Where,

NM = set of News mapping

$M_i \epsilon N$

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

Output Set

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

Where,

m = number of breaking news

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

$threshold = [0,1]$

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.

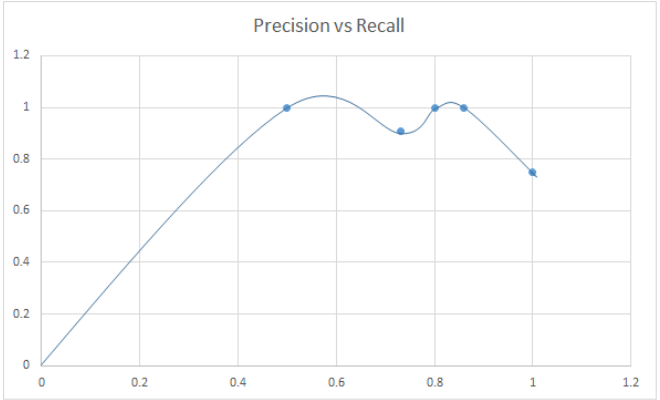
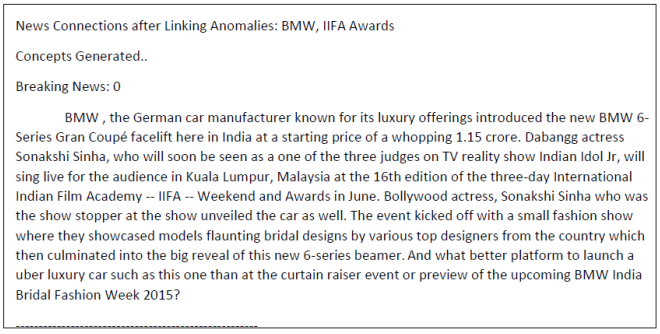


Figure 5a: Precision vs Recall of five news dataset

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.