ISSN:2348-2079

Volume-5 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

A novel feature selection technique for unstructured data analysis

[1]Mr D.Vivek, Assistant Professor, Vivekprasanth87@gmail.com
[2]V.Saranya, UG Scholar, saisaran448@gmail.com
[3]S.Sangavi UG Scholar, sangavisaran11@gmail.com
[4]D.Smitha UG Scholar, smithakrishnadhas@gmail.com
Department of IT, Nandha Engineering College (Autonomous), Erode-52

Abstract - Nowadays social networking has a rapid growth among the users. They post many events which makes difficult to search for a particular event. To avoid this the mining technique is being brought to split the events according to its domain. The summarized form is being changed to attributes. Here temporal database is being used for handling data time. Feature Selection is used for simplification of events. Therefore it could be easier to understand the evolutionary trends in the data set. To evaluate the upcoming proposed method, many experiments are done to produce in qualitative and quantitative manners in social media networking.

I INTRODUCTION

Social Media is an umbrella term that defines the various activities that integrate technology, social interaction, and the construction of words and pictures. It allows to create and share the information, ideas, career interests and the other forms of expression.Social media use web-based technologies, desktop computers and mobile technologies (e.g., smartphones and tablet computers) to create highly interactive platforms through which individuals, communities and organizations can share, co-create, discuss, and modify user-generated content or pre-made content posted online. The term social media is usually used to describe social networking sites such as:Facebook – an online social networking site that allows users to create their personal profiles, share photos and videos, and communicate with other users. Twitter – an internet service that allows users to post "tweets" for their followers to see updates in real-time. LinkedIn – a networking website for the business community that allows users to create professional profiles, post resumes, and communicate with other professionals and job-seekers.Pinterest – an online community that allows users to display photos of items found on the web by "pinning" them and sharing ideas with others[1]. Snapchat – an app on mobile devices that allows users to send and share photos of themselves doing their daily activities.Social media technologies take many different forms including blogs, business networks, enterprise social networks, forums, micro blogs, photo sharing, products/services, bookmarking, social gaming, social networks, video sharing and virtual worlds. Some social media sites have greater potential for content that is posted there to spread virally over social networks. A media event, also known as a pseudo-event, is an event or activity that exists for the sole purpose of media publicity. Media events may centre on a news announcement, a corporate anniversary, a press conference in response to a major media event, or planned events like speeches or demonstrations. Instead of paying for advertising time, a media or pseudo-event seeks to use public relations to gain media and public attention. A news conference is often held when an organization wants members of the press to get an announcement simultaneously. The in-person events may include interviews, questioning, and show-and-tell.

Companies are increasingly using social media monitoring tools to monitor, track, and analyze online conversations on the Web about their brand or products or about related topics of interest. This can be useful in public relations management and advertising campaign tracking, allowing the companies to measure return on investment for their social media ad spending, competitor-auditing, and for public engagement. Tools range from free, basic applications to subscription-based, more in-depth tools. Hootsuite is an example of a social media monitoring and tracking software that companies can use.Social media "mining" is a type

of data mining, a technique of analyzing data to detect patterns. Social media mining is a process of representing, analyzing, and extracting actionable patterns from data collected from people's activities on social media. Social media mining introduces basic concepts and principal algorithms suitable for investigating massive social media data; it discusses theories and methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and mathematics [2][3]. It encompasses the tools to formally represent, measure, model, and mine meaningful patterns from large-scale social media data. Detecting patterns in social media use by data mining is of particular interest to advertisers, major corporations and brands, governments and political parties, among others.

II LITERATURE REVIEW

A. Multi-Modal Event Topic Model For Social Event Analysis

This algorithm is to obtain the evolutionary trends of social media and generate effective event summary details over time. To overcome this a novel technique is introduced, which effectively model social media documents, including the text and with images[1]. To apply this mmETM model, this provides an incremental learning strategy. There is also a drawback found in this model, which data are difficult to find the events as found in an summarized manner.

B. Event Analysis In Social Media Using Clustering Of Heterogeneous Information Network

In this algorithm, they focus on combining multiple types of data from social media in heterogeneous network [2]. This could be either the text of the posts, or the network of users. Here graph based model using users, posts, and concepts extracted from the post content to represent the social media. They also implement to use in cluster posts by various topics and events. This was found as the previous drawback says, data are found in homogeneous network.

C. Matching Words And Pictures

This algorithm focuses on specific case of segmented images with associated text. This in detail can be predicted as auto-annotation and region naming. Auto-annotation might help organize and access large texts[3]Region naming is a process of translating image regions to words. Here we develop about the multi-modal data mining. We study this modal using the Hofmann's hierarchical clustering/aspect model. This is a translation model adapted from statistical machine translation. Here the difficulty raised on measuring the performance of the data.

D. Multiscale Event Detection In Social Media

Event Detection is becoming more important in social media analysis. This raise based on the temporal and spatial resolutions. Here we propose a novel approach towards multiscale event detection. The properties of wavelet transform, which is a well-developed transform, to enable automatic handling of the interaction between temporal and spatial scales [4]. An algorithm is also being used by single graph- based clustering algorithm. They also provide new insights of the influence of noise in the design of event detection algorithms.

E. Social Network Data Analysis For Event Detection

This algorithm concentrate on social network (SN) activity to empower rich models. On minimizing the data properties, we can maximize the total amount of usable data. This is being checked using the model of normal [5] City behaviour which we use to detect abnormal situations (events). On this example, they introduced some applications. The result is a system capable of detecting a wide variety of events with excellent precision. Finally spatio- temporal correlation analysis of events.

F. On-Line New Event Detection And Tracking

The media collection provides users facilities to share and access data, which also demands data management techniques. This event cues to recall people's past memory [6]. This value makes extremely helpful in organizing data. This algorithm provide survey on event enrichment, detection and categorization. This introduce each paradigm and summarize related research efforts and also emerging trends in this area. This utilized as the centre role in social media and organization system. The exponential growth of social requires more scalable, Effective and robust technologies to manage and index them.

G. Understanding Events Through Analysis Of Social Media

This excites an open opportunity to extract social perceptions and obtain its insights

relevant to extract around us[7]. On this they analyse event-centric user generated content on social networks. This aggregates events related to events of interest, along with web resources. This also supports spatial, temporal, thematic and sentiment dimensions. It is unique in its support for user.

H. Mining Relationships Among Interval-Based Events For Classification

In this paper, they argue the hierarchical representation with additional information to achieve a lossless representation. An efficient algorithm called IEMiner is designed to discover temporal patterns from interval-based events [8]. The algorithm employs two techniques to reduce the search space and remove non-promising
Candidates.

I. Detecting Visual Text

This algorithm aim to separate visual text from non-visual text in natural images and their descriptions [9]. This is derived by using computer vision algorithm which improves performance. Atlast it describes, reliably mine visual nouns and adjectives from large corpora and use these in the classification task.

III EXISTING METHODOLOGY

There is a tremendous growth in the social media networking and it is difficult to exactly find and organize certain events from it. In this model, the existing system on novel multi-modal social event tracking and evolution framework obtains the evolutionary trends of the social events and generate effective event summary detail over-time. This provides the correlation between textual and visual modalities to separate the visual and non-visual representative topics. We also adopt an incremental learning strategy denoted as incremental mmETM, which can obtain informative textual and visual topics. Applying this model for social event tracking, we create updating strategy.

There is a drawback found in this existing system, the data that are found in given in a summarized form. Hence it is difficult to find a particular event in this massive social events. The data are in unstructured form.

There also occupies more memory space in the data sets and the searching for a particular event takes more time for the user. Thus, if we summarize this unstructured data into a structured form and provide an event attribute mining with temporal management and the feature selection. This will be easier to search the events. We overcome all these drawbacks.

IV PROPOSED METHODOLOGY

In this proposed system, the drawbacks of existing system is being replaced by the attribute mining in social media. We will explore the tracking performance by considering with different domains like Flickr, Google news and You Tube. The proposed system is being clarified using the following methods.

A. Temporal Database

A temporal database is a database with built-in support for handling data involving time, being related to the slowly changing dimension concept, for example a temporal data model and a temporal version of Structured Query Language (SQL).More specifically the temporal aspects usually include valid time and transaction time. These attributes can be combined to form bitemporal data.Valid time is the time period during which a fact is true in the real world. Transaction time is the time period during which a fact stored in the database was known.Bitemporal data combines both Valid and Transaction Time.Temporal databases support managing and accessing temporal data by providing one or more of the following features:

- A time period datatype, including the ability to represent time periods with no end (infinity or forever)
- The ability to define valid and transaction time period attributes and bitemporal relations
- System-maintained transaction time
- Temporal primary keys, including non-overlapping period constraints
- Temporal constraints, including non-overlapping uniqueness and referential integrity
- Update and deletion of temporal records with automatic splitting and coalescing of time periods
- Temporal queries at current time, time points in the past or future, or over durations.

Valid time is the time for which a fact is true in the real world. A valid time period may be in the past, span the current time, or occur in the future.Transaction time records the time period during which a database entry is accepted as correct. This enables queries that

show the state of the database at a given time. Transaction time periods can only occur in the past or up to the current time. In a transaction time table, records are never deleted. Only new records can be inserted, and existing ones updated by setting their transaction end time to show that they are no longer current.Temporal data is collected to analyze weather patterns and other environmental variables, monitor traffic conditions, study demographic trends, and so on. This data comes from many sources ranging from manual data entry to data collected using observational sensors or generated from simulation models.

B. Feature Selection

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for three reasons:

- Simplification of models to make them easier to interpret by researchers/users shorter training times.

Feature selection techniques should be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

V EXPERIMENT AND RESULT ANALYSIS

The experimental analysis shows a bar graph between the existing and proposed algorithm. In this event, both show a variation according to its analysis designed. In existing system, there is a drawback found which says about the data are found in unstructured form. This makes difficult to find any particular event. There consumes more memory space and more time is used to search a single event. The novel multi-modal event topic model provides those variations and this is the disadvantage.

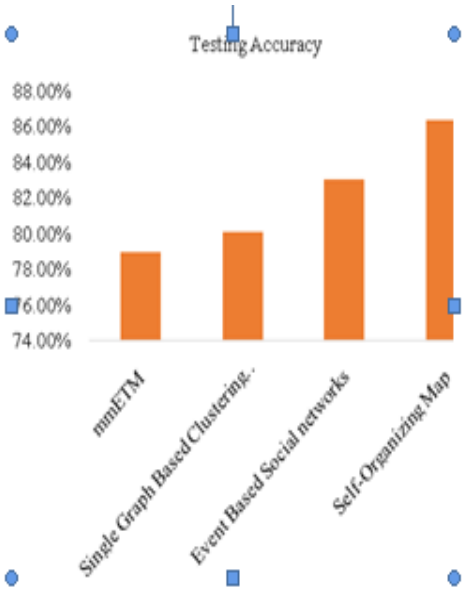

Fig 1: Analysis of various Algorithm

In proposed system, we have used the unsupervised algorithm called Simply Object Model (SOM) algorithm.This algorithm helps to overcome the above disadvantages found. This is helpful for achieving the best results compared to the above algorithm. The data can found in structured form. Easier to find a certain event as data are in attribute mining and domain partition.

VI CONCLUSION

Thus, the existing algorithm is being replaced by the proposed algorithm. The data are found in the

structured form. The summarized data are being converted into events. The attribute mining technique can make easier for the users to search the events. The summarized data found in every resource can be converted to attributes. This also implements the domain partition. This makes the user to find the details on time.

VII REFERENCES

[1] D. Patel, W. Hsu, M. L. Lee, "Mining relationships among interval-based events for classification", Proc. SIGMOD, pp. 393-404, 2008.

[2]. J. Allan, R. Papka, V. Lavrenko, "On-line new event detection and tracking", Proc. SIGIR, pp. 37-45, 1998.

[3]. M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, "Semantic model vectors for complex video event recognition", IEEE Trans. Multimedia, vol. 14, no. 1, pp. 88-101, Feb. 2012.

[4]. T. Zhang, C. Xu, "Cross-domain multi-event tracking via CO-PMHT", ACM Trans. Multimedia Comput. Commun. Appl., vol. 10, no. 4, pp. 31:1-31:19, 2014.

[5]. X. Yang, T. Zhang, C. Xu, M. S. Hossain, "Automatic visual concept learning for social event understanding", IEEE Trans. Multimedia, vol. 17, no. 3, pp. 346-358, Mar. 2015.

[6]. L. Xie, "Discovering meaningful multimedia patterns with audio-visual concepts and associated text", Proc. ICIP, pp. 2383-2386, 2004.

[7]. D. M. Blei, J. D. Lafferty, "Dynamic topic models", Proc. ICML, pp. 113-120, 2006.

[8]. Y. Yang, J. Zhang, J. Carbonell, C. Jin, "Topic-conditioned novelty detection", Proc. KDD, pp. 688-693, 2002.

[9]. J. Makkonen, H. Ahonen-Myka, M. Salmenkivi, "Simple semantics in topic detection and tracking", Inform. Retrieval, vol. 7, no. 3–4, pp. 347-368, 2004.

[10]. N. Diakopoulos, M. Naaman, F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry", Proc. IEEE Symp. Vis. Analytics Sci. Technol., pp. 115-122, 2010-Oct.

[11]. X. Wu, C.-W. Ngo, A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint", IEEE Trans. Multimedia, vol. 10, no. 2, pp. 188-199, Feb. 2008.

[12]. I. Kalamaras, A. Drosou, D. Tzovaras, "Multi-objective optimization for multimodal visualization", IEEE Trans. Multimedia, vol. 16, no. 5, pp. 1460-1472, Aug. 2014.

[13]. X. Chen, A. O. Hero III, S. Savarese, "Multimodal video indexing and retrieval using directed information", IEEE Trans. Multimedia, vol. 14, no. 1, pp. 3-16, Feb. 2012.

[14]. W. Liu, T. Mei, Y. Zhang, "Instant mobile video search with layered audio-video indexing and progressive transmission", IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2242-2255, Dec. 2014.

[15]. C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval", IEEE Trans. Multimedia, vol. 17, no. 3, pp. 370-381, Mar. 2015.