International Journal of Web-Based Learning and Teaching Technologies
Volume 16 • Issue 4 • July-August 2021

A Cognitive Personal Assistant System to Enhance the Individual-Centric Research Capabilities

R. Gowtham, Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

Sanjay S. P., DDepartment of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

Shishir Kumar Shandilya, VIT Bhopal University, India

https://orcid.org/0000-0002-3308-4445

S. Sountharrajan, VIT Bhopal University, India

ABSTRACT

The intelligent personal assistant system is designed to support the individual researchers to enhance their quality of the research through the natural language interface. Specifically, this system automatically provides intrinsic details about the importance of the topic of discussion using the timeline analysis. The results generated by the system help the researchers to understand the preference of the global researchers in the specific research field. This system primarily identifies the core topic of the discussion from the user's presentation. Further, the importance of the topic is calculated based on the research articles published over three decades in the related field. The experimental results confirm that the proposed method accurately identifies whether the research topic the user presented is HOT.

KEYWORDS

HOT Topic Detection, Intelligent Personal Assistant System

1. INTRODUCTION

The Intelligent Personal Assistant systems are software agent designed for the users. Most of these systems offer a natural language user interface to answer the user queries, to make recommendations and to perform a specific operation for the user. Initially, these assistant systems were used by the individuals in their smartphones and tablets to perform fewer routine tasks. Currently, these Intelligent Assistants are being built as part of systems like autonomous cars, computers, and gadgets to offer diverse services. The well-known Intelligent Assistants are Apple Siri, Microsoft Cortana, and Google Echo. Each of these assistants performs the same action but differ in the way they respond to queries or help. The primary objective of these systems is to speed up our day by making or work ease.

Like any other areas, there is a substantial increase in the need for an intelligent personal assistant system for the researchers. It is mainly because of the exponential growth of research publications and availability of the Web resources in recent years. The availability of huge resources gives an opportunity to the researchers to explore various investigations carried out in their field of interest.

DOI: 10.4018/IJWLTT.20210701.oa1

On the other hand, this makes them harder to understand, whether the field of their interest is really a choice of the global research community. It is nearly an impractical task for the researchers to manually check all the scholarly literature published in their relevant field of interest across various sources over a time period. In such cases, the proposed system plays an essential role in finding the right information automatically from the open data sources by extraction necessary information from the vocal presentation of the researchers.

The proposed system consists of three primary modules. The first module converts the vocal presentation of the researcher into a corresponding text representation. The text representation is further preprocessed to remove junk characters, unnecessary white characters and new lines. The main topic of the discussion is identified through the topics detailed in the text document and its context. Finally, the Hot topic detection module detects the importance of the main topic of discussion through the number research articles published in the related field over a time period. The number of articles published in the current year is compared against the threshold value. The topic will be considered as a Hot when the number of publications is above the threshold value. The threshold value considered in this module will be dynamically calculated for every identified topic.

The rest of sections are organized as follows: Section 2 presents an overview of related works. Section 3 exemplifies the overall system architecture. In Section 4, we explained the detailed design and methodologies used in identifying the topic of the discussion is HOT. The experimental results are discussed in Section 5. The conclusion is presented in Section 6.

2. RELATED WORK

This section highlights some of the researches that focused on automatic extraction of the core topics from the text and vocal discussion of the users.

Zheng et al. (Zheng & Li, 2009) developed a method to find HOT topics in their Bulletin Board Systems. This method extract the candidate topics from the various posts. Each of the topic was assessed based on Massive Posts, High Quality Posts, High Cohesion, and Bursting characteristics. These feature values of the topics are given as input to the algorithm to find its energy value and ranked based on it. The topics with highest energy values are considered to as HOT topics.

Thanh et al. (Ho et al., 2014) developed a generic model for the hot topic detection on the social networks. This model identifies the popular and interesting topics in the social networks to recommend the users. This method firstly extracts the data from the user posts across different forum and cleans it by preprocesses. The topics are identified from the preprocessed data and ontology will be manually build based on t he topics to identify the implicit topics. Finally, aging theory will be used to calculate energy levels of each of the topics. The topics with the highest energy levels are termed HOT.

Zhiwen et al. (Yu & Nakamura, 2010) presented an article in survey of research and technological developments in the meeting assistant systems. In this survey authors have detailed methods deployed to extract various information from the meeting's visual and audio data based on the structural features present in the recordings. Also, detailed various approaches used in the automatic speech summarization from the audio recordings, specifically for the user queries.

Adrian et al. (Boteanu et al., 2016) developed an expert system to improve literacy skills of the small children. This system records the discussion between the parent and child while reading a story. Based on the discussion it generates the suggestions with a common sense knowledge base. The recordings are initially transcript into a text for the further processing. All the stop words and common English words are removed from the text to facilitate to I dentify the model topics from it. The question phrases are generated by combining the identified topics with the available edge-information.

Freitas et al. (Freitas et al., 2015) conducted a survey on technological developments in the Smart Meeting Rooms (SMR). The authors have assessed the features and evaluation methods developed over a decade to assist the decision making process in the group discussions. Also, found that the SMRs facilitate effective interactions among the participants through the earlier meeting annotations.

The method of storing live meeting information has been an important factor that is been taken into consideration for the implementation of the proposed work.

Yang et al. (Dubey & Shandilya, 2010a) developed a KeyGraph method to identify hot topics from the news media streamed text data. This method primarily extracts keywords from the text documents and constructs a graph based on keywords co-occurrence in the document. The topics in the graph are identified by partitioning the graph using the community detection method. The influential topic will be identified from the graph communities based on the degree of the node.

Bok et al. (Chaure & Shandilya, 2010) proposed a method to predict near-future hot topics from social media. This method identifies setoff candidate keywords from the social media messages posted at different time intervals using the modified TF-IDF. For each keyword, the prediction score will be calculated based on the user's influence and expertise in the social media those who used the keyword in their posts. The keywords with the highest scores are considered as near-future hot topics.

Zhang et al. (Shandilya & Jain, 2009) proposed a four-stage framework to detect hot topics from the images and shot texts posted on Twitter. In the first stage, the text retrieved from the tweets is preprocessed to remove user specified tags and other special characters. The images retrieved from the tweets are supplied to a deep learning algorithm understand its core semantics. In the third stage, the text and content extracted from the image are combined using enhanced Latent Dirichlet Allocation method to enhance the overall semantics of the retried tweets. Finally, the fuzzy matching of the topic words methods is applied to identify the corresponding hot topics from the tweets.

The method proposed in this paper is compared against other related methods under three different dimensions. Firstly, the methods are assessed based on the capability of taking a decision on the data retrieved from online. This is feature is considered to be important mainly because the method designed to operate on the offline data may use rigged models which in turn limits the method to adapt themselves against variability occurs in the real-time data. Secondly, the methods are assessed based on the semantic feature considered in predicting the topic. Most of the topic identification methods use corpus-based prediction models rather than considering the semantics of the document's content. Adapting semantic based features are considered to be most important for accurately identifying the topics as well as helps methods to resolve the topic ambiguities. Each of the method discussed in this section uses their own topic modelling and ranking algorithms to choose desired topics. Most of the methods rank topics based on the computed index value rather than considering whether the selected topics belong to the legal domain of the documents. This certainly indicates domain specific ranking of topics will more advantages rather than generic indexing, based on this feature the methods are compared finally.

3. SYSTEM ARCHITECTURE

The overall architecture of the system is shown in Figure-1. The proposed system primarily converts the presenter's speech into its corresponding text. The converted text data will be preprocessed to remove the common flaws that occur during the conversion. The preprocessed text will be segmented into a set of valid statements to facilitate to extract triple from each of the sentences. The system constructs a semantic graph based on the relationship between resources of the triples. This graph is utilized by the system to identify the most prominent resources. These resources are further ranked to select the dominant topics of the discussion. Finally, the system concludes that whether the topic of the discussion is actually HOT by analyzing the research works that published in the related fields over a decade time period.

3.1 Speech To Text Conversion and Preprocessing

In this phase, the system records the live audio data of the research presentation using a microphone. The audio data is then converted into corresponding text using Google's speech to text API (Speech API, n.d.). Further, the converted text will be preprocessed to remove the faults ensued during the

Table 1. Comparison of related methods

Method	Real-time data aggregation and dynamic decision	Context based topic identification	Domain specific ranking
Adrian et al. (Boteanu et al., 2016)	No	Yes	Yes
Zheng et al. (Zheng & Li, 2009)	Yes	No	No
Thanh et al. (Ho et al., 2014)	Yes	No	No
Zhang et al. (Dubey & Shandilya, 2010a)	Yes	Yes	No
Bok et al. (Chaure & Shandilya, 2010)	Yes	No	Yes
Yang et al. (Shandilya & Jain, 2009)	No	Yes	No
Proposed method	Yes	Yes	Yes

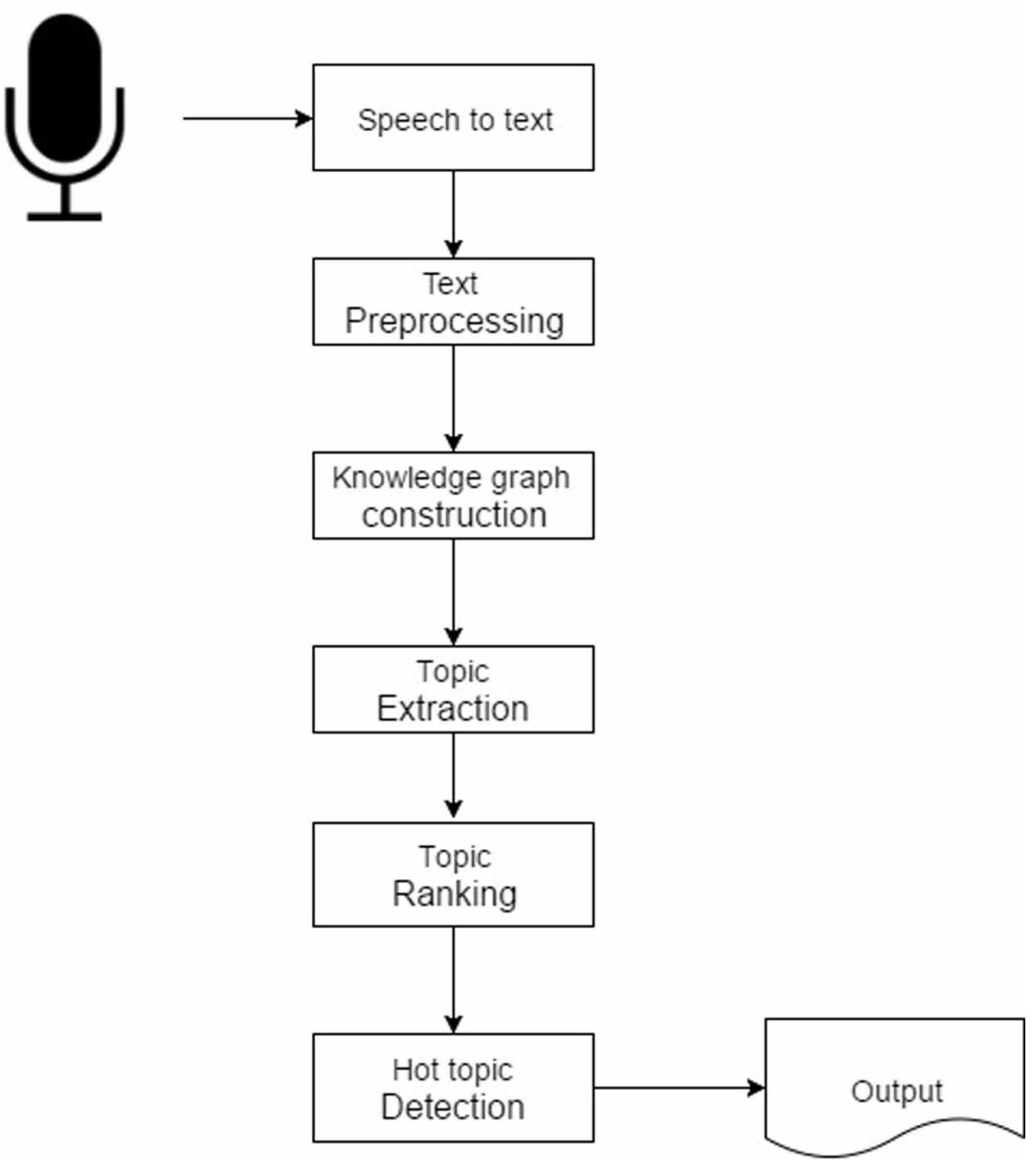
conversion. The sentences in the converted text will be identified and split into a set of single sentences before the preprocessing. In the preprocessing stage, portions of the converted text will be removed that do not contribute in understanding context of the topic being discussed which includes the special characters, stop words, and words with less than four characters as they rarely carry the semantics of the topic. The words with spelling mistakes are replaced with the corrected words that have minimum Damerau Levenshtein distance (Navarro, 2001). The synset words of the incorrect words are retrieved from Wordnet lexical database (Fellbaum, 2012) which in-turn used in the error correction.

3.2 Topic Extraction

This section details the method of extracting the core topic and sub topics of the discussion from the pre-processed sentences. A triple will be constructed from each of the sentences which refer to subject, relation and object. Here, the relation signifies the binary relation between the subject and object which is basically a piece of text linking two entities (Entity-1, relation, Entity-2) (Angeli et al., 2015). The words that represent the subject and object are converted into its root form through the lemmatization. The stemming and lemmatization steps remove the variability in the representation of the word used in the subjects and objects. A dependency graph will be generated from the set of triples generated from the sentences. Each triple will be connected with other triple through the dependency edge to form a dependency graph. The dependency edge is an entity which is common across the set of triples and facilitates to establish a relationship between them. The node with highest in degree links are considered to be a core topic of the discussion. The nodes that directly connected with the core topic of the discussion with reasonable number of in degree links are considered as sub topics of the discussion. The following example shows the sample triples generated from the recordings of the research presentation on the topic Anti-Phishing techniques.

```
1.Phishing     is a     fraudulent scheme
1.user     has     personal information
1.user     has     sensitive financial information
1.fake webpage     capture          sensitive financial
information
1.Phisher      capture      sensitive financial information
1.0     anti-phishing          determine     legitimacy of webpage
```

Figure 1. System Architecture



In the above example, the tab space splits the entities from the relations. Most of these entities represents entities of non-technical domains. According to our method "sensitive financial information" is the entity receives highest in-degrees but the topic of the discussion is Phishing and its counter measures. To resolve this issue, we have consisted only the entities that is part of external domain specific ontologies such as entities from ACM Computing Classification System poly-hierarchy ontology. A word dictionary will be constructed based on the external ontologies. This dictionary helps our system to identify the legal technical entities from the discussion which in-turn helps in finding the main topic of the discussion from the available subject and object set.

3.3 Hot Topic Detection

This section identifies, how important is the main topic of the discussion among the global research community. It is identified by analyzing the scholarly articles published on the topic across various academic journals.

The system primarily constructs a Lustrum Count (LC) map. The LC map stores number of research articles published in the identified topic (detailed in section 3.2) as a set of key and value pairs. In this, the key represents the year of publication and the value represents the number of articles published in the year. When the year of publication was assumed as increment by one it decreases the confidence in their number of publications. In addition, some of the years either have no publications or very less number of publications in the identified topic. To overcome these limitations the key is incremented by lustrums rather than considering every year. The values represent the number of articles published in a specified lustrum.

The elements of LC are retrieved automatically by extracting the necessary information from the articles published in the open access journals for the time period max of three decades. A critical value will be computed for every element in LC using the Equation (1) where the z represents number of articles published in a lustrum.

$$f\left(z\right) = A + B * \tanh\left(\frac{z-\mu}{\sigma}\right) \tag{1}$$

In Equation (1), the publication counts (z) are firstly normalized and further given as an input to the hyperbolic tangent function to map values between the range (-1, 1). In this equation A and B are constants and it helps to find an interception point along the dimension of a number of publications. The A and B values are dynamically determined using the Trust Region algorithm (Voglis & Lagaris, 2004). This algorithm finds the optimal values by considering the bound-constrained nonlinear values of the number of research articles published in the mentioned topic along with its year of publication.

The least value of all the computed critical values is considered as a threshold value as shown in Equation (2). Finally, the topic will be considered as HOT only when the ratio between the threshold value and a number of publications in the current lustrum is less than one as shown in Equation (3).

$$t = \min_{z \in LC} f\left(z\right) \tag{2}$$

$$H = \begin{cases} 1, t < 1 \\ 0, t \geq 1 \end{cases} \tag{3}$$

The lustrums and number of publication on a particular lustrum are plotted in a 2D smoothened line chart. Based on the aforesaid parameters the computed threshold value will also be plotted in the same 2D chart to clearly indicate the importance of a specific topic across years as shown in Appendix-1.

4. IMPLEMENTATION AND EVALUATION

The proposed system is implemented using Python 3.4. The Speech to text conversion module takes the live audio input and converts it into equivalent text using Google Speech API. The topic

extraction model identifies the valid sentences from the converted text using open source software Natural Language Toolkit (NLTK) (Bird, 2006). From each of the sentences the subject, object, and relationship triples are identified using Stanford Open Information Extraction framework (Manning et al., 2014). The inflected words of the subjects and objects are converted into its root form by utilizing the feature of NLTK WordNet wrapper module. The HOT topic detection module constructs the Lustrum Count map by querying the topic of the discussion across various open academic bibliographies and search engines. In this work, we have used the search APIs extended by DOAJ (Directory of Open Access Journals, n.d.), DBLP (DBLP, n.d.), and semanticscholar (Scholar, n.d.) systems to retrieve necessary information regarding the topic of the discussion. The consolidated results about the core topic are presented in the form of 2D smoothened line chart using Matplotlib plotting python library (Matplotlib 2.0, 2012).

4.1 Evaluation

We have manually recorded presentations on selected 96 topics by the final semester, computer science students from our university. We made sure that all the topics recorded were unique and belongs to any one of the fields of computer science. The major and minor topics of each of the presentations were tagged by the members of the respective project groups and its correctness was verified using ACM computing classification system (The ACM Computing Classification System ToC, 2012.). Also, each of the topics was categorized as a hot topic or a non-hot topic by referring various forums and areas of interest of top conferences. Among the 96 topics, 54 of them are identified as hot topics and 41 are identified as non-hot topics manually. We have used four metrics to evaluate the overall accuracy of the system's prediction. The system's prediction is represented as a True Positive (TP) when the topic is classified as Hot and it aligns with the manual prediction. The output is represented as a False Positives (FP) when the system falsely predicts a topic as a hot and contrasts with the manual prediction. The False Negative (FN) represent the false non-hot predictions of the system but contrasts with the manual prediction. The system correctly predicts a topic as non-hot and also it aligns with the manual predictions are represented as a True Negative (FN).

The Precision metric is used to measure the percentage of correctly classified Hot topics over the total number of topics classified as Hot as shown in Equation (4).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

The Recall metrics is used to measure the ratio between the correctly classified Hot topics over the total number correctly classified Hot and non-hot topics.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The accuracy metric measures the deviation of the system's prediction from the total number of referenced manual prediction as shown in Equation (6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

From afore said equations system predicts the importance of the topics with the accuracy of 92.85%, precision of 89.65%, Recall of 96.29%.

A sample experiment result was shown in Figure3. It detail a topic discussed in the area of "cloud computing" was automatically assessed by the system and estimates its dynamic threshold values to find the popularity in the current year. Also, estimates how the global researchers have published in the mentioned topic over the time period of three decades.

4.2 Discussion

The work presented in this paper is limited only for the research meetings and specifically designed for the discussions where only one person presents at a particular point of a time. The system reported few erroneous predictions mainly because of the participation of multiple persons in the discussion. These cross discussions may lead to loss of keywords during the presentation and increase the difficulty of identifying the proper sentences from the converted text. The current system is developed only to identify whether the topic of discussion is Hot or non-hot. In addition, the system can be extended to automatically retrieve the answers for other possible research questions that based on the presentation. The experiment was conducted under the controlled recording environment, the accuracy of the system may drop when the system deployed in the noisy environment. Also, this application requires that the presenter must have good English speaking skill.

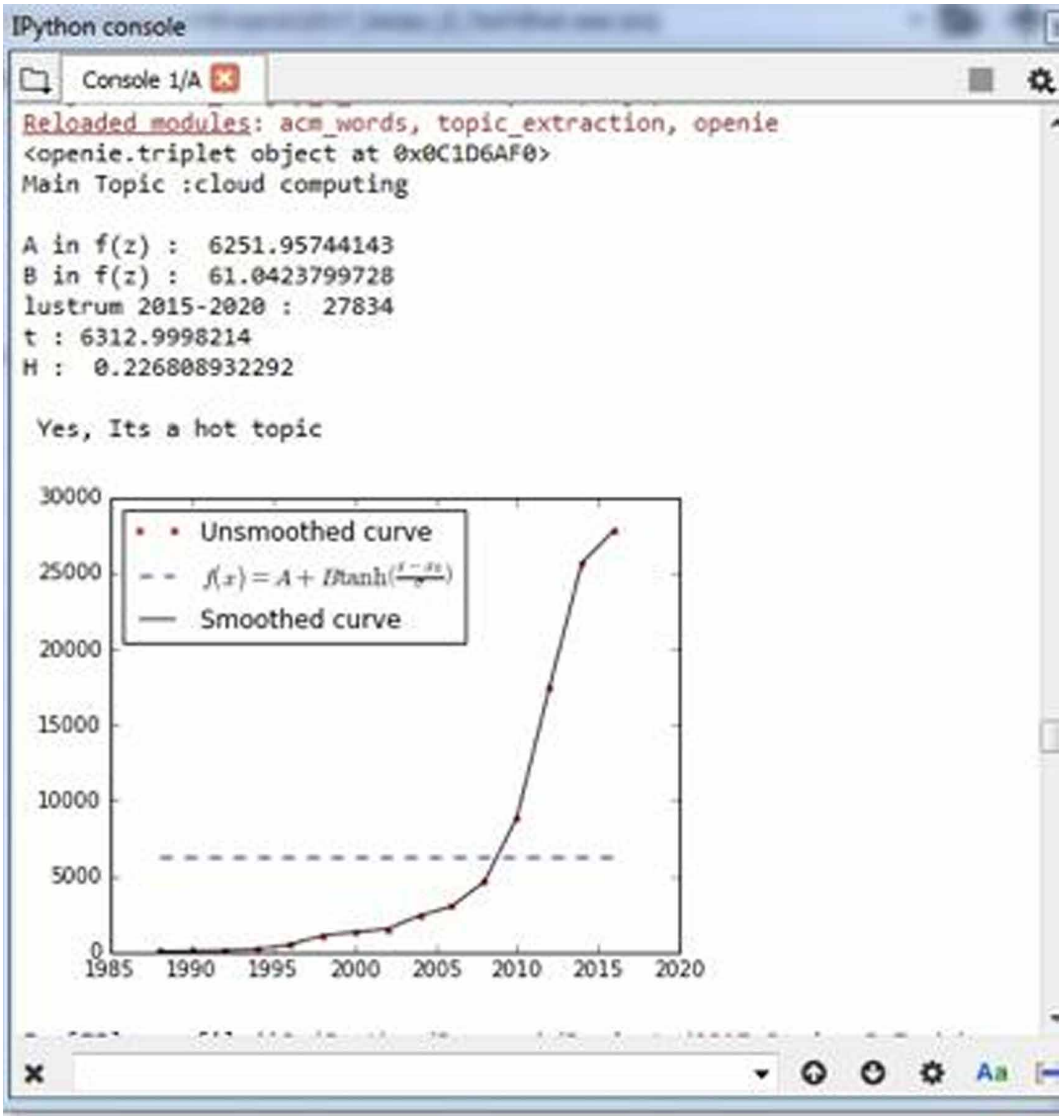
5. CONCLUSION

This Intelligent Meeting Assistant system provides a comprehend information about the research presentations. The systems primarily convert the vocal presentations into corresponding text. The converted text was preprocessed and possible sentences were identified. Each of the identified sentences is further parsed to construct a semantic net representation of the presentation. The concepts in the graph are linked to external concepts defined in the cross domain ontologies. This process in-turn helps to identify the core topic of the discussion based on the in-degree semantic association. All the necessary information about the topic are dynamically retrieved from various data sources in the Internet for maximum of three decade time period. The importance of the topic was finally dynamically identified by comparing the number of publications in the current year against the threshold value. The threshold value was dynamically computed using the hyperbolic tangent

Figure 2. Confusion matrix of test results

	Predicted Hot	Predicted Non-hot
Hot - Topics	52	2
Non-hot Topics	6	35

Figure 3. System output for a topic in "Cloud Computing"



function. The experiment results are shown that the system could correctly predict over 90% of the topics. This system on further development can take the research meetings to a much higher level of perfection through the automated assistance. Also, it helps the presenters to improve their presentation by automatically providing the expert opinions.

REFERENCES

Allan, Carbonell, Doddington, Yamron, & Yang. (1998). *Topic detection and tracking pilot study final report.* Academic Press.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 37-45). ACM.

Angeli, G., Premkumar, M. J., & Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. doi:10.3115/v1/P15-1034

Billsus, D., & Pazzani, M. J. (1999). A personal news agent that talks, learns and explains. In *Proceedings of the third annual conference on Autonomous Agents*, (pp. 268-275). ACM. doi:10.1145/301136.301208

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions.* Association for Computational Linguistics. doi:10.3115/1225403.1225421

Bok, K., Noh, Y., Lim, J., & Yoo, J. (2019). Hot topic prediction considering influence and expertise in social media. *Electronic Commerce Research*, 1–7. doi:10.1007/s10660-018-09327-2

Boteanu, A., Chernova, S., Nunez, D., & Breazeal, C. (2016). Fostering parent–child dialog through automated discussion suggestions. *User Modeling and User-Adapted Interaction*, *26*(5), 393–423. doi:10.1007/s11257-016-9176-8

Chaure, R., & Shandilya, S. K. (2010). Firewall anamolies detection and removal techniques–a survey. *International Journal of Emerging Technologies*, *1*(1), 71–74.

Chen, C., Chen, Y.-T., Sun, Y., & Chen, M. (2003). Life cycle modeling of news events using aging theory. *Machine Learning: ECML*, *2003*, 47–59.

DBLP. (n.d.). *Computer science bibliography*. https://dblp.org/

Directory of Open Access Journals. (n.d.). https://www.doaj.org/doaj?func=home

Dubey, A. K., & Shandilya, S. K. (2010a). A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques. *5th International Conference on Industrial and Information Systems*, 207-212. doi:10.1109/ICIINFS.2010.5578708

Dubey, A. K., & Shandilya, S. K. (2010b). Exploiting need of data mining services in mobile computing environments. *International Conference on Computational Intelligence and Communication Networks, CICN*.

Fellbaum, C. (2012). WordNet. The Encyclopedia of Applied Linguistics. Academic Press.

Freitas, C. F., Barroso, J., & Ramos, C. (2015). A Survey on Smart Meeting Rooms and Open Issues. *International Journal of Smart Home*, *9*(9), 13–20. doi:10.14257/ijsh.2015.9.9.02

Ho, Doan, & Do. (2014). Discovering Hot Topics On Social Network Based On Improving The Aging Theory. *Advances in Computer Science: an International Journal, 3*, 48-53.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL* (pp. 55–60). System Demonstrations. doi:10.3115/v1/P14-5010

Matplotlib 2.0. (n.d.). *A plotting library for the Python*. https://matplotlib.org/

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31-88.

Scholar, S. (n.d.). *An academic search engine for scientific articles*. https://www.semanticscholar.org/

Shandilya, S., Shandilya, S. K., Thakur, T., & Nagar, A. K. (2017). *Handbook of Research on Emerging Technologies for Electrical Power Planning, Analysis, and Optimization*. IGI-USA.

Shandilya, S. K., & Jain, S. (2009). Opinion Extraction & Classification of Reviews from Web Documents. *IEEE International Advance Computing Conference*, 924-927. doi:10.1109/IADCC.2009.4809138

Shandilya, S. K., Shandilya, S., & Nagar, A. (2018). *Advances in Nature-inspired Computing and Applications*. Springer.

Sleator & Temperley. (1995). *Parsing English with a link grammar*. arXiv preprint cmp-lg/9508004.

Speech API. (n.d.). *Speech Recognition, Google Cloud Platform*. Retrieved November 19, 2017, from https://cloud.google.com/speech/

The ACM Computing Classification System ToC. (2012). Association for Computing Machinery. https://www.acm.org/publications/class-2012

Voglis, C., & Lagaris, I. (2004). A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization. *WSEAS Conference*, 17–19.

Yang, S., Sun, Q., Zhou, H., Gong, Z., Zhou, Y., & Huang, J. (2018). A Topic Detection Method Based on KeyGraph and Community Partition. In *Proceedings of the International Conference on Computing and Artificial Intelligence* (pp. 30-34). ACM. doi:10.1145/3194452.3194474

Yu, Z., & Nakamura, Y. (2010). Smart meeting systems: A survey of state-of-the-art and open issues. *ACM Computing Surveys*, *42*(2), 8. doi:10.1145/1667062.1667065

Zhang, C., Lu, S., Zhang, C., Xiao, X., Wang, Q., & Chen, G. (2019). A Novel Hot Topic Detection Framework With Integration of Image and Short Text Information From Twitter. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 9225–9231. doi:10.1109/ACCESS.2018.2886366

Zheng, D., & Li, F. (2009). Hot topic detection on BBS using aging theory. *WISM*, *5854*, 129–138. doi:10.1007/978-3-642-05250-7_14

R. Gowtham is an Assistant Professor in the Department of Computer Science & Engineering at Amrita Vishwa Vidyapeetham, Coimbatore, India. He has obtained his B.E degree from Periyar University, M.E degree from Anna University and Ph.D under Anna University. His research interests are in the areas of Information security and Semantic Web.

Shishir K. Shandilya, Ph.D. (Computer Engineering) and M.Tech (CSE), is an excellent academician and active researcher with proven record of teaching and research. He is a Senior Member of Institute of Electrical and Electronics Engineers (IEEE), USA and was also elected as an executive member of IEEE Industry-Outreach Committee-India. He has received an International Certificate for Teaching and Training from Cambridge University, United Kingdom. Dr. Shandilya has received "Young Scientist Award" for consecutive two years (2005 & 2006) by Indian Science Congress & MP Council of Science & Technology for Computer Engineering.