Tangarathnam Mala & Thekkumburath Variath Geetha

Multi Level Categorization Visualization Based on Gestalt Perception Model

1. Introduction

The number of new documents available is huge and it is very difficult now a days to read all the documents and to understand what useful information the document contains. Therefore automated document visualization techniques are available, which helps the user in analyzing a collection of documents, or topics from a single document without going through all the documents. These techniques are most concerned with the visual aids to reveal trends and relationships in documents leading to new insights. They do not merely illustrate the pre existing relationships in the documents. The goal is to create useful and usable visual representations of language in order to reveal novel trends and relationships that exists in linguistic data. Mind map is one such methodology for browsing and exploring topics and concepts in a collection of documents (Spangler, Kreulen et al. 2002, 1170). Mind map generates a radial graph, binary tree and visualizes the document collection in a high dimensional space. Document keyword emphasis visualization (Manber 1997, 132), Visualization of annotated data in a document (Rogers, Gaizauskas et al. 1997, 338), Visualization of topical and temporal information from text (Mala, Geetha et al. 2006, 839) and visualization of meanings of text (Yeap, Reedy et al. 2005, 883) are some of the recent works in the area of document visualization. Here in our work the emphasis is on document category visualization. The category information of the document is visualized by applying the Gestalt perception model.

In Section 2 we describe various document category visualization techniques. Section 3 describes KNN and SRP classifiers which are used for multi level categorization. Section 4 describes the linguistic perception model generated by applying the Gestalt perception model and the visualization performed over it. Section 5 describes the generation of visual perception models. Section 6 explains the nanool principles of visualization applied over the visual models and section 7 analyses the performance of all visual plots generated and the last section describes the usefulness of the generated plots.

GESTALT THEORY
© 2009 (ISSN 0170-057 X)
Vol. 31, No.1, 43-54

2. Document Category Visualization

Effective document category visualization is performed using many approaches. Self-organizing maps are a neural network based graphical tool that automatically categorizes the web documents into manageable sub-spaces so that users can navigate the map to locate documents of interest (Yang, Chen et al.1999, 258). The Interactive Timeline Viewer is a visualization tool used to depict the variants obtained in a textual collation. A textual collation is a process in which a base text is compared against several comparison texts to identify differences (variants) among them (Monroy, Kochumman et al. 2002, 39). Categorized documents are visualized as icons and user feedback information is also collected and a customized output is produced (Aihara 2000, 139). Document category depends on the amount of fitness of the incoming document and the strength of the category either increases or decreases depending on the relationship of the document to the category. A temporal visual structure visualizes this increase in document category strength (Mala & Geetha 2005, 587). All these previous works take only the category information for visual display. In our work the human perception model is also considered for extracting the category information. Moreover multi level categorization is also visualized using the perceptual model. In our approach we use a set of Tamil newspaper documents for categorization and visualization.

3. Multi Level Categorization

The first stage in the work described in this paper is the categorization of the appropriate set of documents. The set of Tamil documents is categorized through a multi level text categorization process. In this work, since a Tamil corpus is considered, special preprocessing needs to be performed before text categorization. Domain neutral words such as articles (athu, ithu, oru), postpositions (pinnal, aruke) and conjuctions (mattrum, melum) etc are first removed from the documents. Tamil is a highly inflectional language and there is a need to convert all inflected words to root words for feature selection. For this purpose, the words of the documents are first preprocessed using a morphological analyzer named Atcharam, a product of RCILTS- Anna University. The dimensionality of the data space is reduced by considering only relevant features, which are maintained as a dictionary for classifier construction from both training and test documents. A document vector is constructed by considering the number of occurrences of all unique words in the document. The standard normalized TF-IDF is used as the weighting function to assign different weights to words in the dictionary.

As a first step of this categorization work, the KNN algorithm (Han, Karypis et al. 2001, 53; Liao & Vemuri 2002, 51; Baoli, Shinwen et al. 2003, 469-475) is used to categorize the test documents to its respective domains based on the training set. To classify a class-unknown document X, the k-Nearest Neighbor

classifier algorithm ranks the document's neighbors among the training document vectors, and uses the class labels of the *k* most similar (similarity based on cosine value between two document vectors) neighbors to predict the class of the new document. The respective domain for the test documents is thus found out using KNN.

The newly designed Set based, Rank based and Priority based technique is then used for sub domain categorization. In the Set based, Rank based and Priority based document categorization method, a predefined dictionary of domain words, constructed semi automatically and prioritized manually, has been used to improve the performance of the categorization algorithm. The term weight is calculated based on the rank, priority and also based on the index of the word in the dictionaries. The term mapped with the different domain dictionaries will have different weights based on their fitness to that domain. Summing up all the term weights calculated thus contribute to document weight, which portrays the fitness of the document to each domain. The document is then classified to the domain that best fits it. This produces a multi level categorization of the documents.

4. Linguistic Perception Model Based on Gestalt Perception Model

Perception models are psychological and behavioral based models. Gestalt perception laws are based on the belief that humans often perceive more than what our physical senses receive as input (Collins 2005, 9). The main principle of the Gestalt perception model is "any whole is greater than the sum of its parts," which means that the whole has properties that cannot be understood by analyzing it into its individual parts [Max Wertheimer et al.,]. This model is an abstraction based model where groups of elements can be completely replaced by a single element which represents the entire collection. Gestalt laws are based on the factors Proximity, Similarity, Continuity, Closure, Figure/Ground, Surroundedness, Smallness/Area, Symmetry and Pragnanz. We consider the first five laws Proximity, Similarity, Continuity, Closure and symmetry for multi level categorization visualization. The above Gestalt laws were mapped onto the multi category information to form a linguistic based model. The mapping was performed based on statistical linguistic information obtained from multi categorization techniques of KNN and SRP classification approaches. *Proximity* is achieved when documents close to each other are perceived to be of the same category. This gives the category proximity between the documents and helps in visualizing documents near each other as belonging to same category. This information is collected from category weights calculated using SRP classifier which categorizes the document based on this value. The *Similarity* attribute perceives documents having similar qualities to be of the same form. The document term frequency - inverse document frequency values considered in

the SRP classifier for calculating the term weight is considered for looking at similarity aspect. C*ontinuity* is achieved when the document is represented by its occurrence of theme words. The position at which the theme word occurs is considered. Theme words are extracted based on the semi automatic dictionary constructed in SRP classifier. *Closure* can well be represented by plotting the cosine similarity values obtained in KNN classifier using open end closed markers to provide a closed figure for similar documents to appear together. *Symmetry* can be explained by plotting category weights of multi category documents where category information is symmetrical.. Later these are converted into visual models for visualization.

5. Visual Perception Models

Visual perception models are generated by mapping the linguistic model with mathematical models to generate visual interfaces. Proximity talks about viewing nearer objects as belonging to one perceptual group. Here we consider the three category weight information of nine documents. The categories considered were Fire, Plane and Tsunami accident documents. The document weight with respect to each category was calculated using the SRP algorithm and these weights were normalized for plotting.

The weight is calculated using the formula given below:

$$\text{Score} = ((\text{Priority}) * \text{Term Frequency}) + ((\text{Term Frequency} / (\text{Index})) * \text{TFIDF}) * \text{Rank}$$

Where,

Score - Term weight calculated for each term present in the dictionary.
Priority - Priority of the term within the set.
Term Frequency - Number of occurrences of the term.
TF IDF - Term frequency / Total words.
Total Words - Length of the document.
Index - Position of the word in the dictionary.
Rank - Set weight, which gives the significance of a set.

Later the document weight is calculated by adding all the score values of all theme words in the document. For normalized plotting the document category weight information is normalized. Each X(i) is calculated using the formula $X(i)=X(i)/\sum x(i)$ * magnification factor. Similarly Y(i) and Z(i) values were calculated using the above formula and thereby the entire data set gets normalized. The magnification factor was considered to be 100 for simplification. The plot for normalized value of 9 documents is as follows.

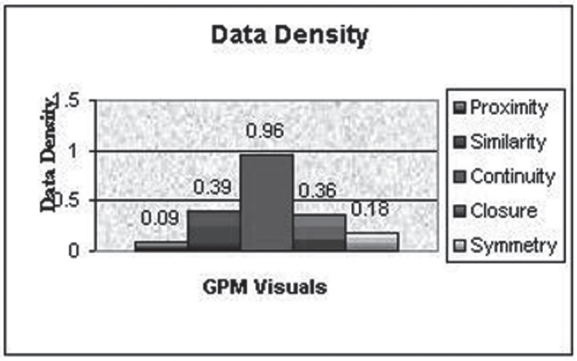

Fig. 1: Proximity

In figure 1 by looking at the data plotted we can group them by their nearness and each group represents a category. In the above case 2 documents are fire documents, 4 belong to plane category and 3 documents belong to tsunami category. This can also be understood by finding the distance between the data points using the formula

$$\mathbf{D = SQRT((X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2)}$$

Where,

D – Distance between two points.

(X_1,Y_1, Z_1) and (X_2,Y_2, Z_2)- Coordinates of two data points in space.

Next the similarity between documents can be viewed using a wavelet approach. Wavelets act as an excellent tool to detect similarity. The document vector formed by considering the TF-IDF values is plotted as a signal. To this document signal a continuous wavelet transformation is applied and the coefficients are collected. The continuous wavelet transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function Ψ. The results of CWT are many wavelet coefficient functions which are function of scale and position. It is given as follows:

$$\mathbf{C(Scale,Position)=\int f(t)\Psi(scale,position,t)dt.}$$

F(t) is the signal to be analyzed, in this case the document vector. In figure 2 the coefficients are plotted for three documents.

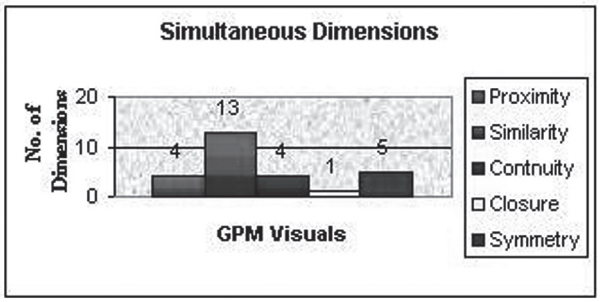

Fig. 2: Similarity

From the above figure it is clearly understood by perception that documents 1 and 2 represented as (a) and (b) are similar and 3 shown as (c) is dissimilar. The next perception principle of Continuity is applied in plotting the theme words. In the SRP classification technique the terms that map the constructed dictionaries of all domains are taken as theme words. The words thus collected which form the basis for multi level categorization are plotted based on their positions.

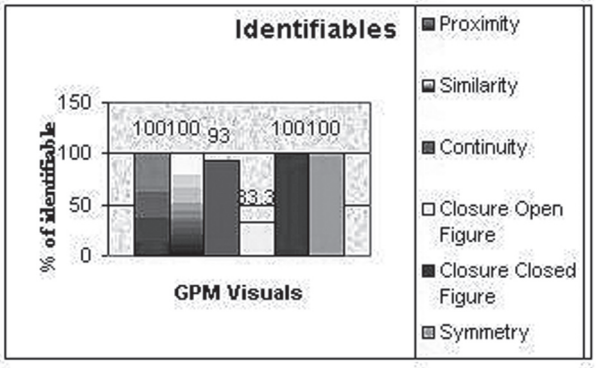

Fig. 3: Continuity

In Figure 3, a document is represented with the theme words and it is perceptually understood that the green line is continuous although it is broken at some places. This visualization represents that the document has words belonging to all the three domains but words of fire domain are more continuously occurring and can be visually perceived as a continuous line even though it is discontinuous at some locations, whereas the other two plots are not continuous and they represent the words belonging to plane and tsunami accidents. From the continuity it is proved that the document belongs to the fire category. The next perception law is Closure.

Document similarity values with respect to a set of documents are calculated and are plotted as scatter plots. A set of 36 documents were taken and are compared with a single document. The cosine similarity values were calculated using the formula

$$\frac{D \times D'}{|D|\ |D'|}$$

where D and D' are the documents to be compared.

The k-Nearest Neighbor classifier algorithm is used to rank the document's neighbors among the training document vectors, and uses the class labels of the k most similar (similarity based on cosine value between two document vectors calculated using the above formula) neighbors to predict the class of the new document. The values are plotted with various shape markers in figure 4. In the visualization generated the data plots with closed markers are easily identifiable points, i.e. those represented as (a) and (d), and the data represented with open ended markers are not easily identifiable, shown as (b) and (c). Especially in scatter plots when the data points cluster in certain locations each individual data cannot be identified clearly when they are represented with open ended markers. This proves the gestalt principle of closure which tells that closed objects are perceived clearly rather than open objects. In the given figure 36 data points were plotted and K value was considered as 4 and the category to which the K nearest neighbors belong to is assigned to the test document, in our case the accident domain is assigned because all the four nearest neighbors belong to accident domain.

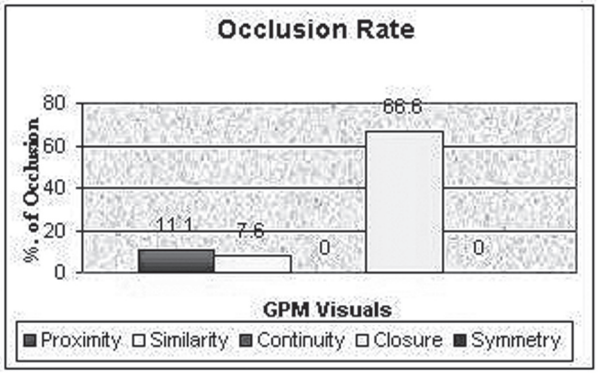

Fig. 4: Closure

The next perception law is symmetry. Symmetry is applicable for multi category documents. Documents that tend to belong to more than one category are multi

category documents. These documents are visually perceived as symmetrical structures when they have equal weightages towards more than one category by applying the SRP algorithm. Multi category documents also have asymmetrical structures when they have same weightings for a set of categories and different weighting for another category. Here in our case in Figure 5 six multi category documents are plotted as pie charts referring to their category weightings for the three categories Tsunami, Plane and Fire accident domains. By the gestalt law of symmetry we perceive symmetrical objects rather than asymmetrical objects. Therefore in the GPM visual generated our eyes tend to perceive all three category weights as symmetrical in the subplots (a), (c), (d) and (e). This means the first, third, fourth and fifth documents have almost equal weightings for all three categories and are multi category documents. The second and sixth documents shown as (b) and (f) subplots have two symmetrical segments and one asymmetrical segment which clearly proves that those two are also multi category documents but they have equal category weightings for two categories and one unequal weighting for another category. Thereby the visualizations tend to show symmetry and asymmetry for multi category documents.

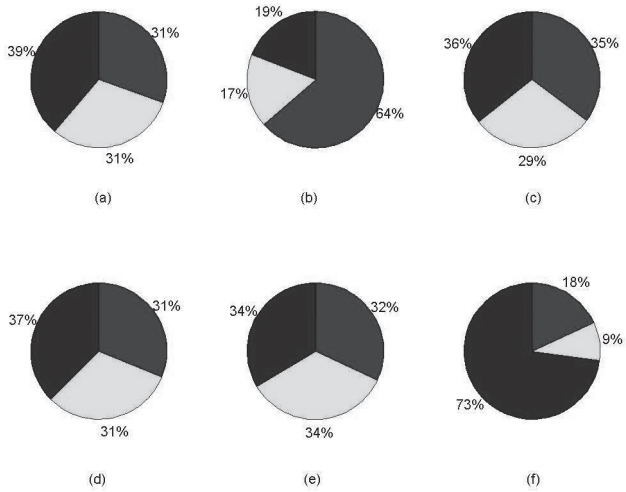

Fig. 5: Symmetry

6. Nanool Principles of Visualization

The VPM is now mapped on to visual parameters. 3D parameters are then projected onto a two-dimensional plane for visual display. The above-generated visual plots are generated for tamil documents. Therefore the visualization performed follows the principles of nanool (Vaarananyanaar, V. 1974, Swaminathaiyer, U. Ve. 1995) a Tamil grammar of the medieval period. Nanool written by Pavanandi has

given ten principles which should be followed by a good literary work. They are followed in our work to produce a good visual output. Multi level categorization information is briefly visualized by using various plots. Eventhough the information is visualized briefly using plots they explain the concept clearly by applying GPM. The visual output provides satisfaction to the user since they are easy to understand. The visual plots are produced by combining aesthetically good visual components like lines, surfaces, meshes and patches. The visual plots are soothing to the eyes since they are viewed with appealing colours. These plots express the insight knowledge such as the relativity of the documents with their categories. Suitable labels are given to the visual plots. Visual plots will not exhaust the user while visualizing since they are not confusing but clear. The visual plots convey the best meaning to the user such as the multi category information. The plots generated are clearly self explaining.

The visual plots of multi level document categorization were plotted following nanool principles. The entire system was developed using JAVA and MATLAB. About 30 Tamil news documents collected from newspapers were used. The documents belonged to Accident, Cinema and Sports categories. The next level of categorization was performed for the accident domain and the documents are once again classified into plane accidents, fire accidents and tsunami. The accuracy of the novel SRP based classifier was found to be 94% for fire accidents, 89% for plane accidents and 87% for tsunami documents. The visual output generated was evaluated based on the efficiency, effectiveness and clarity.

7. Performance Measures and Evaluation

The visual outputs obtained are evaluated based on four factors namely data density, number of simultaneous dimensions displayed, occlusion percentage and number of identifiable points (Bertini & Santucci 2004, 77). The above four metrics are calculated as follows. Data density is given as the ratio of number of data points to the number of pixels. Data density thus gives clearly the amount of data being displayed by effective utilization of space. The number of simultaneous dimension is the number of data attributes displayed at the same time. This metric finds out how efficiently data attributes are displayed. The occlusion percentage is the measure of occluded elements in the visual space, suggesting the reduction of such value as much as possible. This gives the amount of blockage in any visualization. The number of identifiable points is the number of visible data points identifiable in relationship with every other visible data point. This gives the clarity of the data displayed. The above four metrics were calculated for the five visual outputs that were generated (Figure 1 to Figure 5) and graphs are plotted as shown below.

Graphs

From the above graphs the following analysis was performed. The data density in the continuity visualization is good but that of the proximity visualization is very poor. This is because only very few data points are plotted in the proximity visual and can be enhanced by plotting a higher number of data points within the available space. In dimension information the number of data attributes displayed are shown in the graph. The more the dimension displayed the better the visual. Here we can infer that similarity has maximum dimension displayed whereas closure has least dimension displayed. This is because of the fact that the similarity visual takes into consideration the document vector which is multi dimensional but closure looks only at document weight with respect to three categories. The next graph shows the number of identifiable points where in closed figure, proximity, similarity and in symmetry it is maximum and in open figure it is minimum. This is due to the overlapping of points in the open marker figure which is not clearly identifiable. The next graph shows occlusion percentage calculated as number of breakages relative to the whole figure. Occlusion was maximal in the closure visualization and at a minimum in symmetry and continuity. In the symmetry and continuity visuals there is no blockage to what we perceive and understand but in closure there is a block in the elements being perceived due to heavy data collisions.

8. Conclusion and Future Work

Documents are visualized with their category information using various plots. The Gestalt perception model is used to produce a visual perception model which clearly shows the documents' relationships among themselves and with the categories to which they are associated by perception. Gestalt laws help in perceiving the document category information, multicategory documents, multiple level of category information. The degree of categorization and similarity among documents can also be visualized. Nanool principles were applied to generate proper visual outputs. The performance of the system was analyzed based on visual perception parameters. The system can be enhanced by applying more complex mathematical models to the generated visual outputs. Fully automated structures can be developed for visualization. They can be made more interactive.

Summary

This paper describes a visual information system that helps in visualizing the categorization of Tamil news documents applying a Gestalt Perception model. This automated system uses average weighted K nearest neighbor and a newly designed novel set based, rank based, priority based method as classifiers to automatically categorize the documents at multiple levels. The laws of the Gestalt perception model like similarity, proximity,

continuity etc. are then mapped onto the statistical language parameters of these classifiers to obtain a linguistic perception model. Later visual mapping is performed which converts these mapped parameters into various plots. This visual interface helps the user to explore the category information of news items by using statistical information like cosine similarity values, inverse document frequencies, weightage of the document with respect to the category and the words that are related to category information. With these visual displays, the user can visually perceive and evaluate the representations and make an intuitive judgment about the relevance and relationship of the documents to their respective categories without having to read a significant portion of each document. The Gestalt perception model helps in analyzing the documents with multi category information and multi level categorization information by just looking at the visuals generated. These modeling techniques help in the visualization of the categorization information from different perspectives.
Keywords: Gestalt principles, Document visualization, Category visualization, Linguistic perception model.

Zusammenfassung

In diesem Beitrag wird ein visuelles, auf Gestaltwahrnehmung basierendes Informationssystem beschrieben, mit dem die Kategorisierung tamilischer Nachrichtendokumente unterstützt werden kann. Dieses automatisierte System verwendet Gesetzmässigkeiten der Gestaltwahrnehmung wie Ähnlichkeit, Nähe, Kontinuität u.a., die auf statistischen Sprachparametern abgebildet werden, um auf diese Weise ein linguistisches Wahrnehmungmodell zu erhalten. Am Ende steht ein visuelles Interface, das dem Anwender hilft, die vorliegenden Informationen zu kategorisieren. So können mit Hilfe des Interface Repräsentationen visuell wahrgenommen und evaluiert werden, sodass der Anwender die Dokumente intuitiv auf ihre Relevanz und Beziehung zu den entsprechenden Kategorien beurteilen kann, ohne jedes Dokument im Detail lesen zu müssen. Das Gestalt-Wahrnehmungs-Modell ermöglicht die Analyse von Dokumenten sowohl mit multi-kategorieller als auch mit mehrstufiger Information durch einfache visuelle Wahrnehmung der Repräsentationen. Die beschriebenen Darstellungs-Modelle unterstützen die Visualisierung der Kategorisierungsinformation in unterschiedlichen Perspektiven.
Schlüsselwörter: Gestaltprinzipien, Dokumentenvisualisierung, Kategorienvisualisierung, Linguistisches Wahrnehmungsmodell.

References

Baoli, L., Shinwen, Y. & Qin, L. : (2003): An Improved k-Nearest Neighbor Algorithm for Text Categorization, in *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*, June 2003, 469-475. Beijing: Tsinghua University Press.

Bertini, E. & Santucci, G. (2004): Quality metrics for 2D Scatterplot Graphics: Automatically Reducing Visual Clutter, Smart Graphics, 77-89. Canada: Springer.

Collins, C.M. (2005): A Critical Review of Information Visualizations for Natural Language. Doctoral thesis, Toronto, University of Toronto, PhD qualifying exam paper.

Han, E.H., Karypis, G. & Kumar, V. (2001): Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification, in *proceedings of pacific Asia conference on knowledge discovery and data mining,* 53-65. London: Springer-Verlag.

Kenro, A. & Atsuhiro, T. 2000): Accessing Information Space through Text Categorization, in *first International symposium on Advanced Informatics*, March 2000, 139-144. Tokyo, Japan: National centre for Science Information Systems Press.

Liao, Y. & Vemuri, R.V. (2002): Using Text categorization techniques for Intrusion Detection, i*n Proceedings of the 11th USENIX Security Symposium,* 2002, 51-59, Berkeley, CA, USA: USENIX Association.

Mala, T., & Geetha, T.V (2005): Visualization of Biological Patterns in Event Detection and Tracking Based on SRP algorithm, in *Proceedings 9th International IEEE conference on Information Visualization* IV05, July 2005, 587-592. Los Alamitos, CA, USA: IEEE Computer Society.

Mala, T., Geetha, T.V. & Sathish, K. (2006): Topical and Temporal Visualization Using Wavelets, in *proceedings of ninth Pacific Rim International conference on Artificial intelligence*, PRICAI, LNAI 4099, August 2006, 839 – 843. Berlin: Springer.

Manber, U. (1997): The Use of Customized Emphasis in Text Visualization , in *Proceedings of the IEEE Conference on Information Visualization,* IV'97, 1997, 132, Washington, DC, USA: IEEE Computer Society.

Monroy, C., Kochumman, R., Furuta, R. & Urbina, E. (2002): Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Document. *Second international workshop on visual interfaces to digital libraries, ACM/IEEE joint conference on digital libraries,* July 2002, 39-49. Texas: Springer-Verlag.

Rodgers, P., Gaizauskas, R., Humphreys, K. & Gunningham, H. (1997): Visual Execution and Data Visualization in Natural Language Processing, in *proceedings of IEEE conference on Information Visualization*, IV97, 1997, 338, Los Alamitos, CA, USA: IEEE Computer Society.

Spangler, S., Kreulen, J.T. & Lessler, J. (2002): MindMap: Utilizing Multiple Taxonomies and Visualization to Understand a Document Collection, in *Proceedings of 35th Hawaii international conference on System Sciences*, January 2002, 1170-1179, Los Alamitos, CA, USA: IEEE Computer Society.

Swaminathaiyer, U.Ve. (1995): *Pavanandi munivar iyatriya nanool moolamum mayilainathar uraiyum*. Besantnagar, Chennai: Noolnilayam.

Vaarananyanaar, V. (1974): "*Tolkaapiyam-Nannool Ezhuthathikaaram*", Tamil Nadu: Annamalai University, 3rd Edition.

Yang, C.C., Chen, H. & Hong, K. (1999): Visualization tools for Self-Organizing Maps, in *Proceedings of the fourth ACM conference on Digital libraries,* August 1999, 258–259. California: ACM.

Yeap, W.K., Reedy, P., Min, K. & Ho, H. (2005): Visualizing the Meaning of Text, in *Proceedings of the Ninth International Conference on Information Visualization*, IV05, July 2005, 883-888, Los Alamitos, CA, USA: IEEE Computer Society.

Tangarathnam Mala has completed her B. E. (ECE), MCA and obtained M. E. (Multimedia Technology). At present she is working as Lecturer in the Department of Computer Science & Engineering, College of Engineering, Guindy, Anna University, Chennai. She has published articles in reputed international conferences and in refereed international journals.

Address: Department of Computer Science and Engineering, College of Engineering, Guindy Campus, Anna University, Guindy, Chennai-600025, Tamil NAdu, India..

E-Mail: mala@cs.annauniv.edu; malanehru@annauniv.edu

Thekkumburath Variath Geetha, M.E., PhD. is presently Professor in the Department of Computer Science & Engineering, College of Engineering, Guindy, Anna University, Chennai. Her areas of interests include Natural language processing, Intelligent Databases, Distributed Artificial Intelligence and Data Mining. She has published articles in international conferences and in refereed international journals like IEEE transactions.

Address: Department of Computer Science and Engineering, College of Engineering, Guindy Campus, Anna University, Guindy, Chennai-600025, Tamil NAdu, India.

E-Mail: tvgeedir@cs.annauniv.edu; rctamil@annauniv.edu