International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.9-20
http://dx.doi.org/10.14257/ijdta.2016.9.5.02

Navigation through Citation Network Based on Content Similarity Using Cosine Similarity Algorithm

Abdul Ahad[1], Muhammad Fayaz[2] and Abdul Salam Shah[3*]

[1]University of Malakand KPK, Pakistan
[2]JEJU National University, South Korea
[3]SZABIST, Islamabad, Pakistan*
[1]ahadbcs86@gmail.com, [2]hamaz_khan@yahoo.com, [3]shahsalamss@gmail.com*

Abstract

The rate of scientific literature has been increased in the past few decades; new topics and information is added in the form of articles, papers, text documents, web logs, and patents. The growth of information at rapid rate caused a tremendous amount of additions in the current and past knowledge, during this process, new topics emerged, some topics split into many other sub-topics, on the other hand, many topics merge to formed single topic. The selection and search of a topic manually in such a huge amount of information have been found as an expensive and workforce-intensive task. For the emerging need of an automatic process to locate, organize, connect, and make associations among these sources the researchers have proposed different techniques that automatically extract components of the information presented in various formats and organize or structure them. The targeted data which is going to be processed for component extraction might be in the form of text, video or audio. The addition of different algorithms has structured information and grouped similar information into clusters and on the basis of their importance, weighted them. The organized, structured and weighted data is then compared with other structures to find similarity with the use of various algorithms. The semantic patterns can be found by employing visualization techniques that show similarity or relation between topics over time or related to a specific event. In this paper, we have proposed a model based on Cosine Similarity Algorithm for citation network which will answer the questions like, how to connect documents with the help of citation and content similarity and how to visualize and navigate through the document.

***Keywords**: Citation Network, Content Similarity, Component Extraction, Cosine Similarity Algorithm, Data Organization, K-Means Partial Clustering, Navigation, Topic Evolution, TF-IDF technique*

1. Introduction

The scientific literature is expanding with great pace, it is estimated that the literature is doubling in every five years and there are many numbers of new topics which are evolving day by day in science and become very important for future research and growth of science. However, from this huge available published literature, it becomes very difficult for the scientific community to know which topics have been evolved from which other topics, what are relationship types and strengths between topics and some similar questions [1-7]. The current search engines such as Google, Bing, Yahoo, citation indexes: Google Scholar, Cite Seer, ISI Web of Knowledge, and digital libraries: IEEE, ACM, and Springer, do not keep any track of that kind of relationship between the topic

[1] *Corresponding Author

ISSN: 2005-4270 IJDTA

and their origin of emergence and the users are unable to search evolving topics and topic-evolution using the traditional search systems [8].

However, this is very important for scientific community, and are supporting number of tasks such as: 1) when a new researcher wants to start a research on a specific topic, 2) when a scientist is making a literature review of an area, 3) when someone wants to measure context of knowledge diffusion, 4) when someone wants to categorize citations based on topic evolution, and 5) when someone wants to understand the overall area and topic evolution in it, to identify number of trends in scientific community. All of these queries are only possible when one could comprehend the topic evolution in scientific literature. This study will identify the state of the art systems that work on topic evolution and highlight the strengths and weaknesses of each to understand the overall contributions and limitations of existing systems [9].

The topics emerge and evolve over time, which results in huge amount of data that is archived in the form of web documents or at digital repositories, for the understanding of the evolution of these archives they must be properly mined [10-11]. The process of evolution detection is based on various strategies, used on the basis of actors involved in the process 1) type of user 2) type of data 3) nature of evolution. The first actor refers to nature of user *i.e.* researcher or technology enthusiast, the second actor refers to type of data *i.e.* scientific data, blog posts, or data related to technology news and the third type of actor refers to nature of evolution that might result of events or based on time interval. Some techniques based on text mining and similarity finding using citations, authorship network are discussed in following paragraphs [12].

In the text mining, high-quality information is extracted from the text by studying trends and patterns, with the help of different methods like statistical pattern learning. It involves parsing text to transform it into a structured output like spanning trees or vector space model by algorithms like Term Frequency-Inverse Document Frequency (TF-IDF) [13], Inverted Document Frequency (IDF) [13], Suffix Tree Clustering (STC) [14], Vector Space Model (VSMs) [15], these structures organize output such that high-quality information extraction became simple and effective. This information is used by techniques like clustering, documents summarizing, entity extraction and text categorization. The algorithms like K-means partial clustering, PLSA are used to partially cluster objects which are assigned to the center of nearest cluster based on its attributes. These algorithms structure the words based on their respective semantic characteristics [16]. The output from the above techniques is further processed by evaluating the similarity between the objects to show the evolution trends and patterns. The use of different algorithms like Semantics Dependency Distance (SDD) uses Cosine Distance between nodes [17]. The links between these nodes represent similarity relationship, TF-IDF algorithm assigns weight to the objects sharing same weights are considered similar. The citation also helps to find related topics by evaluating cited paper by a publication under consideration [18].

The other techniques consider the publications of authors that are evaluated with the help of different ontological frameworks to help find patterns for evaluating topic evolution in documents; the authors share the common field in their publications. The recent work in this regard resulted in some interesting concepts like authors and co-authors network. It is more likely that a group of authors will publish and reference their past work related to the same area [9].

The related concept to topic evolution is hot topics or burst topic, in this the age of abruptly changing trends, the user wants to be up to date with information as they arrive. Identifying changes in real time data is a non-trivial task for researchers and lots of work is done in this regard. The classification of information based on events and location is as important problem as Topic Evolution Detection itself, location and events refer to venues of the publications [19-20].

One important part of Topic Evolution is a representation of results to the user. Now that we have found the relationship between topics with the help of many algorithms, there is a need to show this information to the user so that they can find useful topic evolution information in an interactive way. Besides some traditional visualization techniques like presenting output in HTML *i.e.* list, tables *etc.* there are other visualization techniques which make use of graphs and charts [15]. Information can be visualized on a timeline that represent temporal characteristics of data. Event-based data can be represented by a timeline graph where changes in the topic will be marked by a landmark [19].

The remaining structure of the paper is organized as the Section 2 presents the literature review, Section 3 contains problem statement, the Section 4 contains proposed model Section 5 and 6 contains experiments and results and finally in Section 7 conclusion and future work is provided.

2. Literature Review

Shubhankar *et al.* [1] highlighted the main goal of ranking and modeling topic evolution, by an efficient algorithm. The topic, evolution has become a challenging task over time for the researchers. They suggested the topic as a summary of its content, and also introduced a unique approach, that assigns the rank to a topic by applying PageRank algorithm, without considering the time of research publication [18]. Furthermore, they have categorized topics based on the set of the topic and closed keyword-set. The closed keyword-set were formed such that; phrases were selected from topics of the publications along with a user-defined minimum support. PageRank algorithm has been applied in iterative passion, to assign authorities to score to the research paper. The authority scores, based on popularity in research community; however, the algorithm also identified hot topics by evaluating them on a timeline and same shown as landmarks. They also tried to find fading topics; they first checked for topic detection, then evaluated landmark topics, and at the end tried to find fading topics. The algorithm proved as most effective and faster when tested on DBLP dataset.

Song *et al.* [9] emphasized the use of text clustering for Topic Detection. The text clustering dominates other algorithms in terms of time, computational complexity and cost. There are many ways to transmit data over a network; still we need methods to avoid noise and irrelevant information. They considered the unstructured and scattered data over the internet as text copra and introduced a two-step algorithm for clustering the text copra. The first step uses C-Process to create overlapping clusters with the help of Canopy Clustering. The Canopy Clustering is usually applied before K-mean algorithm to speed up clustering of large data-set. In the second step named K-means apply rough clustering on result based on the common points between the clusters. The K-Means uses X-Means algorithm. The experiments have proved better performance than, k-means clustering and single pass algorithms, and proved to be more suitable for detection of online topics.

Wu *et al.* [19] discussed CAR (Credible association rule) a new method to relate and track documents. The CAR does not use prior knowledge of categories structure as compared to other automated procedures. The other traditional processes detect related documents based on the topic of documents, with the help of some predefined rules or categorization. This method makes use of term frequency-inverse document frequency (TF-IDF) as feature pre-selection set. TF-IDF is a numerical statistic which shows how words are important to a document. After the feature subset selection, CAR and minimal clique algorithms were applied. These two algorithms use adjacency matrix to produce credible association rules. The refinements, removes noise and common words with the high frequency that are not related to the topic of the document. A high level of reliability, availability and performance are achieved by applying refinements like Inverted Document Frequency (IDF) and quasi-maximal cliques.

Jo *et al.* in [20] proposed algorithms for the topic detection from linked textual corpus using the relationship of topics in terms of the distribution and the link distribution. They algorithm generate a ranked list of the topics the method has shown effective results with arXiv and Citeseer. Jo *et al.* in [21] discovered rich patterns of topic evolution within built-in features of time-stamped documents. Instead of analyzing documents on the specific interval, the method focuses and treats topics separately as they appear over time in chronological order. The information is obtained such that; it qualifies topic as either new or it has some similarity with existing topic. The result was visualized by chronological scanning on a graph known as topic evolution graph. The topological or time restrictions were not considered while building the graph. The nodes of the graph represent topics and the connection between the topic nodes represents the cross-citation relationship. This representation of information projects a huge amount of knowledge about topic evolution. Details about a single topic can be obtained by selecting a seed topic and studying its connections with other nodes. These connections change as time passes, the emergence of new topics adds new nodes to the graph and also change connections between them. The testing was carried with ACM repository.

Jayashri *et al.* [22] discussed retrieval of temporal and event-based knowledge from a huge collection of historical documents. The method uses temporal text mining (TTM) and Topic detection and tracking (TDT) techniques. The TTM extracts important patterns related to time, like collecting term frequency in a sequence of time; it also helps in keeping track of modification in the words with respect to time. In TDT, a clustering problem called Adaptive Resonance Theory (ART) is used, that tracks unknown topics in the system and assign them to previously identified topics. The Evaluation of such information is usually carried at the time of its arrival, which helps to consider temporal properties of the topics. The Evolution is implemented using an incremental algorithm. The experiments helped to discover new trends and identify trends that cannot be obtained from documents if analyzed individually.

Cui *et al.* [23] highlighted topic evolution in text data as an important task. The Text Flow has been introduced; which studies various patterns that appear from various topics. Three-level features selection was conducted namely keywords, critical events, and evolution trends, then these features were visualized via a coherent visualization technique that shows the complex and important relation between them. The Text Flow is an interactive visual analysis tool that help user analyze how and why the correlated topics change over time. First patterns related to merging/splitting are discovered via hierarchical Dirichlet process employed in incremental passion followed by extraction of keyword correlation and critical events. The result can be visualized by three-level directed acyclic graph such that user can deduce various information at any level of consideration. This method helps user visually study relation between two topics that is best as it represents this relationship in an interactive way.

Jin [24] focused on detecting topic in immense information available over the internet. They introduced a new method by combining Suffix Tree Clustering (STC) and Semantic analysis that approaches the problem in two steps. In first step feature selection is done with the help of NLP algorithm, by selecting meaningful words for clustering and the weight is assigned to word using term frequency-inverse document frequenting (TF-IDF). The NLP results in parts of the sentences in the form of the noun, verb and named entity. The result of feature selection is supplied to STC to form clusters where the score is assigned to them. TDT is applied to track topic, focusing on topic drifting. This is an inherent difficulty of topic evolution, which occurs over time as new information emerges. The clustered contents are represented via VSM (Vector Space Model) by selecting only top K words that are added to the vector [15]. The semantic analysis is used to add significance to the topics and the significance can be measured by applying filters to the words and analyzing structure of words under a cluster which share the same meaning. The experiments proved that topics can be tracked effectively.

Zhang *et al.* [25] highlighted that news collection should be structured such that topic detection and clustering become easy. The vector space model (VSM) is one of the easiest and productive methods that can be used for the representation of topics, and the Information gain algorithm is used for feature selection. The features, then ranked by assigning a score to each of them. Those features are selected which have high scored among the feature selection set. The TF-IDF used to represent and score the features. Then K-means employed for partial clustering, each object have assigned to a cluster center where distance based on the cosine distance, which assigns each feature to cluster that is more similar in properties. The K-Means algorithm is not efficient when the dataset is large enough; it is used in conjunction with VSM, which facilitates by representing data in a form which is processed easily by K-Means algorithm. The process was evaluated by use of Topic detection and tracking method and verified that topic detection that is based on K-means outperforms if used for large-scale data.

Yue *et al.* [26] emphasized the importance of technology in finding relevant information in a huge set of fast growing information. A topic detection algorithm based on K-means clustering is proposed by Authors [26]. This method identifies some keywords and assigns weight to it which then helps in detecting topic. Two methods are used to select keywords; the first method selects hot words from the database which is man-made. The second step select words by means of TF-IDF method which makes sure no word with high frequency is left behind, that is related to the document [13]. The method proposed by [26] is a two-step process. In the first step, documents are selected which are related to 14 categories, which are then parsed by well-known Chinese parsing system known as ICTCLAS to obtain components of the document. Text vector is then created from the component by selecting feature words which are represented by vector space model. The weight of featured words is calculated and assigned to them. Document summary is calculated using cosine formula, which is used to find similarity among documents. In the last step documents, clusters are created using k-means algorithm. Various experiments are conducted to prove its efficiency.

Masada *et al.* [27] discussed various ways of connecting and referencing scientific documents that ultimately helps in finding a relationship between scientific documents based on the citation. Instead of text analysis, references are extracted with a model named TERESA. TERESA extends (LDA) Latent Dirichl *et al.* location, which models relation between documents in the form of transitions between them. TERESA focus on the collection of the document as a whole to discover directed relations, thus providing a global view of relationships. Global view of system introduces a problem as the huge amount of computation is required; to overcome this problem (VB) Vibrational Bayesian inference is used. VB accelerates GPUS compatible with NVidia CUDA, thus, hundreds of thread can be executed. As compared to another method which creates multiple local views which are difficult to conceptualize, this method provides a single view of all the relationships. Many iterations of this method will give different projection that shows important information about topic evolution.

Lv *et al.* [28] identified various problems concerned with traditional approaches that consider the huge amount of information while tracking topic as useful information might be skipped. A model is proposed by [28] that make use of similarity based on subtopic and events; it also makes use of time partition to study development history and location characteristics. First considering temporal characteristics of the topic they are partitioned into subtopics clusters based on time slices. The process of partitioning based on temporal characteristics is applied only once. Topics are strongly similar within a cluster. Based on temporal and attribute similarities subtopics are combined into events. Subtopic evolution relation is discovered by keeping in mind the events in which these are combined. Topics might change with time and changing events. This method is applied to a dataset of 500 news articles and proves to be effective.

Shah *et al.* in [29] proposed a model to handle the literature overloading and selecting the most relevant papers published in recent years. They tried to answer the questions like, how many articles are sufficed for a good literature review, and how many past year's literature will be enough to meet the required level for a good literature review?

Chen *et al.* [30] have focused on SDD-PLSA a novel topic detection method. SDD-PLSA is the combination of Semantic Dependency Distance SSD and PLSA. PLSA is proved to be one of the effective and efficient methods in Text Mining. Syntax and semantic information of text are also considered for the effectiveness of method proposed by the Authors [30]. SDD-PLSA method works in two steps; First step uses semantic features of sentences to group them, this is an important step. The second step uses the result of the first step which is a tree structure. Optimal semantic dependency between parent and child node is calculated, these nodes are linked with the help of dependency grammar. The link between a parent node and child node is called dependency link [31]. In Dependency Grammar, a word that represents the meaning of the whole sentence is selected as head word and all other that are dependent on the headwords are decedents. In second step variation of PLSA classifier is applied to the result of the first step. Experiments show that result of SDD-PLSA is more efficient and accurate than PLSA if used alone.

However, these methods mentioned above are immature and not comprehensive, still there is a need of the system that can organize the published material in an informative matter.

3. Problem Statement

Different techniques are used to detect topic evolution trends in literature presented on various sources, however, the result is such that it is difficult to understand and consume. Organization of the complicated result in such meaningful form is ~~as~~ an important task as finding evolution itself.

Usually temporal and topological attributes are ignored or not utilized fully in organizing result-set. Such consideration might result in complex representation, making it harder to understand. As topic evolution occurs over time, temporal attributes are the key to monitoring evolution. Like temporal attributes topological attributes shows how documents are connected to each other thus helps in forming navigation through the result. Among previous research work, some imposes a lack of temporal attributes in documents; as a result, it fails to connect documents in evolution basis, on the other side not imposing any topological attributes makes evolution in topics complicated. In this study documents and topics are connected in increasing time manner to resolve problems in navigation. The additionally strong relationship among documents is introduced, based on resultant similarity value of the document, which make the result more informative.

4. Proposed Model

In this section, a model is proposed for citation network. The primary goal of this model is to provide the solution of following questions.

How to connect documents with the help of citation and content similarity?

Visualize and Navigate through the document in the result?

This approach describes how documents are connected and a navigational structure has created the documents, which are obtained as a result of applying some document similarity algorithms. A measure is introduced to measure how documents may relate strongly, that will help create a navigation among documents.

4.1. Parameters Used

γ Is a positive threshold value, its minimum value represents minimum qualifying value.

α Resulting value of a document after, document similarity value is applied.

(LS). This list will store documents, along with the value of the document that are selected for visualization.

τx Collection of terms collected from documents using TF-IDF technique [13].

4.2. Model Justification

As previous work has shown, citations are one of the most promising measures which can help us find topic evolution. However using citations we get a result which we might get papers that are not strongly related to each other. In this framework content similarity is introduced as another measure, which will be used in conjunction with citations to make the results more useful. Combining these two will help us categories paper in strong and very strong relationship *e.g.* document with the highest value of α (content similarity measure) value is the strongly related document, as it is most similar as well as connected via citation. Cosine Similarity and Term Frequency, Inverse document frequency TF-IDF are used in this model to calculate how similar these two documents are [13]. This method first extracts terms from the document based on their subject and then are compared with Cosine Similarity method.

We select a random start paper, with the help of google scholar we get papers which cite this paper. All those papers are collected manually. Due to many complications and extra processing of attributes, all documents are converted from any other textual format to plain text format using “A-PDF Text Extractor”. “A-PDF Text Extractor” is a powerful open source text extractor tool, which extracts text very efficiently, and ignores any graphical data, as we don’t need graphical data for similarity finding. Second major step in this framework is finding similarity. For document similarity measure we make use of Cosine Similarity, it is a well-known measure widely used in the field of data mining. This measure will give us how documents are similar to each other based on their subject.

An application is developed in C#.NET, which implements Cosine Similarity Algorithm. This application takes two text documents d1 and d2, and process them to calculate similarity. In the very first iteration of this process, d1 is the start paper, while d2 is one of the paper which cites start paper. D1 will retain start paper and compare it with d2 as long as all of the paper in the list are traversed. D1 will then hold the first paper in the list that has the highest similarity value, and compare it with all the papers which cite this paper. This recursive process will continue until all the papers are traversed.

Following are the step which describes how Cosine Similarity Algorithm is approached

- First a random paper is selected as start paper
- All those papers which cite first paper are added to the list of current paper
- Cosine Similarity algorithm is applied to each document in the list. Only those papers are left, who’s α value of exceeds threshold γ value.
- After the list of start, paper is traversed, paper with the highest value of α is selected and list of its cited paper is traversed, this process is continued till all the documents are traversed
- A tree is constructed after all the papers are traversed.

Figure 2: Show above steps, in the form of flow chart, document similarity is a sub process and calculated as follows:

Terms from documents d1 and d2 are extracted

$\tau1$ = TFIDF (d1)

$\tau2$ = TFIDF (d2)

α = Cosine Similarity ($\tau1$, $\tau2$)

Document is only selected if $\alpha \geq \gamma$

If value of α exceeds or is equal to γ then document is selected, and the root document is connected to the child document with an edge, α is assigned as the edge weight.

The process of the proposed model is shown in Figure. 1.

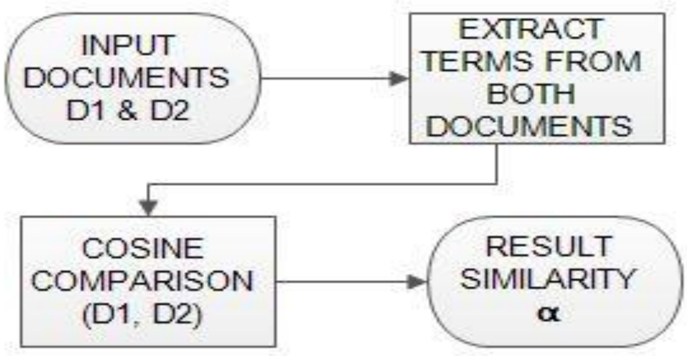

Figure 1. Process of the Proposed Model

5. Experiments

Using proposed model we carried out an experiment on a random set of papers. We selected a random paper [32], it was published in 1994, and was cited by 692 documents which made it a candidate paper to start with. List of the papers was collected using Google Scholar. We applied our model on about 28 papers, up to four level of iterations. During this experiment about 55069, citation edges were traversed, distributes as 2771 citation edges for level 1, 9011 citation edges for level 2 and 43787 edges for level 3. We considered only top 4 citation link from each paper to the papers it cited. Navigation is constructed as depicted in Figure 3.

6. Results and Visualization

The result of the above experiment is displayed in a tree structure, displayed horizontally, on the x-axis we have a time line, here ranging from 1994 to 2010. The root element of the tree is starting paper. Decedents of the root elements are those papers which have qualifying value higher or equal to the threshold value. Documents are added chronologically, as evolution is based on the time interval from earlier to a later time, the edge between documents are numbered with the weight obtained from apply Cosine Similarity Theorem. The edge between document with the highest qualifying value has a distinction from other edges as it is displayed by a bold line with black color, this distinction is also applied to all the edges in child list who has highest qualifying value. This bold line in black is named as trunk which serves as the main navigational link. Document on this trunk are strongly related, this practice can be applied to the document with the second and third highest values as it is represented by a bold line in green. Following are some figures depicting the result of above experiment.

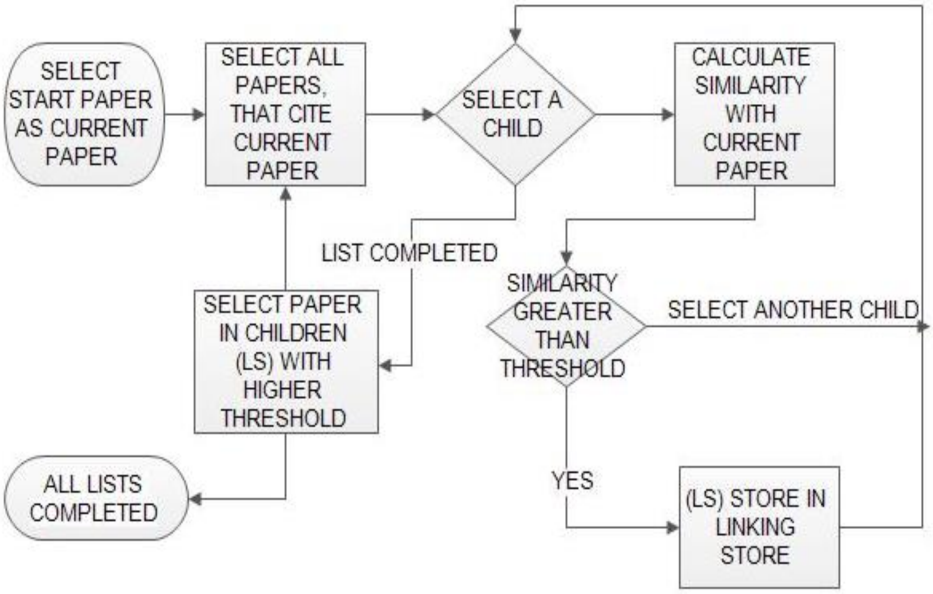
SELECT START PAPER AS CURRENT PAPER
SELECT ALL PAPERS, THAT CITE CURRENT PAPER
SELECT A CHILD
CALCULATE SIMILARITY WITH CURRENT PAPER
LIST COMPLETED
SIMILARITY GREATER THAN THRESHOLD
SELECT ANOTHER CHILD
SELECT PAPER IN CHILDREN (LS) WITH HIGHER THRESHOLD
YES
(LS) STORE IN LINKING STORE
ALL LISTS COMPLETED

Figure 2. Flow Chart

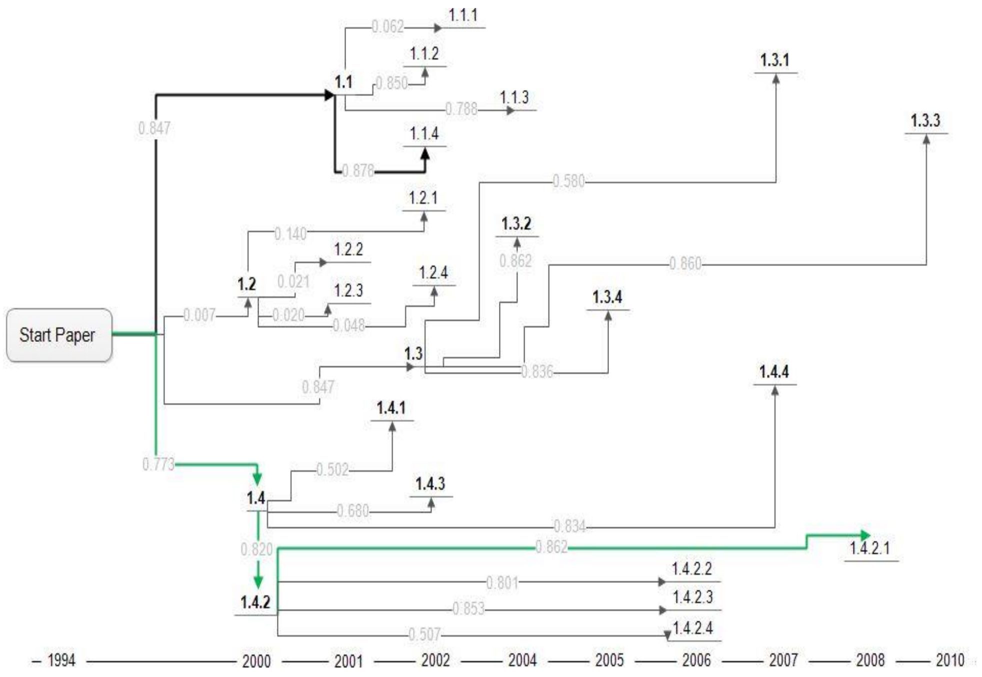

Figure 3. Navigational View

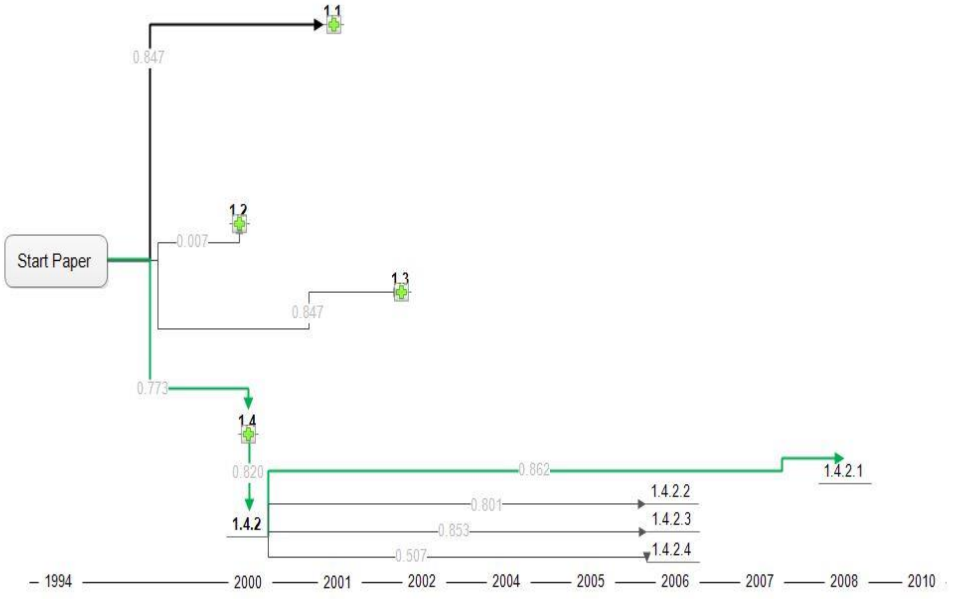

Figure 4. Strong Relation

7. Conclusion and Future Work

In this work, we studied some of the inherent temporal and topological features that can help us present result in more meaningful and descriptive form. In our work, we combined Citation Network with Document Similarity to create a relationship between research papers of the result. First document connection was created using Citation Network, and then that relation was weighted with the help of Cosine Similarity Algorithm, which make use of TF-IDF technique [33]. Link with the highest number of citation and weight is selected as the strong relation, thus navigation is created on that path, as shown in Figure 4. We also evaluated second most strong relation and presented that with another color. Over 55000, Citation links were traversed for a set of 28 documents to form two navigational paths as depicted in Figure 4.

There are a couple of challenges observed in this work, which needs further investigation. First of them is this work only accepts text documents. Work can be done to accept any format of the document, and automatically convert them to a plain text format and then subject to further processing. As in our experiment we process more them 55000 citation links for only 28 research papers, a huge amount of citation edges needs to be traversed, this is also an open area where work can be done. The result shown in this work was just a graphical depiction, in future work it can be translation to some other visualization formats, with which users can interact easily and gain as much information as available.

References

[1] K. Shubhankar, A. P. Singh and V. Pudi, "An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution", Proceedings of the 22nd International Conference on Database and Expert Systems Applications, Toulouse, France, **(2011)**.

[2] J. M. Memon, A. Khan, A. Baig and A. Shah, "A Study of Software Protection Techniques", Innovations Advanced Techniques in Computer and Information Sciences and Engineering, Springer Netherlands, **(2007)**, pp. 249-253.

[3] S. Khawaja, A. Shah and K. Khowaja, "Alternate Paradigm For Navigating The WWW Through Zoomable User Interface", Advances and Innovations in Systems, Computing Sciences and Software Engineering, Springer Netherlands, **(2007)**, pp. 417-420.

[4] A. S. Shah, M. N. A. Khan and A. Shah, "An Appraisal of Off-line Signature Verification Techniques", International Journal of Modern Education and Computer Science (IJMECS), vol. 7, no. 4, **(2015)**, pp. 67-75.

[5] F. Noor, A. Shah and S. A. Khan, "Relation Mining Using Cross Correlation Of Multi Domain Social Networks", In. SAI Intelligent Systems Conference (IntelliSys), 2015, London, **(2015)**, pp. 898-903.

[6] K. Nusratullah, S. A. Khan, A. Shah and W. H. Butt, "Detecting Changes in Context Using Time Series Analysis of Social Network", In. SAI Intelligent Systems Conference (IntelliSys), 2015, London, **(2015)**, pp. 996-1001.

[7] A. Shah, A. Raza, B. Hassan and A. S. Shah, "A Review Of Slicing Techniques In Software Engineering", In: International Conference on Engineering and Technology, Srilanka, **(2015)**, pp. 1-15.

[8] P. H. Adams and C. H. Martell, "Topic Detection and Extraction in Chat", In. Proceedings of the 2008 IEEE International Conference on Semantic Computing, Santa Clara, CA, **(2008)**, pp. 581-588.

[9] Y. Song, J. Du and L. Hou, "A Topic Detection Approach Based on Multi-level Clustering", 2012 31st Chinese Control Conference, Hefei, **(2012)**, pp. 3834-3838.

[10] T. I. Griffiths and M. Steyvers, "Finding Scientific Topics," In: Proceeding of the National Academy of Sciences, vol. 101, no. 1, **(2004)**, pp. 5228-5235.

[11] M. Steyvers, P. Smyth, M. Rosen-Zvi and T. Griffiths, "Probabilistic Author-Topic Models for Information Discovery", In: 10th ACM SIGKDD Conference Knowledge Discovery and Data Mining, Seattle, WA, USA, **(2004),** August 22-25.

[12] Q. Mei and C.Zhai, "Discovery Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining", In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, **(2005),** August 21-24.

[13] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval", Cambridge University Press, New York, NY, USA, **(2008),** pp.118-120.

[14] J. Janruang and S. Guha, "Semantic Suffix Tree Clustering", In. Proc. of 2011 First IRAST International Conference on Data Engineering and Internet Technology (DEIT), **(2011)**, pp.35-40.

[15] P.D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", Journal of Artificial Intelligence Research, vol. 37, **(2010),** pp. 141-188.

[16] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", In: Proc. of the 7th International Conference on World Wide Web, Brisbane, Australia, **(1998)**, April 14-15.

[17] S. Padó and M. Lapata, "Dependency-Based Construction of Semantic Space Models", Computational Linguistics, vol. 33, No. 2, **(2007),** pp.161-199.

[18] C. Wartena and R. Brussee, "Topic Detection by Clustering Keywords", In: Proc. of the 19th International Workshop on Database and Expert Systems Applications, Turin, **(2008)**, pp.54-58, September 1-5.

[19] L. Wu, B. Xiao, Z. Lin and Y. Lu, "A Practical Approach to Topic Detection Based on Credible Association Rule Mining", 2012 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), Bejing, **(2012)**, pp.227-231, September 21-23.

[20] Y. Jo, C. Lagoze and C.L. Giles, "Detecting Research Topics via the Correlation Between the Graphs and Texts", In. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, **(2007)**, pp.370-379.

[21] Y. Jo, J. E. Hopcroft and C. Lagoze, "The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus", In. Proceedings of the 20th International Conference on World Wide Web, New York, NY, USA, **(2011)**, pp.257-266.

[22] M. Jayashri, P. Chitra. "Topic Clustering and Topic Evolution Based On Temporal Parameters", 2012 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, **(2012)**, pp.559-564, April 19-21.

[23] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, H. Qu and X. Tong, "TextFlow: Towards Better Understanding of Evolving Topics in Text", IEEE Transactions on Visualization and Computer Graphics, vol.17, no. 12, **(2011),** pp.2412-2421.

[24] Y. Jin. "A Topic Detection and Tracking method combining NLP with Suffix Tree Clustering", 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE), Hangzhou, vol. 3, **(2012)**, pp. 227-230.

[25] D. Zhang and S. Li. "Topic Detection Based on K-means", 2011 International Conference on Electronics, Communications and Control (ICECC), Ningbo, **(2011)**, pp. 2983-2985.

[26] L. Yue, S. Xiao, X. Lv, T. Wang, "Topic Detection Based On Keyword", 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), Jilin, **(2011)**, pp. 464-467.

[27] T. Masada and A. Takasu, "Extraction of Topic Evolutions from References in Scientific Articles and Its GPU Acceleration", In CIKM, **(2012)**, pp. 1522–1526.

[28] N. Lv, J. Luo, Y. Liu, Q. Wang, Y. Liu and H. Yang, "Analysis of Topic Evolution Based on Subtopic Similarity", 2009. CINC, 09. International Conference on Computational Intelligence and Natural Computing, Wuhan, vol. 2, **(2009)**, pp. 506-509.

[29] A, Shah, K. Khowaja and A. S. Shah "A Model for Handling Overloading of Literature Review Process for Social Science", In Proceeding of International Conference on Advanced Research in Business and Social Sciences 2015, Kuala Lumpur, Malaysia, **(2015)**, pp-335-341.

[30] Y. Chen, Y. Yang, H. Zhang, H. Zhu, and F. Tian, "A Topic Detection Method Based on Semantic Dependency Distance and PLSA", In: Proceeding of 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD), **(2012)**, pp-703-708.

[31] S. M. Krishna and S. D. Bhavani, "An Efficient Approach for Text Clustering Based on Frequent Itemsets", European Journal of Scientific Research, vol. 42, no. 3, **(2010)**, pp. 385-396.

[32] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proceedings of the 20th VLDB Conference Santiago, Chile, **(1994)**.

[33] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichl location", The Journal of Machine Learning Research, vol.3, **(2003)**, pp. 993-1022.

Authors

Abdul Ahad, is currently perusing Ph.D. in Computer Science from University of Malakand, Chakdara, KPK, Pakistan. He received his MS in Computer Science from SZABIST, Islamabad, Pakistan in 2014. He did BS in Computer Science from University of Malakand, Chakdara KPK, Pakistan.

Muhammad Fayaz, is currently perusing Ph.D. in Computer Science from, JEJU National University, South Korea. Before joining the JEJU National University, he has also completed the course work of Ph.D from University of Malakand, Chakdara, KPK, Pakistan. He received MS in Computer Science from SZABIST, Islamabad, Pakistan in 2014. He did MSC from the University of Malakand, KPK, Pakistan in 2011.

Abdul Salam Shah, has completed MS degree in Computer Science from SZABIST, Islamabad, Pakistan in 2016. He did his BS degree in Computer Science from Isra University Hyderabad, Sindh Pakistan in 2012. In addition to his degree, he has completed short courses and diploma certificates in Databases, Machine Learning, Artificial Intelligence, Cybercrime, Cybersecurity, Networking, and Software Engineering. He has published articles in various journals of high repute. He is a young professional and he started his career in the Ministry of Planning, Development and Reforms, Islamabad Pakistan. His research area includes Machine Learning, Artificial Intelligence, Digital Image Processing and Data Mining.

Mr. Shah has contributed in a book titled "Research Methodologies; an Islamic perspectives," International Islamic University Malaysia, November, 2015.