ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

HUMAN INTERACTION PATTERN IDENTIFICATION USING STEMMING IN GENETIC ALGORITHM

[1] S.Queenkirubaananthy, [2] D. Chitra,
[1] PG Scholar, [2] Assistant professor,
[1, 2] Department Of Computer Science,
[1, 2] Government Arts College (Autonomous)
[1, 2] India.

ABSTRACT: Human interaction discovers to the knowledge to (higher) someone. It everything happened in meeting, debate, interview etc. Nature of the person is represented through behavior and mining technique helps to analyze the opinion a person exhibits. Discovering semantic knowledge is significant for understanding and interpreting how people interact in a meeting discussion. Different human interactions, such as proposing an idea, giving comments, and acknowledgements, indicate user intention toward a topic or role in a discussions. To identify the problem of detecting more number of interaction patterns made in the meeting to gain semantic knowledge of meetings by using GA inherit the techniques of STA. Human interaction flow is represented by Tree structure. To identify the human intentions clearly while the people interaction the meeting by using STA technique.

Keywords: [Genetic Algorithm (GA), Stemming Technique Algorithm (STA), Partial Least Square (PLS), Sequential PAttern Discovery using Equivalence classes (SPADE), Latent Semantic Analysis (LSA), Singular Value Decomposition (SVD).]

INTRODUCTION

Image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. The common steps in image processing are image scanning, storing, enhancing and interpretation. It aims in detecting more number of interesting patterns with minimum execution time. So, in this work ABC algorithm is enhanced using Partial Least Square (PLS) mechanism in order to extract frequent interaction among patterns. In order to have powerful algorithm, we have used PLS mechanism in population initialization, so that solutions are generated uniformly within the search space. This helps to generate at least some points in the neighborhood of global solution. Human interaction flow in a discussion session is represented as a tree. Human cooperation is identified by whether the gathering was decently composed or not.

Data mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication.

Related work

M. Rost, L. Barkhuus, H. Cramer, and B. Brown (2013) says Representation and Communication: Challenges in Interpreting Large Social Media Datasets represented according to their algorithm a user who tweets about ‘Crawfish’ is more likely to be in Louisiana, and a user who tweets about ”Clegg” (the British deputy prime minister) in the U.K. The main advantages of this paper are reduction of number of variables, by combining two or more variables into a single factor. For example, performance at running, ball throwing, batting, jumping and weight lifting could be combined into a single factor such as general athletic ability. The Disadvantages are Factor analysis can be only as good as the data allows. In psychology, where researchers often have to rely on less valid and reliable measures such as self-reports, this can be problematic.

R. Ferguson (2012) says The State of Learning Analytics in 2012: A Review and Future Challenges this Algorithms can recognize and provide insight into data and at risk challenges. The main advantages of this paper are ability to write Map Reduce programs in Java, a language which even many non-Computer scientists can learn with sufficient capability to meet powerful data processing needs. The Disadvantages are One input two phase data flow rigid, hard to adapt Does not allow for stateful multiple step processing of records

M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R.Traveler, and K. Smith (2008) says Academic Pathways Study: Processes and Realities Algorithm mainly AVCTP this algorithm is used. The Advantages are it provides an opportunity to understand and assess what the participant intended to do through his or her actions. Gives participants opportunity to correct errors and challenge what are perceived as wrong interpretations. The Disadvantages are Research findings cannot be easily conveyed to the participant. This is most commonly the case in addiction and criminal research. Additionally, findings may be difficult to understand.

C. Moller-Wong and A. Eide (1997) says Understanding freshman engineering student retention through a survey says Before data mining algorithms can be used, a target data set must be assembled. The Advantages are Improved open access to education, including access to full degree programs. Better integration for non-full-time students, particularly in continuing education. The Disadvantages are Ease of cheating. Bias towards tech-savvy students over non-technical students.

J. M. DiMicco and D. R. Millen (2007) says Identity Management: Multiple Presentations of Self in Face book the Advantages are Provision of tools to enable students to independently solve problems. Acquisition of technological skills through practice with tools and computers

The Disadvantages are a synchronic communication hinders fast exchange of question. Danger of procrastination. Unforeseen technical difficulties may impede learning.Managing Identity across Social Networks by M. Vorvoreanu and Q. Clark (2010) says Computers and society, Social issue. The Advantages are in a typical tagging system, there is no explicit information about the meaning or semantics of each tag, and a user can apply new tags to an item as easily as applying older tags. The Disadvantages are Those users of tagging systems tend to notice the current use of "tag terms" within these systems, and thus use existing tags in order to easily form connections to related items

J. Pei et al (2001) by PrefixSpan: Mining Sequential Patterns Efficiently by

Prefix-Projected Pattern Growth In this paper author develop a novel sequential pattern mining method, called PrefixSpan (i.e., Prefix-projected Sequential pattern mining). Its general idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns. To further improve mining efficiency, two kinds of database projections are explored: level-by-level projection and bi-level projection. Moreover, a main-memory-based pseudo-projection technique is developed for saving the cost of projection and speeding up processing when the projected (sub)-database and its associated psuedo-projection processing structure can fit in main memory. Their performance study shows that bi-level projection has better performance when the database is large, and pseudo-projection speeds up the processing substantially when the projected databases can fit in memory. PrefixSpan mines the complete set of patterns and is efficient and runs considerably faster than both Apriori based GSP algorithm and FreeSpan. The Prefix span algorithm and free span algorithm are used. PrefixSpan mines the complete set of patterns and is efficient and runs considerably faster than both Apriori based GSP algorithm and FreeSpan. Among different variations of PrefixSpan, bi-level projection has better performance at disk-based processing and psuedo-projection has the best performance when the projected sequence database can fit in main memory.

The Limitations are GSP always searches in the original database. Many irrelevant sequences have to be scanned and checked, which adds to the unnecessarily heavy cost. FreeSpan cannot gain much from projections, whereas PrefixSpan can cut both the length and the number of sequences in projected databases dramatically.

M. J. Zaki (2001) says SPADE: An Efficient Algorithm for Mining Frequent Sequences represented in this paper author presents SPADE, a new algorithm for fast discovery of Sequential Patterns. The existing solutions to this problem make repeated database scans, and use complex hash structures which have poor locality. SPADE utilizes combinatorial properties to decompose the original problem into smaller sub-problems that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. All sequences are discovered in only three database scans. The SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm is used in this paper. SPADE outperforms the best previous algorithm by a factor of two, and by an order of magnitude with some pre-processed data. It also has linear scalability with respect to the number of input-sequences, and a number of other database parameters. SPADE not only minimizes I/O costs by reducing database scans, but also minimizes computational costs by using efficient search schemes. The vertical id-list based approach is also insensitive to data-skew. An extensive set of experiments shows that SPADE outperforms previous approaches by a factor of two, and by an order of magnitude if they have some additional off-line information. Furthermore, SPADE scales linearly in the database size, and a number of other database parameters. The Limitations are they observed that simple mining of frequent sequence produces an overwhelming number of patterns, many of them trivial or useless.

SYSTEM MODEL

MODULES

1. GENETIC ALGORITHM
2. BAGGING
3. TREE PRUNING
4. LSA
5. STA

MODULE DESCRIPTIONS

1. GENETIC ALGORITHM

Genetic Algorithm is used to optimize the Memory through Human Interaction in

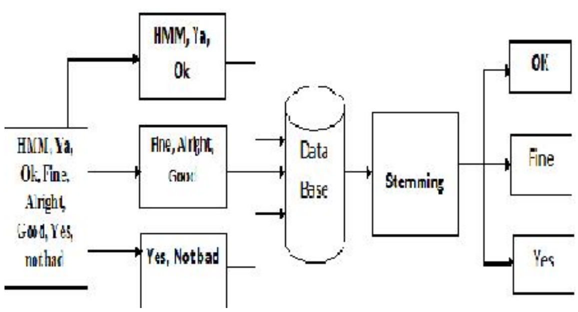
Meeting. A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions.

2. BAGGING

Bagging Builds Multiple Decision Trees-Consensus Prediction. This model represented by number of tree structure. Then it will be give a "Consensus Prediction".

3. TREE PRUNING

Tree pruning is used to remove the irrelevant data in classification processing.
DESIGN



4. LSA

Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns. Words are then compared by taking the cosine of the angle between the two vectors (or the dot product between the normalizations of the two vectors) formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.

5. STEMMING - STA:

Hence STA used to identify the human objective. Stemming is used to split the sentence partition by word segment then it stored in database. Then it will be goes to the final stage for word identification. Here using three types of processing for word identification.

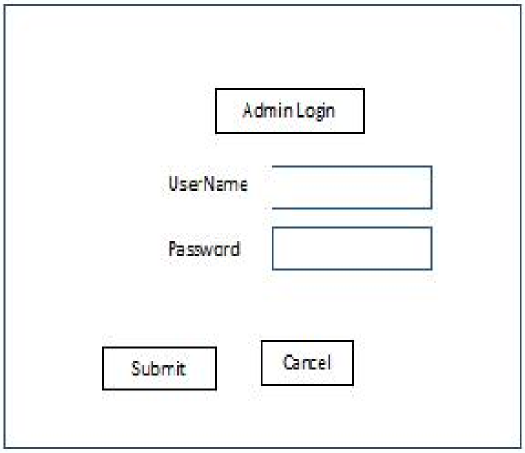
1. Splitting
2. Remove irrelevant data
3. Identify the exact word that is human objective.

SKIN COLOR TONE DETECTION

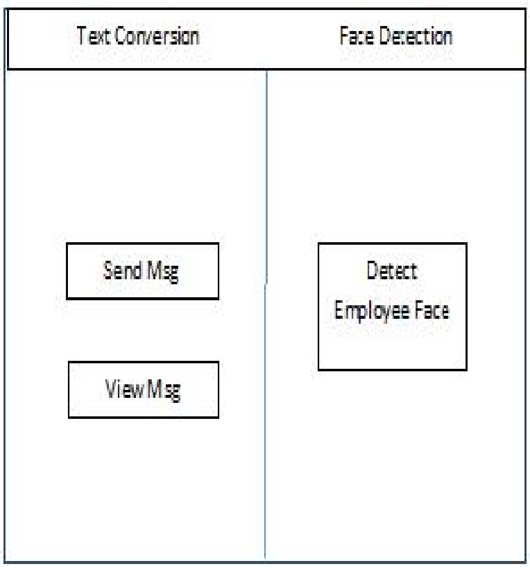
A skin detector typically transforms a given pixel into an appropriate color space and then uses a skin classifier to label the pixel whether it is a skin or a non-skin pixel. A skin classifier defines a decision boundary of the skin color class in the color space. Important challenges in skin detection are to represent the color in way that is invariant or at least insensitive to changes in illumination. Another challenge comes from the fact that many objects in the real world might have skin-tone color. This causes the skin detector to have much false detection in the background. The simplest way to decide whether a pixel is skin color or not, is to explicitly define a boundary. RGB matrix of the given color image converted into different color space to yield distinguishable region of skin or near skin-tone. Color space used for skin detection is HSV. Sobottaka and pitas defined a face location based on HSV. They found that human flesh can be an approximation from a sector out of a hexagon with the constraints: Smin=0.23, Smax=0.68, Hmin=0º and Hmax=50º.

RESULTS

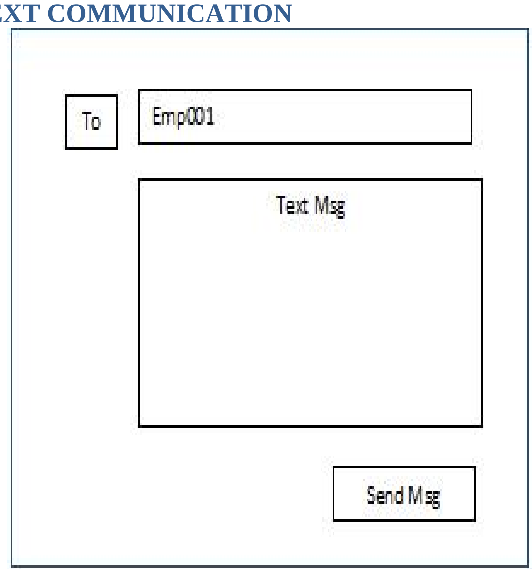
ADMIN SIDE



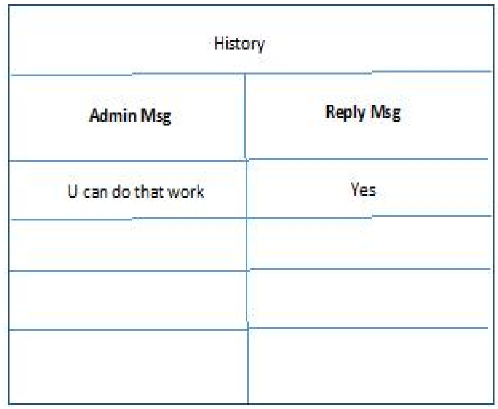
TEXT CONVERSION



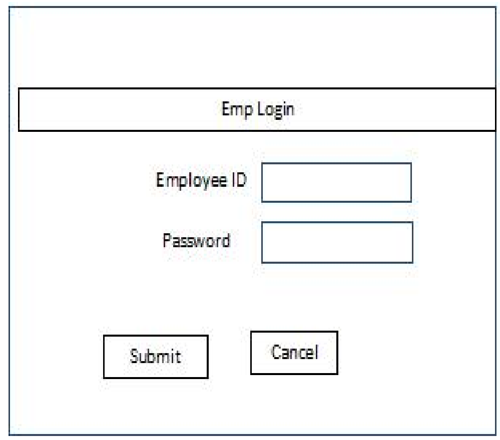
TEXT COMMUNICATION



VIEW ALL HISTORY MESSAGE



EMPLOYEE SIDE



FACE DETECTION IN ADMIN SIDE

CONCLUSIONS

Based on the communications among the people present in the meeting able to retrieve a pattern for each meeting. Mining results can be used for interpreting human interactions in the meetings. Using this STA technique to identify the human intentions clearly while the people interaction the meeting. Using this GA algorithm to identify the problem of detecting more number of interaction patterns made in the meeting to gain semantic knowledge of meetings. As future work, plan to perform clustering based on the interaction patterns to identify the behaviour of each individual in the meeting, thus exploring the involvement of each person in the meeting.

FUTURE WORK

There are several avenues for future research. First will investigate the human interaction. Second will study methods for improving the pattern identification. Third will extend our proposed work. As future work, plan to perform clustering based on the interaction patterns to identify the behaviour of each individual in the meeting, thus exploring the involvement of each person in the meeting.

REFERENCES

[1].J. Pei et al., "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," in Proc. 17th ICDE, Berlin, Germany, 2001.

[2].M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequenes," Mach. Learn., vol. 42, no. 1–2, pp. 31–60, 2001.

[3].A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in Proc. 13th Int. Conf. VLDB, Toronto, ON, Canada, 2004.

[4].Q. Zhang, F. Li, and K. Yi, "Finding frequent items in probabilistic data," in Proc. ACM SIGMOD, Vancouver, BC, Canada, 2008.

[5].C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[6].T. Bernecker, H. P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[7].T. Bernecker, H. P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[8].C. Gao and J. Wang, "Direct mining of discriminative patterns for classifying uncertain data," in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.

[9].H. Chen, W. S. Ku, H. Wang, and M. T. Sun, "Leveraging spatiotemporal redundancy for RFID data cleansing," in Proc. ACM SIGMOD, Indianapolis, IN, USA, 2010.

[10].X. Lian and L. Chen, "Set similarity join on probabilistic data,"in Proc. VLDB, Singapore, 2010.

[11].J. Jestes, F. Li, Z. Yan, and K. Yi, "Probabilistic string similarity joins," in Proc. ACM SIGMOD, Indianapolis, IN, USA, 2010.

[12].L. Wang, R. Cheng, D. Lee, and D. Cheung, "Accelerating probabilistic frequent itemset mining: A model-based approach," in Proc. 19th ACM CIKM, Toronto, ON, Canada, 2010.

[13].M. Muzammal and R. Raman, "Mining sequential patterns from probabilistic databases," in Proc. 15th PAKDD, Shenzhen, China, 2011.

[14].Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in uncertain databases," in Proc. 15th Int. Conf. EDBT, New York, NY, USA, 2012.

[15].Y. Tong, L. Chen, Y. Cheng, and P. S. Yu, "Mining frequent itemsets over uncertain databases," in Proc. VLDB, Istanbul, Turkey, 2012.