WIKI STREAMS: Wikipedia Article Recent Edit Retrieval System using Hierarchical Stream Clustering

Arun Manicka Raja M (✉ arunmanickarajam@gmail.com)

RMK College of Engineering and Technology <https://orcid.org/0000-0001-9904-211X>

Swamynathan Sankaranarayanan

Anna University Chennai

Research Article

Keywords: Wikipedia edits, domain corpus, stream clustering, queries, retrieval

Posted Date: June 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-452931/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

WIKI STREAMS: Wikipedia Article Recent Edit Retrieval System using Hierarchical Stream Clustering

Arun Manicka Raja M. · Swamynathan S.

Abstract Stream analytics, a new paradigm in data analytics, has gained momentum due to the voluminous stream data generation. With the huge increase in the edits performed on Wikipedia topics, it is tedious for the digital knowledge discovery users to find their domain updates immediately. The users need to go through large information and spend more time to find the potential data. There is a need for retrieving the Wikipedia edits based on the meta data of the article edits for later retrieval. Hence, the clustering technique may be employed in order to group the Wikipedia article edits domain wise. Hence, in this paper, hierarchical stream clustering is applied in order to retrieve the edits based on the user interest. Over a period of month, the data from Wikipedia is collected and used as a dataset. Our method is compared with the state-of-the-art clustering system WikiAutoCat and it is observed that the accuracy is improved by 10% and the clustering time is reduced by 20%.

Keywords Wikipedia edits · domain corpus · stream clustering · queries · retrieval

1 Introduction

Web is the collection of heterogeneous data from different sources such as social media, news websites, blogs, online digital libraries and online encyclopedias etc. Social media users post information on various topics. Blogs are created by various users in their interested domains. Online digital libraries provide specific domain

Arun Manicka Raja M.
Assistant Professor, Department of Computer Science and Engineering
RMK College of Engineering and Technology
Chennai, India.
E-mail: arunmanickarajam@gmail.com

Swamynathan S.
Professor, Department of Information Science and Technology
College of Engineering Guindy
Anna University, Chennai, India.
E-mail: swamyns@annauniv.edu

knowledge and research findings. In addition to these information sources, online encyclopedias play a major role in sharing the domain contents to the users. Wikipedia is one of the major online encyclopedias and it provides information on all domains pertaining to multiple topics. Wikipedia articles are referred by most of the online users. These articles are edited by the Wikipedia article editors with a new data on a topic. Wikipedia article edits are an authorized content and these edits are created from different regions round the clock. These edits are continuously generated on various topics and on an average 3 edits are created per second [1]. Since Wikipedia is one of the trusted sources of information, many of the users are interested in knowing the recent updates in their domain from Wikipedia articles.

There is no provision in Wikipedia to maintain the article edits related to relevant domains. It is difficult for the domain specific interested users to immediately refer the recent edit content related to their domain [2]. It is useful if some methods are applied on the edit contents of the articles such as classification, clustering and summarization. The classification of Wikipedia articles is performed using domain ontology-based classifiers. This is done only for few domains and it requires large training data for covering all domains [3]. The clustering of articles is done using various clustering algorithms. It involves grouping of all the domain contents into distinct groups [4]. Summarization of Wikipedia articles is performed using concept model of semantic elements for generating summary on a specific topic using article contents [5].

Among these Wikipedia article-based data management systems, clustering is considered as the prominent method for grouping the article content. Even though the clustering of Wikipedia articles has been done in many existing systems, it is required to implement a system to handle the dynamic Wikipedia edits. It is necessary to estimate the edit similarity and the edit contents have to be maintained in a hierarchical structure to easily access it for query-based retrieval. Since the edits need to be handled as data streams, it is essential to keep the meta data of the streams alone without storing the edit contents in offline. These all create a need to design a system to incorporate the said requirements. Hence, in this paper, a hierarchical stream clustering-based Wikipedia edit retrieval system is proposed. The rest of the paper is organized as follows. Section 2 discusses the works related to stream clustering of Wikipedia articles. Section 3 describes the hierarchical stream clustering and retrieval of Wikipedia article edits. The empirical results are discussed in Section 4. Finally, Section 5 concludes the performance results and observations of the system.

2 Related works

Wikipedia is the largest online platform for providing knowledgeable contents on various topics from different domains. There are many articles available in Wikipedia pertaining to diverse domains. These article contents are updated by the interested users when they add or modify the contents on a specific topic [6]. The edits are rapidly accumulated for various topics. It needs to be collected in-

stantaneously to scan and analyse the content present in it. It requires concurrent processing of data collection using some way of parallel data crawling methods. Multi-threaded based crawlers are used for collecting these Wikipedia edits. The Wikipedia article contents are analysed in various forms [7]. The potential information from these article edits is interesting to the users who look for updated information on their domains. It is necessary to process these article edits to bring some insights from it. The article edits may be grouped, so that users can easily retrieve the recent edits from the grouped data.

Clustering is one of the data mining tasks for grouping the data in an unsupervised method. Many traditional clustering methods are available for grouping the data. These clustering methods are not appropriate to handle the Wikipedia edits. Most of the Wikipedia article clustering have used Wikipedia data repository [8]. The stream clustering is applied on numerical streams and short text messages. It has not been applied on handling the Wikipedia edits. It is needed to apply the stream clustering algorithm for the Wikipedia edit streams to cluster them dynamically. Hai-long et.al [9] conducted a survey on data stream mining algorithms and analysed that stream clustering lacks about handling the text data and need a method to perform it. Mihai et. al analysed the problems of extracting event related information from Wikipedia using clustering of Wikipedia articles [10]. Thomas et.al proposed a system for Wikipedia stream clustering to categorize various article edits performed on Wikipedia. This system failed to focus on handling the streams with some identities for the arriving streams instead of storing the edit contents. Jean et.al developed anytime data stream clustering algorithm using a complex network construction model [11]. This model lacks in handling the drifts in stream data. Chunyong [12] proposed a model for anomaly detection based on data stream clustering. The focus was only on limited data, not efficient when performed on large data.

The clustering process considerably affects the retrieval of contents. Andrei et. al [13] proposed a stream clustering which uses hierarchical structure for summarizing the streams. Dilip et. al [14] introduced stream clustering with time series data to organize the content with temporal information for easy retrieval with user queries. Dean et. al [15] used adaptive hierarchical stream clustering using map reduce for handling large data. Here, the hashing is used for easy retrieval. However, the domain corpuses are not considered.

The variation of results between actual stream clustering and corpus-based stream clustering is throwing light to viable research. A reliable domain content management system shall provide content clustering with recent updates on various topics. From the literature, it is observed that there is a lack of work on creating a retrieval system, especially like Wikipedia kind of edit stream generating platforms. Hence, in this work hierarchical stream clustering based Wikipedia edit retrieval system is developed. The unique contribution of this work is that the live Wikipedia edits are collected instantaneously. Another main contribution of the work is that the streams are not stored for cluster generation, instead, the meta data of the Wikipedia edits are stored in the clusters. The queries are applied on the grouped data and the stream contents are retrieved instantly using

the meta data available in the clusters.

3 Wikipedia edit stream retrieval system

Wikipedia edit Stream Retrieval System (WRS) is shown in figure 1. The system consists of major components such as Wikipedia Edit Stream Data Source, Wikipedia Edit Stream clusterer, and Wikipedia Edit Stream Retriever.

Wikipedia edits are generated for different topics on various domains. The domain specific corpuses pertaining to 9 different domains are given as input for keyword extractor. The keyword extraction is carried out by concurrently scanning the Wikipedia edit streams. The stream clustering involves grouping of the streams based on the keywords present in it. The clustering is performed parallelly on the arriving edit streams. Only the metadata of these clustered edit streams are buffered. These meta data are used for accessing the streams from online Wikipedia. The buffer consists of meta data of the edit streams that are clustered by the hierarchical stream clusterer. The edit stream retriever helps to retrieve the recent edit contents based on the user given query.

Fig. 1 Wikipedia Edit Stream Retrieval System

3.1 Similarity Computation in Hierarchical clustering model

The data are arranged in the hierarchical structure of clusters. The main task in hierarchical clustering is that how data elements are assigned to form the hierarchical structure. In agglomerative hierarchical clustering, the starting of cluster formation is considered as separate data elements. The next steps is proceeded as the grouping of similar data elements to the starting point of the cluster.

Agglomerative clustering assigns to set of data elements O and the continuation of its group $S^0, S^1 \dots S^{n-1}$ to the clusters and it is assigned to every clusters $C_i^k \in S^k$.

The split of the set of data elements S^0 are its distinct elements that is single data cluster where in the number $h(C_i^{(0)}) = 0$, such as $i = 1, 2, \dots, l_0$ come under individual single data cluster $C_i^{(0)}$.

The simple linkage is defined as if D is dissimilarity coefficient, C_1, C_2 are clusters, A_i comes under C_1 and A_j comes under C_2 . The distance of clusters is calculated as follows.

$$d_{SL}(C_1, C_2) = \min_{i,j} d(A_i; A_j)$$

The complete linkage is defined as follows. It determines the distance between the clusters.

$$d_{CL}(C_1, C_2) = \max_{i,j} d(A_i; A_j)$$

The average linkage is defined as follows.

$$d_{AL}(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{A_i \in C_1} \sum_{A_j \in C_2} d(A_i; A_j)$$

n_1 = number of data elements in Cluster 1

n_2 = number of data elements in Cluster 2

In centroid method, the dissimilarity is calculated as the distance between the cluster centroids. The distance coefficient is known as Lance-William formula.

$$d_{LW}(C_1, C_2 \cup C_3) = \frac{n_2}{n_2 + n_3} d_{LW}(C_1, C_2) + \frac{n_3}{n_2 + n_3} d_{LW}(C_1, C_3) - \frac{n_2 n_3}{(n_2 + n_3)} d_{LW}(C_2, C_3)$$

n_1 = number of data elements in Cluster 1

n_2 = number of data elements in Cluster 2

n_3 = number of data elements in Cluster 3

In this proposed work, the dissimilarity of clusters is calculated using complete linkage method. In addition, the fuzzy set concept is applied to distinguish the similar clusters with little dissimilar article edit contents. The mathematical description about fuzzy set is explained as follows.

Given a bunch of articles, $X = \{x_1, \dots, x_n\}$, a fuzzy set S is a subset of X that permit each item in X to have member degree somewhere in the range of 0 and 1. i.e., $FS : X \rightarrow [0, 1]$. The fuzzy set is applied on groups. That is, given arrangement of items, a group is fuzzy arrangement of articles. Such a group is called fuzzy cluster. Thus, a grouping contains different fuzzy bunches of articles.

Given a bunch of items, $o_1, \dots, o_n$, a fuzzy grouping of k fuzzy bunches, $C_1, \dots, C_k$, can be addressed utilizing a segment framework, $M = w_{ij}$ ($1 \leq i \leq n, 1 \leq j \leq k$), where w_{ij} is the participation level of o_i in fuzzy cluster C_j . The segment matrix ought to fulfill the accompanying three prerequisites:

1. For each item, o_i , and group, C_j , $0 \leq w_{ij} \leq 1$. This necessity upholds that a fuzzy bunch is fuzzy set.
2. For each item, o_i , $\sum_{j=1}^k w_{ij} = 1$. This necessity guarantees that each item takes part in grouping identically.

3. For each group, C_j , $0 < n_{wij} < n$. This necessity guarantees that for each bunch, there is at least one item for which the enrollment esteem is nonzero.

Given a bunch of items, $o_1, \dots, o_n$, and a fuzzy bunching C of k groups, $C_1, \dots, C_k$. Let $M = w_{ij}$ ($1 \leq i \leq n, 1 \leq j \leq k$) be the parcel network. Let $c_1, \dots, c_k$ be the focuses of groups $C_1, \dots, C_k$, separately. Focus can be either mean or medoid. Comparability or distance between the focal point of the bunch and an item is task measure which chooses how well article has a place with group. For any article, o_j , and the bunch C_j , if $w_{ij} > 0$, then $dist\ o_i, c_j$ measures conservativeness of item with relating group. As an article has a place with more than one group the sum of distances to the corresponding cluster centers weighted by degrees of membership captures how well the object fits focuses weighted by levels of participation catches how well the item fits the grouping. For an article o_i , the amount of squared blunder (SSE) is $SSE(o_i)$ controls the impact of the levels of enrollment. The bigger the estimation of p , the bigger the impact of the levels of enrollment. Hence, the SSE for a bunch, C_j , is $SSE\ C_j$ and the SSE of the grouping is $SSE(C)$.

4 Results and Analysis

The Wikipedia edits are observed and totally more than 300000 edits were observed in a 1-week period. Since it is tedious to tabulate the entire results, the sample of 19604 Wikipedia edits were considered for the result analysis and these edits are related to 9 domains collected over a period of 6 hours interval. The do-main corpuses used here was collected from various data repository. During the experimentation, each of the domain related words are compared with the edit streams for keyword extraction. The major domains considered in this work are politics, healthcare, business, sports, education, electronics, nature, software and travel. The politics domain comprises of almost 97000 different keywords [17]. Healthcare domain consists of around 60000 keywords [18]. Business domain contains around 600000 related words [19]. Sports domain consists of around 32000 sports related keywords [20]. Education domain consists of nearly 84000 keywords [21]. The keywords for education, nature, software and travel domain are used from Wik-ikipedia hierarchical corpuses [22]. The Stanford POS tagger has been applied for tagging the terms present in the edit streams [19].

The similarity of the stream content is calculated and corpus based keyword tagging is combined for putting the streams into an appropriate cluster. Incorporating the stream omission after the similarity check and considering the Wikipedia article edit ID increases the performance of stream clustering. Equal importance is provided to all edits received at various periods on different topics. Priority is given to the recent Wik-ikipedia edits as the similar content needs to be checked frequently for retrieving it to the users. As content-based similarity depend more on the similar keywords present in the edit streams, the corpus-based key-word extraction and similarity measure is assigned with higher weightage in the applied method. The Hier-archical Stream Clustering (HSC) algorithm is given as follows.

Algorithm: Hierarchical Stream clustering (HSC)

Inputs: streams $S_1, S_2, \dots, S_n$, corpuses $c_1, c_2, \dots, c_k$.

```

Output: clusters  $C_1, C_2 \dots C_m$ .
begin
  foreach  $S_i$  in  $n$ 
    while ( $S_i \neq \text{null}$ ) do
       $K_i = \text{extract-keywords}(S_i)$ 
      foreach  $K_i$  in  $c_k$ 
    foreach  $j$  in  $m$ 
       $\text{Sim}_i = \text{compute-similarity}(K_i, S_i)$ 
    if ( $\text{sim}_i > 0$ )
       $m_i = \text{extract-metadata}(S_i)$ 
       $C_j = \text{hierarchical-clustering}(m_i, S_i)$ 
       $d(C_j, C_{j+1}) = \max_{(x_j \in C_j, x_{j+1} \in C_{j+1})} \{d(x_j, x_{j+1})\}$ 
      if ( $d(C_j, C_{j+1}) = \text{null}$ )
        merge ( $C_j, C_{j+1}$ )
      else
         $C_{m-1} = C_j$ 
         $C_m = C_{j+1}$ 
      end if
    discard  $S_i$ ;
    else
       $\text{others} = \text{clustering}(m_i, S_i)$ 
    end if
  end while
end

```

The clustered Wikipedia edit streams are shown in table 1. Only the streams of 4 sub-domains of 3 domains are listed. The number of streams grouped under each domain with the prominent keywords are mentioned in the table. Each cluster contains the streams with time date information mentioned in it. These information helps in retrieving the streams with topic wise easily for a specific period.

To prove the effectiveness of WRS, the article id (meta data) based cluster representation is analysed since stream data is not stored in clusters. The clusters contain the article title, article id, major keywords and minor clusters. The article id helps in retrieving the edit streams later from the online Wikipedia platform based on the user query. The clusters with the Wikipedia article metadata are illustrated in table 2.

The clusters are generated based on the similarity of the content in the streams. In addition, the main cluster and the sub clusters are formed based on the content received in the streams in a particular interval of time. The cluster generation process is shown for a single domain in figure 2 with various levels. It includes the minor keywords in the outer level and major keywords in the inner level and the cluster in the middle.

4.1 Query Processing

The Hierarchically clustered Wikipedia edits are more informative than the data that are clustered with-out hierarchical clustering. The Wikipedia edit have been

Table 1 Clusters and sub clusters

Cluster	Sub cluster	Number of Wikipedia edit streams	Keywords in streams
Education	Engineering	312	Structural, design, devices, automotive, hardware, aerospace
	Colleges	134	Students, academic, courses, training, research
	Knowledge	288	Cognitive, perception, learning, experience
	Research	442	Discovery, facts, methodology, hypothesis
Sports	Esports	526	Twitch, games, players, teams
	Athletics	453	Sport, track, physical
	Leagues	642	Play, interleague
	Teams	423	Tournament, coach, group
Healthcare	Childbirth	790	Caesarean, gestation
	Patients	643	Outpatient, clinic, medicine, discharge
	Practitioner	521	Doctor, occupational, campaign, medical
	medicine	842	Treatment, physician, healing

Table 2 Clusters with wikipedia article metadata

Cluster	Article Title	Article ID	Major keywords in cluster	Minor keywords in cluster
Education	Pedagogy	924529165	Learning, teaching student	Research, taxonomy, curriculum
Sports	Physical activity	930048317	Metabolic, exercise, energy	Aerobic, muscle, strength
Healthcare	Injury	930649136	disability	Bone, catastrophic
Nature	Supernatural	931039359	Season, naturalist, natural	Universe, rainbow, advent
Politics	Election	930873777	Majority, voters, legitimacy	Democracy, regional, voting
Business	Inventory	929954288	Stock, accounting, sale	Products, defective, demand
Software	Algorithm	931198767	Unambiguous, calculation, logic	Computability, control, memory
Electronics	Sensors	930928002	Range, properties, device	MEMS, processor, accuracy
Travel	Location	930988845	Place, boundary, locality	Geography, site, area

Fig. 2 Hierarchical Cluster Generation for Healthcare Domain

clustered hierarchically without any pre-mentioned number of clusters. These hierarchical cluster results are more deterministic and provides higher retrieval accuracy [24].

The clustered meta data is evaluated for the user query to know the recent edit performed in the last 1-hour period. The result is obtained by parsing the user given query. The time information has been taken from the user query and it is checked with the clustered meta data. Further, the article IDs corresponding to the particular time interval is identified and the number of edits were counted under each cluster. The user query and its cluster contents are shown in table 3.

The most edited Wikipedia topics and domains are identified by parsing the user query and matching it against the number of edits under various topics in each domain. The topics in each domain is identified and it has been shown in table 4.

It is essential to know the countries from where most of the Wikipedia edits are generated. The user given query is parsed for the keyword ‘country’. The IP address and the geolocation information of the edits are identified and then the city name, region name and country name are identified using the IP address.

The most edits happened countries are shown in table 5. The Wikipedia edit information with its originated location (geolocation), country, domain, keyword and wiki edit content is displayed in table 6.

The recent Wikipedia edits are clustered and the cluster wise edits are shown in figure 3. The user given keyword-based Wikipedia edits are retrieved from the clus-

Table 3 Cluster wise Recent Wikipedia Edits

Query	Cluster wise Edits	Total Wikipedia Edits
What are the Wikipedia edits generated in last 1-hour period(05.45 to 06.45 PM IST on 23.12.2019)	Education 458 Healthcare 485 Software 342 Sports 549 Nature 237 Politics 529 Business 368 Electronics 426 Travel 374 Undefined 1058	4826
What are the Wikipedia edits generated in last 10 minutes	Education 46 Healthcare 32 Software 28 Sports 44 Nature 38 Politics 18 Business 22 Electronics 26 Travel 14 Undefined 312	580

Table 4 Most edited Wikipedia Topics and its Clusters

Query	Cluster	Topics
What are the mostly edited topics of Wikipedia in 1-hour period	Politics	Taiwanese legislative election, Jharkhand Legislative Assembly election, Afghan presidential election
	Sports	Albania national youth football team, Rugby World Cup squads, Indonesia national under-23 football team

Table 5 Country Details of Recent Wikipedia Edits from Clusters

Query	Country	Region	City	Cluster	Number of Wikipedia Edits
Which country users are recently generated more edits in last 1-hour.	United States	Maryland,	Rockville,	Sports	346
	Korea	Minnesota	Moorhead		
	Australia	Gyeonggi-do	Yongin		
	Nigeria	Victoria	Melbourne		
	SaudiArabia	Kaduna	Rahama	Politics	164
	India	Ar Riyad	Riyadh		
	Indonesia	Uttar Pradesh	Noida	Nature	256
	Philippines	Jakarta Raya	Jakarta		
		National Capital Region	Mandaluyong City		

Table 6 Query based Retrieved Wikipedia Edit Streams

Query	Geolocation	Country	Cluster	Keyword	Wiki Edit Content
election	80.3.134.183	UK	politics	Voting, republic	Republic—Lithuania]] in 1940 shortly after the Soviet [[occupation of the Baltic states]]; where those who voted received stamps in their passport for voting
	59.89.253.208	India		Law, standard	In many of the countries with weak [[rule of law]], the most common reason why elections do not meet international standards
Geology	130.193.232.154	Iraq	Nature	Rocks, planet	During the [[Roman Empire—Roman]] period, [[Pliny the Elder]] wrote in detail of the many minerals and metals then in practical use – even correctly noting the origin of [[amber]].

ter data. The retrieved results are shown in figure 4 for different user given queries.

Fig. 3 Cluster wise Wikipedia Recent Edits

Various user queries are posed to the cluster data. It is required to retrieve the Wikipedia edit streams based on the user given query. The edit streams are available in the online MediaWiki platform from where the streams need to be retrieved for every query. Since the streams are clustered by retaining the stream meta data, the streams need to be retrieved based on these meta data. Article id

Fig. 4 User Query based Wikipedia Edit Stream Retrieval from Clusters on past data

Table 7 Query Processing Time at various time instances

Query	Query Posed Time	Query Processing Time for Geolocation based queries(in seconds)	Query Processing Time for Most recent edits (in seconds)
1. Last 1-hour edits	7:00 AM	0.4	0.3
	10:00 AM	1	0.8
2. Protest in various countries	12:00 PM	0.6	0.4
	2:00 PM	0.8	0.9
	4:00 PM	0.5	0.5
3. Election in different places	6:00 PM	0.8	1
	8:00 PM	0.9	0.6
	10:00 PM	0.5	0.5

meta data is the prominent information which helps to retrieve the edit stream from online. Wikipedia edit streams cannot be retrieved in a single shot just by passing the article id. It is required to append the article id with the Wikipedia URL. It is also performed in two stages in Wikipedia. These are article id checking and getting the article id related Wikipedia page. This further helps in reaching the edit location and retrieving results back to the users. Since these processes are involved, it is necessary to analyse the time taken for processing the user given query to obtain the Wikipedia edit streams. The time taken for processing the various queries is displayed in table 7. The retrieved edits based on the user given query with the article IDs available in the clusters and the obtained results are shown in table 8.

Table 8 Query based Wikipedia Edit Stream Retrieval using Wikipedia Article IDs in Clusters

Query	Article IDs	Article Title	Wiki edit stream related to Article IDs
Election in different places	932872308	Justin Madders	Before election to parliament he was an employment lawyer, and Leader of the Labour opposition on [[Cheshire West and Chester Council]] and leader of [[Ellesmere Port and Neston]] Borough Council.
	932964228	Jharkhand Legislative Assembly election	—[[Ramachandra Singh]]— [[Kushwaha Shashi Bhushan Mehta]]— [[Alok Kumar Chaurasiya]]— [[Pushpa Devi]]
Politics related recent edits	932513365	Eric Swalwell	''Eric Michael Swalwell'' (born November 16, 1980) is an American politician serving as the [[United States House of Representatives—U.S. Representative]] for [[California's 15th congressional district]] since 2013. A member of the [[Democratic Party (United States)—Democratic Party]], his district covers most of eastern [[Alameda County, California—Alameda County]] and part of central [[Contra Costa County, California—Contra Costa County]].
	933028453	Communist state	Communist states are typically administered by a single, [[Democratic centralism—centralised party apparatus]]. Although some provide the impression of multiple political parties, these are all solely in control by that centralised party.

4.2 Comparison with existing methods

The performance of the proposed Wikipedia edit retrieval system (WRS) has been compared with the already existing system WikiAutoCat [23]. The precision and recall have been measured and significant performance improvement has been observed. The comparative results are tabulated in table 9. The WikiAuto-Cat system performs categorization of Wikipedia articles but the system has been evaluated with the Wikipedia data repository. It has not been evaluated with the streams collected at different time periods. These issues have been tackled in the proposed system and the system has been evaluated with the stream data and achieved better precision and recall in data retrieval from the clusters. The proposed system has been checked against the stream data collected at various time periods. The accuracy has been increased when compared with the already existing Wikipedia categorization systems. The accuracy results are tabulated in table 10.

The precision of WRS is compared with the WikiAutoCat system. The precision is observed for four different datasets. The improved precision values are observed in each dataset. Since the already existing WikiAutoCat system performs the retrieval based on the title and body of the wikipedia article contents

Table 9 Query based Wikipedia Edit Stream Retrieval using Wikipedia Article IDs in Clusters

System	Precision	Recall
Dataset 1		
WikiAutoCat	0.189	0.859
WRS	0.562	0.864
Dataset 2		
WikiAutoCat	0.145	0.671
WRS	0.482	0.685
Dataset 3		
WikiAutoCat	0.137	0.527
WRS	0.576	0.642
Dataset 4		
WikiAutoCat	0.076	0.304
WRS	0.458	0.537

Table 10 Query based Wikipedia Edit Stream Retrieval using Wikipedia Article IDs in Clusters

System	Dataset 1	Dataset 2	Dataset 3	Dataset4
WikiAutoCat	0.981	0.971	0.988	0.992
WRS	0.993	0.984	0.991	0.994

alone, all the relevant articles are not retrieved effectively. But the proposed system WRS uses hierarchical clustering, the number of relevant articles retrieved is high. Thus, it is evident from the graph illustrated in figure 5 that WRS achieves higher precision than WikiAutoCat.

The recall value of WRS is compared with the WikiAutoCat. The recall values are observed for four different datasets. In each dataset, the higher recall is observed for WRS than WikiAutoCat. The higher number of relevant wikipedia article edits are available in the retrieved edits. The recall of WRS and WikiAutoCat is plotted and it is illustrated in figure 6.

Since the proposed system uses effective hierarchical stream clustering, the accuracy of the proposed system is high when compared to the WikiAutoCat. The accuracy of WRS and WikiAutoCat for four different datasets is illustrated in figure 7.

It is required to analyse the clustering time of the system. The number of Wikipedia article edits is cluster under various topics. The time taken for clustering the article edits vary with the number of article edits. The time analysis is shown in the graph figure 8. It is evident from the graph that the proposed Wikipedia edit categorization and retrieval system takes less time when compared to the WikiAutoCat categorization system in the literature. The query time to retrieve the meta data and article-id from the clusters has been evaluated with various matching methods [25] as shown in figure 9. The query time dramatically reduces in hierarchical method (1045 ms) when compared to exhaustive search (5640 ms) and other methods. The query processing time is observed under vari-

Fig. 5 Precision of WikiAutoCat Vs WRS

Fig. 6 Recall of WikiAutoCat Vs WRS

ous time instances for different queries and it is shown in figure 10.

Fig. 7 Accuracy of WikiAutoCat Vs WRS

Fig. 8 Wikipedia Edit Clustering Time analysis with Existing System

5 Conclusion

In this paper, Hierarchical clustering-based Wikipedia edit stream retrieval system is proposed. The re-cent Wikipedia edit streams are observed and domain specific similar keywords are extracted from edit streams using various domain corpuses. The edit streams are clustered under many major and sub categories. The clusters

Fig. 9 Query time to Retrieve Metadata from Clusters using various methods

Fig. 10 Query Processing Time for Wikipedia Edit Stream Retrieval from Online

are made to keep only the meta data of edit streams. Further, the user queries are evaluated and the relevant edit streams are retrieved. The stream clustering has helped to answer the user queries to re-trieve the recent edits. It helps to generate the clusters immediately using the arriving streams. The user queries are appropriately answered and the recent edit streams are effectively retrieved using the cluster data. The WRS system has been compared with the state-of-the-

art Wikipedia clustering system WikiAutoCat and the proposed systems achieved reduced clustering time and improved precision, recall and accuracy results.

6 Funding

Not Applicable.

7 Conflicts of interest

We, the authors of this research paper does not have any conflict of interest to publish in this journal.

8 Availability of data material

The data has been collected from wikipedia using API. We have not stores these data as it is not possible to store the streams. The experiments have been carried out by implement the retrieval system in C sharp and .NET environment as a web application. If required, kindly contact authors to get code repository link.

9 Code Availability

The code has been uploaded in an online repository. If required, kindly contact authors to get code repository link.

10 Author Contributions

All authors contributed to the implementation and paper writing of this work. Material preparation, data collection and analysis were performed by Mr.M. Arun Manicka Raja and Dr.S.Swamynathan.The first draft of the manuscript was written by Mr.M.Arun Manicka Raja and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. <https://meta.wikimedia.org/wiki/WikimedainfiguresWikipedia>
2. Martin-Carreras, T., Kahn, C.E., "Integrating Wikipedia Articles and Images into an Information Resource for Radiology Patients", Journal of Digital Imaging, vol.38,2018, pp.349-353.
3. Kanchana Saengthongpattana, Thepchai Supnithi, Nuanwan Soonthornphisaj, "Ontology-Based Classifiers for Wikipedia Article Quality Classification", The Joint International Symposium on Artificial Intelligence and Natural Language Processing, vol.807, 2017, pp. 68-81.
4. Julian Szymanski, Categorization of Wikipedia Articles with Spectral Clustering, International Conference on Intelligent Data Engineering and Automated Learning, springer, LNCS, vol.6936, 2011, pp. 108-115.

5. Zheng Z., Zhu X. (2012) Summarizing Definition from Wikipedia Articles, *Natural Language Processing and Chinese Computing*, vol.333, 2012, pp.260-271.
- O'Regan G. (2018) Wikipedia. In: *The Innovation in Computing Companion*. Springer, Cham
6. Gandica Y., Lambiotte R., Carletti T., dos Aidos F.S., Carvalho J., "Circadian Patterns on Wikipedia Edits", *Complex networks, Studies in computational intelligence*, springer, vol.644, 2016, pp. 1-22.
7. Karyak J.G., Sisakht F.Y., Abbasi S., Retracted: Clustering of Wikipedia Texts Based on Keywords, *International conference on Computational Science and Its Applications, LNCS*, vol.9790, 2016, pp.513-529.
8. Nguyen, H., Woon, Y. Ng, W. A survey on data stream clustering and classification. *Knowledge information systems*, vol. 45, 2015, pp. 535–569.
9. Georgescu M., Kanhabua N., Krause D., Nejd W., Siersdorfer S., Extracting Event-Related Information from Article Updates in Wikipedia, *Advances in Information Retrieval*, vol. 7814, 2013, pp. 254-266.
10. Thomas steiner, Seth van hooland, Ed summers, MJ no more: Using Concurrent Wikipedia Edit Spikes with Social Network Plausibility Checks for Breaking News Detection, *International world wide web conference*, 2013.
11. Barddal J.P., Gomes H.M., Enembreck F. (2015) A Complex Network-Based Anytime Data Stream Clustering Algorithm, *Neural Information Processing, LNCS*, vol. 9489, 2015, pp.615-622.
12. Yin, C., Zhang, S., Yin, Z. et al. Anomaly detection model based on data stream clustering. *Cluster Comput* 22, 1729–1738 (2019).
13. Olariu A. (2013) Hierarchical Clustering in Improving Microblog Stream Summarization, *Computational Linguistics and Intelligent Text Processing, LNCS*, vol.7817, 2013, pp.424-435.
14. Kavitha V., Hemashree P., Dilip H., Elakkiyarasi K. (2020) Clustering Process with Time Series Data Stream, *Smart Intelligent Computing and Applications, Smart Innovation, Systems and Technologies*, vol 159, pp.335-343.
15. Teffer, D., Srinivasan, R. Ghosh, J. AdaHash: hashing-based scalable, adaptive hierarchical clustering of streaming data on Mapreduce frameworks. *Int J Data Sci Anal* 8, 257–267 (2019).
16. Rebecca Willis. "Taming the Climate? Corpus analysis of politicians' speech on climate change", *Environmental Politics Journal*, Taylor and Francis, vol.6, no.2, 2017 pp.212-23.
17. Hunt D., Harvey K., "Health Communication and Corpus Linguistics: Using Corpus Tools to Analyse Eating Disorder Discourse Online", *Corpora and Discourse Studie*, Palgrave Advances in Language and Linguistics (book), 2015, pp. 134-154.
18. Angkana Tongpoon-Patanasorn, "Constructing an Academic Word List of Business English: A Corpus-Based Approach", *Humanities Social Sciences journal*, vol.34, no.2, 2017, pp.1-31.
19. <https://mailman.uib.no/public/corpora/2015-February/022099.html>
20. AliMozaffari, RaoufMoini, "Academic Words in Education Research Articles: A Corpus Study", *Procedia - Social and Behavioral Sciences*, Elsevier, vol.98, no.6, 2014, pp. 1290-1296.
21. <https://www.english-corpora.org/wiki/>
22. Chen, C., Zhang, J., Chen, X., Xiang, Y., Zhou, W.: 6 million spam tweets: a large ground truth for timely Twitter spam detection. In: *International Conference on Communications (ICC)*. IEEE, 2015.
23. Refaei, N., Hemayed, E.E. Mansour, R. WikiAutoCat: Information Retrieval System for Automatic Categorization of Wikipedia Articles. *Arab J Sci Eng* 43, 8095–8109 (2018).
24. Cong, Z., Fernandez, A., Billhardt, H. et al. Service discovery acceleration with hierarchical clustering. *Inf Syst Front* 17, 799–808 (2015).
25. Kiefer, C., Bernstein, A. (2008). The creation and evaluation of isparql strategies for matchmaking. In: *The Semantic Web: Research and Applications*, (pp. 463–477)

Figures

Figure 1

Wikipedia Edit Stream Retrieval System

Figure 2

Hierarchical Cluster Generation for Healthcare Domain

Wikipedia Recent Edits By category

Figure 3

Cluster wise Wikipedia Recent Edits

Figure 4

User Query based Wikipedia Edit Stream Retrieval from Clusters on past data

Figure 5

Figure 6

Recall of WikiAutoCat Vs WRS

Figure 7

Accuracy of WikiAutoCat Vs WRS

Figure 8

Wikipedia Edit Clustering Time analysis with Existing System

Figure 9

Query time to Retrieve Metadata from Clusters using various methods

Figure 10

Query Processing Time for Wikipedia Edit Stream Retrieval from Online