PICGRAPH: An Extension of Power Iteration Clustering for Inferring Conceptual Relationships from Large Unstructured Datasets

Anoop V.S
IIITM-K, Technopark
Thiruvananthapuram
Kerala

Lekshmi B.G
IIITM-K, Technopark
Thiruvananthapuram
Kerala

ABSTRACT
Data mining and Information extraction is an emerging research area that attempts to find out the hidden and relevant knowledge from the overwhelming amount of information available in structured, semi structured and unstructured forms. Power Iteration Clustering is a powerful clustering algorithm that is proved to be efficient for the clustering of structured data. Large amount of data available in Gmail, micro-blogs etc doesn't have proper structure renowned as unstructured data which covers remarkable percentage of data available on the internet. Due to the unordered structure of the data, the extraction of relevant information from this huge collection is a complex task. This work used the predominant features of Power Iteration Clustering algorithm for the extraction of information and visualization of unstructured data from micro blogging sites.

General Terms
Data Mining, Pattern Recognition, Clustering

Keywords
Information Extraction, Micro blogs, Power Iteration Clustering, Unstructured Data.

1. INTRODUCTION
Advancements of technologies in the field of data representation, processing and storage help to share and communicate information through the internet within fraction of seconds around the world. This helps the people and organization to trust the electronic media to share and store any valuable information with powerful security. Since the digitization of data increases rapidly in the micro-blogging sites, the extraction of useful information from this huge repository of data is a strenuous task. Due to the widespread application of information retrieval and processing, it is a pre-requisite to eliminate the irrelevant information and noise from the data in order to find and link the useful information according to the application. Based on the way of digitization and storage of data in the websites, the data may be structured, semi- structured and unstructured. Hence extracting and processing the required information while ignoring the irrelevant from these heterogeneous collections of repositories is a tedious task. Nowadays, people are more depending on the social media and micro blogging sites as the fastest medium for information sharing and communication. And almost all countries permits the usage of social media and it is easy to develop own blogs and sites for information sharing. The data thus distribute among the micro-blogging sites and social media does not have proper structure while storing and termed as unstructured form of data. Hence a good percentage of datasets available from the websites have unstructured form and which require special attention while processing and finding interesting patterns and related information thus it is an active research topic recently. Due to the unusual representation of data, special characters and symbols, the usual algorithms and techniques are not efficient for the processing. Moreover, the data from the micro blogging sites and social media are lengthier and personal in nature, may be off the topic which have less grammar, in frequent usage of abbreviations and emoticons, unorthodox capitalization, in ordered word sequencing and hash tags and these may have significant role in the information passing [1]. Finding and extracting the relevant data from these assorted datasets requires uprooting the importance of these symbols and emoticons which plays the notable role in the text message. This would be possible with the clustering algorithms which identifies and then groups the similar concepts together and clustering is one of the efficient and proven ways of accomplishing the information retrieval from any complex dataset [2]. Extraction of data involves extraction of entities, events associated to them and the relationship exists among them. The data extracted from the unstructured datasets are later stored into a structured database and the data mining techniques are then applied for discover the new knowledge. The necessary steps in the information extraction process are the linguistics preprocessing which involves a number of linguistics techniques such as tokenization, tagging and lemmatization. Today micro-blogging sites are the powerful tool for fastest communication. Millions of people use this facility to share their views and comments on different topics of their interests. Therefore micro-blogging sites are the very rich source of unstructured data and also data mining on these data are not well studied. Power Iteration Clustering (PIC) [3] is one of the proven clustering algorithms that work well on the structured datasets available in the internet. It is found to be computationally efficient since it reduces the dimensionality of the dataset to single dimensional. Later the same concept has been extended to semi -structured data and the results were found promising. The low dimensional representation, termed as PIC-D, obtained can efficiently used for answering similarity queries like set expansion, automatic set instance acquisition and column classification. This one dimensional embedding acts as cluster indicator which brings the entities of similar kinds together [4]. Here the concept of PIC for the information extraction from the unstructured data collected from various micro-blogging sites has been used. The PIC embedding obtained is further used for the visualization of extracted entities, events and relationship between those entities that helps to assess the entire data, understand the neighborhood and highlight similar entities together with a relationship.

2. LITERATURE REVIEW

Usually when text documents are represented as vectors, the clustering of these documents become challenging due to the large number of terms with high dimension. Processing a high dimensional dataset is a challenging task, as it requires plenty of space to store and high computationally time. In order to handle such problems, high dimensional data are projected from the high dimensional space to low dimensional space using some dimensionality reduction techniques. Later, the clustering algorithms are applied to the resultant datasets which takes less space and computational time than the previous datasets. The efficiency of such algorithms depends on the dimension of the resultant dataset. Due to this, curse of dimensionality poses the tough challenge in the data clustering and information extraction problems.

The existing clustering algorithms directly compute on the weight vectors of n dimension which results in less computational efficiency. With the advancement of research in this area, various techniques are introduced for the purpose of reducing the dimension of the raw data. It is found that, the techniques used for reducing the dimension are basically depend on the form of input data to be used. Some attempts have been made in this area in terms of semantic correlation are frequencies, co-occurrences and graph theory. The naive Bayesian concept is also come to the same category where each term is uniquely assigned to single term cluster [5] using mean computation. But the algorithm performs badly when the degrees of co-occurrence between the terms are high due to the lack of proper definition of the structures of the cluster on such situations. Hierarchical neural network based on Self Organizing Map and Learning Vector Quantization is implemented for the extraction of large similar datasets and it has considerable computational efficiency [6]. Later, Archana.et.al. removed the redundant rules for analyzing online data by introducing new techniques for online data mining [7]. The advantage of this algorithm is that it reduces the irrelevant noise in the data mining process significantly. Various non-linear relationships exists between the terms in the documents are uncovered by the technique introduced by Reshef et al. and is described through Maximal Information Co-efficient (MIC) [8]. This method is found to be more promising method for detecting the dependency and association among terms and documents. But the problem is, the method has failed to satisfy the property of equitability for large datasets. Xinwu [9] put forward an approach to text clustering which combines both the advantages of SOM and K-means clustering algorithm. X. Liu and P. He, H. Wang [10] introduced a frequent term set based clustering (FTSC) which uses frequent term sets for text clustering. This reduces the dimension of the text data efficiently and thus it improves the running speed and accuracy of the clustering algorithm. Spectral clustering method works by determining the correlation between the eigenvectors of each input vector [11]. But the disadvantage of this method is that, the detection of eigenvector from a high dimensional data is computationally costly and usually the embedding obtained from the smallest k eigenvectors may fail due to the noise in the data. The recent advancement in the dimensionality reduction and thus clustering the data effectively is using a Power Iteration Clustering (PIC) which is introduced by Bhavana Dalvi and W. Cohen [4]. It is found to be very simple, scalable and proved to be superior to spectral clustering methods on certain text clustering tasks. The single dimensional embedding obtained from PIC is the linear combination of eigenvectors as obtained from spectral clustering algorithms. In this respect PIC is very unique and different approach for the clustering of text document.

3. PIC AND PIC-D

Power Iteration Clustering (PIC) is a very simple and scalable graph clustering [12] method which finds the one dimensional representation of multidimensional dataset by truncated power iteration clustering method to reduce the computational complexity of matrix multiplication [4]. The embedding created by PIC and spectral clustering is based on the similarity matrix of the raw data and this embedding provides clustering of the same by k-means algorithm. Even though the concept of both the algorithms lie on the same paradigm, the explicit calculation of eigenvectors and thus the time and cost of calculation is replaced by a small number of matrix vector multiplication in PIC. However, the highlights of PIC are its simplicity and scalability. And the basic implementation of PIC may partition the network dataset of huge collection within few seconds without any pre-processing of data. This concept is readily parallelizable and very efficient in terms of time and space.

Algorithm 1: The PIC Algorithm

Input: A row-normalized affinity matrix W and number of clusters K
Pick an initial vector V^0
Repeat

Set $V^{t+1} \leftarrow \frac{WV^t}{||WV^t||}$ and $\delta^{t+1} \leftarrow |V^{t+1} - V^t|$ Increment t

Until $|\delta^{t+1} - \delta^t| \cong 0$
Use K-means to cluster points on V^t

Output: Clusters $C_1, C_2, \ldots C_K$

PIC-D embedding gives the low dimensional embedding of the D partite graph[14] which in turn helps to store and manage multidimensional huge data set. As it uses the PIC concept directly, the 1D embedding can be formed within a short amount of time. The proposed low dimensional representation could represent large D - partite graphs[15] using smaller number of dimensions which enabled similarity queries such as Set Expansion (SE), Automatic Set Instance Acquisition (ASIA) and Column Classification (CC) worked fast on web datasets which are in either structured or in semi - structured forms.

4. PICGRAPH VISUALIZATION USING LOW DIMENSION EMBEDDING

4.1 Dataset used

Here publically available unstructured data from micro-blogging sites and Reuter's dataset are used for this experiment. Digital books and journals, social networks, blogs, contents of emails etc are rich sources of unstructured data. For this experiment, Bollywood movie reviews are used as the dataset.

Algorithm 2: The PIC-D Algorithm

Function Create_PIC-D_Embedding (m, X1,X2…X_D): X_{PIC-D}

Input: m - no: of PIC dimensions per dataset
X - Set of all entities
X_1 - Co-occurrences of X in $dataset_1$
X_2- Co-occurrences of X in $dataset_2$
X_D - Co-occurrences of X in $dataset_1$

Output: $X_{PIC\text{-}D}$: (|X|,D*m) dim. Embedding of X
$X_{PIC\text{-}D} = \varphi$
t = a small positive integer
for i=1:D do
for j=1:m do
V_0 = randomly initialized vector of size |X|*1
V_t = *PIC_Embedding*(X_i, V_0, t)
Append V_t as a new column $X_{PIC\text{-}D}$
end for
end for
end Function

---

which are found to be huge repository of unstructured data. A search API has been used to crawl data from the related websites and constructed a dataset containing more than 5000 text messages. Due to the nature of this micro-blogging service (quick and short messages), people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings, text pre-processing is required for creating accurate dataset. Since named entities present in the datasets are considered for the experiment, removal of all other non-character symbols, URLs and numeric values were necessary from the dataset. As it is found that the online NER tools are not providing enough data from the dataset, the manual labeling was preferred initially to tag the entities.

4.2 Experiment

The data collected from the micro-blogging sites using queries are initially dumped into the text file. Due to the language variations, shortness in the message and ambiguity in the data, pre-processing of the data retrieved from search API are necessary. Manual as well as automatic pre-processing of these text files is required for the creation of datasets for the experimental purpose. As existing NER tagging procedures give poor result due to reduced amount of contextual information in short messages, manual labeling of named entities were preferred for the collected datasets to improve accuracy. Later, these named entities are used for creating the similarity matrix over the collected data set. The similarity matrix obtained is used for creating the low dimensional representation using power iteration clustering which is further used to construct PIC Graph of the entities in the dataset. Fig 1 gives the workflow of the experiment. Fig 2 shows the relationship between entities in the dataset. This relationship is explored through the low dimensional representation which is obtained using PIC algorithm. As the Bollywood movie reviews were used, the named entities identified in the datasets are the names of actor, actress, director, films and year of release. Using these named entities the extraction of any information regarding the Bollywood movie news are possible. The relationships of these entities are explored in Fig. 2.

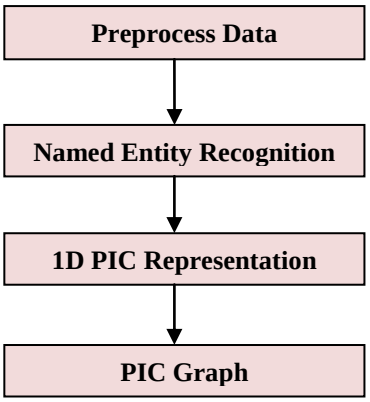

Fig 1: Overall workflow of the experiment

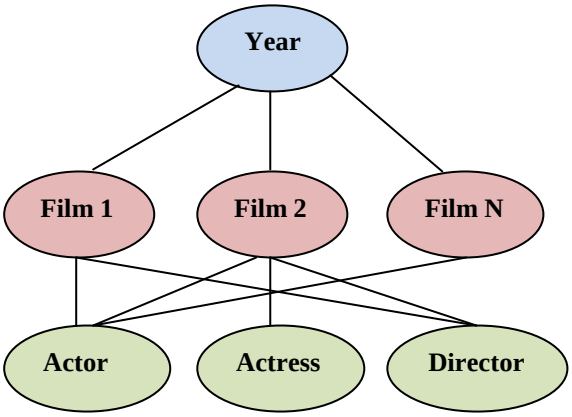

Fig 2: Relationship between entities identified

Fig 3 shows the PIC Graph of the entities in the dataset. The edges in the graph represent their relationship obtained from similarity matrix.

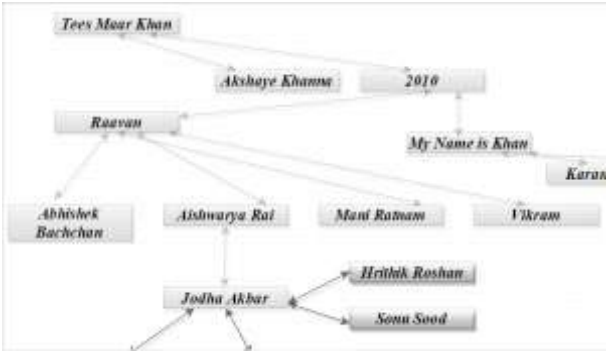

Fig 3: Part of the PICGraph constructed

Fig 4 shows the similar entities in the PIC Graph. To extract the information which is related to the query input can be extracted and visualized using PIC Graph. Here for example, any entity in the dataset can be given as input to the system. Then the graph will highlight those entities which are very close to the queried entity. The Fig 4 shows only the part of the PIC Graph, which shows the related entities in different color.

Fig 4: Related entities of Level 2 are shown in blue colored nodes

The related information can be extracted based upon the level. Lower level gives the high similarity index and higher levels give the lowest similarity index. As the level increase, the amount of related information extracted will also increase. Each level is encountered as the number of edges in the PIC Graph. Level 1 relation gives the entities which are connected using single edge and level 2 gives those entities which are connected using 2 edges from the queried input entity and so on.

5. RESULT AND DISCUSSIONS

The performance of the embedding obtained to visualize the PIC Graph is analyzed using statistical parameters such as precision, recall and accuracy. Table 1 show the low dimensional embedding obtained with the help of PIC algorithm. Each row in the table gives the entity name from the collected datasets and each column value gives the document name in which they are related. The table values reveal that, those entities which relates to the same document have almost similar value in the embedding. Using such a relationship, the PIC Graph relates these entities using an edge and the weight of the edge is the respective table values against each entities.The recall, precision and accuracy are important parameters which can evaluate the performance of the visualization of unstructured dataset. In the below equations, TP indicate True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Accuracy = (TP+TN) / (TP+TN+FP+FN)

The values are calculated based on the level. Each level gives the depth of the graph in which it is related to the other entities. As the level increases, the number of related entities increases and thus accuracy may decreases. Table 2 shows the parameter values up to level 3. The detailed precision and recall for each class is calculated as follows:

Precision and recall for 'Entity' class

No. of levels given: 2

No. of entities returned: 22

True Positive (TP): 18

True Negative (TN): 975

False Positive (FP): 4

False Negative (FN): 3

Precision = 0.81

Recall = 0.85

Accuracy = 0.99

Precision and recall for 'Film' class

No. of levels given: 2

No. of entities returned: 11

True positive (TP): 9

True Negative (TN): 986

False Positive (FP): 2

False Negative (FN): 3

Precision = 0.81

Recall = 0.75

Accuracy = 0.99

Precision and recall for 'Year' class

No. of levels given: 2

No. of entities returned: 23

True Positive (TP): 22

True Negative (TN): 976

False Positive (FP): 1

False Negative (FN): 1

Precision = 0.95

Recall = 0.95

Accuracy = 0.99

The results show that, a low dimensional embedding which can reduce the high dimension of entities in any unstructured dataset is a better form of representation though that the storage space and computational complexity is reduced. To find the similarity matrix, the frequency of each entity in the dataset against the document is calculated and that is the only part which takes some computational time.

Table 1. Low Dimensional Embedding Obtained from this Experiment

Entities	2 states	Chennai Express
Shahrukh Khan	0.0073	0.2563
Arjun Kapoor	0.2055	0.0071
Hrithik Roshan	0.0012	0.2556
Mohit Suri	0.2689	0.0081
Alia Bhutt	0.3456	0.0013
Karan Johar	0.0052	0.3188
Kareena Kapoor	0.0062	0.3558
Aishwarya Rai	0.0096	0.0060
Katrina Kaif	0.0045	0.0078
Kajol	0.0056	0.3810

Table 2. Performance of PICGraph for Information Retrieval

Class	Level	Precision	Recall	Accuracy

Entity	1	.83	.80	.99
	2	.79	.74	.98
	3	.75	.79	.96
Film	1	.82	.85	.99
	2	.81	.75	.96
	3	.76	.72	.95
Year	1	.83	.89	.99
	2	.81	.78	.98
	3	.79	.76	.96

The embedding thus obtained is promising representation, with that; the visualization of entities in the unstructured dataset became easy. And thus the similar entities are tied together in a single graph and hence the information retrieval becomes far better than any other graph clustering method.

6. CONCLUSION

The embedding thus obtained is a promising representation, using which the visualization of entities in the unstructured dataset became easy and scalable. Thus the similar entities are tied together in a single graph and hence the information retrieval becomes far better than any other graph clustering method. Experiments with real and synthetic dataset shows interesting results which made the authors think to extend this for future research. As future work, it is planned to extend this experiment into a big data paradigm so that the huge volume and velocity of the data can be considered. A real time Twitter streaming and PICGraph construction is also planned as a future work. Besides, the authors may add more interactive graph visualization for easy interpretation of the results**.**

7. ACKNOWLEDGEMENTS

The authors would like to thank all the researchers and staff members of Data Engineering Lab at IIITM-K, Thiruvananthapuram for their valuable suggestions and feedback. They also thank all the anonymous reviewers for the constructive suggestions that helped to improve the quality of this paper.

8. REFERENCES

[1] Nahm, Un Yong, and Raymond J. Mooney. Text mining with information extraction. AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Vol. 1. 2002.

[2] Karanikas, Haralampos, Christos Tjortjis, and Babis Theodoulidis. An approach to text mining using information extraction. Proc. Knowledge Management Theory Applications orkshop,(KMTA). 2000.

[3] Lin, Frank, and William W. Cohen. Power iteration clustering. In Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.

[4] Dalvi, Bhavana, and W. Cohen. Very fast similarity queries on semi structured data from the web. SDM. 2013.K. Elissa

[5] Toshniwal, Durga, and Rishiraj Saha Roy. Clustering Unstructured Text Documents Using Naive Bayesian Concept and Shape Pattern Matching. IJACT: International Journal of Advancements in Computing Technology. 2009.

[6] Lu, Yen-ling, and Chin-Shyurng Fahn. Hierarchical artificial neural networks for recognizing high similar large data sets. International Conference on Machine Learning and Cybernetics. IEEE, 2007.Archana Singh, Megha Chaudhary, Dr (Prof.) Ajay Rana, Gaurav Dubey, Online Mining of data to Generate Association Rule Mining in Large Databases , International Conference on Recent Trends in Information Systems. 2011

[7] David N. Reshef et al.,Detecting Novel Associations in Large Data Sets. Science AAAS. 2011

[8] L. Xinwu. Research on Text Clustering Algorithm Based on k-means and SOM. International Symposium on Intelligent Information Technology Application Workshops 2008

[9] X. Liu, P. He, H. Wang. The Research of Text Clustering Algorithms Based on Frequent Term Sets. Proc. International Conference on Machine Learning and Cybernetics 2005

[10] Fiedler, Miroslav. An Algebraic Approach to Connectivity off Graphs. Recent advances in graph theory: proceedings of the Symposium held in Prague, June 1975

[11] Schaeffer, Satu Elisa. "Graph clustering." Computer Science Review 1.1 (2007): 27-64.

[12] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." Neural Networks, IEEE Transactions on 16.3 (2005): 645-678.

[13] Dhillon, Inderjit S. "Co-clustering documents and words using bipartite spectral graph partitioning." Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.

[14] Hartuv, Erez, and Ron Shamir. "A clustering algorithm based on graph connectivity." Information processing letters 76.4 (2000): 175-181.