Exploiting Emergence of New Topics via Anamoly Detection: A Survey

Miss. S. V. Saswade
ME II Computer
VPCOE, Baramati
Pune University (MH), India.
Pune, India
shweta.saswade0@gmail.com

Prof. S. S. Nandgaonkar
Assistant Professor
VPCOE, Baramati
Pune University (MH), India.
Pune, India
sushma.nandgaonkar@gmail.com

Abstract- Detecting and generating new concepts has attracted much attention in data mining era, nowadays. The emergence of new topics in news data is a big challenge. The problem can be extended as *"finding breaking news"*. Years ago the emergence of new stories were detected and followed up by domain experts. But manually reading stories and concluding the misbehaviors is a critical and time consuming task. Further mapping these misbehaviors to various stories needs excellent knowledge about the news and old concepts. So automatically modeling breaking news has much interest in data mining. The anomalies in news published in newspapers are the basic clues for concluding the emergence of a new story(s). The anomalies are the keywords or phrases which doesn't match the whole concept of the news. These anomalies then processed and mapped to the stories where these keywords and phrases doesn't behave as anomalies. After mapping these anomalies one can conclude that these mapped topic by anomaly linking can generate a new concept which eventually can be modeled as emerging story. We survey some techniques which can be used to efficiently model the new concept. News Classification, Anomaly Detection, Concept Detection and Generation are some of those techniques which collectively can be the basics of modeling breaking news. We further discussed some data sources which can process and used as input stories or news for modeling emergence of new stories.

Keyword- *News Classification, Anomaly Detection, Anomaly Linking, Concept Detection, Concept Generation*

*****

1. Introduction

As massive news data is increasing real-time, the new concepts are getting added to the web. The attention is getting towards the new topics which can be discovered by linking some of the previously discussed or published data. Social media platforms have arguably evolved far be yond passive facilitation of online social interactions. Rapid analysis of information content in online social media streams (news articles, blogs, tweets etc.) is the need of the hour as it allows business and government bodies to understand public opinion about products and policies. In most of these settings, data points appear as a stream of high dimensional feature vectors. Guided by real-world industrial deployment scenarios, we revisit the problem of online learning of topics from streaming social media content. On one hand, the topics need to be dynamically adapted to the statistics of incoming data points, and on the other hand, early detection of rising new trends is important in many applications. Literature proposes an online nonnegative matrix factorizations framework to capture the evolution and emergence of themes in unstructured text under a novel temporal regularization framework. They develop scalable optimization algorithms for our framework, propose a new set of evaluation metrics, and report promising empirical results on traditional TDT tasks as well as streaming Twitter data. Previous systems are able to rapidly capture emerging themes, track existing topics over time while maintaining temporal consistency and continuity in user views, and can be explicitly configured to bind the amount of information being presented to the user [7].

Here in our survey study we organized the paper as follows. Section 2 describes various data sources from which we can have the input data. In Section 3 we surveyed some text classification techniques used for news classification. Section 4 describes different method for outlier/anomaly detection. In Section 5 and 6 we have analyzed concept detection and generation methods respectively. Section 7 deals with the challenges faced during the outlier detection while in section 8 we have discussed the application of the discovering emerging topics via link anomaly detection.

Figure 1: Various concepts in daily news data

ISSN: 2321-8169
4063 - 4069

2. Data Sources

Now a day there are n number of data sources are available for sentiment analysis. Customer's opinion is a major criterion for increasing the growth of the company and to improve the quality of the service. The different data sources are social media, news articles, review sites, blogs, datasets, etc [5]

2.1 Social Media

Twitter is a popular micro blogging service that enables the users to send and read short text messages commonly known as tweets. Twitter one of the fastest growing web sites in the world. People post tweets for a variety of purposes, including daily chatter, conversations, sharing information/URLs and reporting news, defining a continuous real-time status stream about every argument. People re-tweet on the given tweets.

Figure 2: News data in Twitter Tweet

2.2 News Articles

The websites like www.googlenews.co.in, www.abpmajha.com, www.aajtak.com and www.lokmat.com, www.bhaskar.com has news articles that allow users or readers to comment. This helps in recording the opinions of the people in issues that are of current relevance and importance, like politics, corruption etc


Port Macquarie News

Nancy's moving on a high-tech course

Miniguide: why you want one

Figure 3: Example of news on a news paper

2.3 News magazines

A news magazine is a typed, printed, and published piece of paper, magazine or a radio or television program, usually weekly, showing articles on current events. News magazines generally go more in-depth into stories than newspapers or television programs, trying to give the reader an understanding of the important events, rather than just the facts.

Figure 4: News collection in magazines

2.4 Blogs

A web log is called as blog it is a personal webpage on which particulars can write their likes, dislikes, opinions, hyperlinks to various sites etc. daily. Tweeter is one of the popular micro blogging service in which user creates status messages in a limited word count which called as tweets. The tweeter will get flooded while the elections were going on. Tweets can also use as data source for sentiment classification. Many of the blogs contain the issues; product information's recopies etc. so blogs used for the data source of the sentiment analysis [8]

2.5 Datasets

A corpus of newswire stories recently made available by Reuters, Ltd. Details about the collection and how to obtain it can be found at Reuter's home page for corpora. There is also a mailing list for discussions about the collection. I have written, along with Yiming Yang, Tony Rose, and Fan Li, a JMLR paper, describing the collection and defining a corrected version of the collection, RCV1-v2. Two formatted versions of RCV1-v2, and other useful files, are available as online appendices to that paper.

3. News Classification

News text classification or news categorization is a problem in library science, information science and computer science. The task is to assign a news to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of news

 ISSN: 2321-8169
4063 - 4069

is used mainly in information science and computer science. The problems are overlapping, however, and there is therefore also interdisciplinary research on news classification [1].

3.1 k-means classification

Clustering is the division of data into groups of similar objects. Among many different clustering algorithms, K-means is one of the simplest and most popular. K-means is not capable of dealing with non-convex shapes. Partitioned clustering approach [2].

a) Each cluster is associated with a centroid (center point)

b) Each point is assigned to the cluster with the closest centroid

c) Number of clusters, K, must be specified

The basic algorithm is very simple

1: Select K points as the initial centroids.

2: repeat

3: from K clusters by assigning all points to the closest centroid.

4: Recomputed the centroid o F each cluster.

5: Until the centroid do not change.

Text categorization problems are usually linearly separable. If the classes are linearly separable, then they are convex as well. This justifies the use of K-means clustering as a simple baseline, because it generates hyper spherical clusters that are convex, able to cover the whole vector space of presented points, and relatively balanced.

3.2 K- Medoid

The k-Medoids: in k-medoids algorithm, Rather than calculate the mean of the items in each cluster, a representative item, or medoid, is chosen for each cluster at each iteration.

The k-medoids algorithm can be summarized as follows:

1. Choose k objects at random to be the initial cluster medoids.
2. Assign each object to the cluster associated with the closest medoid.
3. Recalculate the positions of the k medoids.
4. Repeat Steps 2 and 3 until the medoids become fixed.

Step 3 could be performed by calculating for each object i from scratch at each iteration. However, many objects remain in the same cluster from one iteration of the algorithm to the next. Improvements in speed can be obtained by adjusting the sums whenever an object leaves or enters a cluster. Step 2 can also be made more efficient in terms of speed, for larger values of k. For each object, an array of the other objects, sorted on distance, is maintained. The closest medoid can be found by scanning through this array until a medoid is found, rather than comparing the distance of every medoid.

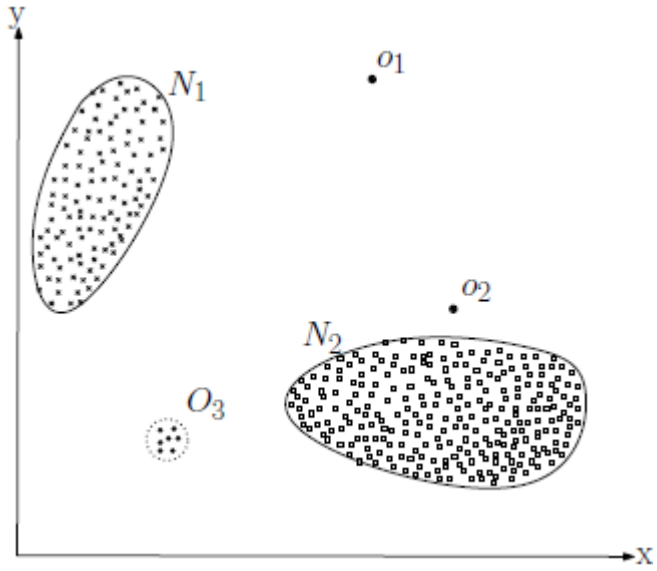

Figure 5: Basic Structure of text Classification

4. Outlier/Anomaly Detection

The data mining partitional algorithm CLARANS, is an optimized derivative of the k-medoids algorithm and can handle outlier detection which is achieved as a by-product of the clustering process. It applies a random but bounded heuristic search to find an optimal clustering by only searching a random selection of cluster updates. It requires two user-specified parameters, the value of k and the number of cluster updates to randomly select. Rather than searching the entire data set for the optimal medoid it tests a pre-specified number of potential medoids and selects the first medoid it tests which improves the cluster quality [8].

4.1. Proximity-based Techniques

K-Nearest Neighbor (k-NN) algorithm for outlier detection but all calculate the nearest neighbors of a record using a suitable distance calculation metric such as Euclidean distance or Mahalanobis distance. Euclidean distance is simply the vector distance whereas the Mahalanobis distance calculates the distance from a point to the centroid (μ) defined by correlated attributes given by the Covariance matrix (**C**). Mahalanobis distance is computationally expensive to calculate for large high dimensional data sets compared to the Euclidean distance as it requires a pass through the entire data set to identify the attribute correlations [8].

4.2. K-medoid

A very similar partitioning algorithm is the k-medoids algorithm or PAM (Partition around Medoids) which represents each cluster using an actual point and a

radius rather than a prototype (average) point and a radius. Bolton & Hand use a k-medoids type approach they call Peer Group Analysis for fraud detection. K-medoids is robust to outliers as it does not use optimization to solve the vector placement problem but rather uses actual data points to represent cluster centers. K-medoids is less susceptible to local minima than standard k-means during training where k-means often converges to poor quality clusters. It is also data-order independent unlike standard k-means where the order of the input data affects the positioning of the cluster centers and Bradley shows that k-medoids provides better class separation than k-means and hence is better suited to a novelty recognition task due to the improved separation capabilities. However, k-means outperforms k-medoids and can handle larger data sets more efficiently [8].

4.3. Partitional Clustering

The data mining partitional algorithm CLARANS, is an optimized derivative of the k-medoids algorithm and can handle outlier detection which is achieved as a by-product of the clustering process. It applies a random but bounded heuristic search to find an optimal clustering by only searching a random selection of cluster updates. It requires two user-specified parameters, the value of k and the number of cluster updates to randomly select. Rather than searching the entire data set for the optimal medoid it tests a pre-specified number of potential medoids and selects the first medoid it tests which improves the cluster quality [8].

Here we discussed some methods for discovering topics

Name of model	Usage
Hidden Markov Model	Detection and tracking of Events.
Temporal Text Mining	Discover, Extract, and Summarize evolutionary Theme patterns automatically.
Change Finder	Detecting change points in Time-series data.
Topic Dynamics	Detection and analysis of bursts
Finite Mixture Model	Discover topic trends and analyze their dynamics

Table 1: Model and their Usage

5. Concept Detection

News concepts are described by a concept language [4]. A concept to be recognized is a phrase of the concept language. Concept descriptions and source code are parsed. The concept recognition problem becomes the problem of establishing correspondences, as in machine translation, between a parse tree of the concept description language and the parse tree of the code.

A new formalism is proposed to see the problem as a stochastic syntax-directed translation. Translation rules are pairs of rewriting rules and have associated a probability that can be set initially to uniform values for all the possible alternatives. Matching of concept representations and source code representations involves alignment that is again performed using a dynamic programming algorithm that compares feature vectors of concept descriptions, and source code[4].

The proposed concept description language, models insertions as wild characters does not allow any deletions from the pattern. The comparison and selection granularity is at the statement level. Comparison of a concept description language statement with a source code statement is achieved by comparing feature vectors.

The use of a statistical formalism allows a score to be assigned to every match that is attempted. Incomplete or imperfect matching is also possible leaving to the software engineer the final decision on the similar candidates proposed by the matcher. A way of dynamically updating matching probabilities as new data are observed is also suggested in this paper. Concept-to-code matching [4] is under testing and optimization. It has been implemented using the REFINE environment and supports plan localization in C programs.

6. Concept Generation

Concepts means for providing function and give an indication how the function can be achieved.

Different methods of Concept Generation:

6.1 Morphological Method

This technique uses the functions identified to foster ideas. It is powerful method that can be used formally or informally as part of everyday thinking [5].

It includes 2 steps

Step 1 - Developing Concepts for Each Function.

The goal is to find as many concepts as possible that can provide each function identified in the decomposition. If there is a function with only one conceptual idea, this function must be re-examined. Situations explaining the lack of more concepts. The designer has made a fundamental assumption that is domain knowledge is limited. Keep the concepts as abstract as possible and at the same level of abstraction for better comparison of developed concepts.

Step 2 - Combining Concepts:

Combine these individual concepts into overall concepts to meet all the functional requirements.

Select one concept for each function and combine those into a single design.

Pitfalls:

1. This method may generate too many ideas.
2. It erroneously assumes that each function of the design is independent and that each function satisfies only one function.
3. The results may not make any sense.

Concept generation process is the time that sketches and begin useful.

Only way to design an object with any complexity is to use sketches to extend the short-term memory. Sketches made in the design notebook provide a clear record of the development of the concept and the product.

6.2 Logical Methods for Concept Generation

The Theory of Inventive Machines, TRIZ: Developed by Genrikh Altshuller in Soviet Union in the 1950s based on patterns found in patented ideas [5]. The goal is to find the major contradiction that is making the problem hard to solve, then use inventive ideas for overcoming the contraindication. With TRIZ, we can systematically innovate; we don't have to wait for an inspiration or use the trial and error common to other methods.

Axiomatic Design: This design is used to make the design process logical [6].

1. Axiom is to maintain the independence and then change in a specific design parameter should have an effect only on a single function.
2. Axiom is to minimize the information content of the design.

The simplest design has the highest probability of success and is the best alternative.

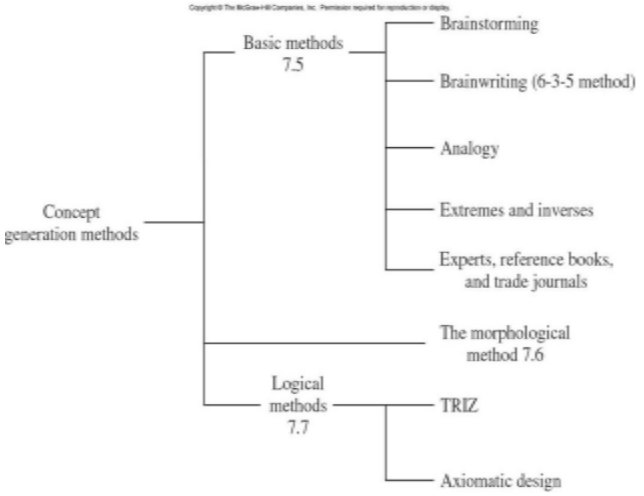

Figure 6: Basic Methods for Concept Generation

7. Challenges for Outlier Detection for Discovering Emerging Topics

Emerging topic detection is ongoing research field it deals with the electronic text. While dealing with this text many pitfalls come in to the picture. These Different pitfalls are listed as follows

1. Natural language processing will have problems if the language contains grammatical errors, ambiguity, co-references so it will give influence on anomaly/outlier detection.
2. The internet contains the authenticated and spam electronic text so for efficient outlier detection the spam text should be removed before processing of the data. This is called detection of the fake and spam comments. This can be done by detecting duplicates, by identifying outliers.
3. In some cases noun words can be also considered as the outlier words but verbs and adjectives can also use as the outlier word which are complicated for identify.
4. Now a day's English shortcuts are come in picture so many of the users comment in the free text style format, he can use the shortcuts, numerical words, abbreviation. For example *"gud"* for *"good"*, *"b4"* for *"before"*, "5in" for *"fine"* etc. so to work with this words lot of work should be needed.
5. The Major challenge in the topic detection is the domain dependent behavior of the user's language. One outlier set may be give good output in one domain as compared to another domain.
6. Now a day's style of comment changes to smileys, images and videos so this will be challenging task to decode it and analyses the opinions of the users in tweet data.
7. The short comes in classification filtering while dealing with the most popular concept. For efficient news classification result of this drawback should be removed. The risk of filter bubble give irrelevant news sets and it result to the false summary of news.
8. The ongoing research should be present to improve the user friendliness and efficiency in the field of emerging topic detection.

8. Applications of Outlier Linking for Discovering Emerging Topics

Under the umbrella of discovering emerging topic using outlier detection and linking there are certain areas which attract the researchers in today's world. The data mining and forensic communities are showing their interest in concept detection and generation.

1) **Concept Detection:** Variety of data uploaded on web day-to-day. Manipulating this data is serious task. This data contains number of topics which may be area of interest.
2) **Concept Generation:** Generating a new concept by concluding the input data is a critical task. The data mining experts working hard on this area. Given a set of words or phrases it's very difficult to generate a

concept. The emerging topic discovery can be applied to this problem.

3) **News/ Blogs/ Tweets Classification:** The data on the social websites is a large input to the data mining. But this data need to be to classified as to make effective use in era of data mining
4) **Outlier Detection:** Detecting misbehaviors in a text document is finding its applications in a vast areas such as data mining,

9. Conclusion

In this paper we have studied the survey of outlier/Anomaly linking for discovering emerging topics in news datasets. We further analyzed different techniques used for topic classification, Outlier Detection, Concept Detection and Concept generations. Further we surveyed challenges of these techniques in details and the applications areas where these methods can be used to improve the mining processes. Anomaly detection is a prominent field of the data mining used to extract the essential knowledge from a enormous amount of users comments, sentiments, reviews, feedback, news, blogs and tweets etc. Now days it is founded that the discovering news topics or trend is challenging task and has much importance in data mining fields. Further it has application sin forensic analysis to determine the new stories around a topic in interest. So it is always be a ongoing research field for future researchers.

Acknowledgement

I express great many thanks to Prof. S. S. Nandgaonkar and Department staff for their great effort of supervising and leading me, to accomplish this fine work. To college and department staff, they were a great source of support and encouragement. To my friends and family, for their warm, kind encourages and loves. To every person who gave me something too light along my pathway. I thanks for believing in me.

References

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

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

[3] 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. 1498–1512. Sept. 2010.

[4] 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.

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

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

[7] 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=10.1145/2124295.2124376 http://doi.acm.org/10.1145/2124295.2124376.

[8] Victoria J. Hodge, “A survey of outlier Detection Methodologies”, Kluwer Academic Publisher, Netherlands, 2004

[9] 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.

[10] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y.: 1998, ‘Topic detection and tracking pilot study: Final report’. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

[11] Arning, A., Agrawal, R., and Raghavan, P.: 1996, ‘A Linear Method for Deviation Detection in Large Databases’. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 164–169.

[12] Baker, L. D., Hofmann, T., McCallum, A. K., and Yang, Y.: 1999, ‘A Hierarchical Probabilistic Model for Novelty Detection in Text’. In: Unpublished manuscript. (Submitted to NIPS’99.).

[13] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons, 3 edition.

[14] Beale, R. and Jackson, T.: 1990, Neural Computing: An Introduction. Bristol, U.K. and Philadelphia, PA: Institute of Physics Publishing.

[15] Bishop, C. M.: 1994, ‘Novelty detection and Neural Network validation’. In: Proceedings of the IEE Conference on Vision, Image and Signal Processing. pp. 217–222.

[16] Adamov, R. “Literature review on software metrics”, Zurich: Institute for Informatics der Univercity Zurich, 1987.

[17] Baker S. B, “On Finding Duplication and Near-Duplication in Large Software Systems” In

 ISSN: 2321-8169
4063 - 4069

Proceedings of the Working Conference on Reverse Engineering 1995, Toronto ON. July 1995

[18] Biggerstaff, T., Mitbander, B., Webster, D., "Program Understanding and the Concept Assignment Problem", Communications of the ACM, May 1994, Vol. 37, No.5, pp. 73-83.

[19] P. Brown et. al. "Class-Based n-gram Models of natural Language", Journal of Computational Linguistics, Vol. 18, No.4, December 1992, pp.467-479.

[20] Buss, E., et. al. "Investigating Reverse Engineering Technologies for the CAS Program Understanding Project", IBM Systems Journal, Vol. 33, No. 3, 1994, pp. 477-500.

[21] G. Canfora., A. Cimitile., U. Carlini., "A Logic-Based Approach to Reverse Engineering Tools Production" Transactions of Software Engineering, Vol.18, No. 12, December 1992, pp. 1053-1063.

[22] Chikofsky, E.J. and Cross, J.H. II, "Reverse Engineering and Design Recovery: A Taxonomy," IEEE Software, Jan. 1990, pp. 13 - 17.

[23] Church, K., Helfman, I., "Dotplot: a program for exploring self-similarity in millions of lines of text and code", J. Computational and Graphical Statistics 2, 2 June 1993, pp. 153-174.

Authors

Miss. Saswade Shweta V received his 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, Data Classification, and Natural Language Processing.