International Journal of LatestResearch in Engineering and Management" (IJLREM)
ISSN: 2456-0766
www.ijlrem.org Volume 3 Issue 9 ǁ September 2019 ǁ PP 01-07

Event Management and Clustering

Navninderjit Singh
Dept. of Commerce
Punjabi University
Patiala, India

Gurvinder Pal Singh
CT University
Sidhwan
Ludhiana, India

Abstract: Various challenges are emerging for information retrieval as a result of the fast-growing amount of data. The traditional query-driven retrieval system is not so appropriate and efficient so a system is required which automatically process the information, summarize the events and group the associated events. This paper presents a system for event detection and clustering which detects the incoming new events and group them into clusters according to their relativity using Vector space model. The selection of origin document is made for a new event of its own type and the related events are grouped under that origin which results in a cluster. The time window is used to control the cluster size, and idea of threshold is used to control the number of origin documents indefinitely. Our system also provides considerable savings over re-clustering because a significant amount of re-clustering is required whenever new insertions or old deletions are made in the system.
Keywords: event, origin, cluster, management, system

1. Introduction

As we know, in this fast-growing and changing world where a large number of events are happening, it is very difficult to track all the changes manually. Query driven systems fail unless the user knows the type and amount of information needed. Meaning thereby data collection based on recent data motivated queries is usually inadequate to collect various related events [13]. So, an efficient method that inevitably identifies various events in the arriving documents and clubs them into clusters, is needed so that there is no need to track the whole database.

This developed system finds and recovers data which is significant to user queries. Textual data like tweets, various articles, journals, blogs, posts, etc. are considered documents. All such documents are hosted in a database (DB). The management of these documents totally differs from DBMS (Database Management System). The developed system works with the metrics of relevance and similarity. On the other hand DBMS is based on Boolean evaluation [2].

The task of detecting a new event is to determine new events in a given number of documents [4]. One cluster holds documents that are relevant and similar. The process of making clusters is called clustering [3]. A well-known clustering hypothesis for information retrieval was introduced by van Rijsbergen in his book, which states that "Closely associated documents tend to be relevant to the same request". The mentioned proposition underlies the generation of clusters with relevant documents [12] [11]. Our proposed approach is implemented with the expectation to increase the performance of the data organization and retrieval [7].

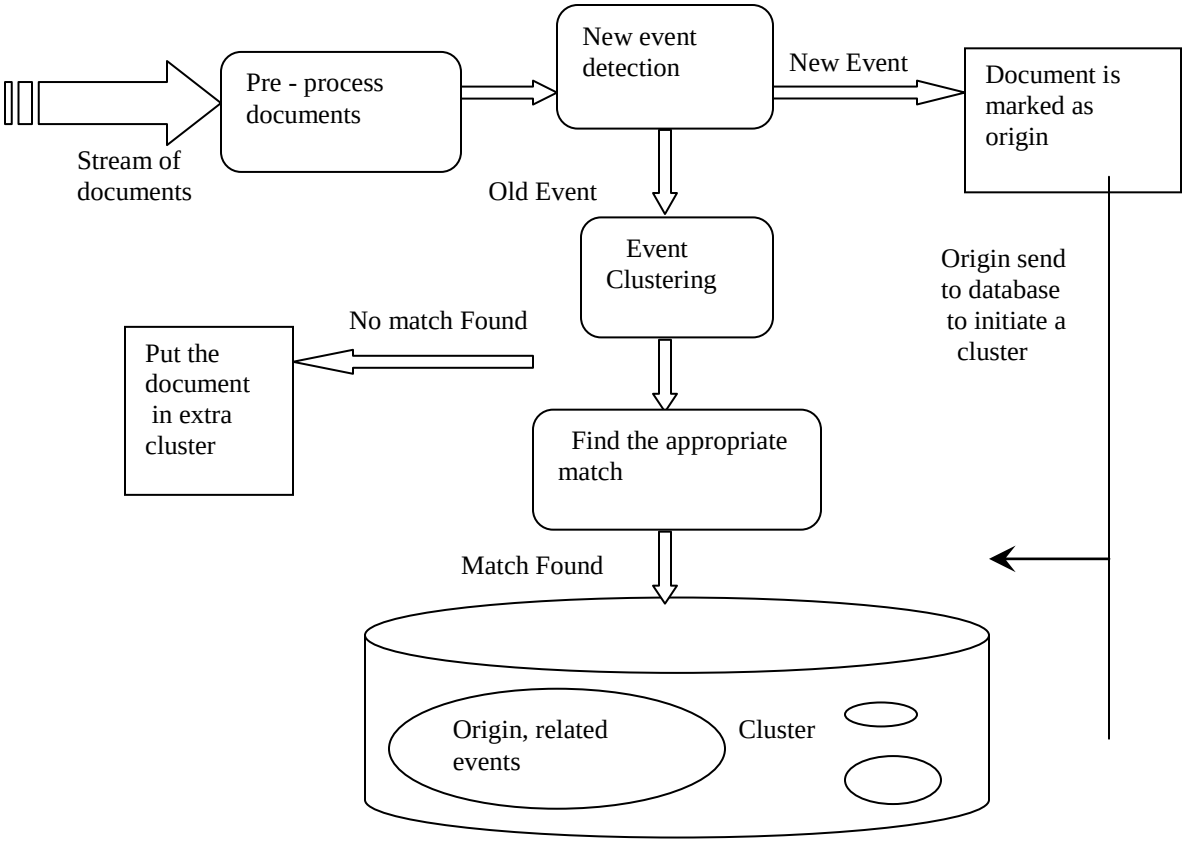

Figure 1. Event Detection and Clustering Process

As the database is usually very large, our approach generally makes the data access on representative documents. Among the existing models for representative document we are using vector space model [10]. Figure 1 illustrates the steps followed. The collection of documents characterizes document-x-term matrix. This matrix is represented as DT matrix. Before representing a document in matrix form, preprocessing of document is done. The stop words and tags etc. are removed, word-stemming is also done, in which suffixes are removed from the word roots, for increasing the scheme performance [1].

The columns of the DT matrix correspond to terms and rows correspond to a document. The construction of the DT matrix could be done by hand or some indexing technique could be used [5, 9, 11]. By definition of Can and Özkarahan [2]: "A DT matrix portrays the document DB D ={d_1, d_2, …, dm}. It also portrays the index terms T={t_1, t_2, ….t_n} . The matrix C presents documents. The entries c_{ij} (1≤ i, j ≤ m) specify any document could be selected with the given probability of d_i from d_j." In this representation, modeling of document is done on m-dim by n-terms.

In the developed system, the concept of origin is used. Origin is the first document in the cluster that detects the event of some new type and initiates the cluster. Afterward, the new upcoming events being detected are grouped according to their combinational properties with respect to the already existing origin documents and similar kinds of events.

We add the concept of Time Window (TW) to prevent the production of oversized clusters. This will ensure to provide a fair chance to all the documents to be an origin of the cluster. The basic idea used

is to process only those documents, which come under a predefined Time Window. This technique puts a limit on the number of terms to be processed. This could be justified as the terms, which are appearing close together on the stream of data, are more likely to represent same kind of events than the documents.

For the selection of origin documents for clustering, we put a threshold on the documents, to make the decision of considering that particular document as an origin document or not. The threshold value differs from document to document. The threshold is being applied to control the number of origin documents otherwise all incoming documents will grow as origin and which is not favorable condition.

The problem of re-clustering is also handled in the developed system. The cluster conserves information for all the updates and modifications on all the addition of original documents. It also stores records on all deleted documents. The motivation of the given algorithm is to avoid unnecessary re-clusterings. Instead of re-cluster all the documents because of any modification, only the documents which fall under the flagged clusters, are re–clustered with this algorithm.

2. Concept of the Algorithm

The first step in this approach is to make a DT matrix from the document database. Rows represent documents and columns represent terms. The entries of DT matrix, d_{ij} ($1 <= j <= n$, $1 <= i <= m$) specifies the index term j (t_j) in document i (d_i). We consider the binary indexing that is the entries d_{ij} in the DT matrix will be either 0 or 1, where 0 signifies the term does not exist and 1 signifies that the term exists in document d_i.

A DT matrix can be represented as:

	$Term_1$	$Term_2$	$Term_3$	$Term_m$	
Doc_1	d_{11}	d_{12}	d_{13}.	. . . . d_{1m}	
Doc_2	d_{21}	d_{22}	d_{23}	. . . . . d_{2m}	
Doc_3	d_{31}	d_{32}	d_{33}	. . . . . d_{3m}	$= DT_{nxm}$
	.	.	.	.	
	.	.	.	.	
	.	.	.	.	
Doc_n	d_{n1} d_{n2}		d_{n3}	. . . . . d_{nm}	

An example of DT matrix is given as follows:

TERMS
T1: eat (ing)
T2: home
T3: bab(y, ies, y's)
T4: sweet(s)
T5: child(ren's)
T6: health(y)

DOCUMENTS
D1: babies and children's home
D2: baby and child will remain healthy at home
D3: she is very sweet
D4: children eat sweets
D5: a child is eating sweets, which are healthy enough.

	T_1	T_2	T_3	T_4	T_5	T_6	
D_1	1	1	0	0	1	0	
D_2	1	1	0 1	1	0		
D_3	0	0	0	0	0	1	$= DT_{5X6}$
D_4	0	1	1	0	0	1	
D_5	0	1	1	1	0	1	

Now from this DT matrix, a C matrix will be generated. The matrix C shows the association among documents calculated using the 2 step probability test. If we consider E_{ik} shows the event of choosing a term t_k as of document d_i is still at 1st step and let E'_{jk} shows the event of choosing the document d_j as of term t_k at the 2nd step, then the probability $P(E_{ik}, E'_{jk})$ is represented as $P(E_{ik}) \times P(E'_{jk})$ [6]. E_{ik} and E'_{jk} represented as:

$$E_{ik} = d_{ik} \times \left(\sum_{h=1}^{n} d_{ih} \right)^{-1} \qquad E'_{jk} = d_{jk} \times \left(\sum_{h=1}^{n} d_{jh} \right)^{-1}$$

The probability matrices PT and PD derived from matrix DT are as follows:

$$PT = \begin{pmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 1/4 & 1/4 & 1/4 & 0 & 1/4 \end{pmatrix} \quad PD = \begin{pmatrix} 1/2 & 1/4 & 0 & 0 & 1/2 & 0 \\ 1/2 & 1/4 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 1/4 & 1/2 & 0 & 0 & 1/3 \\ 0 & 1/4 & 1/2 & 1/2 & 0 & 1/3 \end{pmatrix}$$

The matrix C is constructed by multiplying 2 matrices PT and PD. The resultant matrix for given example will be:

$$C = PT \times PD = \begin{pmatrix} 0.417 & 0.417 & 0.000 & 0.083 & 0.083 \\ 0.313 & 0.438 & 0.000 & 0.083 & 0.188 \\ 0.000 & 0.000 & 0.333 & 0.333 & 0.333 \\ 0.083 & 0.083 & 0.111 & 0.361 & 0.361 \\ 0.063 & 0.188 & 0.083 & 0.271 & 0.396 \end{pmatrix}$$

One element of the matrix C_{ij} can be calculated as:

$$C_{ij} = \sum_{k=1}^{n} E_{ik} \times E'_{jk}$$

$$= \sum_{k=1}^{n} (\text{probability of choosing } t_k \text{ as of } d_i) \times (\text{probability of choosing } d_j \text{ from } t_k)$$

This equation is presented as follows:

$$C_{ij} = \theta_i \times \sum_{k=1}^{n} d_{ik} \times d_{jk} \times \phi_k \qquad 1 <= i , j <= m$$

The θ_i is inverse of the ith row sum and ϕ_k inverse of the k^{th} column sum as represented below:

$$\theta_i = \left(\sum_{j=1}^{n} d_{ij} \right)^{-1} \quad 1<= i <=m \qquad \text{and} \qquad \phi_k = \left(\sum d_{jk} \right)^{-1} \quad 1<= k <=n$$

3.1 Algorithm U1: For Clustering

For each new incoming document d*i*, do the following:

1. Put the oldest document in the database to the d_j i.e.
   d_j = oldest document
2. Determine some value for Time Window.
3. If the difference between the arrival times of current document d*i* and oldest document d*j* is greater than predetermined Time Window (TW) i.e. if di_{time} - dj_{time} > TW

Then repeat the following until di_{time} - dj_{time} < TW

   3a. Delete the oldest document d_j.
   3b. Check whether d_j is an origin document or not
      If yes, make the next document the origin of that cluster.
   3c. Make the oldest document in that Time Window, the d_j.
      i.e. d_j = oldest document in TW.

4. Calculate the origin strength value OS_i for the document d_i as follows:

$$OS_i = \rho_i \cdot \sigma_i \sum_{k=1}^{n} d_{ik}$$

where ρ_i presents the percent of nonrelevance of document d_i to other documents. Whereas σ_i presents the percent relevance with other documents in the database. ρ_i and σ_i are stated below:

ρ_i = decoupling coefficient of $d_i = c_{ii}$

and σ_i = coupling coefficient of $d_i = 1 - \rho_i$

5. Calculate the Threshold value Thr for the document d_i as follows:
   Thr = $\sum$ OS*i* / No. of Documents in the TW
   di ε TW
6. If origin strength value comes out to be greater than threshold calculations i.e.

if OS*i* > Thr then

   6a. Mark the document as a new event, make it an origin and starts a new document.

Else

   6b. Mark the document as an old event and do the following:
      6b1. If any cluster that exists
         Calculate C_{ij} to find out that to which cluster this event belongs.
      6b2. Else
         Put the document in Un - Matched Cluster.

3.2 Algorithm U2: For re-clustering after when some insertions and deletions have been made.

1. Calculate θ_i (inverse row sum values) and ϕ_i (inverse column sum values), ρ_i
2. Calculate origin strength values of all the documents currently present.
3. Sort documents by their origin strength value.
4. Start selecting documents as origin in order of decreasing origin strength.
5. Sort documents by their document number.
6. For the first cluster going to be initiated, go to the last step.
7. For each document in the sorted list do the following:
   7a. If this document has become a non – origin document whereas it was an origin the document then flag the cluster containing this document.
   7b. If this document has become an origin document whereas it was not an origin the document then flag the cluster containing this document.
8. Intended for every flagged cluster in the 7th step, re - cluster all documents using Algorithm 1.
9. Cluster all the new documents.

4. Conclusion

In this study, a new system for event detection and clustering has been developed. When a new event is detected then either an origin document of its own type is made for this new event and the related new events are grouped under that origin which results in a cluster or the system labels it as per its relativity and categorizes it into one existing cluster. To prevent the clusters grow unlimited, the concept of Time Window is used. We avail the concept of threshold to control the number of origin documents indefinitely. When insertions and deletions are done to the data a considerable re-clustering is required and for this to happen our system provides considerable savings over re-clustering.

5. References

[1] K. Kaur, X. Xiaojiang Du and K. Nygard, "Enhanced routing in Heterogeneous Sensor Networks", IEEE Computation World'09, pp. 569-574, Athens, Greece, Nov. 15-20, 2009.

[2] Lauren Evanoff, Nicole Hatch, Gagneja K.K., "Home Network Security: Beginner vs Advanced", ICWN, Las Vegas, USA, July 27-30, 2015.

[3] Gagneja K.K. and Nygard K., "Heuristic Clustering with Secured Routing in Heterogeneous Sensor Networks", IEEE SECON, New Orleans, USA, pages 51-58, June 24-26, 2013.

[4] Gagneja K.K., "Knowing the Ransomware and Building Defense Against it - Specific to HealthCare Institutes", IEEE MobiSecServ, Miami, USA, pp. 1-5, Feb. 11-12, 2017.

[5] Gagneja K.K., "Secure Communication Scheme for Wireless Sensor Networks to maintain Anonymity", IEEE ICNC, Anaheim, California, USA, pp. 1142-1147, Feb. 16-19, 2015.

[6] Gagneja K.K., "Pairwise Post Deployment Key Management Scheme for Heterogeneous Sensor Networks", 13th IEEE WoWMoM 2012, San Francisco, California, USA, pages 1-2, June 25-28, 2012.

[7] Gagneja K.K., "Global Perspective of Security Breaches in Facebook", FECS, Las Vegas, USA, July 21-24, 2014.

[8] Gagneja K.K., "Pairwise Key Distribution Scheme for Two-Tier Sensor Networks", IEEE ICNC, Honolulu, Hawaii, USA, pp 1081-1086, Feb. 3-6, 2014.

[9] Gagneja K., Nygard K., "Energy Efficient Approach with Integrated Key Management Scheme for Wireless Sensor Networks", ACM MOBIHOC, Bangalore, India, pp 13-18, July 29, 2013.

[10] Gagneja K.K., Nygard K., "A QoS based Heuristics for Clustering in Two-Tier Sensor Networks", IEEE FedCSIS 2012, Wroclaw, Poland, pages 779-784, Sept. 9-12, 2012.

[11] K. K. Gagneja, K. E. Nygard and N. Singh, "Tabu-Voronoi Clustering Heuristics with Key Management Scheme for Heterogeneous Sensor Networks", IEEE ICUFN 2012, Phuket, Thailand, pages 46-51, July 4-6, 2012.

[12] Gagneja K.K., Nygard K., "Key Management Scheme for Routing in Clustered Heterogeneous Sensor Networks", IEEE NTMS 2012, Security Track, Istanbul, Turkey, pp. 1-5, 7-10 May, 2012.

[13] Runia Max, Gagneja K.K., "Raspberry Pi Webserver", ESA, Las Vegas, USA, July 27-30, 2015.

[14] A. S. Gagneja and K. K. Gagneja, "Incident Response through Behavioral Science: An Industrial Approach," 2015 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2015, pp. 36-41.

[15] Tirado E., Turpin B., Beltz C., Roshon P., Judge R., Gagneja K., "A New Distributed Brute-Force Password Cracking Technique", Future Network Systems and Security, FNSS Communications in Computer and Information Science, vol. 878, pp 117-127, 2018

[16] Caleb Riggs, Tanner Douglas and Kanwal Gagneja, "Image Mapping through Metadata," Third International Conference on Security of Smart Cities, Industrial Control System and Communications (SSIC), Shanghai, China, 2018, pp. 1-8.

[17] Keely Hill, Gagneja K.K., "Concept network design for a young Mars science station and Trans-planetary communication", IEEE MobiSecServ 2018, Miami, FL, USA, Feb. 24-25, 2018.

[18] Javier Campos, Slater Colteryahn, Gagneja Kanwal, "IPv6 transmission over BLE Using Raspberry PI 3", International Conference on Computing, Networking and Communications, Wireless Networks (ICNC'18 WN), March, 2018, pp. 200-204.

[19] Gagneja K., Jaimes L.G., "Computational Security and the Economics of Password Hacking", Future Network Systems and Security. FNSS 2017. Communications in Computer and Information Science, vol. 759, pp. 30-40, Springer, 2017.

[20] Gagneja K.K. Ranganathan P., Boughosn S., Loree P. and Nygard K., "Limiting Transmit Power of Antennas in Heterogeneous Sensor Networks", IEEE EIT2012, IUPUI Indianapolis, IN, USA, pages 1-4, May 6-8, 2012.

[21] C. Riggs, J. Patel and K. Gagneja, "IoT Device Discovery for Incidence Response," 2019 Fifth Conference on Mobile and Secure Services (MobiSecServ), Miami Beach, FL, USA, 2019, pp. 1-8.

[22] S. Godwin, B. Glendenning and K. Gagneja, "Future Security of Smart Speaker and IoT Smart Home Devices," 2019 Fifth Conference on Mobile and Secure Services (MobiSecServ), Miami Beach, FL, USA, 2019, pp. 1-6.

[23] Keely Hill, Kanwalinderjit Kaur Gagneja, Navninderjit Singh, "LoRa PHY Range Tests and Software Decoding - Physical Layer Security", 6th IEEE International Conference on Signal Processing and Integrated Networks (SPIN 2019), 7 - 8 March 2019.

[24] Alexandro Riuz, Carloas Machdo, Kanwal Gagneja, Navninderjit Singh, "Messaging App uses IRC Servers and any Available Channel", 6th IEEE International Conference on Signal Processing and Integrated Networks (SPIN 2019), 7 - 8 March 2019.

[25] Nica Ameeno, Kalib Sherry, Kanwal Gagneja, "Using Machine Learning to detect the File Compression or Encryption", AJCS, 2019.

[26] Broderick Wolff, Alexander Hughes, Xin Wang, Kanwal Gagneja, "The Nature of Phishing and Payload Delivery", AJCS, 2019.