftDom, RightDom, SrchIndx)

Precondition: LeftDom and RightDom are initialized appropriately And SrchIndx is one more than the last time Getsplit was called with these domains for this expression.


Input
- LeftDom: Left expr's domain with Bot and Top values
- RightDom: right expr's domain with Bot and Top values

Output
- Split: a value the divides a domain of values into two sub domains.

BEGIN

--- Compute the current search point ---
- srchPt = (1/2, 1/4, 3/4, 1/8, 3/8 ...)
- Try to equally split the left and right expression's

ELSE
END IF
RETURN split
END GetSplit

Parallel Execution method
This method consists of three methods:
1. Common test case generation:
2. Individual path test case identifier
3. Parallel test case executor

Here common test case generation method used to identify the interval from all independent paths and divide the interval based on number of common interval used in every individual path. Individual path test case identifier generates some unique value to each interval on independent path. In the parallel executor, each independent path executed by separate processor.

Each processor parallely fetches the resource from resource pool if resource is available. If resource is not available at a time of fetching the resources then processor go to waiting state. Once processor completing their task then resource is released and then transfer to resource pool.

Evaluation
A comparative evaluation has been made between the proposed technique and the Existing technique. The following areas are used to compare with existing techniques:
1) Number of test cases
2) Reduction percentage of test cases
3) Compilation time

The evaluation is described using one example.

Example:
The function value takes three marks as input such as mark1, Mark2, mark3 and returns some total mark for student depending upon the performance.

Source code
int value(mark1,mark2,mark3)
{
    int total;
    Total=0;
    If(mark1<mark2)
    {
        Mark3=16;
        If (mark1<mark3)
Mark3=10;
Total=mark1+10;
Else
Total=mark1+5;
Else
{
mark3=16;
total=mark1+mark2+mark3;
}
return (total); }

Conclusion

Limitation of the proposed technique is not applicable where there is no common variable in the independent path. The new proposed technique that achieved greater reduction percentage of the test cases and execution time by using parallelism. Furthermore, for compilation, it has been found that the new technique is the least time-consuming among the existing technique. Based on the analysis done, the proposed method can be considered a superior technique from all others available in current literatures.

References

7. Offutt A. Jefferson, J. Pan and J. M. Voas "Procedures for reducing the size of coverage based test".
ABSTRACT: Text mining should be useful for anticipating new technologies and new uses for existing technologies, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature. The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically. Perhaps for this reason, there has been little work in text data mining to date, and most people who have talked about it have either conflated it with information access or have not made use of text directly to discover therefore unknown information. In this paper we will first define data mining, information access, and corpus-based computational linguistics, and then discuss the relationship of these to text data mining. I describe examples of what I consider to be real text data mining efforts and briefly outline our recent ideas about how to pursue exploratory data analysis over text.

Keywords: Data mining, Text data mining (TDM), Data mining process.

Introduction

Technological innovation often proceeds by applying advances made in one field to a separate arena. Once the innovation is implemented, the transfer of knowledge may appear obvious or even inevitable, but without the benefit of hindsight it is surprisingly difficult to identify. Specific technologies that are ripe for transfer [1]. The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM), also known as knowledge discovery in databases, and information archeology.

Additionally, there are some disagreements about what actually constitutes data mining. It turns out that ``mining'' is not a very good metaphor for what people in the field actually do. Mining implies extracting precious nuggets of ore from otherwise worthless rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets [2], usually for the purposes of decision making. Part of what I wish to argue here is that in the case of text, it can be interesting to take the mining-nuggets metaphor seriously.

TDM Vs Information Access

This led us to revise the query, to assess whether one or more of these packaging technologies might plausibly. It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known). The goal of information access is to help users find documents that satisfy their information needs. The standard procedure is akin to looking for needles in a needle stack - the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information.

The problem is one of homing in on what is currently of interest to the user. As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested does not imply that a new discovery has been made: the information had to have already been known to the author of the text; otherwise the author could not have written it down. I have observed that many people, when asked about text data mining, assume it should have something to do with ``making things easier to find on the web''. For example, the description of the KDD-97 panel on Data Mining and the Web stated. Two challenges are predominant for data mining on the Web [3].

The first goal is to help users in finding useful information on the Web and in discovering knowledge about a domain that is represented by a collection of Web-documents. The second goal is to analyze the transactions run in a Web-based system, be it to optimize the system or to find information about the clients using the system. This search-centric view misses the point that we might actually want to
treat the information in the web as a large knowledge base from which we can extract new, never-before encountered information.

On the other hand, the results of certain types of text processing can yield tools that indirectly aid in the information access process. Examples include text clustering to create thematic overviews of text collections automatically generating term associations to aid in query expansion and using co-citation analysis to find general topics within a collection or identify central web pages. Aside from providing tools to aid in the standard information access process, I think text data mining can contribute along another dimension. In future I hope to see information access systems supplemented with tools for exploratory data analysis.

**TDM and Computational Linguistics**

If we extrapolate from data mining (as practiced) on numerical data to data mining from text collections, we discover that there already exists a field engaged in text data mining: corpus-based computational linguistics. Empirical computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within natural language processing, such as part-of-speech tagging, word sense disambiguation, and bilingual dictionary creation.

It is certainly of interest to a computational linguist that the words "prices, prescription, and patent" are highly likely to co-occur with the medical sense of "drug" while "abuse, paraphernalia, and illicit" are likely to co-occur with the illegal drug sense of this word. However, the kinds of patterns found and used in computational linguistics are not likely to be what the general business community hopes for when they use the term text data mining.

Within the computational linguistics framework, efforts in automatic augmentation of existing lexical structures seem to fit the data-mining-as-ore-extraction metaphor. Examples include automatic augmentation of Word Net relations by identifying lexico-syntactic patterns that unambiguously indicate those relations, and automatic acquisition of sub categorization data from large text corpora. However, these serve the specific needs of computational linguistics and are not applicable to a broader audience.

**TDM and Category Metadata**

Some researchers have claimed that text categorization should be considered text data mining. Although analogies can be found in the data mining literature (e.g., referring to classification of astronomical phenomena as data mining), I believe when applied to text categorization this is a misnomer. Text categorization is a boiling down of the specific content of a document into one (or more) of a set of pre-defined labels. This does not lead to discovery of new information; presumably the person who wrote the document knew what it was about. Rather, it produces a compact summary of something that is already known.

However, there are two recent areas of inquiry that make use of text categorization and do seem to fit within the conceptual framework of discovery of trends and patterns within textual data for more general purpose usage.

One body of work uses text category labels (associated with Reuters newswire) to find unexpected patterns among text articles. The main approach is to compare distributions of category assignments within subsets of the document collection. For instance, distributions of commodities in country C1 are compared against those of country C2 to see if interesting or unexpected trends can be found. Extending this idea, one country's export trends might be compared against those of a set of countries that are seen as an economic unit.

Another effort is that of the DARPA Topic Detection and Tracking initiative. While several of the tasks within this initiative are standard text analysis problems (such as categorization and segmentation), there is an interesting task called On-line New Event Detection, whose input is a stream of news stories in chronological order, and whose output is a yes/no decision for each story, made at the time the story arrives, indicating whether the story is the first reference to a newly occurring event. In other words, the system must detect the first instance of what will become a series of reports on some important topic.

The reason I consider this examples - using multiple occurrences of text categories to detect trends or patterns - to be "real" data mining is that they use text metadata to tell us something about the world, outside of the text collection itself. The computational linguistics applications tell us about how to improve
language analysis, but they do not discover more widely usable information. The various contrasts made above are summarized in Table 1.

<table>
<thead>
<tr>
<th>Finding patterns</th>
<th>Finding Nuggets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-textual data</td>
<td>Novel</td>
</tr>
<tr>
<td>Textual data</td>
<td>Non-Novel</td>
</tr>
<tr>
<td>Standard data mining</td>
<td>?</td>
</tr>
<tr>
<td>Computational linguistics</td>
<td>Database queries</td>
</tr>
<tr>
<td>Real TDM</td>
<td>Information retrieval</td>
</tr>
</tbody>
</table>

**Table 1 A classification of data mining and text data mining applications.**

**Text Data Mining as Exploratory Data Analysis**

Another way to view text data mining is as a process of exploratory data analysis that leads to the discovery of heretofore unknown information, or to answers to questions for which the answer is not currently known. The idea here is to use text for discovery in a more direct manner.

**Using Text to Form Hypotheses about Disease**

For more than a decade, Don Swanson has eloquently argued why it is plausible to expect new information to be derivable from text collections: experts can only read a small subset of what is published in their fields and are often unaware of developments in related fields. Thus it should be possible to find useful linkages between information in related literatures, if the authors of those literatures rarely refer to one another's work. Swanson has shown how chains of causal implication within the medical literature [7] can lead to hypotheses for causes of rare diseases, some of which have received supporting experimental evidence. For example, when investigating causes of migraine headaches, he extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high levels of magnesium inhibit SCD
- migraine patients have high platelet agreeability
- magnesium can suppress platelet agreeability

These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links. The hypothesis has to be tested via non-textual means, but the important point is that a new, potentially plausible medical hypothesis was derived from a combination of text fragments and the explorer's medical expertise. (According to swanson91, subsequent study found support for the magnesium-migraine hypothesis. This approach has been only partially automated. There is, of course, a potential for combinatorial explosion of potentially valid links. Beeferman98 has developed a flexible interface and analysis tool for exploring certain kinds of chains of links among lexical relations within Word Net [8] However, sophisticated new algorithms are needed for helping in the pruning process, since a good pruning algorithm will want to take into account various kinds of semantic constraints. This may be an interesting area of investigation for computational linguists.

**Using Text to Uncover Social Impact**

Switching to an entirely different domain, consider a recent effort to determine the effects of publicly financed research on industrial advances. After years of preliminary studies and building special purpose tools, the authors found that the technology industry relies more heavily than ever on government-sponsored research results.

Further narrowing its focus, the study set aside patents given to schools and governments and zeroed in on those awarded to industry. For 2,841 patents issued in 1993 and 1994, it examined the peak year of literature references, 1988, and found 5,217 citations to science papers. Of these, it found that 73.3 percent had been written at public institutions - universities, government labs and other public agencies,
both in the United States and abroad. Thus a heterogeneous mix of operations was required to conduct complex analyses over large text collections.

These operations included:
1. Retrieval of articles from a particular collection (patents) within a particular date range.
2. Identification of the citation pool (articles cited by the patents).
3. Bracketing of this pool by date, creating a new subset of articles.
4. Computation of the percentage of articles that remain after bracketing.
5. Joining these results with those of other collections to identify the publishers of articles in the pool.
6. Elimination of redundant articles.
7. Elimination of articles based on an attribute type (author nationality).
8. Location of full-text versions of the articles.
9. Extraction of a special attribute from the full text (the acknowledgement of funding).
10. Classification of this attribute (by institution type).

The LINDI Project

The objectives of the LINDI project [9] are to investigate how researchers can use large text collections in the discovery of new important information, and to build software systems to help support this process. The main tools for discovering new information are of two types: support for issuing sequences of queries and related operations across text collections, and tightly coupled statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents. Both sets of tools make use of attributes associated specifically with text collections and their metadata. Thus the broadening, narrowing, and linking of relations seen in the patent example should be tightly integrated with analysis and interpretation tools as needed in the biomedical example.

Most information access systems require the user to execute and keep track of tactical moves, often distracting from the thought-intensive aspects of the problem. The LINDI interface provides a facility for users to build and so reuse sequences of query operations via a drag-and-drop interface. These allow the user to repeat the same sequence of actions for different queries

- Transformation, i.e., applying an operation to an item and returning a transformed item (such as extracting a feature).
- Ranking, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same cardinality.
- Selection, i.e., applying an operation to a set of items and returning a (possibly) reordered set of items with the same or smaller cardinality.
- Reduction, i.e., applying an operation to one or more sets of items to yield a singleton result (e.g., to compute percentages and averages).

This system will allow maintenance of several different types of history including history of commands issued, history of strategies employed, and history of hypotheses tested. For the history view, we plan to use a spreadsheet’ layout as well as a variation on a “slide sorter” view which Visage uses for presentation creation but not for history retention. These include facilities that refer to metadata structure, allowing, for example, query terms to be expanded by terms one level above or below them in a subject hierarchy. Once a successful set of strategies has been devised, they can be re-used by other researchers and (with luck) by an automated version of the system. The intent is to build up enough strategies that the system will begin to be used as an assistant or advisor, ranking hypotheses according to projected importance and plausibility.

Thus the emphasis of this system is to help automate the tedious parts of the text manipulation process and to integrate underlying computationally-driven text analysis with human-guided decision making within exploratory data analysis over text.
A hypothetical sequence of operations for the exploration of gene function within a biomedical text collection, where the functions of genes A, B, and C are known, and commonalities are sought to hypothesize the function of the unknown gene. The mapping operation imposes a rank ordering on the selected keywords. The final operation is a selection of only those documents that contain at least one of the top-ranked keywords and that contain mentions of all three known genes.

Conclusion

For almost a decade the computational linguistics community has viewed large text collections as a resource to be tapped in order to produce better text analysis algorithms. In this paper, we have attempted to suggest a new emphasis: the use of large online text collections to discover new facts and trends about the world itself. We suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results.

References

ABSTRACT: In a series of natural language processing, the initial step is parsing the sentence in a particular language based on the grammar in order to help understanding the meaning of sentence. 'Parsing' is the term used to describe the process of automatically building syntactic analyses of a sentence in terms of a given grammar and lexicon. The resulting syntactic analyses may be used as input to a process of semantic interpretation. 'Parsing' is also used to include both syntactic and semantic analysis. "Parse selection is the process of deciding which of a number of possible parses (trees) is the most likely for a given sentence". This paper briefly describes an overview of various parsing technique in natural language processing.

Keywords: Natural Language Processing, Parsing, semantics, lexicon, semantic analysis, Grammar.

Introduction
Parsing is the process of analyzing a sentence for its structure, content, and meaning, this process uncover the structure, articulate the constituents, and the relation between the constituents of the input sentence. Parsing is the basic task in natural language processing and it is also a basis of all natural language applications such as machine learning, question answering, and information retrieval. Each individual word in a sentence is called 'token'. Parsing process is divided into three steps.

i. Input the sentence
ii. Decomposing sentence into tokens.
iii. Tagging Part-Of-Speech (POS)

Parsing Mathematical Expressions
Parsing the set of all mathematical expressions is beyond the power of a true regular expression. The reason is that these can contain arbitrarily deep nested pairs of parentheses. For example, consider the expression:

\[(2 + (3 * (7 - 4)))\]

The structure of the arithmetical expression is effectively a tree in fig 1:

```
+  /
  \  2*
    / \
   3-  \\
    /  \\
   7 4
```

Fig. 1 Tree of an arithmetic expression

The tree structure generated as the result of running a CFG (Context Free Grammar) parser is called a parse tree.

Parsing Techniques
A parsing technique uses grammatical rules to search for a combination of rules that describe the structure of the input sentence. Fig. 2 shows the general grammatical rules.
Generally, parsing technique are divided into two types, they are:

- Top-down parsing
- Bottom-up parsing

**Top-down Parsing**

When the parser starts constructing the parse tree from the start symbol and then tries to transform the start symbol to the input, it is called **top-down parsing**. A top-down parser starts from the starting rule and rewrite it step by step into symbols that match words in the input sentence.

This is an example to illustrate the relationship between words and their structure. For example, take the sentence,

"The dogs cried"

\( S \rightarrow \text{NP, VP} \)
\( \text{NP} \rightarrow \text{ART, N} \)
\( \text{VP} \rightarrow \text{V, NP} \)
\( \text{NP} \rightarrow \text{ART, ADJ, N} \)
\( \text{VP} \rightarrow \text{V} \)
\( \text{NP} \rightarrow \text{ADJ, N} \)

In top-down parsing, parse tree constructed based on a leftmost derivation of the input (which is read left-to-right). The parser begins with the symbol S, and then applies productions, at each step deriving a sequence of terminals and non-terminals, until finally arriving input program. The parse tree of above sentence is depicted in fig. 3 as follows.

![Fig. 3 Top-down parsing](image)

We can use either depth-first-strategy or breadth-first-strategy to implement top-down parsing. Two properties of top-down parser is easy to understand and easy to customize.

**Bottom-up Parsing**

Bottom-up parsing starts with an input symbols and tries to construct parse tree up to the start symbol. Consider the sentence,

"The dogs cried"

\( S \rightarrow \text{NP (ART the) (N dogs)) (VP cried)} \)

The following fig. 4 shows parse tree representation using bottom-up parsing.

![Fig. 4 Parse tree using Bottom-up parsing](image)

It is more powerful and used by most parser generators. Shift-reduce is the common bottom-up technique.

One big disadvantage is the fact that bottom-up parsing does not support left/right information flow. (For example, checking symbol definitions.)
Other Parsing Techniques avoids ambiguity

Parsing is used to solve various complex NLP problems such as conversational dialogues and text summarization. It is different from ‘shallow parsing’ in that it yields more expressive structural representations which directly capture long-distance dependencies and underlying predicate-argument structures. There are three main types of parse tree structures – constituency parsing, dependency parsing and shallow parsing.

Constituency parsing

Constituent-based grammars used to analyze and determine components which a sentence is composed of. There are usually several rules for different types of phrases based on the type of components they can contain, and this can be used to build a parse tree. The non-terminals nodes in the tree are types of phrases and the terminal nodes are the words in the sentence which are constituents of the phrase.

For example, consider the following sentence and its constituency parse tree fig 5 below.

“Harry met Sejal”

![Fig. 5 Parse tree – Constituency Parsing](image)

Dependency parsing

The dependency-based grammar based on the notion that linguistic units. Words are connected to each other by directed links (one-to-one mappings) between words which signify their dependencies. The resulting parse tree representation is a labeled directed graph where the nodes are the lexical tokens and the labeled edges show dependency relationships between the heads and their dependents. The labels on the edges indicate the grammatical role of the dependent.

For eg, consider the sentence

“Harry met Segal”.

![Fig. 6 Dependency parse tree](image)

Shallow parsing

Text chunking, also referred to as shallow parsing, is a task that follows Part-Of-Speech Tagging and that adds more structure to the sentence. The result is a grouping of the words in “chunks”. Shallow parsing or chunking is a process dividing a text into syntactically related group. Here’s a example:

“My dog likes his food”

POS tagging output :
My/PRP$ dog/NN likes/VBZ his/PRP$ food/NN/

Chunking output :
[NP My dog] [VP likes] [NP his food]
Conclusion

Parsing is a search problem which may be implemented with many control strategies. Broadly it is implemented with either top-down or bottom-up approaches, each have problems. By combining these two we can solve some problems, but not all. Both are left recursive and also syntactic ambiguity. To avoid ambiguity we have constituent parsing, dependency parsing and shallow parsing works by defining the functional relationship between the words.

References
2. www.tutorialspoint.com/compiler_design/compiler_design_types_of_parsing.html
4. J. Kimball 1973 Seven principles of surface structure parsing in natural language Cognition 2
ABSTRACT: The primary purpose of this paper concentrates on to find and describes the challenges faced in the project arrangement. I have observed that Project arrangement is an new managerial method to advance productivity and performance. But there are some factors like undefined goals, changing scope, improper risk management and impossible deadlines which makes this difficult to implement and act as per schedule. The paper shows how some factors affects project management and effects on projects efficiency. Software project arrangement problem (SPAP) is one of the important problem faced by the software project managers. To develop a flexible and effective model for software project planning, in this paper i have tried to do the survey of several methodologies and results yielded are elaborated.

Keywords: Software engineering, project management, software project resources, project scheduling.

Introduction

Project management is the planning, organizing and managing the effort to accomplish a successful project. A project is a one-time activity that produces a specific output and or outcome, for example, a building or a major new computer system. This is in contrast to a program, an ongoing process, such as a quality control program, an activity to manage a number of multiple projects together.

Project Arrangement Components

Core processes are usually interrelated and intertwined with each other and need to be performed in the same order on the majority of projects. They include three essential procedures for developing the project plan:

a. Planning and Defining the Scope: The first thing to do is to develop a clear scope statement as the project's foundation. The whole project will be built upon this and after that you can subdivide the major project objectives into clear and manageable deliverables.

b. Developing the Schedule: This requires a series of several procedures in order to create the project schedule that involves: Identifying and cataloguing the activities that must be performed to produce the various project deliverables, analyzing the sequence of activities and documenting any interactivity dependencies. Estimating all the work periods needed and how long it will take to complete individual activities. It is also useful to design a Milestone List, which lists the most important deliverables as way points, which will help you to evaluate the progress of a project.

c. Resource Planning: This includes the determination of resources (people, equipment, materials, etc.) and the quantities necessary to carry out the project activities, the development of a cost approximation of the resources required to complete the project and the cost budgeting and the allocation of the overall cost estimation to individual work packages.

The facilitating processes, are general interactions among the planning processes that are dependent on the nature of each project. Although these processes are performed intermittently and in no particular sequence during project planning and execution, they are not optional and include six additional components of a project plan:

1. Quality planning: You should identify the quality standards relevant to the project and determine how to achieve them.

2. Organizational Planning – WBS (Work Breakdown Structure): Here, you should identify, document and assign project roles and responsibilities among the staff and set the reporting relations. The next step is Staff Acquisition, or getting the needed human resources assigned to and working on the project. A useful tip is to form clear and manageable work packages that correspond to the
individual work performed by each of the staff and create a document which will serve as a reference point for managing project progress.

3. Setting a Communications plan: This involves configuring the information and communications needs of all the stakeholders (who needs what information, when will they need it and how they will get it).

4. Risk Management Planning, is one of the most important components of a project plan and decides the approach and plan for risk management in a project. In other words, it is the safety net of each project and involves the following processes: Identifying key risks likely to affect the project and documenting the characteristics of each. Performing a Qualitative Risk Analysis of the project’s risks and conditions in order to prioritize their effects on the project objectives. Running a Quantitative Risk Analysis, measuring the probability and impact of each of the risks and estimating their impact on the project’s objectives. Develop a Risk Response Plan, building up a mechanism to strengthen contingencies and reduce any threats to the project’s objectives from risk.

5. Procurement Planning: Defining what to procure, how much to procure and when.


Problems Faced In Project Arrangement

Fig. 1 Challenges Your Organization Faces in Managing Projects

In my research project managers also identified 'Training of Project Sponsors' fifty-four percent (54%) of the time as something that would benefit them most to improve their ability to manage a project. This item saw a significant increase over the prior year by 11 percentage points. As well they said they would benefit from 'Communication Skills' and 'Coaching Skills' (34% each).

Fig. 2 Beneficial Items to Improve Your Ability to Manage a Project
The research also showed the following major that are facing organizations when managing projects:

- A lack of project management skills
- Project not linked to organizational goals
- Project does not include all stakeholder needs
- Loss of control due to lack of detail in project plan
- Conflict among project team members
- Lack of senior management support/buy-in

All of these challenges point to the need for greater strategic perspective within organizations when it comes to project management. Projects must be strategically aligned to support the organizations corporate strategy if they are to survive the ever-changing priorities of the organization. Here are some recommendations on what project managers can do to manage these challenges.

**Shifting Organizational Priorities**

The ever-changing nature of our economies and organizations creates uncertainty on organizational priorities. One of the most frustrating experiences a project manager can suffer is managing within this environment – while the project is being implemented. We refer to this type of project change as Strategically-Driven Scope Creep. This can seriously impact the entire project.

There are a number of reasons why Strategically-Driven Scope Creep can occur:

1. The ever-changing nature of the organization may change the project’s goal, deliverable, budget, or timeline. For example, a project that is critical today is suddenly not as important tomorrow. And the project manager is told to put it aside for awhile and work on other, newer priorities.
2. A change in top management may be accompanied by a change in priorities and even in the direction of expansion and other efforts. For example, a project that progressing at a reasonable pace is suddenly thrust forward. A shorter time line is forced on the project, yet all deliverables, as originally defined, must be met.
3. A new President or even department manager may render a project meaningless in the light of new direction, new services, new products, new technology, etc. Projects are accelerated, slowed down, cancelled, maintained but with different deliverables, etc.
4. Even without the influence of change, management may change their own course. If we could all get into the brains of our senior management group, this challenge would be easy to manage.

Unfortunately, the project manager has little control over these occurrences. However, the project hopefully has a clear process for managing project changes. This Change Management process is designed to assist in managing this type of major, externally influenced change. Here are steps to help reduce the impact of these major changes:

1. Immediately inform the team...avoid adopting an anti-management attitude when your project’s goal, deliverables or even its continued life is threatened. Explain why the change is occurring, and then use the change management process to manage the change.
2. Concentrate on executing the change, rather than on less productive and negative activities...do everything you can to channel your team's energy into the new effort.
3. Revise your schedule and budget...only after the change request has been approved.
4. Revise your project plan...and gain your team's commitment to the new schedule.
5. Attach the signed Change Request Form to the Project Scope Statement. The final success of a project is measured against the Project Scope Statement plus all attached, signed Change Request Forms.
6. If the project has been cancelled, complete a close-out evaluation report on the project to date.

These projects have a funny habit of re-appearing again in one or two years. And although it may be a new team assigned to re-start the project, possibly with different deliverables, if the documentation from this effort has been retained, it will assist this new effort enormously because the knowledge will not be lost.

**Lack of Clarity in the Scope of the Project**

The Scope Statement, once completed, becomes the basis for future project decisions including, in particular, the criteria used to determine if the project (or phase of the project) has been completed successfully. Essentially, this document will form the basis for an agreement between the project team, the
project customer and the project sponsor, by identifying both the project goal and major project deliverables.

So, why is there a lack of clarity in the scope of the project? We coach many projects in crisis every year. During the initial auditing phase of the project, we ask to review how the project was scoped out. Invariably, we receive piles of documents, files, etcetera, yet there is often nothing that succinctly identifies the scope of the project.

To bring these projects back on track, the project team must go back and define the scope of their project as well as gain a clear understanding of the stakeholders’ needs. By defining scope through the planning process, there becomes greater control and the need for project change is minimized. This further reduces the likelihood of risk. Once developed by the project team, the Scope Statement is reviewed and approved by the Sponsor. Then clarity on the scope of the project is obtained.

As the project progresses, one may find that the scope of the project has changed. To address this it is important to apply the Change Management Process. Requests for change are added as an addendum to the Scope Statement. However, the Scope Statement itself is not re-written.

Project success is measured by the ability to have achieved all the requirements as outlined in the Scope Statement, with all approved Project Change Requests.

**Project changes not well managed**

The approved Scope Statement and Project Plan initiate a Request for Change when there is a need to modify the agreed-upon project scope as defined. Scope changes often require adjustments to cost, time, quality, risk or other project deliverables. These changes must be put back through the planning process, updating plans as needed and notifying stakeholders as deemed appropriate. Corrective action is needed to bring expected future project performance into line with the project plan.

The term "Scope Change" is often used. It refers to anything that will now be different to what had originally been agreed-upon in the original Project Scope Statement and subsequently the Project Plan. However, there are issues that the Project Manager and/or team will face in the management of Project Change, which will lead to an inability to manage project change. Some of these issues are:

1. The Project Team does not recognize that a change has occurred. They permit schedule and budget variances to occur. Anything that alters the project planned dates, budget, deliverables, or customer expectations must be documented and approved before implementation. Do not go crazy with change requests. Generally, if there is a slippage in the schedule or budget, track these. When it becomes apparent that you will not be able to catch up on these slippages and therefore a milestone will be missed, that is when a Change Request should be completed and approved.

2. The Project Team does not handle change well. They get upset when the sponsor initiates a change rather than submit it back as a change request.

   The only thing that we know for certain is that the Sponsor will often change their mind for reasons that we do not necessarily understand. Nevertheless, each change can be reflected back to the Scope Statement and therefore given back to them in the form of a Change Request. Often, when they realize the impact that this change will have on the project, they decide not to proceed.

3. The Project Team doesn’t provide options to the sponsor. This may lead to rejection of the change request. If the Sponsor is backed into a corner and told that there is only one option and that is to approve the requested change, then the team must accept the consequences of a “No” response. However, if they provide options to the Sponsor. Furthermore, let them know the consequences of not approving any option. How will this impact the project? The more data you can provide to your sponsor, the more likely they will give you some level of approval.

**Project risk not assessed or managed**

Risk Management is the process of identifying, analyzing and responding to the risk factors throughout the life of a project in order to ensure that the scope deliverables are successfully realized. Essentially, the purpose of risk management is to:

- Identify possible risks.
- Reduce or allocate risks.
- Provide a rational basis for better decision making in regards to all risks.
- Plan appropriate responses to the eventual outcome of a risk.
Proper risk management implies that we can control possible future events. Of course we cannot, but it forces us to think about what might prevent success, based on our own experiences. And therefore, what can we do to reduce risk likelihood or to manage the risk, should it occur. It is proactive rather than reactive.

Proper risk management will reduce the likelihood of an event occurring and the magnitude of its impact. Projects incorporate a Risk Management Process to reduce crisis management. There will always be some things that will occur but most of these, through sound risk management, can be managed.

Assessing and managing risks is the best weapon you have against project catastrophes. By evaluating your plan for potential risks and developing strategies to address them, you’ll improve your chances of a successful, if not perfect, project.

Risk Management is continuous. It is not done just at the beginning of a project; rather it is done throughout a project to ensure that high priority risks are aggressively managed and that all risks are cost-effectively managed throughout the project. As well, it provides management at all levels with the information required to make informed decisions on issues critical to project success.

Training of project sponsors

Training project sponsors on their role in ensuring project success helps the project manager and increases the likelihood of project excellence. This type of training provides sponsors with an understanding of what the expectation is, on their part, to help manage project issues, to approve project scope and to provide timely approvals. When project sponsors are on side, resources are more likely to be allocated in accordance with project importance, as they tend to continuously use a set of standard criteria to prioritize projects.

While sponsors will not be managing the day-to-day details of the project they need to understand their role in ensuring success. For this reason, BIA offers an a series of executive training courses in Project Management that includes training for project sponsors. The training comprehensively covers the issues and concerns around project plan implementation such as the setting of criteria for signing off on Scope Statement approval.

Solutions to the problems

1. Budget

When allocated resources start to shrink, you're faced with re-juggling your project to stretch fewer resources further. To accommodate such changes, address the project objectives, features and timelines with your executive management. You might end up doing more with fewer resources, but before updating your project plan get your management’s buy-in on how you accomplish that.

2. Scope

Managing scope creep is a key talent for IT project managers. Best practice is to include a process for handling scope changes in the initial project plan. This way, you have a better chance to address necessary changes in timeline and budget. If you didn’t put a scope change clause in your original plan, you still need to negotiate time and resources.

3. Timeline

If new business conditions make your project more strategic, you might be asked to accelerate your delivery times. Before making promises, discuss and negotiate resource allocation (dollar and human), so you can realistically make the shorter timeline. The opposite problem may arise too—an unforeseen situation downgrades your project and you get more time. Then you face the challenge of keeping your people engaged on your project that is no longer as strategic as when you started.

4. Team composition

For long-running projects, your project team can change, with people leaving for other projects or new jobs. It's frustrating, but you must recalibrate. Review the unfinished tasks that departing individuals were assigned (easy enough if you're using an online project management software) and look for new people to fill the gaps. If you’re a manager and have the right skills, don’t be tempted to step in yourself. Your job is to manage the project, not execute it. Consider looking outside of your organization to address the unexpected need.

Change takes many interrelated forms, and it's not necessarily unreasonable. At the beginning of a project, no crystal ball can tell you and your decision-makers how new circumstances can affect the project. You just have to be ready for anything.
Conclusion

There is a need for organizations to link their projects back to their corporate strategies and to train project sponsors to increase the likelihood that projects will survive shifting organizational priorities. Project Management Offices and project managers must understand how each project contributes to achieving corporate goals which should be found in the strategic plan. With a process in place to manage change and risk, project managers are better able to cope and deal with these challenges. Similarly, training of Project Sponsors will help to bring them on side so that resources are allocated in accordance with project importance.

References
PERSONALIZED RECOMMENDATION IN E-LEARNING SYSTEM - A SEMANTIC APPROACH

Gomathi. B
Department of Computer Science,
Madurai Sivakasi Nadars Pioneer Meenakshi Women’s college, Poovanthi

ABSTRACT: The proposal of various electronic learning contents, e.g. remote education or virtual classrooms, has given a powerful impact to the E-Learning techniques. However, there still remain several hard and complicated problems unsolved. Especially when compared with e-commerce and medicine, the problems of the recommenders in E-Learning system have not been fully solved out. In this paper, a novel personalized recommendation system (PRS) for E-Learning is designed. The proposed PRS system using the semantic relationship between learning materials and the learner’s need, system can select the suitable learning object (LO) based on Lingo3G clustering algorithms. Subsequently, the PRS adopts rules for selecting the LO based on ontology rules. Depending on the specific ontology knowledge with domain rules, semantic mapping and intelligent reasoning techniques are applied to generate certain semantic related recommending items by using Protégé. Experimental results show that the PRS performs with Page rank and HITS algorithm.

Keywords: Semantic Web, Web Mining, E-learning, Personalization, Recommendation

Introduction
E-learning is a form of electronic teaching that enables people to learn anytime and anywhere. The objective of an online personalization system is to provide users with the information they require, without asking them explicitly [1].

Electronic learning refers to any learning with the aid of information and communication technology (ICT) such as online learning, webinars, discussion forums, chats, computer-based learning and educational materials on videos. (Wang, 2004) The key advantage of e-learning is flexibility, convenience, and the ability to work at any place at any time where a network connection is available.

There number of techniques available to retrieve the content of e-learning. In this era of web-enabled technologies, E-learning has emerged as a boon to building knowledge skills amongst learners [2]. In a personalized E-Learning system, user customize the learning environment based on personal choices. In a general search process, a hyperlink which is having maximum number of hits will get displayed first. For making a personalized system history of every user need to be saved in the form of user logs [3].

![Fig.1 Search component Architecture for Personalized E-learning](image)

Literature Survey
In this era of web-enabled technologies, E-learning has emerged as a boon to building knowledge skills amongst learners [5]. A clustering search engine called Lingo used only snippets to cluster the documents. This method takes less time to cluster the documents. This method focus on clustering all documents using by applying semantic similarity between words and then by applying modified lingo algorithm in less time and produce good quality[5]. The comparison between HITS[6] and Page rank algorithm is done after clustering the LO. XML, URL[12], SPARQL languages are used to retrieve the documents. Ontology is constructed semantically by using Protégé ontology editor.[12]
Web Mining Techniques

Web mining gives an innovative idea for scientific research and pushing web technology to toward making the meaningful information and exploits some data mining techniques to automatically retrieve valuable information from the World Wide Web. It makes an environment where the information available on the web can be easily semantically interpreted. Web mining assembles more feature to built web personalize interaction and customizing a web site according to the requirements of users, obtaining advantage of the knowledge attained from the study of the user’s browsing behavior. However, the existing heterogeneous environment and usually a large server can’t guarantee the reliability and making for unstructured information, only for structured environment.

Semantic Web

Current web pages based on Hypertext Markup Language (HTML) technology display only the text and images with links to original web pages only and do not identify their meaning. But semantic architectures use tags for every words in the page is always enable system programmers to find relationships between them by semantic rules and ontological data.

Implementation and Optimization

Preprocessing
- White space removal
- Stemming
- Stopword Removal

Clustering Algorithms

Several clustering algorithms are used in personalization, but based on its efficiency Lingo3G Clustering algorithms recommended in this approach. The steps for clustering are
- Step 1: Preprocessing
- Step 2: Semantic similarity calculation
- Step 3: Frequent phrase extraction
- Step 4: Cluster label induction
- Step 5: Cluster content discovery

Lingo3G are open source clustering algorithms available in the many frameworks like Carrot2 framework.

Analysis of algorithms

HITS algorithm

From here on, we translate everything into mathematical language. We associate to each page $i$ any numbers: an authority weight $a_i$, and a hub weight $h_i$. We consider pages with a higher $a_i$ number as being better authorities, and pages with a higher $h_i$ number as being better hubs. Given the weights $\{a_i\}$ and $\{h_i\}$ of all the nodes in $S_Q$, we dynamically update the weights as follows:
HITS, like Page and Brin’s PageRank, is an iterative algorithm based on the linkage of the documents on the web. However it does have some major differences:

It is query dependent, that is, the (Hubs and Authority) scores resulting from the link analysis rank are influenced by the search terms; To begin the ranking, and we consider two types of updates: Authority Update Rule and Hub Update Rule. In order to calculate the hub/authority scores of each node, repeated iterations of the Authority Update Rule and the Hub Update Rule are applied. A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule.

**Page rank algorithm**

The PageRank of a page A is given as follows:

\[ PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right) \]

PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

1. PR(Tn) - Each page has a notion of its own self-importance. That’s “PR(T1)” for the first page in the web all the way up to “PR(Tn)” for the last page
2. C(Tn) - Each page spreads its vote out evenly amongst all of it’s outgoing links. The count, or number, of outgoing links for page 1 is “C(T1)”, “C(Tn)” for page n, and so on for all pages.
3. PR(Tn)/C(Tn) - so if our page (page A) has a back link from page “n” the share of the vote page A will get is “PR(Tn)/C(Tn)”
4. d(… - All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is “damped down” by multiplying it by 0.85 (the factor “d”)
5. (1 - d) - The (1 – d) bit at the beginning is a bit of probability math magic so the “sum of all web pages' Page Ranks will be one”: it adds in the bit lost by the d(… It also means that if a page has no links to it (no back links) even then it will still get a small PR of 0.15 (i.e. 1 – 0.85). (Aside: the Google paper says “the sum of all pages” but they mean the “the normalized sum” – otherwise known as “the average” to you and me.
6. Lets take the simplest example network: two pages, each pointing to the other:
1. Each page has one outgoing link (the outgoing count is 1, i.e. \( C(A) = 1 \) and \( C(B) = 1 \)).
2. where \( n \) is the total number of pages \( p \) connects to and \( i \) is a page which \( p \) connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking pages
3. Normalization is performed
4. The final hub-authority scores of nodes are determined after infinite repetitions of the algorithm every time. As directly and iteratively applying the Hub Update Rule and Authority Update Rule leads to diverging values in a page, it is necessary to normalize the matrix values after every iteration. Thus the values obtained from this process will eventually converge and increase.

Experimental results

Fig.4.Clustering

![Clustering Diagram]

Fig.5.Ontology creation

![Ontology Diagram]
Conclusion and Future Enhancement

Web mining uses data mining technologies to find required information from structured or unstructured or semi-structured web documents. Internet is a major resource in performing research and education related activities for students. There are various attempts to model the integration and interoperability of web pages for an efficient knowledge management. Our proposed searching mechanism consists of LO as values. The personalization is done through the page rank algorithms HITS and Page rank algorithm. Based on the query we can use both.

In Future, systematically we can choose which ranking algorithm is suitable for E-learning environment.

References

1. Yuewu Dong, Jiangtao Li, "Personalized Distance Education System Based on Web Mining", International Conference on Education and Information Technology (ICEIT 2010), IEEE, 2010
6. Saeko Nomura Satoshi Oyama Tetsuo Hayamizu Toru Ishida, "Analysis and Improvement of HITS Algorithm for Detecting Web Communities", Department of Social Informatics, Kyoto University Honmachi Yoshida Sakyo-ku, Kyoto, 606-8501 Japan@kus.kyoto-u.ac.jp, ishida@i.kyotou.ac.jp
A PROBE ON TEXT MINING TOOLS, ALGORITHMS AND METRICS

P.Mahizharuvi & S.Ashika Parvin

1Department of Computer Science, Sri Meenakshi College, Madurai
2Department of Computer Science, Madurai Sivakasi Nadars Pioneer Meenakshi Womens College, Poovanthi

ABSTRACT: Today's world scenario is something homogeneous for people who are drastically flourishing in the cyber world as it grows much more expeditious. There arises a quandary of categorizing, analyzing, summarizing and discovering of data that are given by users in gregarious media like Facebook, Twitter, E-mails etc., in the form of text. Text mining is essential if constructive cognizance needs to dig out from heaps of text. However, where to commence, what are the favourite implements, which techniques and algorithms are utilized and what metrics can be used along with them are discussed. This paper endeavours to explain the available implements for text mining and algorithms with its performance metrics which gives a smooth kick-start to the incipient researchers and practitioners in the field of text mining.

Keywords: Categorizing, Analyzing, Summarizing, Discovering, Metrics, Text Mining.

Introduction

Due to the rapid development of hardware technology, there is a desideratum for storing large data as its capacity increases immensely. As exponentially the data grows every day, the users need variants of data in which they utilize textual data hugely rather than the numerical one. Hence there comes the role of developing techniques that are applied to the textual data which are different from the numerical data. The mining of these kinds of textual data is called Text mining.[1]

Text mining is flourishing nowadays. There is an elevating demand to build better and efficient solutions that are capable of dealing with some sticky situations. There is no bridge between the utilizing needs and the implements available in the industry. Text mining is an integrated ground concerning the retrieval of erudition, understanding of the text, extraction of information, clustering and categorization, visualization of data, machine learning in data science and mining of data.

Text mining is a confronting task since it deals with unstructured and fuzzy data. Utilizing text mining the possibility of analyzing and structuring large colossal sets of documents that apply statistical and computational linguistics technologies essential nowadays. The pattern and hidden information of the text can be elucidated utilizing Text Mining [2]. The subsistence of data is ubiquitous in the form of text where we are unable to store all the available data in the database. From time to time the analyzing of data is done that is not backed up in synergistic databases preferably it can be extracted from the firm’s website, emails or reports. Indeed, the technique utilized for structured data backed up in the database cannot be implemented in an unstructured data. Due to the complexity of an unstructured data, to extract information from them, more robust methods are required. Consequently, to analyze an unstructured data, i.e. Text, we obligate text mining implements with better performance.

A standard text mining task includes a categorization of text, clustering of text, extraction of text, engendered of granular taxonomies, analysis of text, summarization of documents, and entity cognition modelling[3]. Simple text mining implements are inadequate to handle text, so we require some well-equipped implements with high-performance rate that can analyze, categorize, summarize and discover text. Text mining can be utilized on any ground for business perspicacity, software process analysis, gregarious media analysis, biomedical analysis, sentiment analysis and even for security analysis. Thus text mining has extensive utilization in different grounds, and one needs to ken about the implements that are subsisting unambiguous algorithms and the metrics utilized by the implements to assist so that gaining of knowledge becomes more facile and more expeditious.

The overall goal and purpose are to convert text into data, for analyzing or classifying using categorization, information retrieval, lexical analysis, pattern recognition, tagging/annotation, information extraction using text analytics. Discovering of textual data built on the approach of probability and statistics can be recognized by the statistical functional words in the sample data, i.e. Text and whether or not the text is suspicious is identified utilizing powerful mining algorithm and tool. The term summarization refers to the automated summarizing of a sizably voluminous document into a crisp one, as opposed to the manipulation done manually. The utilization of text mining algorithm and implements can explore the quandaries faced during each process of text mining.
Certainly text mining algorithms are remedy for a class of complex data; they perform analytical calculation and also have automatic reasoning responsibilities based on statistical learning techniques. The set of rules of text mining algorithms can be expressed inside a tool in a right time and in a properly-described formal language for calculating a metric. Metric recognition using text mining algorithm is essential for categorising, analyzing, summarizing and discovering of data.

For the text mining system development, algorithm used by the implements plays a vital role in comparison of metric evaluation, but selecting congruous measures and metrics are not a simple task. The measures and metrics are perpetually developing in many domains, and it is consequential to consider the characteristics of the domain into account, then the metrics are decidable[6].

Motivation

Text mining [5] is the course of collaborating the information from a massive database by analyzing its relations, rules, and patterns from the text. The critical point is that rearranging the mined data together to produce a novel hypothesis that can be further developed by means of experimentations. The utilizer’s search for a relevant document for their requirements, but simultaneously irrelevant documents is also fetched along with them. To resolve the above problem, the texts are mined so that the utilizer gets what they want by excluding the other kinds of stuff that are irrelevant to them. Text mining [6] is classified as text summarization, text categorization, text analytics and knowledge discovery.

Due to increment in the availability of information in digital form, there occurs a quandary of intricacy in accessing them in an efficient way. So text categorization acquires a wide ocular perceiver on it because the paramountcy of categorizing the textual data makes the utilizer to access their required information facilely. The application of text categorization varies from the indexing of the document according to the lexical and the extraction of a document. Text categorization is the intersection of machine learning and information retrieval which assimilates cognizance from texts as well as from text mining [7]. The limber development of textual information in online, there is a desideratum for categorizing of texts according to the utilizer’s requisites, we require a technique to relegate incipient facts and discover intriguing information available online.

The concept of "Text Mining" is commonly used for characterizing a large amount of textual data, pattern recognition of the text and extraction of knowledge from it. The terms “text analytics” and “text mining” are basically transposable. They both have methods, tools, and applications in common. Their difference stalks from the people who is using what. Data miners use "Text Mining", and the organizations and individuals use "Text Analytics" in various domains to pave the way for business development methods that make a lot of difference in its growth [8]. If structured data is immensely colossal, then unstructured data is sizably voluminous. It is generally accepted that analyzing textual data cannot be done 100% as structured data as 70% of them are unstructured and fuzzy. The gain of an organization is insufficient because of the huge amount of structured data. In order to overcome the barrier between structured and unstructured data potentially we need well-equipped tools and techniques for finding the solution. As a solution Text Analytics - extracts information from the conversation of language/ framework using 5-W's ("Who, "Where," “When", “What” and “Why") conversation and I-H ("how" people feel, and the conversation comes to light). This is the reason why man is going to be replaced by the machine [9].

Machines are needed more than a human if the size of the data is astronomically immense. The time consumed by a human to read, understand and analyze the data is far more than a machine. In order to engender an outline, we have to identify the most dominant pieces of information from the document, omitting impolite information and minimizing particulars, and pull together into a compact coherent report, we havethe need of an automated tool[22] that has paved the way for extensive research in automatic text summarization [10]. Automatic text summarization can produce a precise and fluent summary of the preserving content that may be structured or unstructured. Man fails over machines as they yield far better results than human works [11].

Knowledge discovery is the process of extracting useful information that is spoken openly inside the text mining document, which is an efficacious approach for mining of text on the basis of erudition to be discovered. Information Retrieval is a technique that can accommodate new mining process for text. On the other hand, if an unstructured data resides in a document instead of conceptual knowledge, Information Retrieval system is just a supplementary technique for extracting that information to make over those unstructured data into a structured data. Then the extracted data are identified by the use of traditional cognizance mining tools to obtain the conceptual patterns [12].
Justification

The vital role of text mining is to dig out information from textual resources and enable them to utilize whereatext mining tool deals with the functions of categorizing/relegation, retrieving, Natural Language Processing (NLP). In order to relegate and discover patterns from various documents, some of machine learning tools can collaborate with text mining tools [10].

Text Categorization

For handling and organizing an enormous amount of amorphous textual data that are yielded from online, text categorization plays a vital role in this work. This technique is habituated to relegate new stories to obtain exciting news that is available on WWW to match the user’s needs. Text Categorization is the process of transferring a document into a definite number of categories. The initial step in the categorizing of text is the transformation of documents which contains strings of characters into an illustration for the cognition algorithm and the relegate task. The precision of the modern text relegate systems rivals with an amalgamation of information retrieval technology and machine learning technology [13]. There are many types of implements available in the market some of them are given below in table 1.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Outline</th>
<th>Features</th>
<th>Licence</th>
<th>Algorithm supported</th>
<th>Measurement and Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discover Text</td>
<td>Combines flexible and adaptive software algorithm to provide reliable large-scale analyses</td>
<td>Classification via manual training and automation, Filter by metadata and threshold classification</td>
<td>Proprietary</td>
<td>Sentiment By Term algorithm, Social Sentiment Analysis [29] [30]</td>
<td>Joint probability, inter-rater correlation, concordance correlation coefficient - Inter-Rater Reliability</td>
</tr>
<tr>
<td>Clustify</td>
<td>First software to offer Real-Time Predictive Coding.</td>
<td>Relevance of documents, Consistency Checking.</td>
<td>Proprietary</td>
<td>Continuous active learning, Logistic regression, Nearest neighbor, Naïve Bayes [33]</td>
<td>Prevalence, Precision, Recall</td>
</tr>
<tr>
<td>UClassify Corpus</td>
<td>Free machine learning web services to classify text.</td>
<td>Multinomial Naïve Bayesian classifier</td>
<td>Online</td>
<td>Naïve Bayes [34], Statistical machine translation (SMT) [35]</td>
<td>Accuracy, Macro-precision, Macro-Recall</td>
</tr>
</tbody>
</table>
A vast amount of unstructured text and visual content to empower your product for image classification.

Deep Learning

Online

Long Short Term Memory (LSTM) algorithms, Convolutional Neural Networks (Convnets) [36], Gradient-Based Learning [37], DeCAF — A Deep Convolutional Activation Feature [38][39]

MSRC-21

Table 1: Tools and Metrics for Text Categorization

Text Analytics

Text Analytics is the course of translating amorphous textual data into essential data for analysis, to quantify views and responsibility of the customer to offer search facilities, sentimental analysis, and text modelling to fortify fact-predicated conclusions. It uses many linguistic, statistical and machine learning techniques for the retrieval of information from an amorphous text data and the techniques for obtaining patterns, evaluating and interpreting the yield. The analysis takes into account the keywords, meanings, and tags from extensive data that are obtained from a firm in the form of different file formats. Text analytics go in hand in hand with text mining. Utilizing text analytics implements in table 2 gives you valuable information from data that aren't facilely quantified in any other way. It turns the unstructured cerebrations of customers into structured data that can be utilized by the business.
It is a Mixture of qualitative, quantitative methods for text analytics. Provalis, QDA Miner 5, PROSUITE, Uni-variate keyword frequency analysis. Proprietary Principal Coordinates Analysis, Principal Components Analysis (PCA), Correspondence Analysis, Diversity indices, unique patent-pending algorithm[51], Multiple Regression Analysis, Naive Bayes, k-means [52]

Dispersion Range, variance and standard deviation

Tool to build information extracting systems, NLP systems, and Text analysis. Interpreted NLP++ execution, Rich integrated GUI toolset, Hierarchical knowledge base management systems. Proprietary Named Entity Recognition(NER), MCL Clustering, Graph Packing Algorithm [53], Automatic Tokenization,[54]

Evaluation metrics

Proprietary

Tool to build information extracting systems, NLP systems, and Text analysis.

Tool to build information extracting systems, NLP systems, and Text analysis.

Tool to build information extracting systems, NLP systems, and Text analysis.

Table 2: Tools and Metrics for Text Analytics

Text summarization

A concise and fluent summary of text by preserving its crucial information and overall construct is called Text Summarization. A wide range of text summarization techniques has been developed in recent years in several domains. As a human, we are unable to summarize an astronomically large piece of text by reading it entirely which is facilely done by a machine with the availability of the implement highlighting the critical content. A machine lacks language capability which can be facilely handled utilizing an automatic summarizing tool that yields result far better than the human works. There are many implements available in the market as given in table 3.
Great Summary | Document Summarization | Great Summary Subscriptions | Online | Linear regression, Random forest, Support vector machines, Graph-Based Algorithms [60], Keyphrase Extraction Algorithm[KEA],LEXRan k, Gradual NLP algorithm[61] | ROUGE
---|---|---|---|---|---
Topic Marks | Advanced summarizing tools found on the Internet | Summarize text documents and display only the most essential points and save on their cloud-managed servers. | Online | Hamming distance, Levenshtein distance and Needleman-Wunch distance, Ontology Matching [62] | Precision, Recall, F-measure

Table 3: Tools and Metrics for Text Summarization

Knowledge Discovery
The extraction of information from a vast database that is implicitly accessible, anteriorly indefinite and latently subsidiary for the utilizers is kenned as knowledge revelation. Text mining is the revelation of fascinating erudition in text documents. It is a challenging issue to find precise erudition (or features) in text documents to avail users to find what they operate. Knowledge revelation can be efficaciously used, and update discovered patterns and apply it to the field of text mining. There are many tools available in the market as given in table 4.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Synopsis</th>
<th>Features</th>
<th>Licence</th>
<th>Algorithms supported</th>
<th>Measurement and Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEKA</td>
<td>Implement to discover unknown data from repository and file systems with simple to utilize options and GUI</td>
<td>WEKA GUI Chooser</td>
<td>Open source</td>
<td>Naive Bayes, Linear Regression, k-Nearest Neighbors, Ensembles, One Rule, Random Forest [63], Learning Vector Quantization (LVQ), Clonal Selection Algorithm (CLONALG) [64], Decision trees, Instance-based classifiers, Support Vector machines, Multi-layer Perceptrons, Logistic regression[65]</td>
<td>Statistical accuracy metrics, Decision support accuracy metrics, Precision, Recall and F1 Score</td>
</tr>
<tr>
<td>WITNESS MINER</td>
<td>Collection of data structures and algorithms are written specifically for the tasks in KD</td>
<td>Uses modern heuristic techniques.</td>
<td>Open source</td>
<td>Rule-based Algorithms[68] Decision trees, Clustering, Feature Subset Selection, Rule Induction[69]</td>
<td>Inter-rater reliability</td>
</tr>
<tr>
<td>DBMINER</td>
<td>Interactive mining of multiple-level knowledge in large relational databases</td>
<td>Meta-rule guided knowledge mining.</td>
<td>Open source</td>
<td>k-means, Decision tree [66], OLAP [67]</td>
<td>Capability, Learnability/ Usability, Interoperability, and Flexibility</td>
</tr>
</tbody>
</table>

Table 4: Tools and Metrics of Knowledge Discovery

Metric Analysis
Metrics are used for measuring the manners, activities of its supporting organization and needs of the utilizers. The metrics of text mining tools and its backing algorithms are examined broadly and are tabulated. In fig1, Comparative study of metrics among text mining tools is visualized. Indeed, new metrics are placing their importance in the market than the existing metrics. Some of the metrics from the tabulations are discussed here.
Inter-rater Reliability

Inter-rater Reliability, inter-rater agreement, or concordance, is the degree of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges. Different statistics are appropriate for different types of measurement. Some options are joint-probability, inter-rater correlation, concordance correlation coefficient.

Similarity Measures

Similarity measure or Similarity function is an absolute-valued function that quantifies the similarity between the two data. Frey and Dueck suggest defining a similarity measure

\[ s(x, y) = -\| x - y \|_2 \]

Where \( \| x - y \|_2 \) quantifies the similarity between the data

**Precision(P)**

Precision in text mining context, is the fraction of recovered textual file that are related to the query:

\[ \text{Precision}(P) = \frac{|\{\text{related textual file}\} \cap \{\text{recovered textual file}\}|}{|\{\text{recovered textual file}\}|} \]

**Recall(R)**

Recall in text mining context, is the fraction of the related textual file that is successfully recovered.

\[ \text{Recall}(R) = \frac{|\{\text{related textual file}\} \cap \{\text{recovered textual file}\}|}{|\{\text{related textual file}\}|} \]

**F-Measure**

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

\[ F = 2 \times \frac{P \times R}{P + R} \]

**Discrete Metric**

The discrete topology is the most exceptional topology—it cannot be subdivided further, each one appears as a singleton set.

\[ \rho(x, y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y \end{cases} \]
In this case, \( X, \rho \) is called a discrete metric space or a space of isolated points.

**Euclidean Metric**

Euclidean metric is the "ordinary" straight-line distance between two points is given by

\[
d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.
\]

**Taxi-Cab Metric**

It is the sum of absolute differences of the line segment between the points onto the Cartesian coordinates. More formally,

\[
d_1(p, q) = \|p - q\|_1 = \sum_{i=1}^{n}|p_i - q_i|,
\]

Where \((p, q)\) are vectors

**Kendall Tau Metric**

The Kendall tau ranking distance between two lists \( \tau_1 \) and \( \tau_2 \)

\[
K(\tau_1, \tau_2) = |\{(i,j) : i < j, (\tau_1(i) < \tau_1(j) \land \tau_2(i) > \tau_2(j)) \lor (\tau_1(i) > \tau_1(j) \land \tau_2(i) < \tau_2(j))\}|.
\]

Where \( \tau_1(i) \) and \( \tau_2(i) \) are the rankings of the element \( i \) in \( \tau_1 \) and \( \tau_2 \) respectively.

**Spearman's rank correlation coefficient**

Spearman's rank correlation coefficient or Spearman's rho, a measure of rank correlation is denoted by the Greek letter \( \rho \) (rho). It assesses how well the relationship between two variables can be described using a monotonic function.

\[
\rho_s = \rho_{rg_X, rg_Y} = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}
\]

Where \( \rho \) denotes rank variables. \( \text{cov}(rg_X, rg_Y) \) is the covariance of the rank variables \( \sigma_{rg_X} \) and \( \sigma_{rg_Y} \) are the standard deviations of the rank variables

**Pseudo_F**

It is calculated as the ratio of the squared distance between the centre of each cluster to the geometrical centre of the whole dataset to the number of clusters minus one, multiplied by the number of observations in each cluster. This number is then divided by the ratio of the squared distance between each point and centroid of the cluster to the total number of observations less the number of clusters.

**Dispersion**

Dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.

**Variance**

Variance is the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value. The variance of a random variable \( X \) is the expected value of the squared deviation from the mean of \( X, \) \( \delta = E[X] \)

\[
\text{VAR}(X) = E[(X-\delta)^2]
\]

**Standard Deviation (SD)**

SD quantifies the amount of disparity among data values. A low SD be likely to be close to the mean, while a high SD spreads over a broader range of values.
Range

The range of a set of data is the difference between the largest and smallest values

\[ s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}}. \]

Click-through rate

It is the proportion of utilizers who click on a specific link to the number of total utilizers who view a page, email, or advertisement.

\[ \text{CTR} = \frac{\text{Number of click-throughs}}{\text{Number of impressions}} \times 100(\%) \]

Digimind

It evaluates how many people actually read a post or mention.

DPMO

In process improvement efforts, defects per million opportunities or DPMO (or nonconformities per million opportunities (NPMO)) is a measure of process performance. It is defined as

\[ DPMO = \frac{1,000,000 \times \text{number of defects}}{\text{number of units} \times \text{number of Defects opportunities per unit}} \]

ROUGE

Recall-Oriented Understory for Gisting Evaluation (ROUGE) is a set of metrics used in NLP as an automatic summarization and machine translation software.

KPI

A Key Performance Indicator (KPI) is a metric that reveals how efficiently a firm is achieving its business key objectives. Organizations use KPI 'sto evaluate their success at reaching targets. Once you've selected your key business metrics, you will want to track these KPIs in a real-time reporting tool.

Conclusion

There are many text mining implements that are available in the market to enhance the conventional research on text mining. Some of them are unable to handle an enormous amount of unstructured data since most of the data is in the form of text. Virtually every utilization works with text, there is a desideratum for text mining implements with a complete interpretation along with text mining algorithms can provide a better performance rate, available in the form of open source, online and proprietary. This paper has given different ideas from different papers and websites for useful text mining with the help of tools that are available with well supported algorithm and the measures and metrics supported by the tool to achieve high-performance rate. The insight of metrics and measures discussed in this paper immensely guide the new researchers and practitioners to select the tool according to their requirements.

Future Work

Future work concerns the more in-depth analysis of best tool in text mining in the sense of its supported algorithm and metrics. New proposals can try different methodology will be undertaken using different datasets.

References


4. https://www.igi-global.com

5. Berry Michael W., Automatic Discovery of Similar Words, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer


CONVEX ANALYSIS OF FLOOD AND REGIONAL FREQUENCY IN HADOOP ANALYSIS

SARANYA .M
Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women’s College Poovanthi

ABSTRACT: Floods is one of the irremediable calamities. They are always dangerous causing heavy damage at the surroundings due to heavy rainfall which cannot be figured out. In some cases the behaviour of the streams and their link with other streams are examined and the GIS tool is used to predict the floods. Safety measures are not taken into consideration for flood. The data must be collected and analyzed for particular area and various attributes are taken for analysis purpose. Here analysis of risk is based on certain attributes to check whether the flood may cross the limit or not. It has been a challenge in predicting the floods that cause damage i.e mostly in the crowded areas. There are some difficulties in physics based models in demonstration the information about the absolute rainfall prediction. In this paper, we analyze the rainfall datasets based upon the threshold value. Big data is used to examine large datasets of rainfall and to predict the occurrence of floods. If the rainfall crosses the threshold value then the flood is predicted and an alert message is sent to the nearby locations. Map Reduce technique is used to give the exact reduced data from the large amount of input datasets.

Keywords: rainfall datasets; threshold value; map reduce

Introduction
Floods are common natural disasters in the world. Each year they cause considerable damage to people’s lives and properties. In spring 1973, the lower Saint John River in the Fredericton area (New Brunswick, Canada) experienced its worst ever recorded flooding. Since 1973 other floods have left another three people dead and caused more than $68.9 million in damage. The system neither directly display the areas affected by flooding, nor show the difference between two flood events. Based on the water levels, it is hard for users to directly determine which houses, roads, and structures will be affected by the predicted flooding. To deal with this problem, it is necessary to visualize the output from hydrological modeling in a Geographic Information System (GIS)[1] The analysis of drainage system is performed. The behaviour of the stream and their relationship with other streams are also analyzed. Remotely sensed data is used to portray about the stream that is buried underneath which is in connection with other drainage networks. Geographical Information System tool is additionally used to provide analyzing the data that helps us to find the areas where floods are likely to occur. [2] The sudden floods that occur rapidly which cause huge destruction to the surroundings. The risk map is generated for the rainfall based on certain parameters. Digital Elevation Model(DEM) is applied to Geographical Information System(GIS) to describe about the rainfall and the various parameters are calculated which helps to get rid of the floods that are about to occur in that particular region. [4] The wet spa model which is the mixture of altitude, soil and the data about the use of land that predicts floods with the parameters of the rainfall. The data of the soil is represented in GIS format and the use of the land is received from the remote sensed images. The obtained data is processed to determine the floods. [5] Heavy rainfall that causes floods in rivers and the nearby areas that is in charge of regular day to day life. An approach is used to automatically view the floods in the particular area by using cyber surveillance systems and image process methods to receive the immediate flooding results[6]

Fig.1
The GMR (Google map reduce) was invented by Google back in their earlier days so they could usefully index all the rich textural and structural information they were collecting, and then present meaningful and actionable results to users. MapReduce (you map the operation out to all of those servers and then you reduce the results back into a single result set), is a software paradigm for processing a large data set in a distributed parallel way. Since Google’s MapReduce and Google file system (GFS) are proprietary, an open source MapReduce software project, Hadoop, was launched to provide similar capabilities of the Google’s MapReduce platform by using thousands of cluster nodes.

Hadoop cluster architecture

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Job Tracker, TaskTracker, Name Node and Data Node. A slave or worker node acts as both a Data Node and Task Tracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications. Hadoop requires Java Runtime Environment (JRE). The standard start-up and shutdown scripts require Secure Shell to be set up between nodes in the cluster. In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing file-system corruption and reducing loss of data.

Proposed Work

In the existing work there is no automatic system to predict the floods in advance and no proper alert system is maintained. In the proposed work, the data must be collected and analyzed for particular area and various attributes such as the rainfall data that has occurred previous years, surface of the soil etc are taken for analysis purpose. Sensors are important elements in the Flood Observatory System. Further studies on wireless sensor technology will be best to replace the current sensors. Precise and accurate detection of water level will improve the data collection system for the monitoring station. The flood alert information’s can be displayed on LED display boards for road users and for safety reasons could be placed at strategic locations. Such Here analysis of risk is based on these attributes to check whether the flood may cross the limit or not. It has been a challenge in predicting the floods that cause damage i. e., mostly in the jam-packed areas locations which helps us to get can get pre warning of flood menace and is a full preset system. information’s should be in real time and transmitted wirelessly from the measured location. A possible means of power supply for the sensors and centralized control unit is via solar cells.

- Sensors are important elements in the Flood
- Observatory System. Further studies on wireless
- sensor technology will be best to replace the current
- Sensors.
- Precise and accurate detection of water

Data and Methodology

The rainfall dataset is collected for the particular area where the analysis process must be performed to figure out whether there are possibilities of flood occurrence in that region. Analysis of risk is
based on various attributes to determine whether the floods may cross the limit or not. Map Reduce technique is used to present the exact reduced data from the huge quantity of input datasets. This algorithm divides the input datasets into separate pieces of data which are further processed simultaneously to produce the desired reduced data.

Conclusion

Floods which are one of the irremediable calamities destroy the duration of life for many people who are situated around. The certain measures for safety must be considered for floods that mostly occurs in the heavily jam-packed areas. The Map Reduce technique is well suited for springy the exact reduced data from the vast quantity of rainfall datasets which are given as input.

Reference

1. EARLY WARNING AND MAPPING FOR FLOOD DISASTERS