Survey of Trending Techniques for Detection of Emerging Topics in Computer Science within Social Media

Pooja Ajwani* , Harshal A. Arolkar**

*Assistant Professor, GLS University, Ahmedabad, Gujarat, India – 380006.

**Associate Professor, GLS University, Ahmedabad, Gujarat, India – 380006.

E-mail: pooja.ajwani@glsuniversity.ac.in, harshal.arolkar@glsuniversity.ac.in

Abstract: With the advent of Internet there has been a significant and exponential growth in information available to users. The availability of resources like smart mobile phone, low cost data plans and improvement in mobile communication infrastructure has further increased the reach and availability of information. The Internet allowed creation of websites and applications that significantly kept on adding data. The data generated through these websites can be structured (relational database), unstructured (digital images, video, audio files) or semi-structured (word document). The growth of Internet and WWW services gave user a liberty to create his own data, which then can be shared with the world. The development of User-Generated Content (UGC) [2] such as blogs, wikis, forums, tweets, discussions, posts, chats, podcasts, advertisements and other form of media led to the shift of information exchange from media conglomerates to individual user. With this huge amount of data, we address the problem of trending the emerging topics. Identify trending of these emerging topics allows us to know the probable trend of computer science research topics or other relevant research topics in future.

***Keyword**: Trend Analysis, Computer Science, Emerging Topic Detection, Information Retrieval, Information filtering.*

*****

I. INTRODUCTION

With the immense use of web 2.0, the user generated content has emerged to be the standard form of media publishing [6]. UGC reconstructed the way of dissemination and accessibility of information. This leads to the shifting informational power from traditional news sources to individuals [10]. UGC emerged from various applications available on the Internet including social networking, blogging, wikis, digital videos, forums, user-reviews, non-commercial (open-source) software, podcasts, and advertisements. This UGC leads to the evolution of large volume of data.

The UGC has now emerged as a field of research, this study of user content helps in analysing the human behaviour as well as helps in many decision support systems. This may include study of comments, tweets, recommendations systems, blogs and other online contents.

User Generated Content changes rapidly and increasing exponentially day by day which demands a system for the fast and accurate information retrieval and recommendation. As a result of which an important branch of UGC research emerged for the detection of popularity trends in UGC activities **(posting, viewing, downloading)**. Understanding of these systems helps in the implementation of fast information retrieval system, recommendation system, advertising and marketing.

Various other research studies have also been done for UGC in categories like - i) User-generated content (narrower sense) examples: Wikis, Blogs, Reviews, Public short messages, Profiles, Photos, Audio, Video platform (YouTube), ii) User-generated structure examples: Tags, Links, iii) User-generated complex objects examples: Sxipper maps, ActiveTags, TagExtractors, iv) Functionality examples: Yahoo Pipes, Popfly, Greasemonkey, ActiveTags mashups [12].

There are many popular websites which provide a platform to users for the exchange of their knowledge, views, ideas, information and data. Amongst all such websites that allows us to upload video content, YouTube has proven to be the most quickly populated ever since it was launched in the year 2005 [1]. YouTube contents are available in almost 76 different languages covering around 95% of total Internet population [W4]. It has nearly 1.3 billion active users per month [W3] and it is considered to be the second most visited website in the world by 2016. People around the world upload near about 300 hours of videos in every single minute. The repository created by these uploads by the year 2015 was around 500 petabytes.

Another popular micro blogging service that enables the users to share their views through tweets is Twitter. As of, third quarter of 2017, this micro blogging service averaged at 330 million monthly active users. [W5]

Due to the availability of large amount of data; searching for the precise contents was the most difficult task for the naive user [10]. To help such naïve users find required contents UGC sources introduced their own trending applications. YouTube's Trending, Twitter's Trendistic [W2], Google's Google Trends [W1] are few examples of trending applications used for showcasing most emerging topics. The method or approach used by each of the UGC sources differs in metrics. For example some of the metrics used for judging the popularity of the video may be view count, comments, ratings, favourites, tweets.

Understanding trending is necessary for the implementation of actual and fast information retrieval, for building robust recommendation system or for directed marketing and advertisements [5].

In this paper we have studied the algorithms implemented by the social media for the purpose of trending.

II. LITERATURE REVIEW

In this section we have described few of the recent papers that have been published in the domain of trend analysis. A large amount of the work has been done in the area of data analysis on social media but not much has been done on trend analysis.

In 2014, Lun-Chi et. al. [6], proposed a system for the trend analysis of information and communication technology using YouTube videos. They analysed the text caption of Youtube videos over a period of time. They investigated the evolution of topics through a time series-splitting framework. For the identification of the topics in each period, they adopted the Latent Dirichlet Allocation (LDA) model. The similarity between the two documents was determined by frequency-inverse document frequency, which involves counting the frequency of two terms occurring in both documents. The authors used Wikimedia Commons as a knowledge resource to establish a set of technology vocabulary. This vocabulary was then compared and merged with the n-grams in the YouTube captions to acquire a corpus of technological terms. Before applying LDA on sub datasets, a time series-splitting framework was applied, in which documents were first sorted by time and then clustered by topic. Perplexity was then computed to determine the optimal number of topics for each period and then thereby generated topic groups for multiple periods. As a result they plotted trend charts depicting the association among sets of topics of multiple periods and also showed the strength of each topic. The trend analysis for ICT through text caption analysis done by the authors is commendable. The said technique though would fail to give proper trends if the text caption of the YouTube video does not exactly give the idea of the content of the YouTube video.

In 2013, Amar et. al. [1], analysed the polarity trends of public sentiments on Youtube by analysing the comments. They collected more than 4 million YouTube comments to provide a prediction of possible sentiment scores for 26 weeks in to future with a confidence interval of 95% by using Weka forecasting tool. Python scripts using the YouTube API were used to extract comment information about each video. They created the training set consisting of 5000 positive and 5000 negative movie reviews. The comments for each keyword collected served as the test data for classification. For each comment, the polarity/sentiment of each word was calculated by calculating the number of times the word appears in the positive and negative dictionaries. For each word in the comment they calculated the positive as well as negative polarity. Positive polarity is the number of times the word appears in the positive dictionary divided by the total number times it appear in both the positive and negative dictionaries. In the same way they calculated the negative polarity. On the basis of the polarity of each word in the comment, the comment was then classified as positive or negative. Naïve Bayes classification technique was used by the authors for analysis of the trends in sentiments. They then aggregated the sentiments for comments on weekly basis and calculated the mean sentiment for each week using the statistical tool R. For the 26 weeks forecasting, authors selected Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) metrics of the Weka. They selected the default confidence level of 95% . A 95% confidence level means that 95% of the true target values fall within the interval. This study showed the variations in the sentiments of the people time to time by analysing the comments they uploaded on the YouTube videos.

In 2013, Flavio Figueiredo [3], predicted the popularity trends for the user generated videos, especially YouTube. The key parameter for his prediction was based on the analysis of the referrers, incoming links, which capture the information on incoming links to videos. The YouTube provides the cumulative popularity time series as well as Significant Discovery Events, which represents the top 10 referrers of each video that attracts more views to the video. The author implemented the time series clustering algorithm, called K- Spectral Clustering (KSC) for trend discovery. For predicting trends, the author classified the features analysed, which can be derived from the datasets, that are grouped in to three classes namely content features (Video category, Upload date, Video age, Time window size), link features (Referrer first date, Referrer number of views) and popularity features (number of views, number of comments, number of favourites, Average change rate in number of views, Average change rate in number of comments, Average change rate in number of favourites, Peak fraction). The author learned the prediction model from a training set of pre-labelled videos i.e. videos whose popularity trends are known. Then to evaluate the accuracy of that prediction model they used F1 and Macro-F1 metrics using a 5-fold cross validation.

In 2012, Novita Sari et. al.[8], proposed a system for trend prediction computer science research topics using Extreme Learning Machine. The purpose was to generate trend prediction research topics in the field of computer science with the help of ELM and analyse the accuracy of predicted results of ELM. In this study they gathered the data from scientific journals, IEEE Computer Society submitted from 1996 to 2011. Then they grouped the journals by taxonomic computer science based on the IEEE Computer Society. Firstly they perform the training and testing of data sets. Testing process was used to develop the ELM while the testing process was used to evaluate the ability of ELM. Then the ranking of the resultant ELM research topics were done using Growth rate of a term’s hit count. The performance analysis of ELM was obtained by using an average precision which measures the ability of the ranking methods. As a result of this study they showed that prediction research topics using ELM has high accuracy with average precision of 0.7782 for uptrend topics prediction and 0.8954 for downstream topics prediction. They also concluded that ELM performance in terms of speed and testing accuracy is better than the performance of linear prediction and backpropagation.

In 2011, Sitaram et. al [11], conducted an intensive study of trending topics on Twitter and provided a theoretical basis for the formation, persistence and decay of trends. For this they collected 16.32 million tweets on 3361 different topics over a course of 40 days in September-October 2010. They picked 20 minutes as the duration of a timestamp after evaluating time lengths, to optimize the discovery of new trends. They derived a stochastic model to explain growth of trending topics and showed that it leads to a lognormal distribution, which is validated by their empirical results. They also found that most topics do not trend for long and for those that are long-trending; their persistence obeys a geometric distribution. They also discovered that the number of followers and tweet-rate of users are not the attributes that cause trends rather retweets by other users proves to be an important in determining trends.

In 2011, Colorado et. al. [2], implemented the user generated content - emerging topic detection system using YouTube as a case study. They examined the emerging topics in 24 hour time period from 2.2 million YouTube video posts between 3/5/2011 and 3/19/2011. They developed the ETD (Emerging Topic Detection) system for examining the top emerging terms. This system was based on the Cataldi et. al.'s system for ETD.

In 2011, Shiva Prasad Kasiviswanathan et. al.[9], proposed a system for emerging topic detection using Dictionary Learning. They developed an approach which was based on sparse coding. They validated this approach on several datasets from broadcast news, news, groups and Twitter. Firstly they formulated the task of detecting novel signals in streaming data sets as a sparse signal representation problem. They derive a function for the same, which appeared to be well-suited for sparse high-dimensional datasets. This approach is very scalable and well-suited for streaming datasets as it computes the sparse coding for each data point only once, instead of an iterative batch update scheme. For optimizations of problems that appears in formulation, they used practical alternating direction method (ADM). The clustering algorithm used for clustering the novel documents was Spherical K-Means. This dictionary learning formulation combines the ideas from robustness, sparsity, and non-negative matrix factorization for analysis of streaming text. This algorithm alternates between a “detection stage” and “dictionary learning stage”.

In 2010, James et. al. [4], studied and explained the YouTube recommendation system. They presented their video recommendation system, which delivers the personalized sets of videos to sign in users based on their previous activity on the YouTube site. They generated a set of recommended videos by using a user's personal activity (watched, favorite, liked videos) as seeds and expanded the set of videos by traversing a co-visitation based graph of videos. To compute the personalized recommendation they combine the related videos association rules with a user’s personal activity on the site. They also studied the Click Through Rate (CTR) metrics over the period of 3 weeks from the “browse” pages and compared recommendations to other algorithmically generated video sets and found that co-visitation based recommendation performs at 207% of the baseline, Most viewed page averaged over the entire period while Top Favourite and Top Rated perform at similar levels or below the Most Viewed baseline.

In 2010, Mario Cataldi et. al[11], developed a system for emerging topic detection on Twitter based on temporal and social terms evaluation. They collected the dataset from Twitter and provided a new method to extract the emerging topics by analysing in real time the emerging terms expressed by the community. Firstly they extracted and formalized the tweets as a vector of terms with their relative frequencies. Then they define a directed graph of the active authors and calculated their authority through Page-Rank algorithm. Then for each term they model a life cycle according to an aging theory that leverages the user’s authority in order to study its usage in specified time interval. After that they selected a set of emerging terms by ranking the keywords depending on their life status. Finally they created a navigable topic graph which links the extracted emerging terms with their relative co-occurant term to obtain a set of emerging topics.

Outcome of survey

Title of the Paper	Year	Researcher	Parameter Selected	Algorithm Used for emerging topic detection	Outcomes
Information and Communication Technology Trend Analysis Using YouTube Video Based on Latent Dirichlet Allocation Model	2014	Lun-Chi Chen, Hao-Hsun Tesng, I-En Liao	Text caption of YouTube videos	Latent Dirichlet Allocation (LDA) model	Plotted trend charts depicting the association among sets of topics of multiple periods
Polarity Trend Analysis of Public Sentiment on YouTube	2013	Amar Krishna , Joseph Zambreno and Sandeep Krishnan	Comments of the YouTube videos	Naïve Bayes classification technique	Prediction of possible sentiments scores for 26 weeks in to future with a confidence interval of 95%
On the Prediction of Popularity of Trends and Hits for User Generated Videos	2013	Flavio Figueiredo	Referrers (incoming links)	Time series clustering algorithm, called K-Spectral Clustering (KSC)	Macro F1 metrics results are when 1% of video's lifespan and this result increase with the monitoring period.
Trend Prediction for Computer Science Research Topics Using Extreme Learning Machine	2012	Novita Sari, Suharjito Suharjito, Agus Widodo	Scientific Journals, IEEE	MATLAB and Extreme Learning Machine	Accuracy of ELM, 0.7782 for up trends topic prediction and 0.8954 for downstream topics prediction.
Trends in Social Media : Persistence and Decay	2011	Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, Chunyan Wang	Tweets	Develop stochastic model	Retweets by other users proves to be an important in determining trends.
What’s Trending? Mining Topical Trends in UGC Systems with YouTube as a Case Study	2011	Colorado Reed, Todd Elvers , Padmini Srinivasan	Topic on YouTube (text caption)	ETD (Emerging Topic Detection) system	Top emerging topics of particular time period.
Emerging Topic Detection using Dictionary Learning	2011	Shiv Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, Vikas Sindhwani	Broadcast news, news, Twitter	Dictionary Learning	Emerging topics detected in particular interval of time.
The YouTube Video Recommendation System	2010	James Davidson , Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet	User's personal activity (watched, favourite, liked videos)	Recommendations are generated through a series of Map Reduces computations	Co-visitation based recommen-dation performs at 207% of the baseline
Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation	2010	Mario Cataldi, Luigi Di Caro, Claudio Schifanella	Twitter dataset	Navigable Topic Graph	Top emerging topics of particular time period.

III. CONCLUSION

After the survey we can say that most of the algorithms are using inverse document frequency and term frequency as a major logic for identifying trends and emerging topics. Further they have developed various machine learning systems.

One of the major observation of the survey is that most of the work on YouTube has been done using parameters like text caption, comments, subscriptions, likes and dislikes and researcher seems not to have concentrated on the actual contents of the video for trend analysis. Thus there is a further scope of research using content as major character for analysis in this domain.

REFERENCES

[1]. Amar Krishna , Joseph Zambreno and Sandeep Krishnan: "Polarity Trend Analysis of Public Sentiment on YouTube" , The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

[2]. Colorado Reed, Todd Elvers , Padmini Srinivasan : "What's Trending? Mining Topical Trends in UGC Systems with YouTube as a Case Study". Proceedings of the 11th International Workshop on Multimedia Data Mining (MDMKDD'11).

[3]. Flavio Figueiredo, "On the Prediction of Popularity of Trends and Hits for User Generated Videos". Web Search and Data Mining'13 ACM, February 4–8, 2013, Rome, Italy.

[4]. James Davidson , Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, "The YouTube Video Recommendation System", September 26–30, 2010, Barcelona, Spain, ACM 978-1-60558-906-0/10/09

[5]. Jose L. Hurtado, Ankur Agarwal, Xingquan Zhu, "Topic discovery and future trend forecasting for textx", Journal of Big Data

[6]. Lun-Chi Chen, Hao-Hsun Tesng, I-En Liao: "Information and Communication Technology Trend Analysis Using YouTube Video Based on Latent Dirichlet Allocation Model". ISBN: 978-1-61804-313-9

[7]. Mario Cataldi, Luigi Di Caro, Claudio Schifanella, "Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation", Proceedings of the 10 th International Workshop on Multimedia Data Mining (MDMKDD'10), July 25, ACM

[8]. Novita Sari, Suharjito Suharjito, Agus Widodo, "Trend Prediction for Computer Science Research Topics Using Extreme Learning Machine". Proceedings of International Conference on Advance Science and Contemporary Engineering 2012 (ICASCE 2012), Elsevier.

[9]. Shiv Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, Vikas Sindhwani, "Emerging Topic Detection using Dictionary Learning", Conference on Information and Knowledge Management", October 24-28, 2011, Glasgow, Scotland, UK, ACM 978-1-4503-0717-8/11/10

[10]. Shumeet Baluja, Rohan Seth, D. Sivakumar, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, Mohamed Aly, Yushi Jing, "Video Suggestion and Discovery for YouTube: Taking Random Walks Through the View Graph ", International World Wide Web Conference Committee (IW3C2).

[11]. Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, Chunyan Wang, "Trends in Social Media : Persistence and Decay",

[12]. Stephan Hagemann, Gottfried Vossen "Categorizing User-Generated Content (extended abstract)".

Web References

W1. Google Trend, https://www.google.co.in/trends/
W2. Twitter Trendisctic, http://trendistic.com/
W3. YouTube Company Statics, http://www.statisticbrain.com/youtube-statistics/, 2016
W4.YouTube Statics, https://www.youtube.com/yt/press/statistics.html, 2016
W5. Twitter monthly active users https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-user/, 2017