Abstract

Interactions between human beings take place often in meeting discussions. Semantic knowledge of meetings can be exposed by discovering interaction patterns from such meetings. Human interaction flow in a discussion session is used to extract the frequent pattern interaction. This paper proposed an Enhanced Principal Component Analysis (EPCA) with Partial Ancestral Graph (PAG) meet method to mine frequent interaction among patterns. The experimental result shows that, the proposed method extracts several frequent interesting patterns that are useful for the interpretation of human behaviour in meeting discussions with less execution time.

1. Introduction

Data mining is an imperative system for finding unique data, which is broadly received in different fields such as bioinformatics, advertising and security. Knowledge Discovery in Database is the methodology of finding unique patterns from vast information sets concerning the routines at the gathering of counterfeit consciousness, machine learning, measurements and database frameworks [1]. The methodology of information mining is to concentrate learning from a dataset in a human justifiable structure. Frequent patterns are item sets, subsequence, or substructures that show up in information set with recurrence at a least user specified limit. Discovering continuous examples assumes a vital part in mining affiliations, connections, and numerous other intriguing connections among information. In addition, it helps in information indexing, arrangement, grouping, and other
information mining of undertakings too. Accordingly, successive mining has turned into a critical information mining errand and a centered subject in information mining examination. Human cooperation figures out that the gathering was decently composed or not. It is considered as one of the primary issues in the gatherings [2]. The interaction flow is shortly termed as semantic knowledge. This shows the pattern of the knowledge. The formation of well-defined dictionary for the relevant meetings is one among the challenging problem which is based on the people’s interaction in the meetings, forums, discussions and so on. There are several performance metrics to determine the interesting pattern in such meetings which includes execution time, number of frequent sub-trees etc. This research work focussed on detecting more number of interesting patterns with reduced execution time.

The rest of this paper is organized as follows: In section 2, previous work is discussed. Section 3 describes the methodology used and experimental results are discussed in Section 4. Finally, the work is concluded in Section 5.

2. Related Works

The success rate of a decision made in a meeting can be got by mining the patterns from human interactions that took place in the meeting. In Xin Chen et al., [3], authors developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. Authors focused on students' Twitter posts to understand issues and problems in their educational experiences. There have been several works done in discovering human behavior patterns by using stochastic techniques. Jayagopi et al., [4] presented a systematic study on dominance modeling in group meetings from fully automatic nonverbal activity cues, in a multi-camera, multi-microphone setting. The authors investigated efficient audio and visual activity cues for the characterization of dominant behavior, analyzing single and joint modalities. Unsupervised and supervised approaches for dominance modeling were investigated. Konopnicki et al., [5] presented a way of analyzing social media conversional data in order to better understand customers. Ultimately, authors goal was to analyze customer behavior as it is expressed in free-form conversations and extract from it commercially valuable information about the customer. Anolli.L et al., [6] has used T-pattern for interaction analysis. Salamin et al., [7] proposed an approach for the automatic recognition of roles in conversational broadcast data, in particular, news and talk shows. The approach makes use of behavioral evidence extracted from speaker turns and applies conditional random fields to infer the roles played by different individuals.

Wai Nang Chan et al., [8] analyzed the speaker discrimination power of vocal source related features, in comparison to the conventional vocal tract related features. The vocal source features, named wavelet octave coefficients of residues (WOCOR), are extracted by pitch-synchronous wavelet transform of the linear predictive (LP) residual signals. Using a series of controlled experiments, it is shown that WOCOR is less sensitive to spoken content than the conventional MFCC features and thus more discriminative when the amount of training data is limited. Jing He et al., [9] focused in domain-driven classification method that takes advantage of multiple criteria and multiple constraint-level programming for intelligent credit scoring. The method involves credit scoring to produce a set of customers' scores that allows the classification results actionable and controllable by human interaction during the scoring process.

Chris Ding et al., [10] analyzed that the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace. Uma et al., [11] used Weighted Interesting Pattern
mining algorithm for discovering patterns by calculating weight confidence and similar patterns are grouped by using similar weights.

3. Methodology

3.1 Dimensionality Reduction

3.1.1 Pre-processing
Since a database consists of a large volume of data collected from heterogeneous sources of data sets, the data will be inconsistent and noisy. Complex data analysis will take very long time to process this original complete data set. So, in order to prepare the data in a suitable and reduced format, some method of preprocessing has to be applied so that useful knowledge can be extracted from it easily. First, remove the words void of semantic content such as ‘and’, ‘the’, ‘of’, etc. using the created list of stop words and then perform stemming using Porter Stemming Algorithm which is a process for removing commoner morphological and inflexional endings from words in English [12]. Now the resultant reduced data set will further be reduced in order to increase the accuracy of results and redundancy of computation.

3.1.2 Principal Component Analysis (PCA)
It is a technique from statistics for simplifying a data set. It was developed by Pearson (1901) and Hotelling (1933) [13]. The aim of the method is to compress the data, by reducing the number of dimensions, without much loss of information [14].
It is recommended as an exploratory tool to uncover unknown trends in the data. The technique has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. The method is mostly used as a tool in exploratory data analysis and for making predictive models.
It is completely nonparametric: any data set can be plugged in and an answer comes out, requiring no parameters to tweak and no regard for how the data was recorded. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences.

Steps to perform Principal Component Analysis:
1. Organize data as an m×n matrix, where m is the number of measurement types and n is the number of samples
2. Subtract off the mean for each measurement type
3. Calculate covariance matrix
4. Calculate the eigenvectors and eigenvalues of the covariance matrix

3.1.3 Enhanced Principal Component Analysis (EPCA)
In human interaction, patterns can be triggered or influenced by multiple interactions. The extent of influence can be significantly dependent on the weight/rank of the person triggering that interaction. Such interaction dataset contains confusing / irrelevant words in sentences. Hence there is a need for a simple, non-parametric method of extracting relevant information from confusing datasets. Hence in this research work, an Enhanced Principal Component Analysis is proposed that offers reduction of complex data set to a lower dimension to reveal the hidden/ simplified dynamics that often lie behind it.
There exist a few linear combinations of the variables that account for most of the variation present in the data that can be eliminated by using the Enhanced Principal Component Analysis method. In PCA, X is denoted as \( X = (X_1, X_2, X_3, \ldots, X_p)^T \) the vector containing all measurements for the number of variables represented as p. The meeting dataset consisted of n vectors \( X_1, X_2, X_3, \ldots, X_p \) in a space of p dimensions, where n is the number of samples. Mathematically, the EPCA problem consists in finding a subspace of dimension K of the original space which maximizes the dispersion of the points projected onto that subspace. The solution to the above said optimization problem is specified by the eigenvectors corresponding to the K largest eigen values of the covariance matrix of the sample

\[
E = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^T, \text{ where } \mu \text{ is the mean vector of the sample.}
\]

For various reasons, it is common to start by standardizing the data. With the initial standardization, the enhanced principal components obtained are linear combinations of the original variables, and the coefficients of these linear combinations are given by the elements of the eigenvectors of the usual correlation matrix based on Pearson’s correlation coefficient r. In EPCA more importance is given to observations whose values are more important. Also, the correlation coefficient is sensitive to the presence of outliers and noise in the data. The ranks of the observations are used [15]. In the meeting dataset ranking the observations for each conversation from 1 (highest rank) to n (lowest rank) is taken. The Pearson’s correlation coefficient of the ranked data is thus obtained using the Spearman’s rank correlation coefficient \( r_s \), which is given by the expression

\[
r_s = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2 \sum_{i=1}^{n} (Q_i - \bar{Q})^2}}
\]

where R and Q are the average ranks. However, for computational purposes, a more convenient expression which assumes there are no ties is

\[
r_s = 1 - \frac{6 \sum_{i=1}^{n} (R_i - Q_i)^2}{n^3 - n}
\]

It is clear from this rewritten form of \( r_s \) that the calculation of the distance between two ranks in Spearman’s coefficient is given by

\[
D_i^2 = (R_i - Q_i)^2,
\]

which does not take rank importance into account, because if \( (R_i - Q_i) \) is, for instance, (1, 3) or (n-2, n), the contribution is the same. The following alternative distance measure is proposed:

\[
WD_i^2 = (R_i - Q_i)^2 \left( (n - R_i + 1) + (n - Q_i + 1) \right) = D_i^2 \left( 2n + 2 - R_i - Q_i \right)
\]

The first term of this product is \( D_i^2 \), exactly as in Spearman’s coefficient, and represents the distance between \( R_i \) and \( Q_i \); the second term is a linear weighting function which represents both the importance of \( R_i \) and \( Q_i \). Here, the proposed alternative distance measure incorporates the rank measure. Since there may be lot of conversations, ranking is the technique used to identify and sort the conversations. Here i refer to the initial conversation and increments by 1 up to n number of conversations. This process is carried out until all the conversations are ranked. Also, the weighted rank measure of correlation is obtained using

\[
r_w = 1 - \frac{6 \sum_{i=1}^{n} (R_i - Q_i)^2 \left( 2n + 2 - R_i - Q_i \right)}{n^4 + n^3 - n^2 - n}
\]
which yields values between -1 and +1. The calculation of the distance between two ranks $R_i$ and $Q_i$ is given by $W_1D_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)$ where the second term of the product is a linear weighting function which represents the importance of $R_i$ and $Q_i$. Hence, the distance measure is

$$W_1D_i^2 = (R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2$$

which reflects more than $W_1D_i^2$ the higher importance of agreement on top ranks. It is common to define rank correlation coefficients, such as Spearman’s, as a linear function of the distance between the two vectors of ranks. In this research, this corresponds to define a coefficient of the form

$$A + B \sum_{i=1}^{n} (R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2$$

where the conversations are such that it takes values between -1 and 1. In order to find A and B, we will start by doing a specific data transformation and then compute the Pearson’s coefficient on the transformed data. The expression obtained is exactly of the form, from where the constants A and B follow. The transformation consists in substituting the value of observation i in the first variable by the value of $R_i = R_i(2n + 2 - R_i)$, where $R_i$ is the rank of that observation. It is clear from above that the computation of the new correlation coefficient is equivalent to do a data transformation to each variable as $R_i = R_i(2n + 2 - R_i)$ and then compute the Pearson’s correlation coefficient. $R_i$ represents the rank of each observation value; usually the smallest value has rank 1, the second smallest rank 2, and so on. Once when the EPCA completes the reduction of deviated discussion contents, the PAG meet takes the pre-processed meeting contents to construct the PAG graph.

### 3.2 Frequent Pattern Mining (FPM)

FPM is a tree based mining method which will discover frequent patterns of human interaction from meeting discussions. For each tree in tree dataset, isomorphic trees are generated with the number of occurrences. The support of each tree will be calculated and the trees whose supports are larger than the threshold value are selected. If many trees are isomorphic, one of them is selected and the others are discarded and the frequent trees are given as output [16]. FPM will generate more frequent patterns but will miss some important frequent pattern as it does not capture all triggering relations.

### 3.3 Partial Ancestral Graph (PAG) meet

PAG meet will find temporal frequent interactions. It will generate a set of all frequent nodes and then expand these nodes with new root, new level, new node and new edge. Sub-graphs containing siblings may not be connected without the presence of their common ancestor in a tree. So, if a common ancestor is not frequent, tree based mining method will fail to mine them as frequent pattern. Also it will not distinguish multiple interactions but PAG meet will identify such multiple interactions and extracted both temporal and triggering relations [17].

### 4. Experimental Results

Dataset is collected from meetings that are conveyed through on-line forums and documented discussions from various data sources. Table 1 shows the categorisation of the dataset.
S. Uma and Dr. J. Suguna :: Human Interaction Pattern Mining Using Enhanced Principal Component Analysis

The above selected datasets are focused on the topics namely corruption, democracy, defense expenditure, social expenditure, backwardness, women quota bill, politics, communal and religious reservations, retirement age, elections etc. The datasets have certain amount of irrelevant data and hence EPCA mechanism is used to remove such irrelevant data and then PAG meet is applied to generate interaction flow in meetings. From the figures 4.1, 4.2 and 4.3 it can be observed that the conversations in the meeting can be grouped legitimately using the EPCA algorithm.

### Table 4.1: Dataset Information

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Number of documents</th>
<th>Maximum number of conversations per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini dataset</td>
<td>2000</td>
<td>8</td>
</tr>
<tr>
<td>Large dataset</td>
<td>2000</td>
<td>250</td>
</tr>
<tr>
<td>Very large dataset</td>
<td>2000</td>
<td>5000</td>
</tr>
</tbody>
</table>

Figure 4.1: Analysis of EPCA in mini dataset

Figure 4.2: Analysis of EPCA in large dataset
Table 4.2 shows the detection percentage of interactions made during the sessions. From the observation, it is clear that the proposed EPCA-PAG meet method detects more number of interactions in the sessions when compared to FPM and PAG meet. The same is flashed in figure 4.4.

<table>
<thead>
<tr>
<th>Interactions</th>
<th>FPM</th>
<th>PAG meet</th>
<th>EPCA-PAG meet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 4.3 shows the execution time of the three mechanisms namely FPM, PAG meet and EPCA followed with PAG Meet over the support threshold. It can be clearly understood that the proposed EPCA-PAG meet consumes less execution time than the other mechanisms and it is noteworthy that when support threshold value increases, the execution time decreases and the same is flashed in figure 4.5.
Detecting more number of interactions gains advantage towards discovering frequent pattern mining. From table 4.4 it is shown that the number of discovered frequent sub trees is increased in EPCA-PAG meet than the other mechanisms and it is exposed in figure 4.6.

Table 4.4: Support Threshold Vs Number of Discovered Frequent Sub trees

<table>
<thead>
<tr>
<th>Support Threshold</th>
<th>FPM</th>
<th>PAG meet</th>
<th>EPCA-PAG meet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8200</td>
<td>9560</td>
<td>11210</td>
</tr>
<tr>
<td>2</td>
<td>3800</td>
<td>4300</td>
<td>4650</td>
</tr>
<tr>
<td>3</td>
<td>2300</td>
<td>3750</td>
<td>4101</td>
</tr>
<tr>
<td>4</td>
<td>1902</td>
<td>2400</td>
<td>3802</td>
</tr>
<tr>
<td>5</td>
<td>1704</td>
<td>2208</td>
<td>3607</td>
</tr>
<tr>
<td>6</td>
<td>1203</td>
<td>2008</td>
<td>3207</td>
</tr>
<tr>
<td>7</td>
<td>1101</td>
<td>1905</td>
<td>2804</td>
</tr>
<tr>
<td>8</td>
<td>902</td>
<td>1803</td>
<td>2789</td>
</tr>
<tr>
<td>9</td>
<td>753</td>
<td>1653</td>
<td>2605</td>
</tr>
<tr>
<td>10</td>
<td>401</td>
<td>1205</td>
<td>2105</td>
</tr>
</tbody>
</table>
5. Conclusion

This paper proposed an Enhanced Principal Component Analysis method which reduced the complex dataset to a lower dimension and removed the deviated discussions in the meetings. Then the frequent interaction patterns from meetings were revealed by applying PAG meet. In this work, more number of frequent interactions was detected by EPCA-PAG meet method with less execution time when compared to FPM and PAG meet mechanisms. It was also found that, when the data sets are large in number, more number of interactions were given away by the proposed method. In future, the work can be carried out by optimizing the dataset with bio-inspired computing paradigms.

References

S. Uma and Dr. J. Suguna :: Human Interaction Pattern Mining Using Enhanced Principal Component Analysis