# Pattern Mining for Human Interactions in Group Discussions using Tree-based Mining Algorithm

Samreen Sadaf Qazi F.
ME(Computer)-II
Vishwakarma Institute of IT, Pune

Prof. Leena A. Deshpande
Assistant Professor
Vishwakarma Institute of IT, Pune

**Abstract-** Human-human interaction is one of the most important characteristics of any group activity. Group activities could be anything like meetings, group discussions, debates, panel discussions, presentations, and sports, etc. Research in such field generally includes: gesture, gaze, haptic, and the speech. But the human style of interactions and communication poses lots of challenges for researchers from technological perspective.

To discover the frequent patterns from human-human interactions, we are going to use tree-based pattern mining algorithm. Patterns will be mined from the text file containing captured contents of face-to-face group discussions. For this, discovery of higher-level semantic knowledge is necessary for interpreting and understanding of whether the meeting discussion was meaningful and how people interact in a discussion and what were the major opinions (frequent patterns) of the discussion. Human interactions like asking an opinion, proposing an idea, giving comments, and expressing a negative/positive opinion regarding a topic indicate user intention towards a topic or role in a discussion. These user intentions are categorized using N-gram technique. This tree-based pattern mining algorithm with n-gram technique could be useful in interpreting the user behavior in group activity, like determining the flow of interactions, frequent interactions, and relationships between different types of interactions. The discovered patterns could be utilized to check whether a discussion is efficient and to make strategies for future group activities**.**

**Index Terms-** frequent pattern mining, tree-based pattern mining, interaction pattern, human-human interactions

## 1 INTRODUCTION

For many years, frequent pattern mining is an important topic for researchers in the field of data mining. Frequent patterns are patterns that appear in a data set frequently, such as itemsets, subsequences, or substructures. An itemset A (or subsequence, or substructure) is said frequent if it satisfies the predetermined minimum support count, where support count of A is the number of occurrences of from the total number of itemsets in a dataset.

A human interaction in any group activity has attracted much research in the fields of image and speech processing, human-computer interaction, human-human interaction and computer vision. Work related with finding frequent patterns in human interactions is done considering human actions, gestures, and gaze. Patterns can be mined from meeting, group discussion, debate, interviews, and sports. Now a days, meeting data, sports data is increasing immensely. Because of that analysis of such data is getting more importance. Discovering meaningful information from interaction of any group activity record is difficult because these records do not contain any structural information.

Human interaction is an important characteristic of any group activity that can be analyzed. Group activity could be anything: it could be meeting, group discussion, panel discussion, debate, sports, etc. In meetings status is discussed, new decisions are made, alternatives are considered, details of topic are explained, information is presented, new ideas are generated and strategies are made. Therefore, such group activities contain a large amount of information that is often not formally and properly documented. Capturing all of this information has been a research topic in several communities over the past decade. Patterns can be mined from group discussions and debates to analyze them as what the majority of views are. Research in the field of human-human interactions may include research on gesture, gaze, speech etc. The human style of communication poses lots of challenges in front of researchers. Extracting knowledge from human interactions could be used for meetings by discovering patterns of interaction, conclusion etc. It could also be used in sports for making winning strategies. Manual hand-recorded notes have several drawbacks. It is time consuming; also it requires extra focus of any member taking hand-written notes involved in discussion. Moreover, it could be biased because of human nature.

## 2 RELATED WORK

Now a days, meeting data, sports data is increasing immensely. Because of that analysis of such data is getting more importance. Discovering meaningful information from interaction of any group activity record is difficult because these records do not contain any structural information.

Time series and sequential data mining is one of the most important problems from 10 challenges identified in paper [1]. Clustering, classification, and trend prediction of these data is an important open research topic. Another problem identified in [1] is mining complex data in the form of graphs.

Human interaction in meetings or any group activity has attracted much research work in the fields like image processing, speech processing, human computer interaction and computer vision. Mining higher level knowledge is important for understanding the interaction contents and it can be used for indexing interaction contents [2].

Interaction between two persons in a video from TV show is recognized in [3]. It deals with four interaction classes: handshakes, high fives, hugs, and kisses. A mining method for extracting the interaction patterns from multimodal interaction is proposed in [8]. It extracts the simultaneously occurring patterns such as gaze and speech, which is set of LOOK and SPEAK events.

In paper [4], authors have used mining algorithm for product session data for improving efficiencies. They have proposed fuzzy mining algorithm which transforms every quantitative value into fuzzy set which includes linguistic terms. Then it calculates scalar cardinality of every linguistic term for all transaction data. After that fuzzy association rules are mined. Paper [5] focuses discovering higher-level knowledge about human interaction from perspective of data mining. Interaction tree pattern mining algorithm is designed to analyze tree structures and extract interaction flow patterns. Authors of [5] have adopted a multimodal method to infer human interactions based on various features like gestures, attention, speech tone, and speaking time, interaction occasion, and information about the previous interactions.

Mining frequent tree pattern is an important open research area. Previous research studies highly suggest the pattern growth method for efficient pattern mining. Authors of [6] have developed a pattern growth method for mining frequent tree patterns. Two algorithms, Chopper and XSpanner, have been devised, from which XSpanner algorithm is faster than Chopper. These two algorithms perform better than TreeMinerV of M.J.Zaki, Efficiently mining frequent trees in a forest of KDD02. In that, Chopper consists of two separate phases (i) mining sequential patterns and (ii) the extraction of frequent tree patterns. It generates and tests all possible tree patterns of the database. XSpanner algorithm combines these two phases of Chopper algorithm.

In paper [7], the problem of mining sub-trees in a forest(database) of rooted, labeled and ordered tree is formulated. And TREEMINER algorithm is presented to discover all frequent subtrees in a forest that uses a new data structure called scope-list. Tree mining algorithm is also applied to analyse RNA structure and phylogenetic data sets from bioinformatics domain. Authors of [7] have developed a framework for nonredundant pattern generation. There work allows to discover all subtrees in a forest of tree database and all subtrees in a single large tree.

Tree based pattern mining algorithm is generally used in analyzing tree structures and in extracting the interaction patterns. Tree-based pattern mining algorithm is suggested in paper [5] and is used for finding patterns of human interactions in meeting.

The above frequent pattern mining algorithms may often generate an enormous number of frequent patterns that may contain a lot of redundancy. And this can create a challenge on understanding, visualizing and analyzing the generated patterns. Therefore, authors of [11] have proposed an algorithm called MinRPset to generate minimum representative patterns that best approximate the all other patterns. It produces the smallest solution for a given problem. But MinRPset is very time-consuming and space-consuming too. So, authors have developed another algorithm to overcome this problem and proposed FlexRPset which allows users to trade-off between efficiency and result-size.

## 3 PROPOSED SYSTEM

A smarter pattern mining system for human interactions in group discussion is designed which classifies the input and extracts the frequent patterns from human interactions. That can be used to ensure the performance analysis of human interactions and to make winning strategies.

### 3.1 Methodology

This system will take text file containing conversation of group discussion as an input and will generate frequent patterns of interaction as output. It will then parse the input interaction, identify the roles (persons) involved the discussion and assign the class label to persons' statements. System will generate interaction tree and subtrees of it.

A tree is used to represent the flow of interaction, this will be acyclic connected graph maintained by TD, a Tree Data structure. Tree is denoted as T = (V, E) where, V = {V1,V2,…..,Vn}, is the set of vertices representing individual interactions E = {( Vi,Vj) | Vi,Vj∈ V, Vi ≠ Vj}, is the set of edges connecting the interactions Vi is parent of Vj node (Vj responses toVi). Subtree support count is calculated to detect the frequent pattern from interaction tree. Isomorphic subtrees are detected and after keeping one of them other isomorphic subtrees are discarded.

Given two trees T1 = (V1,E1) and T2 = (V2,E2). If tps(T1) ≠ tps(T2) and through exchanging the places of siblings on T1 or T2, tps(T1) = tps(T2), then T1 and T2 are isomorphic trees.

Tree-based pattern mining algorithm used in paper [5] for meeting purpose is used here to find frequent patterns from interactions. Tree structure is used to represent an interaction flow in a session and string constructed by depth first search tree traversal will represent the tree formally. Then isomorphic trees are identified to detect the same tree structure pattern in interaction tree.

Each tree will denote one session of interactions that begins with spontaneous interaction like propose or ask opinion. Following Fig. represents an example of tree representation of interaction.

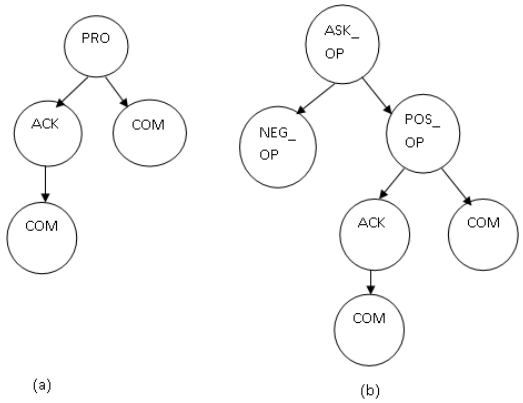


Fig.1 : Example of an interaction flow tree

Patterns are frequent tree structures or subtrees in the tree database. Patterns are discovered from the database based on the support of tree T in the tree database TD. From Fig.(b) pattern like ASK_OP-POS_OP-(ACK*COM) could be found.

Following are the steps involved in mining patterns from human interactions.

- Pre-processing of input content (text file containing conversation)
- Identify the roles (persons involved in the interaction)
- Categorization (assigning the labels)
- Identify the flow (who follows the whom)
- Construct interaction tree
- Generate subtrees of interaction tree
- Detect isomorphism and calculate the support count of subtrees
- Find the frequent patterns and conclusion of interaction

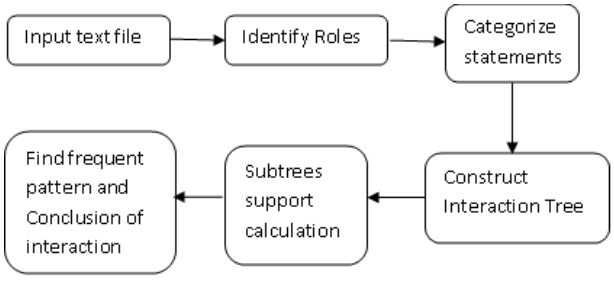


Fig.2 : System Architecture

### 3.2 Preprocessing

Some pre-processing is required before generating interaction tree from discussion. Following three preprocessing steps are performed:

- Tokenization:
  In this, word boundaries are identified. In most languages, it is done with respect to spaces and punctuation marks.
- Stopwords Elimination:
  Stopwords affect the performance of studies because they occur very frequently in texts. Therefore, to overcome this effect stopwords elimination is performed.
- Stemming:
  In stemming, prefixes and suffixes are removed to derive the stemmed words for comparing different words. For each input sentence conflation algorithm is applied with Porter Stemming algorithm.

### 3.3 N-gram Algorithm

For assigning the class labels to each statement n-gram technique is used. Four-grams are assigned the highest weightage among tri-grams, bi-grams and uni-gram because if four consecutive words of statement and of any class are matched then it is highly possible that the statement belongs to that category. Following are the steps involved in categorization of statements in ASK_OPINION, PROPOSAL, POSITIVE_OPINION, NEGATIVE_OPINION, COMMENT, and ACKNOWLEGE.

- Initialize fourGramTokens, triGramTokens, biGramTokens, and uniGramTokens lists for each class label ASK_OPINION, POSITIVE_OPINION, NEGATIVE_OPINION, COMMENT, ACKNOWLEDGE
- Initialize weightage constants for fourGram=20, triGram=10, biGram=5, uniGram=2
- If a fourGramToken is found in the sentence then associate particular class label along with its weightage to the statement
- Do similar for triGramToken, biGramToken, and uniGramToken
- Assign class label having maximum weightage to the input sentence

### 3.4 Tree Generation Algorithm

For each statement do

- If class label of statement is ASK_OPINION or PROPOSAL then create new root node for tree
- Else if class label of statement is POSITIVE_OPINION or NEGATIVE_OPINION then add child to last opinion node
- Else if class label of statement is COMMENT or ACKNOWLEDGE and if last opinion is null then add child to root node otherwise add child to last opinion

After tree generation subtrees support is calculated using the subtree support calculation algorithm used in

[5]. In that isomorphic subtrees are also detected and after keeping one of them all other isomorphic subtrees are discarded.

## 4 DATASETS AND RESULTS

Our work involves pattern detection from small interactions consisting 10 group discussions on cricket and 5 discussions on general topics. Each discussion had four to six participants. Their interaction in text file is given as an input to our system and they are categorized using n-gram technique. Patterns are mined using tree based pattern mining algorithm of [5]. One sample interaction is given below:

| | |
|---|---|
| Vince: | Do you think the toss winning team should bat first? |
| Eric: | Yes I agree with you. |
| Olivia: | No, I think they should bowl first. |
| Tresa: | Yes, they should bowl first as the outfield looks wet. |
| Audry: | Yeah, also there will be good bounce on the pitch. |

It is categorized as follows:

| | | |
|---|---|---|
| Vince: | ASK_OPINION | Do you think the toss winning team should bat first? |
| Eric: | ACKNOWLEDGE | Yes I agree with you. |
| Olivia: | NEGATIVE_OPINION | No, I think they should bowl first. |
| Tresa: | ACKNOWLEDGE | Yes, they should bowl first as the outfield looks wet. |
| Audry: | ACKNOWLEDGE | Yeah, also there will be good bounce on the pitch. |

Following tree is generated using our tree generation algorithm mentioned above having two subtrees with subtree support count of 0.5.

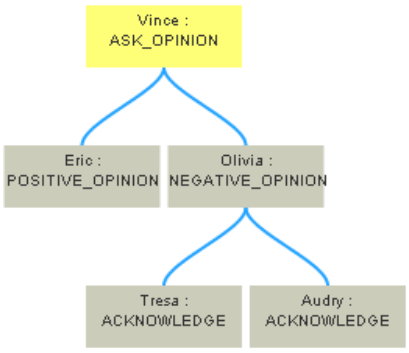


Fig. 3a: Interaction Tree

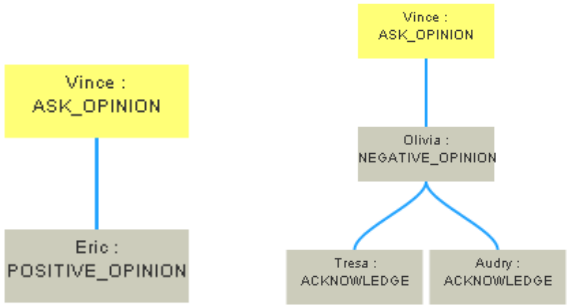


Fig. 3b: Interaction Subtrees

N-gram technique used here assigns the labels to the statements. From the first group of interactions of cricket, 86% of statements of 10 interactions have been correctly categorized. From the second group of interactions, 83% of the statements have been correctly categorized.

Based on the subtrees support count and the subtree string code generated percentage of positive and negative opinions are calculated. Conclusion of first group of interactions consists of 23% positive opinions and 77% negative opinions and conclusion of second group of interactions consists of 59% positive opinions and 41 of negative opinions.

## CONCLUSION AND FUTURE SCOPE

This pattern mining system will discover the frequent patterns from human interactions in group discussions. Mined frequent patterns can be used in various applications for predicting human interactions, making strategies, checking the fruitfulness of discussions, and so on. Manual work of finding conclusion and patterns of discussion can be saved.

## ACKNOWLEDGMENT

I would like to thank my guide Prof. Mrs. L.A.Deshpande for her valuable guidance. Also I wish to thank our Head of Department, Prof. S.R. Sakhare and P.G. Co-ordinator Prof. Mrs. M.P.Mali for their suggestions.

## REFERENCES

[1] Q. Yang and X. Wu, "10 Challenging Problems in Data Mining Research", *Intl J. Information Technology and Decision Making*, Vol. 5, no. 4, pp.597-604, 2006.

[2] W. Geyer, H. Richter, and G.D. Abowd, "Towards a Smarter Meeting Record - Capture and Access of Meetings Revisited", *Multimedia Tools and Applications*, Vol. 27, no. 3, pp. 393-410, 2005.

[3] Alonso Patron-Perez, Ian Reid, "Structured Learning of Human Interactions in TV Shows", *IEEE Transactions On Pattern Analysis And*

*Machine Intelligence*, Vol. 34, NO. 12, December 2012.

[4] Nikhil Gaikwad, Gaurav Chavan, Twinkle Samal, Avinash Sonule, "Implementation of Mining Algorithms for Improving Efficiencies in Product Session Data", *International Conference on Engineering Technology and Science-(ICETS14)*, Vol. 3, Special Issue 1, February 2014.

[5] Zhiwen Yu, Zhiyong Yu, Xingshe Zhou, Christian Becker, Yuichi Nakamura, "Tree-Based Mining for Discovering Patterns of Human Interaction in Meetings", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, no. 4, April 2012.

[6] C. Wang, M. Hong, J. Pei, H. Zhou, W. Wang, and B. Shi, "Efficient Pattern-Growth Methods for Frequent Tree Pattern Mining", *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD 04)*, pp. 441-451, 2004.

[7] M.J. Zaki, "Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications", *IEEE Trans. on Knowledge and Data Eng.*, Vol. 17, no. 8, pp. 1021-1035, Aug. 2005.

[8] Tomoyuki Morita, Yasushi Hirano ,Yasuyuki Sumi, "A Pattern Mining Method for Interpretation of Interaction", *ICMI05*, ACM 2005, October 46, 2005, Trento, Italy, 1-59593-028-0/05/0010.

[9] N. Zhong, Y. Li, Sheng-Tang W., "Effective Pattern Discovery for Text Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, no. 1, January 2012.

[10] D. Karthika, R. RangaRaj, "A Graph-Based Interaction Pattern Discovery for Human Meetings", *International Journal of Advances in Computer Science and Technology*, Vol. 2, No.8, August 2013, ISSN 2320 2602.

[11] Guimei Liu, Haojun Zhang, and Limsoon Wong, "A Flexible Approach to Finding Representative Pattern Sets", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 26, no. 7, July 2014.

[12] P. UmaMaheswari, Dr. M. Rajaram, "A Novel Approach for Mining Association Rules on Sports Data using Principal Component Analysis: For Cricket match perspective", *IEEE International Advance Computing Conference (IACC 2009)* Patiala, India, 6-7 March 2009.

[13] K. Antony ArokiaDurai Raj, Panchapakesan Padma, "Application of Association Rule Mining: A Case Study on Team India", *2013 International Conference on Computer Communication and Informatics (ICCCI-2013)*, Jan. 04 06, 2013, Coimbatore, INDIA.

[14] T. Karthikeyan, N. Ravikumar, "A Survey on Association Rule Mining", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, Issue 1, January 2014.

[15] Tulips Angel Thankachan, Dr.Kumudha Raimond, "A Survey on Classification and Rule Extraction Techniques for Datamining", *IOSR Journal of Computer Engineering (IOSR-JCE)*, e-ISSN: 2278-0661, p- ISSN: 2278- 8727, Vol. 8, Issue 5, Jan. - Feb. 2013, PP 75-78.

[16] Jiawei Han, Micheline Kamber, "Data Mining- Concepts and Techniques", *Elsevier Inc.*, Indian ISBN: 978-81-312-0535-8, 2006.

[17] Manisha Girotra, Kanika Nagpal, Saloni Minocha, Neha Sharma, "Comparative Survey on Association Rule Mining Algorithms", *International Journal of Computer Applications* (0975 8887) Vol. 84 No 10, December 2013.

[18] P. Maheswari, R. Manavalan, "Tree-Based Mining with Sentiment Analysis for Discovering Patterns of Human Interaction in Meetings", *International Journal of Scientific Engineering and Technology*, ISSN : 2277-1581, Vol. No.2, Issue No.9, pp : 866-871, 1 Sept. 2013.

[19] Qi Han, Junfei Guo, Hinrich Schutze ,"CodeX: Combining an SVM Classifier and Character N-gram Language Models for Sentiment Analysis on Twitter Text", Institute for Nature Language Processing, University of Stuttgart, Stuttgart, Germany.

[20] Shamila Nasreen, Muhammad Awais Azam, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey", *The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)*, Procedia Computer Science 37, pp. 109 – 116, Elsevier 2014.

[21] Mereen Thomas Vadakkel, "An Approach To Understand Human Behaviour Pattern", *International Journal of Computer Trends and Technology (IJCTT)*, Vol. 12 no.1, Jun 2014