Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210072

Topic Detection Using Multiple Semantic Spider Hunting Algorithm

E.Elakiya[a,1], R.Kanagaraj[b] and N. Rajkumar [c]
[a,1]*Assistant Professor, Dept of CSE, E.G.S. Pillay Engineering College, Nagapattinam*
[b]*Assistant Professor (SL.Gr), Dept CSE, Sri Ramakrishna Engineering College, Coimbatore*
[c]*Professor & Head, Dept of IT, Hindusthan College of Engineering and Technology, Coimbatore*

Abstract. In every moment, there is a huge capacity of data and information communicated through social network. Analyzing huge amounts of text data is very tedious, time consuming, expensive and manual sorting leads to mistakes and inconsistency. Document dispensation phase is still not accomplished of extracting data as a human reader. Furthermore the significance of content in the text may also differ from one reader to another. The proposed Multiple Spider Hunting Algorithm has been used to diminish the time complexity in compare with single spider move with multiple spiders. The construction of spider is dynamic depends on the volume of a corpus. In some case tokens may related to more than one topic and there is a need to detect Topic on semantic way. Multiple Semantic Spider Hunting Algorithm is proposed based on the semantics among terms and association can be drawn between words using semantic lexicons. Topic or lists of opinions are generated from the knowledge graph. News articles are gathered from five dissimilar topics such as sports, business, education, tourism and media. Usefulness of the proposed algorithms have been calculated based on the factors precision, recall, f-measure, accuracy, true positive, false positive and topic detection percentage. Multiple Semantic Spider Hunting Algorithm produced good result. Topic detection percentage of Spider Hunting Algorithm has been compared to other algorithms Naïve bayes, Neural Network, Decision tree and Particle Swarm Optimization. Spider Hunting Algorithm produced more than 90% precise detection of topic and subtopic.

Keywords. Topic Detection, Sub topic Detection, Multiple Spider Model, Multiple semantic Spider Hunting

1. Introduction

This electronic era deals with large volumes of unstructured text every day in the form of E-mail, social media post, customer feedback, reviews and other information. there is a huge capacity of data and information communicated through social network and its really challenging to recognize what is important from huge text data. Analyzing huge amounts of text data is very tedious, time consuming, expensive and manual sorting also difficult. Document dispensation phase is still not accomplished of extracting data as a human reader. Topic discovery strategies is used to mine important patterns from text data sets. These topics once recognized can be like trend analysis

[1]Elakiya.E, Assistant Professor, Dept of IT , E.G.S. Pillay Engineering College, Nagapattinam ;
E-mail: er.elakiyaelamaran@gmail.com

document summarization, recommender systems, information navigation etc. With the emerging dimensions of text corpus fashioned on web and focused digital documents, topic detection has twisted into a serious maneuver for browsing, summarizing and clustering the documents. Topic detection is accomplished using both supervised and unsupervised machine learning algorithms like Cluster, Naïve Bayes, Neural Network, and Support Vector Machine etc. Both algorithms have convinced aids and inconvenience also. In the previous investigation, Topic detection technique is useful to certain precise parts like Sentiment Analysis, Social Network, Twitter, Chatting, Medical, Clinical, Micro blog and Community detection. Each investigation paper discussed so for give importance purely single specialized area not additional.

2. Related Work

Corpus are nourish as a contribution to the ERT framework. The main stage expansion hunts the text which contains any Acronyms, Short Forms, Polysemes, Mis-spelling, Icons or Abbreviations. This stage enlarges the short text content and advancing to removal phase. The next stage eliminates the prefix and suffix of the footings and the non- keyword footings. The production of this stage is only keywords and origin footings. The last stage changes nonstop word group into the list of arguments called signs and kept in a database (Elakiya 2017).Aggarwal (2012) Topic detection using text gathering changes toward the optimum cluster midpoint according to the culture amount by adjusting its mass vector. Clustering the recovery outcomes and crisp the content of clusters is investigated Chieh-Jen (2012).Microblog Hot Topic Detection using Particle Swarm Optimization applied to continuous problem and domain variable is not finite Huifang Yugang (2016). Community Detection using Ant Colony Optimization have a dependent sequence of random decisions Ruochen Liu (2019).

In the Spider Hunting Algorithm, bag of words are only maintained and it does not focus on the bag of related words. So topic detection may not cent percent accurate in some scenarios, to overcome this issue move for Enhanced Spider Hunting Approach. (Elakiya 2018). Corpus contains additional amount of pages then the amount of paragraphs and sentences also enlarged quickly. Emerging cluster for every sentence and dispensation the clusters to detect topic will revenue more time. To decrease the time complexity rather than using single spider go with multiple spiders. This spider scheme can run numerous procedures in parallel to each other professionally (Elakiya 2018).

3. Topic and Subtopic Detection using Word Net

The flow of topic detection process is given below Collect the input corpus. Preprocess the corpus using ERT Framework Generate the token list Depends on the relationship among terms, association can be drawn between words using word net Graph represents all the terms and their relationships. Detect topic and subtopic based on the relation and their frequencies.

3.1. Word Net

Word Net is a collection of lexical database. It consists of group of English words into a cluster of synonyms called Synset. It is a combination of dictionary and thesaurus and word net is mainly used for text analysis (i.e. text summarization and text categorization).

Scenario 1

Tendulkar born 24 April 1973 is a previous Indian cricketer and a earlier captain, observed as one of the greatest batsmen of all time.
In the Scenario1, all sentences are preprocessed and extract keywords. Form clusters for each sentence based on the extracted keywords. Each cluster is processed and trained with the topic model and produces the Topic and Subtopic.
Topic percentage=(Number of Sentence belong to particular topic)/(Total number of Sentence)*100

Figure 1. Word Collections for Paragraph1 in scenario 1

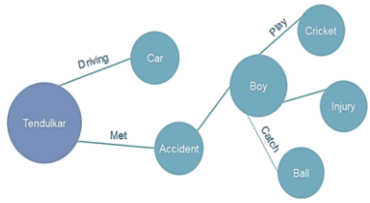

Figure 2. Knowledge Graph for

Tendulkar → Accident →Boy →Injured

Sports→ Cricket→ Tendulkar

4. Multiple Semantic Spider Hunting Algorithm

Multiple spider hunting algorithm detects topics based on the text content in the corpus, it does not find the topic on semantic based approach, some topic detection based on the text content is correct but in some scenarios the text may relevant to two or more topics. In these cases, there is a need for semantic search to detect topics.Each cluster is processed with the topic model and produces the Topic and Subtopic. All the clusters are mapped with the Topic

5. Opinion Generation

In some cases there is no effective keyword to finalize the Topics. If the two topics are equal priority then generate only opinions not Topics.

Scenario 3

A boy would like to buy one apple; he went to shopping and searching for apple. The boy spends more time to found the expected apple and finally brought the desired apple. In the Scenario 3, The same word apple belongs to two topics Fruit and Technology.

5.1 Opinion Analyze

While generating opinions, there is a need to analyze the opinions either it is a false positive or false negative. The opinion analyze is very helpful to reduce the number of generated opinions. Opinion Analyze classify each opinion based on the maximum relevancy with the content. Precision, Recall and F-score measures are computed for surveying the nature of Topic Detection

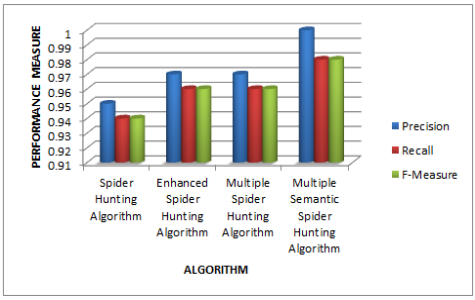

Figure 3. Precision, Recall, F-Measure Vs Various Spider Hunting Algorithms

Table 1. Precision, Recall and F-Measure Values

	Precision	**Recall**	**F-Measure**
SHA	0.95	0.94	0.94
ESHA	0.97	0.96	0.96
MSHA	0.97	0.96	0.96
MSSHA	1.00	0.98	0.98

Table 2. True Positive and False Positive Values

	True Positive	False Positive
SHA	0.94	0.06
ESHA	0.96	0.04
MSHA	0.96	0.04
MSSHA	0.98	0.02

5.2 True Positive & True Negative

Sensitivity measures the extent of positives that are effectively recognized using true positive and false positive.

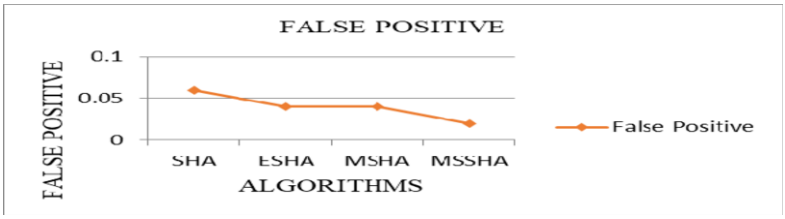

Figure 4. False Positive Vs Various Spider Hunting Algorithms

6. Conclusion

A novel method for a topic detection system was presented. Preprocessing is done by using proposed Expansion Removal and Tokenization (ERT) Framework. Designed Spider model and Multiple Spider model with any number topic and subtopic. Trained the model using BBC News Dataset and found emerging top keywords. Topic and subtopic detection is performed by using proposed SHA, ESHA, MSHA and MSSHA. Based on the experimental results, it has been clear that MSSHA achieves more than 90% precise detection of topic and subtopic. In case of non-effective keywords, Opinion is generated and analyzed.

References

[1] Elakiya E & Rajkumar N 2019, In Text Mining: Detection of Topic and Sub-Topic using Multiple Spider Hunting Model, Journal of Ambient Intelligence and Humanized Computing, Springer.
[2] Metin Turan & Coskun Sönmez 2015. Automatize Document Topic and Subtopic Detection with Support of a Corpus, In Procedia - Social and Behavioral Sciences vol.177, pp. 169 – 177.
[3] Chieh-Jen Wang, Yung-WeiLin, Ming-Feng Tsai & Hsin-Hsi Chen 2012. Mining subtopics from different aspects for diversifying search results Springer Science, Business Media New York, pp. 452-483.
[4] Hyui Geon Yoona, Hyungjun Kima, Chang Ouk Kima & Min Songb 2016.Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling, Journal of Informetrics vol.10, 634–644.
[5] Elakiya E & Rajkumar N 2018.Efficient Detection of Text Data Topic and Tracking System Based on Enhanced Spider Hunting Technique – ESH, Taga Journal of Graphic Technology.
[6] Mujin Kim, Youngjin Park & Janghyeok Yoon 2016, Generating patent development maps for technology monitoring using semantic patent-topic analysis, Computers & Industrial Engineering vol.98, pp. 289–299.
[7] Nenkova, A & McKeown, K 2012. A Survey of Text Summarization Techniques, Mining Text Data.Springer US, pp. 43-76.
[8] Elakiya E & Rajkumar N 2018 .Topic Detection using Spider Hunting Algorithm, Journal of Computational and Theoretical Nanoscience
[9] Martin Riedl & Chris Biemann 2012. Text Segmentation with Topic Models JLCL 2012 – Band vol.27(1), pp. 47-69.
[10] Aggarwal, Charu C & Cheng Xiang Zhai 2012.A survey of text clustering algorithm, In Mining Text Data, pp. 77-128. Springer US.
[11] Wu, Q, Zhang, C, Hong, Q & Chen, L 2014, Topic evolution based on LDA and HMM and its application in stem cell research, Journal of Information Science, vol. 40(5), pp. 611–620.
[12] Le, Q & Mikolov, T 2014, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1188–1196.
[13] Aksoy, F, Can & Kocberber, S 2011, 'Novelty detection for topic tracking Journal of the American Society for Information Science and Technology', pp. 777–779.
[14] Bhanuse, S, Shailesh, D & Kamble, Sandeep, M 2016, Text Mining using metadata for generation of side information' Procedia Computer Science, vol. 78, pp. 807-814.
[15] Caroline Langlet & Chloé Clavel 2016, Grounding the detection of the user's likes and dislikes on the topic structure of human-agent interactions Knowledge-Based Systems, vol. 106, pp. 116–124.
[16] Scott Jensena, Xiaozhong Liub, Yingying Yuc & Stasa Milojevicd 2016,.Generation of topic evolution trees from heterogeneous bibliographic networks .Journal of Informetrics vol. 10, pp. 606–621.
[17] Ruochen Liu, Jiangdi Liu, Manman He 2019, A multi-objective ant colony optimization with decomposition for community detection in complex networks, Transactions of the institute and measurement and control vol. 41, pp. 2521–2534.
[18] Huifang ma, Yugang Ji, Xiaohong Li, Runan zhou 2016.A microblog hot topic detection algorithm based on discrete particle swarm optimization Pacific Rim International Conference on Trends in Artificial Intelligence pp. 271–282.
[19] Elakiya E & Rajkumar N 2017, Preprocessing using ERT Model, IEEE Xplore.

[20] K. Nanagasabapathy et.al; , Validation system using smartphone luminescence, IEEE International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Pages: 235 – 239, 6-7 July 2017, Kannur, India

[21] V. D. Ambeth Kumar et.al;, Cloud enabled media streaming using Amazon Web Services, IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Pages: 195 – 198, 2-4 Aug. 2017, Vel Tech University, Chennai, India (DOI: 10.1109/icstm.2017.8089150)

[22] B. Aravindh et.al; , A novel graphical authentication system for secure banking systems, IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Pages: 177 – 183, 2-4 Aug. 2017,

[23] V.D.Ambeth Kumar et.al, (2016), An Efficient Security System for Data base Management from Illegal Access, IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), SSN Engineering College, Chennai, India, 23-25 March, 2016

[24] V.D.Ambeth Kumar (2017), Efficient Routing for Low Rate Wireless Network a Novel Approach, International Journal of Image Mining, Vol. 2, Nos. 3/4, 2017, 2017

[25] K. Sabarinathan et.al ., " Machine Maintenance Using Augmented Reality", 3rd International Conference on Communication and Electronics Systems (ICCES), 2018. (DOI: 10.1109/CESYS.2018.8723900)

[26] R. Subha Shini et.al., " Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation",IEEE International Conference on Inventive Computation Technologies (ICICT), 2021, DOI: 10.1109/ICICT50816.2021.9358764