Trends in Engineering and Technology (NCTET-2K17)
International Journal of Advanced Engineering Research and Science (IJAERS) *Special Issue-3*
https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

Big data Curation: Enhanced Information Retrieval System

K.Naresh, A.BasiReddy, S.Swarnalatha

Asst Prof, Dept of CSE, Sri Venkateswara College of Engineering, Tirupati, Andhrapradesh, India.
Asst Prof, Dept of CSE, Sri Venkateswara College of Engineering, Tirupati, Andhrapradesh, India.
Asst Prof, Dept of CSE, Sri Venkateswara College of Engineering, Tirupati, Andhrapradesh, India.

Abstract

In the modern days of big data, the curation of data has become more and more important, especially for handling high volume and complex data systems. With data volumes growing exponentially, along with the increasing variety and heterogeneity of data sources, acquiring the data you may need for analysis has become a costly and time-consuming process. Multiple data sets from various sources must first be processed and connected before they can be used by big data analytics tools. Publication and Presentation of data analytics are also very important. However, traditional data curation systems are not designed for this purpose and there is no consideration on the chronological values. Another limitation is that they are usually designed for programmers, not for the ordinary users. In this paper, we propose Chronological Big data Curation system. In the proposed system, acquisition and care of data are processed on the basis of relations between specific topics and chronological order to ensure that data maintains its value over time. The system is implemented and experimental results show the goodness of the proposed system.

Index Terms—Information Retrieval, Big data, Curation, Chronological Order

I. INTRODUCTION

WITH the emergence and advancement of web technologies, the vast amount of different types of data are rapidly generated and the amount of information increase significantly. Consumers are flooded with data and information in this big data era, but it is not easy to identify valuable data in the overload of information and knowledge. Finding valuable information from the huge amount of data is getting more important, and many nations and companies are investing time and money for acquisition and analysis of data. One of the important issues in the era of big data is to build an informative and trusty information system that is able to satisfy the demands of savvy users, since the efforts and cost of searching, understanding and identifying the central topic in the retrieving service is significant. Existing and the importance and interests relevant to the core of events change over time. As various events and/or accidents get tangled with others, it becomes more difficult to understand the implicit and underlying meaning of the story associatedwith search data . One approach to alleviating this problem is to utilize the collective intelligence. Wikipedia assists users to be able to understand various topics comprehensively through the user's participation. In Wikipedia, a number of small knowledge are collected, and repetitively modify a topic to construct the collective intelligence. However, in the case of collective intelligence, the number of participants and quantity of knowledge will decide the quality of the result. For example, the English Wikipedia includes even academic vocabulary and neologisms by the active participation of a number of users . But, it is not true in other languages. Korean Wikipedia does not have enough information on most of the pages due to the low participation. For the collective intelligence to be useful, a load of amending works and time of the users are required to create correct information. With low participation the reliability of individual information is low, and the information can be utilized for spin control or propaganda. Therefore the elaborate and exquisite methodology to acquire, manage, analyze and reuse the data is required. Data curation is the end-to-end process of creating good data through the identification and formation of resources with long-time value. In information technology, it refers mainly to the management of data throughout its lifecycle, from creation and initial storage to the time when it is archived for future research and analysis, or becomes obsolete and is deleted. In this paper, we proposed the Chronological Big Data Curation system that provides relevant core information in chronological order to assist comprehensive understanding of a specific event or accident that the user attempts to search in the big data environment. The proposed model collects event or accident data from various sources and analyzes the relevant information in chronological order. Comprehensive information is provided by modeling the specific events or

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

knowledge over time and presented by visualization tools. As a result, users can reduce repetitive search tasks for understanding a specific event or knowledge. Reusability of information is also increased by the standardization of the results.

The remainder of this paper is organized as follows.

Table-I
The Features of collected data

	propagation	Document length	Degree of reliability	Extraction
Sns	High	Short	Low	Topic
Blog	Low	Low	Medium	Releation
News	High	Medium	High	Relation

II. RELATED WORK

A. Big data

Big data is defined by three Vs and they are Volume, Velocity, Variety, and it has been very important and useful in achieving precious values such as supporting decision making, finding new insight, and optimizing a process . Big data is a high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization. There are considerable efforts to use big data in many sectors such as the economy, culture, and security because they can provide a range of values that were not able to be obtained using existing data. The movie streaming company Netflix produced "House of Cards" based on the analysis of various information such as the video streaming information, user's evaluation, and search information . Conclusively the drama was successful and won great popularity and awards. During the US presidential election in 2008 and 2013, the presidential camp of the US present Barack Obama analyzed the big data of voters and established customized strategy. This is another example of using big data in practice. Lastly building an epidemic map based on the user's search query by Google is another case of utilization. Since the importance and value of big data are apparent, many studies have been done to find better-utilizing methods. One of the important technical break-through in processing big data is Hadoop, which is MapReduce-based distributed parallel processing system. Hadoop uses map/reduce algorithms for Hadoop Distribution File System (HDFS) for data storage and processing With emerging large volume of data, the information retrieval system also gains attentions from many research groups. Traditional systems have intrinsic limitations with the large volume of data Collective intelligence alleviates these problems to some extent. Wikipedia is a typical example of collective intelligence. The collective intelligence has an advantage that the high-quality contents can be created based on the knowledge of a number of the participants. However, it also has limitations that the user's participation is indispensable, and it can have an error by the users or receive attacks . If the total number of participation is limited, the quality of the content can be compromised. Or the participants can create distorted information on purpose to misinformation.

B. Digital Curation

Curation means assisting audiences to understand the topic by collecting relevant works and providing the in interpretation of the works. Digital curation means performing the curation on digital information. It also means adding new values by storing the digital data throughout the life cycle. Digital curation is originally studied in the e-Science field and has been limited to use to manage the research information efficiently in the art or academic disciplines. However, to manage explosively increasing data and provide the required information, digital curation extends the concept to the whole digital information beyond the purpose of research data storage. There are a number of the definitions of digital curation. The representative definition is the operation that organizes various information on the topic and relevance, provides the organized information to the users with high readability and accessibility, and makes the data possible to reuse . The representative digital curation service is Pinterest . It commenced the service in the USA and expanding the service globally. Currently, there are more than 20 million users. It offers a various scraptype function. A user can curate, organize, share and transfer favorite images on the web. There are various types of digital curation services. Firstly there are curation news services that provide selected news by bloggers; secondly new types of community services such as iamday and facebook timeline that allows people to share news service, SNS(Social Network Service), URL(Uniform Resource Locator) and videos; and thirdly a service performing the placement of similar contents such as Scoop it. A new study has been made on how to find useful information from the News . The digital curation center in the UK studies the method to manage data, and support related research and education. The life cycle model proposed by Digital Curation Centre (DCC) uses the steps presented in Table II. Not all digital curations follow the steps defined in Table II. Another interesting study result to expand digital curation to the whole digital information system takes a different approach. But, general digital

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

curation process consists of collection/store, selection, use, and preservation steps. Our research scope consists of the cycle defined by DCC.

III. PROPOSED METHOD

The entire process of Chronological Big data Curation is summarized in Figure 1. The process of

Table -II
DIGITAL CURATION SEQUENTIAL ACTIONS

Sequence	Description
Conceptualise	Conceive and plan the creation of data, including capture method and storage options.
Create and Receive	Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added nat the time of creation. Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.
Appraise and Select	Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements
Ingest	Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements
Preservation Action	Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats
Store	Store the data in a secure manner adhering to relevant standards.
Access,Use and Reuse	Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.
Transform	Create new data from the original, for example - By migration into a different format. - By creating a subset

the data life cycleof acquiring, storing, using, selecting, preserving of digital

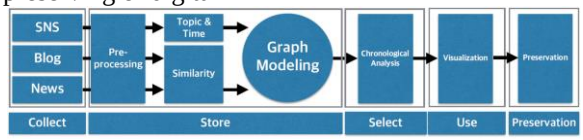


Fig: Chronological Big data Curation Process

curation is explained. Data from various sources is collected, a graph structure is modeled to analyze it according to the time relationship and correlation between them, and the result is reused.

A. Data Acquisition Process

In data collection process, along with the data crawling from each data source, preprocessing such as extracting the date, topic, and relations is performed to use the data effectively in the later stage. In our experiments, data are collected from three different sources. They are SNS, blog and news media. Features that we collected from different data sources are in Table I. Topics and data creation time are extracted from SNS data, and also relationships between different data sources .

1) SNS Data: SNS data is the most important and representative data of our analysis due to its large volume. Topics and creation time are extracted from SNS data. They are mostly unstructured text data, and pre-processing is required to extract the topics by creation time. It is difficult to extract the exact data creation time since it requires high dimensional text mining techniques. Therefore in this paper, the approximate creation time of the topics is estimated on the assumption that data spreads rapidly. For estimating the approximate creation time, it is assumed that data on the events and accidents relevant to the topic appear most frequently around the actual time that it happens. Additionally due to the format of SNS, which is similar to a microblog, each keyword may represent a whole document. Based on the above observations, score (or weight) of each word based on creation time is formulated as in Equation 1.

Scoretotal = Scorestd Scoretfidf (1)

Scorestd calculates the occurrence of a topic using a standardized distribution of scoring. Scoretfidf is the weight using Term Frequency-Inverted Document Frequency (TF-IDF). TFIDF calculates weight using frequency of a word and reverses frequency of words in the entire documents. If specified words appear more frequent, then they gain more weights. On the other hand, if words repeatedly appear

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

throughout the document, the weights decrease. The selected words for

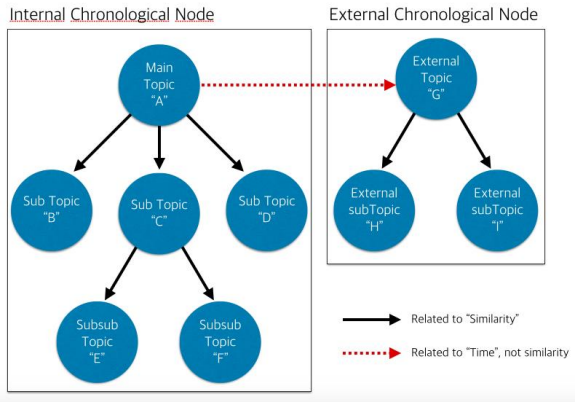


Fig. 2. Chronological Big data Curation Graph

computation are the ones that strongly related to the topic of a document.

2) Blog Data and News Media Data: The blog contents tend to be an opinion about the topic after the actual events or accidents occurred rather than describing the actual event. Thus, blogs have slower spread speed than SNS. Therefore the relationship between the creation time of a blog and the actual time of the topic event is less significant. Due to this observation, blog creation time is not considered as an important factor. However often, a blog has a considerable number of words in a document, and a blogger expresses own opinion clearly about a topic, unlike SNS . Therefore ablog document has many keywords relevant to the topic, and the information that can explain the relation between topics. News data are collected in real-time, and they are usually

based on the actual facts. In many news data standardized keywords appear repeatedly and these characteristics make news data to be appropriate for extracting the relations between topics like we did in blog data.

B. Graph Modeling

The graph is used as the main data structure for our analysis. Nodes in a graph represent topics and links are used to model relevant relations. The topics are extracted from the preprocessing process, and the relations between topics are defined by the similarity and co-occurrence values. Links in a graph are created on the basis of relationships which are computed from time sequences of the topics. Relationship analysis between nodes is based on keywords and their cooccurrences in a document. Two words are assumed to be closely related if they appear concurrently in a document from the blog and news data. If there is a strong relationship between two topics, they are divided as the main topic and a sub-topic based on the time sequences. The relationship between topics is represented as a graph structure to expand to Map/Reduce for parallel processing.

1) Topic: Topic, or theme, is used to describe the core information structure of specified events that can be represented represented in a chronological order. Topics can have sub-topics according to the creation time of the event, and the sub-topics also can include other topics. For each main topic, sub-topics are represented hierarchically using a directed graph structure. The graph consists of directed links like in Figure 2. The directionality means the sequence of time relation and link value means the time value.

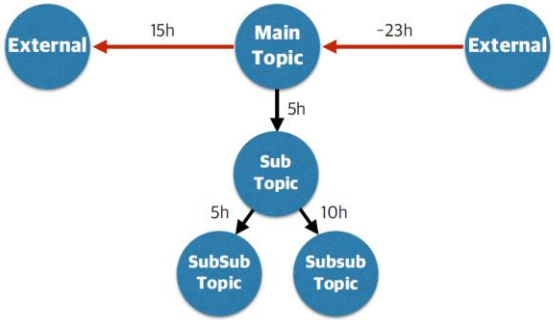
Root node A in Figure 2 represents the main topic, and other lower-level nodes are sub-topics. This topic structure is useful to understand the hierarchical structure of topics. It is possible to analyze which leaf topics are caused by the main topic through the analysis of the graph. Figure 2 shows that node C has two sub-topics F and G, and ultimately the main topic A has an implicit relationship with F and G. This graph also shows that topic C occurred after the topic A, and followed by topic F and G.

2) Relationship and Time Line: There are two types of linksmin the relation graph. The first one is a time relation link which is directional, and its value means the time difference between two topics. The other one is a similarity link which is nondirectional, and it represents “the number of documents that include any particular words concurrently” from the blog and news data using the modified TF-IDF. Figure 3a presents the time relationship among five topics. The integer value means the time difference between the starting point and the end point of the link. For example, Figure 3a presents that the topic B occurred 25 hours after the topic A, and the topic D occurred 1 hour after the topic A. After the time relation link is constructed, the similarity relation between links is calculated. We explained that there is similarity relation between two topics occurred in a document concurrently. The similarity relation graph is presented in Figure 3b. To quantify the relation, we calculated the weight of each keyword using the co-occurrence from the blog and the news. The high co-occurrence means that the topics are closely related as they occurred together across many documents.

3) Chgronological Analysis: As shown in Figure 2, the configured graph consists of internal and external nodes. The internal node represents the set of topics which are related to a series of chronological events and the external node means that the set of topics are indirectly related. The parent nodes mean a topic occurred before the child nodes and the root node represents the main topic. Although the similarity relation between the external node and the main topic is low, it shows that both events occur on the similar time period. A number of exposed nodes is based on the maximum cut values of the time and similarity relation vectors. The threshold value is used to control the number of nodes in a graph using similarity relation value. If the article related to our analysis event appeared long after the first one, the relationship between both topics may not be

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

significant. Similarly, if the similarity relation value between two topics is high, the relationship between both topics may be significant even two topic events occurred long after. To satisfy the both cases, the nodes exposed to the user will be identified by dividing the graph at the point where the similarity is minimum and the time is maximum using the maximum cut threshold value



(a) XML structure

4) Storage and Reuse: Once a graph is constructed, all theb information in a graph is stored using a standardized format.n As a result, the availability and re-usability of data increase and it are also easier to maintain and reuse in the future. Themanalysis result is stored in the tree structure to be able to use in the chronological analysis step, and the eXtensible Markup Language (XML) format is used to express the tree structure appropriately

IV. EXPERIMENT

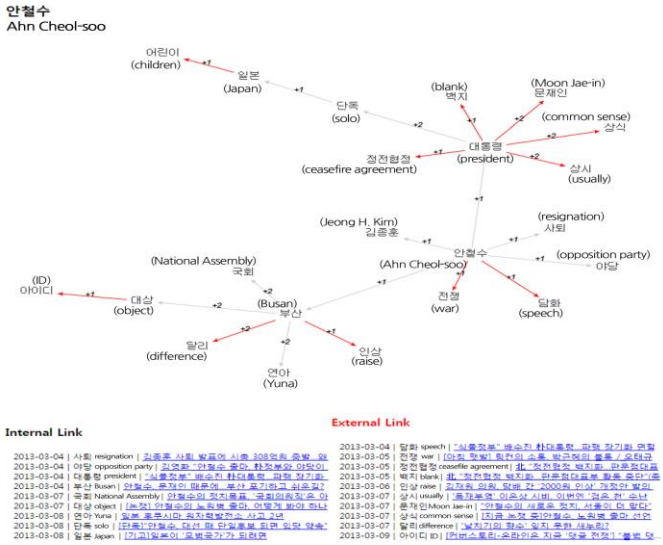
A. Dataset

Twitter, Naver blogs and Naver news for one month period are collected for experiments. The number of total Twitter data is one and a half million, and the total size is about 30GB. 600 topics were extracted from the Twitter and a graph is constructed using this data. We also collected data from 22,692 Naver blogs, And 16,288 news articles during a month in march 2013 were also collected using the Twitter topic. Blog data has 328 words per an article, and news data has 254 words per an article on average. The graph was constructed using the Naver blog and the Naver news data, and there were 123,894 links in the graph. In this paper, we limited the range of data sets to SNS, blogs, and news, but it can be expanded easily by adding the preprocessing module for either extracting topics or calculating the link to collect data from more various sources.

B. Implementation

Komoran , Korean word extractor, is used to extract terms from data sources. Creation date and keywords were used in extracting the topics. Relevant blog and news data were collected using the keywords extracted from the Twitter data. The co-occurrences of words in a document are found from blog and news data, and two pairs of TF-IDF of topics were calculated. The value of the similarity relation link becomes between 0 and 1. The time relation link is generated by the creation time of the topic extracted from the Twitter data. The time relation link value is initially computed by the differences of document creation time and normalized, and final its value would be between 0 and 1. The graph is constructed using the topics, similarity, and time relation. The internal and external nodes are constructed by the chronological analysis. If a similarity relation link has a value above a specific threshold, then these nodes are included as a parent node and a child one. This group is the internal chronological group. The other nodes that did not exceed a specific threshold value were classified as the external chronological group. Below figure presents the result of searching the Korean famous politician, ‘Chulsoo Ahn’. ‘Chulsoo Ahn’ is the main topic, and ‘Jonghoon Kim’ is the sub-topic. ‘Chulsoo Ahn’ ran as

.



. Fig. Search result for keyword ‘Chulsoo Ahn’

an independent presidential candidate, and ‘Jonghoon Kim’ was cited as a counterpart candidate from a ruling party. Later the candidate ‘Chulsoo Ahn’ ‘resigned’ from the presidential candidate to empower the ‘opposition’ leader. Figure below presents the search result of ‘Hacking’ incidents of Korean bank ‘Nonghyup’. One of the major suspicions is that ‘North Korea’ maneuver from behind the scene. And people say “This hacking incident is related to the ‘Cheonan ship’ incident”. The ‘Cheonan ship’ was destroyed by a torpedo of North Korea on March 26, killing 46 South Korean Navy soldiers. And one external link of North Korea is ‘Kim Kwanjin’ who was the Minister of Defense of Korea on the accident day. There are some wrong topics. But almost all of them are related to ‘Hacking’. Below Figure presents a different result compared to other keywords. ‘Whiteday’ is one of the

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

Korean cultures that a boy presents candies to a girl while confessing his love on the 14th of March. 'Whiteday' had a sub-topic 'candy'. Although there was a semantic similarity relation between those two keywords, it is hard to see that they have a timebased relationship. On entering various search keywords, the proposed Chronological Big data Curation performs robustly for the most of the cases, and it helps ordinary people to understand the topic in chronological order. However, for general topics or one-time event that does not last long we found that our proposed system did not show any better results over the other ones.

C. Evaluation

In this paper, we performed a comparative analysis between our proposed method and a traditional one for the evaluation

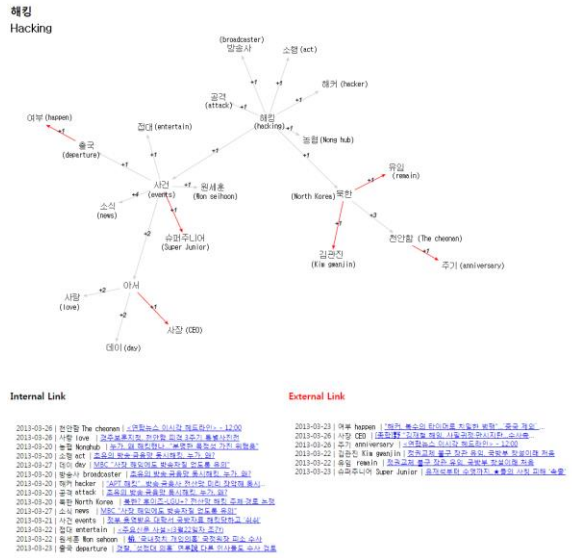


Fig. Search result for keyword 'Whiteday'

Fig. Search result for keyword 'Hacking'

Fig. Graph structure in chronological big data curation

purpose. The question is that "Is it possible to comprehensively understand the event using a query?" We compare our method against the Naver news portal, which is the news site most widely used in Korea. Figure show the results of Naver news search on entering search key word 'Chulsoo Ahn' and Hacking'. As shown in these figures, it is very difficult and

Rank	Date	Headline(KOR)	Headline(ENG)
1	3.31	안철수 "안면치 절세"...가시발길 노원병	Ahn chulsoo says "It is challenging"... Election district ,Nowon-gu second
2	3.31	박원순 "안철수 신당 참여? 그 말은 너무 큰 선물"	Par wonsoon says "do Ahn chulsoo join a new party? The answer is too great a gift"
3	3.31	이야기나누는 안철수-홍합진보당 정책홍 후보	Ahan chulsoo is talking to Jeong taehwung
4	3.31	대선평가보고서 7일 완료, 문재인·안철수 책임론 재점화?	Presidential Election Assessment Report complete at march 7
5	3.31	안철수-허준영 백빙...홍일 조기투표' 최대 변수	Election that is a close-run, Ahn chulsoo and Hur junyoung, Maximum variable is "Early voting in holiday"
6	3.31	"안철수, 허준영과 16%P 격이 1위"	Ahn chulsoo is first ranks compared to the Hur junyoung
7	3.31	안사하는 허준영과 안철수	Ahn chulsoo and hur junyoung are greeted.
8	3.31	악수하는 안철수-김지선	Ahn chulsoo, Kim jisun, shaking hands
9	3.31	안철수, 지지율 제고·단일화 딜레마 고심	Ahn chulsoo, It is struggling to improve and unify poll dilemma.
10	3.31	출산격돌에게 진사하는 무소속 안철수 후보	Independent candidate Ahn chulsoo to greet hikers.

(a) Naver news result. keyword: 'Chulsoo Ahn'

Rank	Date	Headline(KOR)	Headline(ENG)
1	3.31	'원세훈 지시사항' 폭로, 전산망 의원실 해킹 당했다	Exposing 'The instructions of Won sehoon", Hacking the Office of member of National Assembly
2	3.31	[방송·은행 전산망 마비해킹]	Hacking Paralyzed The Broadcaster and Bank Network
3	3.31	"전산망 해킹 활용 중국 뿐는 국내 것" 정부 '웃음거리'	The government become a laugh, utilized IT for hack is China IP, but it is domestic IP
4	3.31	"YTN 단순장애 아냐"...해킹 가능성 !	Not true network problem in YTN is simple, the probability of Hacking increase
5	3.31	방송금융 전산망 마비 '악성코드 해킹' 추정(종합2보)	The assumption: "Malware based hacking" causes paralyzing the broadcaster and bank network
6	3.31	전산망 마비 원인으로 추정되는 해킹 유형, APT란?	What is APT? the type of hacking causes paralyzing the network
7	3.31	방송사 마비사태 고도의 해킹이...대체 누가?	Who hacked Broadcaster Network with high technology?
8	3.31	'에버노트도 해킹..연쇄 해킹 공포에 떨다'	Serial Hacking frighten the company even Evernote ...
9	3.31	[Why] '해킹, 범죄의 판자도 운전자도 죽일 수 있다'	[Why] The distant people can be killed by Hacking
10	3.31	2차 해킹? 술두릉 보고 놀란 YTN	YTN Alarmism : worry about a secondary hacking

(b) Naver news result. keyword: 'Hacking'

Table-III

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

Comparing chronikal bigdatacuration with naver news retrieval systems

	C.Bigdatacuration		Naver news	
	Unique word (%)	Date Varience (D)	Unique word (%)	Date variene (D)
K@5	0.5423	7.6	0.3582	1.2
K@10	0.5193	10.5	0.3755	2.7
K@15	0.5107	12.4	0.3495	3.1

almost impossible for ordinary people to understand what it says. Although it was sorted in order of accuracy, the search results in the first page represented only the articles of the latest day, March 31 in our example. Articles of only two days earlier, March 29, showed up at the 13th page. It is very hard to understand the event in the big picture. It is even worse for the example of the second case. Naver results on the search key word 'Hacking' are hard to understand what happened in relation to 'Hacking'. The first ranked article has been retrieved not from North Korea's hacking attack in March, but the hacking war between the USA and China. On the other hand, in Figure 6 which is the result of the analysis of our proposed method, it is easy to understand the whole story in a big picture on the search keyword 'Hacking'. For a more accurate evaluation, a comparative analysis was performed between the traditional news retrieval system and our curation method. The rational metrics were how much comprehensive understanding is possible? To assess our approach, we used a number of unique words in the articles. Table III shows the results of comparing the proposed method and Naver news retrieval system. On entering a search keyword, each of the search results shows; 1) Unique Word ratio: the ratio of the proper nouns that can identify the topic for the news, 2) Date Variance: Average date range of the creation time of the articles shown in the final result. And 3)K@n is the ratio of the number of exposed news articles over Naver search results.

V. CONCLUSION

In this paper, we propose the Chronological Big Data Curation system that provides relevant core information in a chronological order. The goal of the system is to assist system users to understand the specific event or accident comprehensively in a chronological order on the assumption that the news or blog articles on the specific event appeared over time. Graph data structure is used to establish relations among the articles related to the specified event or accident.Nodes in a graph represent the specified event and links represent two characteristics. The first one is about similarity on the theme of the event that we want to analyze, and the other one is about time-line relationship. Based on the metrics of similarity and chronological order, the main article which really represents the event is placed on the root node. All the directly related articles are placed as child nodes and it repeats recursively.We also propose the notion of internal and external nodes in the proposed graph structure to support important and informative articles but not directly related. All the analysis results are also presented in a graph format to assist users to understand the event in the big picture visually. The proposed method strictly follows the standard data curation methodology. All the data curation activities such as data acquisition, annotation, maintenance, and storage for use conform to the standards. Publication and presentation of the analysis results are maintained such that the value of the data and analysis results are maintained over time. The proposed method is implemented and tested for manydifferent cases over a huge data set and it is proven that our system overwhelmingly shows better results over the traditional systems. The limitation of our approach is word ambiguity. Each topic is not considered as the semantic factor. So, it is hard to distinguish by a difference in their really meaning(i. e. hacking in march or general hacking). In our experiments, the dataset has a narrow range. Therefore, word ambiguity could not find, in the case of large dataset more than about 1 year, we should find how to deal with this ambiguity. For future research, we may look into the case of events that happened one time and the issue does not last long. Our proposed system does not show any better results on those cases. A reasonable inference on this issue is that the significance of the chronological order is very small in these cases as the relationship among articles which appeared in the different time period is not strong even though they talk about the similar top. Also, we used a very simple algorithm in a graph cutting and a more sophisticated one like PageRank might improve the quality of the analysis.

REFERENCES

[1] P. Campbell, "Editorial on special issue on big data: Community cleverness required," Nature, vol. 455, no. 7209, p. 1, 2008.

[2] D. T. Nguyen and J. E. Jung, "Real-time event detection for online behavioral analysis of big social data," Future Generation Computer Systems, 2016.

[3] S. Agrawal, S. Chaudhuri, and G. Das, "Dbxplorer: A system for keyword-based search over relational databases," in Data Engineering,
2002.Proceedings.18th International Conference on. IEEE, 2002, pp.5–16.

[4] T. John, "What is semantic search and how it works with google search," Techulator, Tech. Rep., 7 2012.

[5] M. Zarro and C. Hall, "Pinterest: Social collecting for# linking# using# sharing," in Proceedings of the 12th ACM/IEEE-CS joint conference Digital Libraries. ACM, 2012, pp. 417–418.

https://dx.doi.org/10.22161/ijaers/nctet.2017.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

[6] T.-Y. Liu, "Learning to rank for information retrieval," Foundations and Trends in Information Retrieval, vol. 3, no. 3, pp. 225–331, 2009.
[7] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996, pp. 4–11.
[8] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarizationmand tracking in news and blog corpora." in AAAI spring symposium: Computational approaches to analyzing weblogs, vol. 100107, 2006.
[9] S. Ye and S. F. Wu, Measuring message propagation and social influence on Twitter. com. Springer, 2010.
[10] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," Available at SSRN 1313405, 2008.
[11] G. Orlova and Y. G. Dorfman, "Finding maximum cut in a graph," Engineering Cybernetics, vol. 10, no. 3, pp. 502–506, 1972.
[12] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (xml)," World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/RECxml-19980210, vol. 16, 1998.
[26]S. Soft, "Komoran 2.0, https://github.com/shineware/komoran-2.0,"2014.

Authors:

A. BasiReddy received his Bachelor degree and Master degree in Computer science and engineering from Jawaharlal Nehru Technological University, Anantapur., He was a Assistant professor at Sri Venkateswara college of engineering from 2011 to Till now ,Andhrapradesh

K.Naresh received his Bachelor degree and Master degree in Computer science from Jawaharlal Nehru Technological University, Anantapur., in 2014. He was a Assistant professor at Sri Venkateswara college of engineering from 2015 to Till now ,Andhrapradesh

S.Swarnalatha received her Bachelor degree and Master degree in Computer science and engineering from Jawaharlal Nehru Technological University, Anantapur in 2014. She was a Assistant professor at Sri Venkateswara college of engineering from 2015 to till now ,Andhrapradesh