Identifying Emerging Topics Using Link Anomaly Detection in Social Media

Shweta V. Saswade[1], **Prof. S. S. Nandgaonkar**[2]

[1]Student of ME-II, Department of Computer Engineering,
VPCOE, Baramati, Savitribai Phule Pune University Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering,
VPCOE, Baramati, Savitribai Phule Pune University Maharashtra, India

Abstract: ***In data mining era, detecting and generating new concepts has attracted much attention. To discover, the emergence of new topics in news data is a biggest challenge in data mining. This problem can be enlarged as discovering a new concept. Few years ago, domain experts detected emergence of new stories. But it is very critical and time consuming task to read stories and concluding misbehaviors manually. In addition, mapping these misbehaviors to various stories requires excellent knowledge about the old concepts and news. Also automatically modeling a new concept has much importance in data mining. The outliers in news are the basic clues for concluding the emergence of a new story. The outliers are the keywords which doesn't match the whole concept of the news. These outliers are mapped to the stories where this keyword does not behave as outliers. After mapping these outliers, anomaly linking can generate a new concept which can be modeled as emerging story. News Classification, Anomaly Detection, Concept Detection and Generation techniques can be used to efficiently model the new concept.***

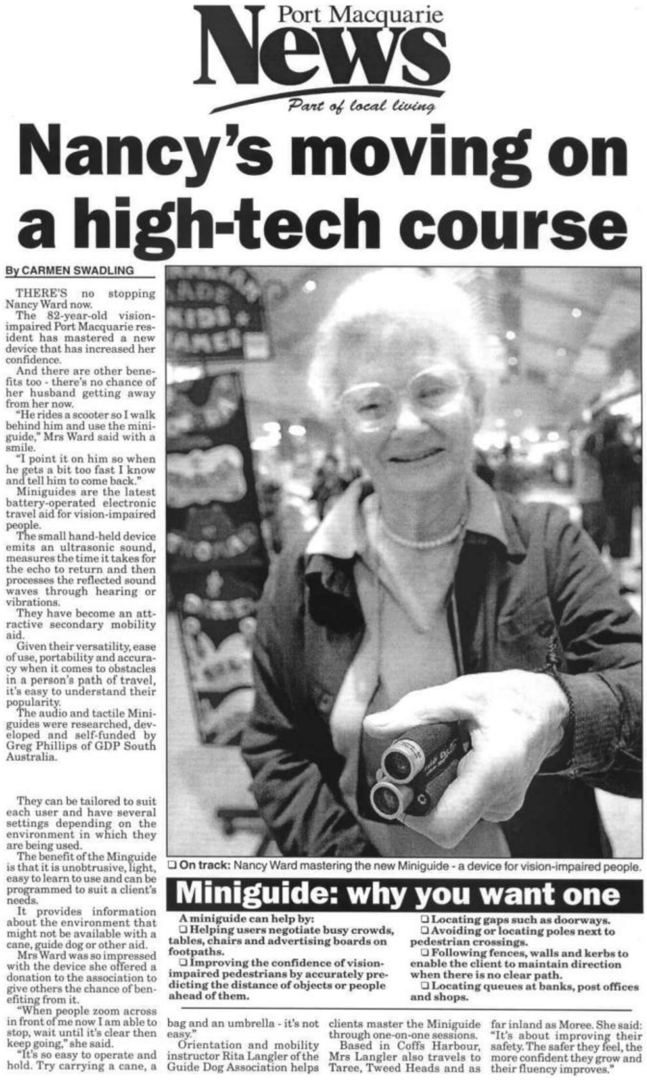
Keywords: News Classification, Anomaly Detection, Anomaly Linking, Concept Detection, Concept Generation.

1. Introduction

The news data is growing tremendously in real-time so the new concepts are getting added to the web. The focus is towards the new topics which can be discovered by mapping some of the previously discussed or published data. Social media platforms have evolved far beyond passive facilitation of online social interactions. It is the need of an hour to analyze the information content in online social media (news articles, blogs, tweets etc.). It allows business to understand public opinion about policies and products. In most of these cases, data points appear as a stream of high dimensional feature vectors. We revisit the problem of online learning of topics from social media content in real-world industrial deployment scenarios. On one hand, the statistics of incoming data points is adapted by the topics dynamically and on the other h and, early detection of new trends is important in many applications.

Previous methods propose online nonnegative matrix factorizations framework. This framework is mainly used to capture the evolution and emergence of themes in unstructured text under a novel temporal regularization framework. An optimization algorithm is developed for this framework and also for streaming Twitter data.

Emerging themes are rapidly captured by the previous system. Also previous system can track the existing topics over time while maintaining temporal consistency and can be explicitly configured to bind the amount of information being presented to the user.



Port Macquarie News

Part of local living

Nancy's moving on a high-tech course

By CARMEN SWADLING

THERE'S no stopping Nancy Ward now.

The 82-year-old vision-impaired Port Macquarie resident has mastered a new device that has increased her confidence.

And there are other benefits too - there's no chance of her husband getting away from her now.

"He rides a scooter so I walk behind him and use the miniguide," Mrs Ward said with a smile.

"I point it on him so when he gets a bit too fast I know and tell him to come back."

Miniguides are the latest battery-operated electronic travel aid for vision-impaired people.

The small hand-held device emits an ultrasonic sound, measures the time it takes for the echo to return and then processes the reflected sound waves through hearing or vibrations.

They have become an attractive secondary mobility aid.

Given their versatility, ease of use, portability and accuracy when it comes to obstacles in a person's path of travel, it's easy to understand their popularity.

The audio and tactile Miniguides were researched, developed and self-funded by Greg Phillips of GDP South Australia.

They can be tailored to suit each user and have several settings depending on the environment in which they are being used.

The benefit of the Minguide is that it is unobtrusive, light, easy to learn to use and can be programmed to suit a client's needs.

It provides information about the environment that might not be available with a cane, guide dog or other aid.

Mrs Ward was so impressed with the device she offered a donation to the association to give others the chance of benefiting from it.

"When people zoom across in front of me now I am able to stop, wait until it's clear then keep going," she said.

"It's so easy to operate and hold. Try carrying a cane, a bag and an umbrella - it's not easy."

Orientation and mobility instructor Rita Langler of the Guide Dog Association helps clients master the Miniguide through one-on-one sessions.

Based in Coffs Harbour, Mrs Langler also travels to Taree, Tweed Heads and as far inland as Moree. She said: "It's about improving their safety. The safer they feel, the more confident they grow and their fluency improves."

❑ **On track:** Nancy Ward mastering the new Miniguide - a device for vision-impaired people.

Miniguide: why you want one

A miniguide can help by:

❑ **Helping users negotiate busy crowds, tables, chairs and advertising boards on footpaths.**

❑ **Improving the confidence of vision-impaired pedestrians by accurately predicting the distance of objects or people ahead of them.**

❑ **Locating gaps such as doorways.**

❑ **Avoiding or locating poles next to pedestrian crossings.**

❑ **Following fences, walls and kerbs to enable the client to maintain direction when there is no clear path.**

❑ **Locating queues at banks, post offices and shops.**

Fig. 1: Example of news on a newspaper

2. Related Work

A. Classification using K-means Clustering

An extensive survey of different clustering technique is provided by B.G.Obula Reddy, Maligela Ussenaiah [2]. Generally clustering algorithms can be classified into hierarchical clustering methods, partitioning clustering methods, density-based clustering methods, grid based clustering methods

Clustering is the grouping of data objects in which data objects from the same cluster are more similar to each other than objects from different clusters.

1. Partition methods [2]: - Partitioning methods are divided into two types, one is centroid and other is medoids algorithms. Each cluster is represented centroid algorithm using the gravity Centre of the instances. In medoid algorithms each cluster is represented by using instances which is closest to gravity Centre. K-means is the well-known centroid algorithm. The k-means algorithm, partitions the data set into k subsets such that all points in a given subset are closest to the same center. In k-medoids algorithm, instead of calculating the mean of the items in each cluster, a representative item, or medoid, is chosen for each cluster at each iteration.

Advantages:

K-means is scalable and efficient in processing large data sets.

Disadvantages:

Clusters form in different sizes and densities.

2. Hierarchical methods [2]: - A hierarchical method divides the given data set into smaller subsets in hierarchical manner. It group the data instances into a tree of clusters. It is divided into two subcategories one is agglomerative method and other is divisive method. In agglomerative method, clusters are formed in a bottom-up fashion until all data instances belong to the same cluster. In divisive method, the data set is split into smaller cluster in a top-down fashion until each of clusters contains only one instance. Both algorithms can be represented by dendrograms.

Advantages:

This method addresses the scalability problem and improves the quality of clustering results.

Disadvantages:

Suffer from identifying only convex or spherical clusters of uniform size.

3. Density based methods [2]: - This algorithm try to find clusters based on density of data points in a region. The idea of this algorithm is that for each instance of a cluster, the neighborhood of a given radius containing at least a minimum number of instances. DBSCAN is one of the most well-known density based clustering algorithm. It separate data points into three classes.

1) Core points: - These points are at the interior of a cluster.
2) Border points: -These are the points that are not a core points.
3) Noise point: - A point which is not core point or border point is referred as Noise Point.

Advantages:

For low dimensional data, DBSCAN formula performs with high efficiency and effectively.

Disadvantages:

The algorithm is not partitionable for multiprocessing systems.

4. Grid based methods [2]: - This algorithm first quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. The dense cells are connected to form the clusters. Dense cells mean the cells that contain more than certain number of points. Following are some grid-based clustering algorithms.

Wave Cluster, Statistical Information Grid-based method STING. STING divides the spatial area into several levels of rectangular cells in order to form a hierarchical structure.

Advantages:

The grid-based computation is query-independent and it provide an efficiency.

Disadvantages:

Boundaries of all clusters are either horizontal or vertical, and no diagonal boundary is chosen.

B. Anomaly Detection or Outlier Detection

VARUN CHANDOLA, ARINDAM BANERJEE, VIPIN KUMAR [3] presented a broad review of anomaly detection technique in 2009.

1. Classification based Anomaly Detection technique [3]: - It classifies the instances into one of the classes by learning the set of labeled data instances. This technique includes the two phases, training and testing phase. The training phase learns a classifier using the labeled training data. The testing phase classifies a test instances as normal or anomaly. Based on labels, this category is divided into multiclass and one-class anomaly detection technique. The training data contains labeled instances with multiple normal classes in multiclass technique. In One-class classification based techniques, all training instances have only one class label.

Advantages:

4) The multi-class techniques use powerful algorithms that can distinguish between instances belonging to different classes.
5) The testing phase of this technique is fast.

Disadvantages:

(1) Multi-class classification based techniques depend on availability of accurate labels for various normal classes, which is often not possible.
(2) This techniques assign a label to each instance which is not useful when anomaly score is considered for the instances.

2. Nearest neighbor based Anomaly Detection technique [3]: - These anomaly detection techniques require a distance or similarity measure which is defined between two data instances. Distance between two data instances can be measured in different ways. Euclidean distance is used for continuous attributes whereas for categorical attributes, simple matching coefficient is used.

Advantages:

(1) They are unsupervised in nature and do not make any assumptions regarding the generative distribution for the data. Instead, they are strictly data driven.
(2) In terms of missed anomalies, semi-supervised techniques perform better than unsupervised therefore the likelihood of an anomaly to form a close neighborhood is very low.

Disadvantages:

(1) If the data instances do not have enough close neighbors then the technique fails to label them correctly in unsupervised techniques.

3. Clustering based Anomaly Detection techniques [3]:- Clustering is used to group similar data instances into clusters. Clustering and anomaly detection techniques are different from each other but most of the clustering based anomaly detection techniques have been developed.

Advantages:

1) These techniques can operate in an unsupervised mode.
2) This techniques can be used for complex data types by simply plugging the clustering algorithm.
3) The testing phase is fast.

Disadvantages:

(1) Performance is highly dependent on the effectiveness of clustering algorithm.

4. Statistical Anomaly Detection techniques [3]: - This technique fit a statistical model to the given data and then applies a statistical inference to determine an unseen instance to this model. Instances that have a low probability are declared as anomalies. Both parametric and non-parametric techniques have been applied to fit a statistical model.

Advantages:

1) Statistical techniques provide a convenient solution for anomaly detection.
2) The anomaly score provided is based on confidence interval, which can be used while making a decision regarding any test instance.
3) Statistical techniques can operate in an unsupervised setting without any need for labeled training data.

Disadvantages:

(1)The key disadvantage of this technique is that they depend on the data is generated from a particular distribution. Sometimes this assumption does not hold true for high dimensional real data sets.

C. Concept Generation
Concept generation has been studied for number of purposes over the recent years and several algorithms have been proposed.

1. Concept generation technique is suitable for search domains, suggested by Callan in 1993 [6]. He observed that search goal descriptions are usually not monolithic, but rather consist of sub expressions, each describing a goal. They proposed a set of heuristics for decomposing goal specifications into their constituent parts, in order to use them as concepts.

Concept generation approaches:

1.1 Data-driven, in these class new features is created by combining existing features in various ways. Feedback is taken from the learned concept then it is used to suggest plausible concept combinations.

1.2 Analytical, this class uses domain theory to deduce appropriate new features. Create complex features using information about the domain, in one step.

2. Fawcett [6] proposed a hybrid theory of feature generation in 1991. In this technique, features can be derived from abstractions and combinations of abstractions of the domain theory. Abstractions are created by using a hybrid of data-driven (bottom-up) and theory-driven (top-down) approaches.

3. Implementation Details

News data is in unstructured format so that it required breaking down in into structured format. To detect anomalies is always suffered from ambiguity caused by synonyms and homonyms. Classification of news into fix amount of known categories is difficult and time consuming task and is required to manage. Linking anomalies for concept detection and generation requires extensive domain knowledge and it is also time-consuming process for large amounts of data.

3.1 Architecture

1) Preprocessing:
In this phase, removal of high frequency words, stop-words and non-words is performed. This is done by comparing the input text with and list of stop-words and non-words. After removing stop-words and non-words, stemming operation is performed. The advantage of this process is that non-significant words are removed.

2) News Classification:
News classification is a problem in library science, information science and computer science. The main task of this module is to assign news to one or more classes or categories. This may be done” manually” or algorithmically.

In information science and computer science, classification of documents is done by algorithmically while in library science it is done by manually.

Because of problems of overlapping, there is an interdisciplinary research on news classification [2].For classification of News data; K-means algorithm is used.

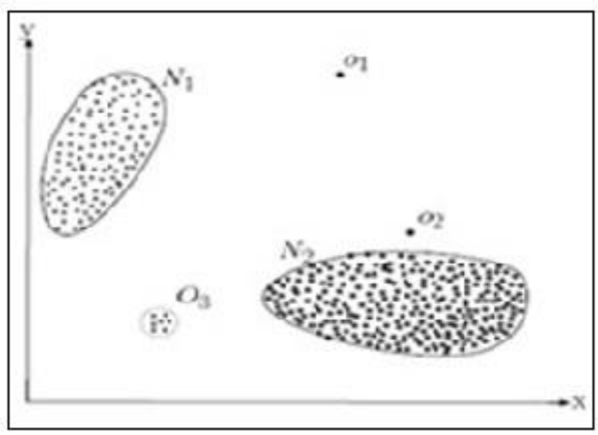


Fig. 2: Basic structure of News Classification

Division of information into groups of similar objects is termed as clustering. There are many different clustering algorithms among them K-means is one of the simplest and most popular. But it is not capable of dealing with non-convex shapes. This is nothing but partitioned clustering approach [4].

In this method, seeding is performed and k numbers of clusters are formed initially. After that each point is assigned to the cluster with the closest centroid.

K-means algorithm

a. Place K points into the space as the initial centroids.
b. Assign each point to the closest centroid.
c. When all points are assigned, recalculate the centroid.
d. Repeat the process b and c until the centroid do not change

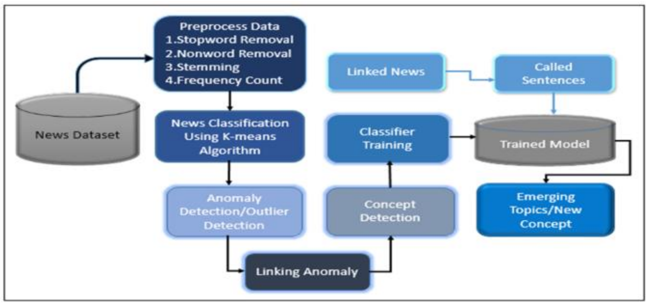


Figure 3: System Architecture

News data categorization problems are usually linearly separable. If the classes are linearly separable, then they are convex so that K-means clustering is a simple baseline. As a result of, it generates hyper spherical clusters that are convex. It covers the whole vector space of presented points and relatively balanced.

3) Outlier Detection:

Anomaly detection refers to the problem of finding pattern in the data that do not conform to expected behavior. LDA algorithm is used to find outliers between the given news dataset. LDA split out the words in the document using certain probabilities and performs topic modelling. Documents are produced using following steps.

Decide on the number of words N the document will have

1) Select a topic mixture for the document using a Dirichlet distribution where a fixed set of 2 topics.
2) Generate each word in the document by picking a topic according to multinomial distribution.

4) Linking Anomaly:

The anomalies are detected from each news class. That anomaly is mapped to the news which has same set of tokens but is not identified as anomalies. The TF-IDF values of all tokens are calculated in the mapped news which contains anomaly as well as the News class which has the same set of words. It is a numerical evaluation of how important a word is to a document in a corpus. It calculates possibility value of the number of times a word appears in the document. After that taking mean of all tokens in the mapped news. The mean is taken as threshold value. If token tf-idf value is greater than threshold then it is considered as most probable token and these tokens are ranked for further processing.

5) Concept Detection:
Concept descriptions are parsed for concept detection. LDA algorithm is used for the concept detection technique. The main goal of LDA algorithm is to discover the topics from a collection of documents. For outlier detection purpose, the number of topics are kept constant i.e. t=2. Out of these two topics, we label the larger one as concept and the other as anomalies.

6) Concept Generation: Concepts generation means building new features using old or current stories. Concept means providing function and indication of how the function is to be

achieved.

Different methods of Concept Generation:

1. Morphological Method: This method uses the functions to identify the foster ideas. It is can be used formally or informally as part of everyday thinking. It is powerful technique [7].

It includes two steps.

In first step, Concepts are developed for each function. Find many concepts as possible that can provide each function identified in the decomposition. The function must be reexamined, if functions have only one conceptual idea. This condition explains the lack of more concepts. The designer has made an assumption that is domain knowledge is limited.

The concepts are kept as an abstract as possible for better comparison of developed concepts.

In second step, concepts are combined. These individual concepts are combined into overall concepts to meet all the functional requirements. One concept is selected for each function and combined into a single design.

2. Logical Methods for Concept Generation: It is based on patterns found in patented ideas [7]. The purpose of this method is to find the major contradiction that is making the problem hard to solve, and then use inventive ideas for overcoming the contradiction. With TRIZ, we can systematically change into new and also we do not have to wait for a trial and error which is common to other methods. Axiomatic Design: To make the design process logical, axiomatic design is used [8].

1. It maintains the independence and then change in a specific design parameter should have an effect only on a single function.
2. It minimizes the information content of the design.

In concept generation phase, vector classifier is used for training the most probable tokens which are generated in linking anomaly phase. These tokens are learned in Training model. The sentences in the mapped news classes are classified against the training model and then the probability of most probable tokens with each sentences are calculated. The mean of all probabilities are taken as a threshold. The sentences which have probability greater than the threshold are considered as an emerging topic or a new concept.

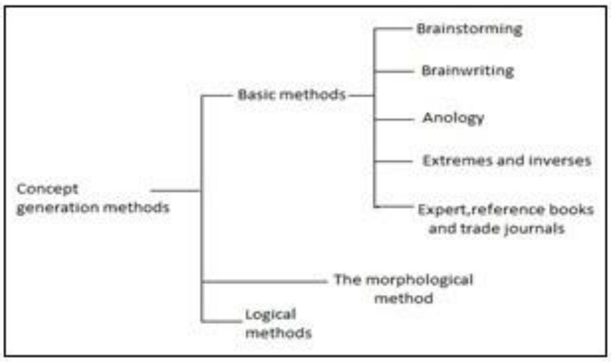


Figure 4: Basic methods of Concept Generation

3.2 Relevant Mathematics

Input Set

N= {N_i; 0 < i <n}
Where,
N= set of News
n= number of News

P_N = {P_N ; 0 < i <n}
Where,
P_N = set of Preprocessed News
n = number of News

Processing Set

T = {T_i ; 0 < i <t}
Where,
T = set of tokens
t = number of token

TN = {TN_{ij}, $\exists TN_{ij}$ |T_i ϵ T, N_j ϵ N}
Where,
TN = TN matrix
TN_{ij} = number of Occurrences

C_i = {C_i; 0 < i < k}
Where,
C_i = Set of classes
K = number of News Classes

$O_N = \{O_{N_i}; 0 < i < m\}$, Set of outliers
Where,
O_{N_i} = distance (O_{N_i}, O_{N}) ≥ threshold
$O_{\{N\}}$ = number of tokens in Outlier Set.

$S_i m_{ij}$ = {(C_i, N_i), $\exists N_i$ |$dist(N_i, C_i) \le threshold$}
Where,
C_i ϵ C ,
N_i ϵ N,
$threshold = [0,1]$

$NM = \{NM_{ij}; 0 < j < n\}$

Where,

NM = set of News mapping

$M_i \epsilon \mathrm{N}$

$\exists M_{ij} | O_{N_i} \neq O_{N_j}, D_{N_i} \epsilon \mathrm{N_i}, \mathrm{N_j}$

Output Set

$BN = \{BN_i, 0 < i < m\}$ Set of breaking news.

Where,

m = number of breaking news

$\exists S_{N_i} | T \epsilon \mathrm{N_i} \geq \mathrm{threshold}$

$threshold = [0,1]$

4. Dataset and Result

1) The corpus of news stories is available at Reuters, Ltd. Reuters home page gives details about the collection and how to obtain it. Reuter.com brings latest news around the world. Mailing list is also available for discussions about the collection.

Table 1

Dataset	*#News*	*#Link should form*	*#Link*	*#Correct Link*	*Precision*	*Recall*
1	50	20	20	10	0.5	1
2	70	40	30	30	1	0.75
3	90	50	50	40	0.8	1
4	160	70	70	60	0.86	1
5	210	110	110	80	0.73	1

Five data sets are collected from google news. Each dataset is associated with a list of news, links are found during anomaly linking. News that is related to each other and organizes a new concept, that news is used for anomaly linking. Main goal of this system is to detect and generate the emergence of topics. In table 1, the number of news collected for each dataset. In dataset1, two links are formed by system. Out of these two, one link is correct.

The Fig.5a shows the result of the link anomaly detection. The result varies as the number of linking increases or decreases in the news dataset. If the news is related to each other then it will give better result. Precision depends on correct link found in the news dataset. Recall depends on link found in the system.

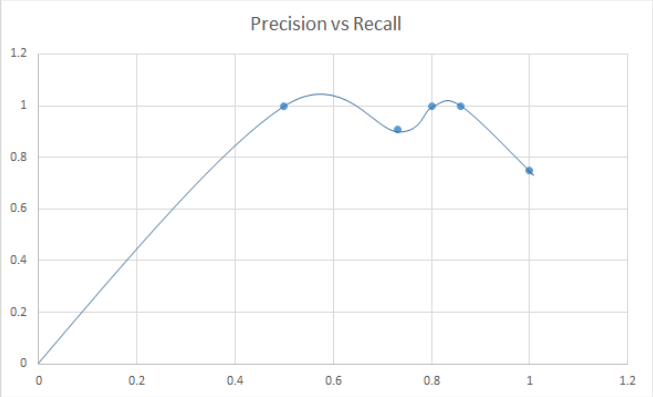
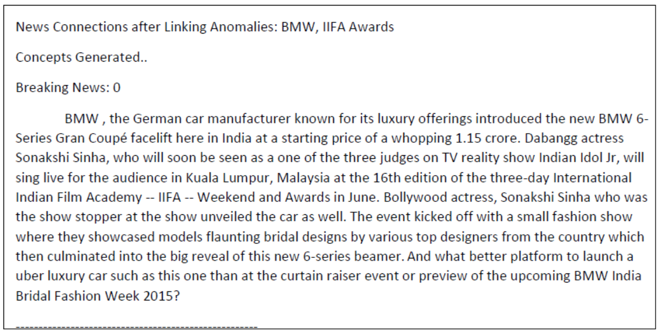


Figure 5a: Precision vs Recall of five news dataset

2) After preprocessing phase, news dataset is classified into k number of clusters. Anomalies are detected from each news class. After mapping of these anomalies to the news class, a new concept is generated.

Figure 5c shows that the system combined two news that is "BMW car launched by Sonakshi Sinha" and "Sonakshi Sinha sing in IIFA Awards 2015," by using linking anomaly method and generate an emerging topics or new concept.



News Connections after Linking Anomalies: BMW, IIFA Awards

Concepts Generated..

Breaking News: 0

BMW , the German car manufacturer known for its luxury offerings introduced the new BMW 6-Series Gran Coupé facelift here in India at a starting price of a whopping 1.15 crore. Dabangg actress Sonakshi Sinha, who will soon be seen as a one of the three judges on TV reality show Indian Idol Jr, will sing live for the audience in Kuala Lumpur, Malaysia at the 16th edition of the three-day International Indian Film Academy -- IIFA -- Weekend and Awards in June. Bollywood actress, Sonakshi Sinha who was the show stopper at the show unveiled the car as well. The event kicked off with a small fashion show where they showcased models flaunting bridal designs by various top designers from the country which then culminated into the big reveal of this new 6-series beamer. And what better platform to launch a uber luxury car such as this one than at the curtain raiser event or preview of the upcoming BMW India Bridal Fashion Week 2015?

Figure 5c: News Generated from Dataset 2

5. Conclusion

Recently it is found that the discovering of news topics is challenging task and has much importance in data mining fields. In this paper, a new approach is used to detect the emergence of topics in a social network stream. The basic idea is to focus on anomaly detection in news class. Anomalies are detected and then mapped to the news class. After mapping these anomalies, a new concept is generated.

Further it has application in forensic analysis to determine the new stories around a topic. So it will always be research field for future researchers.

References

[1] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.

[2] B.G. Obula Reddy, Dr. Maligela Ussenaiah, "Literature Survey on Clustering Techniques," IOSR Journal of Computer Engineering, Volume 3, pp 01-12.

[3] VARUN CHANDOLA, ARINDAM BANERJEE, VIPIN KUMAR, "Anomaly Detection: A Survey," A modified version of this technical report will appear in ACM Computing Surveys, September 2009.

[4] Artur Silie, Lovro Zmak, Bojana Dalbelo, Marie-Francine Moens, "Comparing Document Classification using K-means Clustering".

[5] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 14981512. Sept.2010.

[6] K.A. Kontogiannis, R. Demori, M. Galler, M. Bernstein,"Pattern matching for Clone and Concept Detection," Automated Software Engineering Volume 3, pp 77-108, 1996.

[7] Genrikh Altshuller, "Concept Generation," Soviet patent investigator, 1950.

[8] Prof. Nam Suh, "Axiomatic Design for Concept Generation," MIT.

[9] Ankan Saha and Vikas Sindhwani: 2012,“ Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization," In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12).ACM, New York, NY, USA,693-702.DOI=10.1145/2124295.2124376http://doi.acm.org/10.1145/2124295.2124376.

[10] Victoria J. Hodge, "A survey of outlier Detection Methodologies," Kluwer Academic Publisher, Netherlands, 2004.

[11] Aha, D. W. and Bankert, R. B.: 1994, "Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison", In: Proceedings of the AAAI-94, Workshop on Case-Based Reasoning.

Authors

Shweta Saswade received her B.E. degree in Information Technology from University of Pune in 2012. She is currently working toward the M.E. Degree in Computer Engineering from University of Pune, Pune. Her research interests lies in Data Mining, Information Retrieval, and Data Classification.