IMMM 2020

The Tenth International Conference on Advances in Information Mining and Management

ISBN: 978-1-61208-806-8

September 27th – October 1st, 2020

IMMM 2020 Editors

Dirk Labudde, Hochschule Mittweida, Germany

Michael Spreanger, Hochschule Mittweida, Germany

IMMM 2020

Foreword

The Tenth International Conference on Advances in Information Mining and Management (IMMM 2020), held between September 27 – October 1st, 2020, continued a series of academic and industrial events focusing on advances in all aspects related to information mining, management, and use.

The amount of information and its complexity makes it difficult for our society to take advantage of the distributed knowledge value. Knowledge, text, speech, picture, data, opinion, and other forms of information representation, as well as the large spectrum of different potential sources (sensors, bio, geographic, health, etc.) led to the development of special mining techniques, mechanisms support, applications and enabling tools. However, the variety of information semantics, the dynamic of information update and the rapid change in user needs are challenging aspects when gathering and analyzing information.

We take here the opportunity to warmly thank all the members of the IMMM 2020 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to IMMM 2020. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the IMMM 2020 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that IMMM 2020 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of information mining and management.

IMMM 2020 Chairs:

IMMM 2020 Publicity Chair
Mar Parra, Universitat Politecnica de Valencia, Spain

IMMM 2020

COMMITTEE

INFOCOMP 2020 Publicity Chair

Mar Parra, Universitat Politecnica de Valencia, Spain

INFOCOMP 2020 Technical Program Committee

Vicki H. Allan, Utah State University, USA
Ghada Almashaqbeh, NuCypher, USA
Daniel Andresen, Kansas State University, USA
Marc Baaden, CNRS, France
Raymond Bair, Argonne National Laboratory / University of Chicago, USA
Jacob Balma, Hewlett Packard Enterprise Company, USA
Bernhard Bandow, GWDG - Göttingen, Germany
Christine Bassem, Wellesley College, USA
Raoudha Ben Djemaa, MIRACL, Sfax, Tunisia
Abbas Bradai, University of Poitiers, France
Stephanie Brink, Lawrence Livermore National Laboratory, USA
Hans-Joachim Bungartz, Technische Universität München (TUM) - Garching, Germany
Paolo Burgio, University of Modena and Reggio Emilia, Italy
Xiao-Chuan Cai, University of Colorado Boulder, USA
Nicola Calabretta, Eindhoven University of Technology, Netherlands
Enrico Casella, University of Kentucky, USA
Jian Chang, Bournemouth University, UK
Jieyang Chen, Oak Ridge National Laboratory, USA
Albert M. K. Cheng, University of Houston, USA
Enrique Chirivella Pérez, Universitat de Valencia, Spain
Noelia Correia, Center for Electronics Opto-Electronics and Telecommunications (CEOT) | University of Algarve, Portugal
Kei Davis, Los Alamos National Laboratory, USA
Tiziano De Matteis, ETH Zurich, Switzerland
Daniele De Sensi, ETH Zurich, Switzerland
Iman Faraji, Nvidia Inc., Canada
Josué Feliu, Universitat Politècnica de València, Spain
Francesco Fraternali, University of California, San Diego, USA
Hans-Hermann Frese, Gesellschaft für Informatik e.V., Germany
Steffen Frey, Visualization Research Center - University of Stuttgart, Germany
Marco Furini, University of Modena and Reggio Emilia, Italy
Jason Ge, Snark AI Inc, USA
Alfred Geiger, T-Systems Solutions for Research GmbH, Germany
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Franca Giannini, IMATI-CNR, Italy
Barbara Guidi, University of Pisa, Italy
Önder Gürcan, CEA LIST, France

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Randomized Sampling Algorithm based on Triangle for Community Extraction in Graphs

Yanting Li
School of Information Science and Engineering
Shaoguan University
China
Email:yanting8015@sgu.edu.cn

Ying Huo
School of Information Science and Engineering
Shaoguan University
China
Email:huoying@sgu.edu.cn

***Abstract*—The techniques of sampling play a vital role in extracting communities from graphs. Most of sampling algoritms mainly take the advantage of the degree distribution or the weight of edge. However, it may lead to huge consumption of memory usage and computation time because of the complicated structure of graphs and the noise inherent of algorithm. We propose a novel randomized sampling algorithm, which is a triangle-based sampling algorithm. A random value R_v colors each node uniformly. The edge $e_{ij} = (v_i, v_j)$ is a monochromatic edge if node v_i and v_j receive the same random value R_v. The third edge of triangle Δ_T will be sampled if two edges of the triangle Δ_T are sampled. The triangle Δ_T that formed by three monochromatic edges is considered as the smallest sampling unit in graph G. An extracted community contains at least one triangle. Overviewing, experimental results demonstrate that the proposed algorithm extracts sufficiently dense subgraphs and significantly reduces computation time compared to the reservior sampling algorithm and graph priority sampling algorithm.**

Keywords–Triangle; Monocchromatic edge; Dense subgraph; Pattern extraction; Randomized sampling.

I. INTRODUCTION

The triangle, as the smallest dense subgraph, has gained more and more attention in studying graphs. Triangle is one of the basic shapes of complete subgraphs. Either directed edge or undirected edge, a triangle is the shortest complete cycle and the smallest non-trivial clique. Triangle is applicable to various measurements of network analysis, such as clustering coefficient, transitivity analysis and triangular connectivity. Both transitivity coefficient and clustering coefficient are widely used metrics in the study of social network analysis. Therefore, triangle has emerged as a crucial building block of graphs, thematic structure identification of graphs, node clustering and link classification, etc.

A streaming algorithm has been proposed for counting and sampling triangles from a given graph G [5]. It is one-pass streaming algorithm based on neighborhood sampling. Initially, an edge e_{ij} in the graph stream [6] [31] is randomly extracted. The edge e_{jk} that share the same node n_j with the edge e_{ij} is then extracted. The edge e_{jk} is extracted by employing the neighborhood relationship of edge e_{ij}. By extracting the edges e_{ij} and e_{jk}, the nodes n_i, n_j and n_k can be considered as a potential triangle $T_{(ijk)}$ if the node n_k is the common neighbor node of n_i and n_j. Besides, for studying the characteristic of graphs, a novel notion of dense subgraph named H^*-graph is proposed in [17]. The H^*-graph represents the core of graphs, and extends to encompass the neighborhood among all nodes of the core. Accordingly, an external-memory algorithm for maximal clique enumeration (ExtMCE) is also proposed by employing the H^*-graph for memory usage bounding. Such that, the degree rank of all h-nodes in graph can be computed. The memory usage can be estimated and controlled by using the ExtMCE algorithm. Furthermore, the core of a graph is represented by the subgraph with maximal order of the graph [41], such as k-core of a graph is a core of order k, where $deg_{(v)} \geq k$. The core number of node v is the highest order of the core within maximum value of k [4]. A massive graph is partitioned into a set of subgraphs with smaller size by employing the decomposition algorithm. The characteristic of core graph can be captured after the pruning of sparse components.

To study the characteristic of entire graph, however, becomes expensive due to the increasingly huge size of graph with complicated structure. Numerous sampling techniques have been proposed for extracting essential portions with significant characteristic of graphs. Graph sampling addresses the issue of seeking dense subgraphs, which represent similar properties to the original graph [13] [19] [32] [34]. The reliability of graph sampling techniques is validated by how closely the combinatorial properties of subgraphs simulate the original graph [14]. The interaction network of all proteins that confined to the mitochondria is a real world example. The protein-protein interaction network may not represent the entire network, but can reveal valuable insight into communication or biological process within a defined sphere. Hence, extracting partial of the interaction network of all proteins as the investigation samples is an efficient way of focusing on the sampling property of the network [30].

We propose a novel randomized sampling algorithm, which is based on node coloring for extracting functional unit denoted as Δ_T. A triangle $\Delta_T = \{v_i, v_j, v_k\}$ is considered as the smallest sampling unit, which constitute the community of G. Initially, a random value R_v is uniformly given to each node of a triangle Δ_T in graph G. The edge $e_{ij} = (v_i, v_j)$ is monochromatic if both its two endpoints v_i and v_j receive the same random value R_v. A random value R_v is considered as a color to a node. The third edge is sampled if two edges of a traingle Δ_T are sampled. With the set of monochromatic edges is sampled, all triangles formed by the set of monochromatic

 ISBN: 978-1-61208-806-8

edges can be extracted. Considering triangle as the smallest sampling unit is to prune the search by seeking a few number of densely interconnected communities, which best represent the frequently occurred characteristic of graph G. The random value R_v is a positive value, where $0 \leq R_v \leq |n|$. The $|n|$ is the total number of nodes in G without known beforehand. The random value R_v is unique and unrepeatable for every node. We assume that the range of random values has finite expectations and variances mathematically. We gain the generation of R_v as follow:

$$X_n + 1 = (\frac{{X_n}^2}{10^s})(mod 10^{2s})$$

$$R_v + 1 = \frac{X_n + 1}{10^{2s}}$$

where $(X_n + 1)$ is an iterative operator, and $(R_v + 1)$ is the random value R_v that needs to be generated every time. The s is the shifting of X_n square metre for generating new R_v.

The proposed algorithm aims at extracting communities from graphs. Each edge in graph G is selected based on the uniformly coloring of nodes with probability denoted by P_r, where $0 < P_r < 1$. The color is randomly given to each node, which is a real integer number denoted by R_v. A triangle Δ_T is the smallest sampling unit. A community of graph G contains at least one triangle Δ_T. Therefore, the more triangles anchor in an extracted community, the denser of the ommunity with avaliable insight characteristic of graph G.

This paper is organized as follow. Section I introduces the background of community extraction in graphs, and states the importance of community extraction in network analysis. Section II introduces various approaches of relevant researches. Section III describes the mainframe of randomized sampling algorithm in detail. The experimential results for verifying efficiency of the proposed method are concluded in Section IV. Section V concludes the whole paper.

II. Related Works

A simple and available technique named random sampling has been proposed [8] for scaling the massive graph G into small subgraphs and randomly select samples from the set of subgraphs of G. It is an unbiased sampling technique. Every individual sample is labelled with a random number. The individual sample is randomly extracted from the given matrix according to its labelled random number. The probability of each individual sample being selected at any stage is the same. A systematic sampling method involves the selection of individual samples from an ordered sampling matrix [11]. In this approach, progression through the sample list circles to the top once when the end of the sample list has passed. The sampling procedure begins from selecting an individual sample in the sample list randomly. Every k^{th} individual sample in the sample list is selected where k indicates the sampling interval value. Furthermore, a multistage sampling can be referred in [1] if the matrix data is too expensive to be sampled. This approach is a complicated form of cluster sampling [23]. The clusters are constructed in the given graph G at the first stage. The second stage is to decide the available individual sample in the cluster. All individual samples in the matrix data are appropriately listed. The Random Node Sampling algorithm is a node selection based algorithm [21] [38]. The sampling begins from a given distributed degree of node in a completely self-contained graph G. All nodes of the graph G are given with probability p. The degree distribution of node is denoted by P_k. Extrapolating from the subgraphs to the property of graph G if the randomly extracted subgraph samples have the same characteristic of probability distribution. Alternatively, The TIES algorithm [2] is the total induced edge sampling algorithm. The potential nodes are randomly selected with graph induction based on edge. The degree of node in G is computed. Then, the set of favor nodes with high degrees are selected. The edge e_{ij} is picked up from G at random. The two nodes v_j and v_j are added to the sampled node set in each iteration. The algorithm stops adding nodes to the sampled node set if the fraction $\varnothing$ of nodes are collected. The graph induction process begins once all edges of graph G are traversed. Once all edges that connecting the nodes in the sampled node set, the induced graph is formed. The induced graph holds similar characteristic and structure to the original graph G. A similar sampling approach based on the degree distribution for random graphs is proposed in [38]. The probability of selecting a node is identical for all nodes where $p_i = p$ for all i are considered initial case. The number of connections influences the probability of sampling a node with certain degree is a further sampling scheme for uncorrelated graphs. Therefore, the connectivity of a node denpends on the degrees of its neighbor nodes in the same subgraph. Minne Li, Dongsheng Li and Siqi Shen et al. propose the *DSS algorithm* [27]. All sampling processes is parallelly executed by calculating the exact size of subsample in each partitioning.

The random walk technique is frequently applied onto crawling websites for extracting useful data from the web. The proposal by Bowen and Steve et al. [43] addresses the issue of collecting samples from a graph by adopting random walk. This method achieves the reconstruction of a priori unknown graph. Besides, random walk is a random process technique, in which models the traverse path of a graph by mathematical space. Hence, it is appliable to graph sampling algorithm [7] [35]. An m-dimensional random walk sampling algorithm named Frontier Sampling has been proposed in [7]. The algorithm begins from a set of selected nodes in which preserve the crucial characteristic of regular random walk technique. All nodes of graph G is visited with proportional probability to their degrees. The joint distribution of frontier sampling is similar to the uniform sampling method [40]. Other two graph sampling algorithms, the Rejection-controlled Metropolis-Hashings Algorithm and the Generalized Maximum-degree random walk algorithm, has been proposed in [35]. The Rejection-controlled Metropolis-Hashing Algorithm is a modified Metropolis-Hashing algorithm [40] [44] in which parameterized with an acceptance function α where $0 \leq \alpha \leq 1$. The modified acceptance function improves the acceptance ratio of the original algorithm. Initially, the algorithm begins sampling from a root node v. It stops sampling if the condition of the node v is not satified. Otherwise, node v is selected from its neighbor node w at random. Then, a uniform random value q is generated where $q \in [0, 1]$. The neighbor node w is selected if the uniform random value $q \leq (\frac{d_v}{d_w})^\alpha$. Likewise, the procedure iterates till

the last node with satified parameter in graph G is sampled. Similarly, the generic sampling framework [3] is also based on Metropolis-Hastings algorithm. The notion of the generic sampling is to sample the interesting subgraph patterns without enumerating the entire set of candidate frequent patterns. All candidate patterns form a partial order based on the subgraph relationship. Then, the subgraph samples are returned when the partial order converges to a desired stationary distribution. The function of interestingness of the subgraphs in the sample space determines the stationary distribution selection. The output space sampling is scalable and parallelizable. Besides, the Generalized Maximum-degree random walk algorithm is proposed for unbiased graph sampling. A controlled paramater C is applied onto the original maximum degree algorithm [9] [45] in which C is a nonnegative integer. Similar to the original maximum degree algorithm, the Generalized Maximum-degree random walk algorithm adds $(C - d_v)$ self-loops onto node v of graph G if $d_v < C$ where d_v indicates the degree of node v. Therefore, the walk of GMD algorithm equals to the traditional random walk if $d_v \geq C$. Or the next node that chosen by the GMD walk with probability $\frac{1}{C}$ is the neighbor node of node v. Such that, the round of sampling iteration can be reduced. The graph sampling by random walk begins from a given node and randomly follow the out-connection of the given node [26]. This sampling technique is biased towards the sub-structure of nodes with high connectivity occurrence in a graph. Community extraction is another important method in studying graphs [18] [25]. The partial clustering of nodes in G is computed based on recognizing matrix column similarity. According to the distribution characteristic of graph data, an approach of Horvitz-Thompson estimation to T-stage snowball sampling is proposed by L. C. Zhang and M. Patone [24].

Succinct representation of community in graph is one of the most important techniques in our study. The sampling approach of random multiple snowball with cohen is proposed in [33]. A node is randomly chosen as seed. The neighbor nodes of a root node is selected with the same probability P_c. The process iterates until the set of desirable number of nodes are sampled. Community plays a crucial role in characterizing large-scale complex graphs. A link-tracing sampling algorithm consists of two steps: the set of nodes with shortest path to the set of root nodes is sampled by approximating personalized PageRank vectors, and connect to unvisited neighbor nodes in a new community based on PageRank vectors [28]. The framework of biologicalrelationships are represented by different graph layers It is expected to retain as much information as possible. Didier, Brun and Baudot et al. propose multiplex-modularity approaches to detect communities from multiple graphs [12]. It achieves the recovery of communities more accurately annotated than aggregated counterparts.

III. Randomized Sampling Algorithm

Triangle plays a vital role in both clustering coefficient and transitivity analysis. A triangle consists of three fully connected nodes $\{v_1, v_2, v_3\}$. Either directed edge or undirected edge, a triangle Δ_T can be described as:

$$\Delta_{T_{123}} = \{(v_1, v_2), (v_2, v_3), (v_1, v_3)\}$$

The triangle listing algoirthm is for counting the total number of triangles in a given graph G [37]. The input graph G is iteratively partitioned into a set of subgraphs and stored in the main memory. All triangles in each local subgraph are listed. It is an I/O-efficient in-memory algorithm for extracting subgraphs from G by using a mainframe with limited memory. To list all triangles in a graph could be a huge consumption of memory space. Thus, a decomposition algorithm named truss decomposition [22] is proposed. Similar with k-core decomposition algorithm, truss decomposition algorithm partitions a graph into subgraphs hierarchically. The *k-truss* is defined as the cores of graph G in which every edge of a core must be contained in at least $(k-2)$ triangles. The *k-truss* strengthens each edge in a core by at least $(k-2)$ strong ties. Other notions of subgraphs, such as *k-plex* [36] and *quasi-clique* [20], are proposed for analyzing the social networks. The *k-plex* looses the degree of every node in a clique of c nodes from $(c-1)$ to $(c-k)$. The *quasi-clique* can be considered as a relaxation on the density [16] or the degree [20]. However, the computation of all the above cohesive subgraphs is NP-hard for it could be scattered all over the entire graph, or may overlap largely with eath other. In addition to node-based sampling approaches, the edge-based wedge sampling applies onto estimating the number of triangles in graphs [10].

The approach we proposed considers triangle as the samllest sampling unit due to the full connection among three nodes of a triangle. Contrarily, differ from node-based sampling algorithm [21] and edge based sampling algorithm, the density of extracted subgraph by adopting triangular structure can be moderate. The highly dense components represent the core of the graph G. The key idea of the *Randomized Sampling algorithm* is to color all nodes for extracting a set of monochromatic edges from a graph G. The third edge will be sampled if other two edges of a triangle Δ_T are sampled. A triangle with three monochromatic edges is the smallest sample unit in graph G. The extracted communities of graph G must contain at least one triangle Δ_T. We apply the proposed approach onto social network analysis, as well as the technique of ease that output a sample based on its "closeness" to the original sample [39], or the security system based on biometrics for fingerprint recognition [15].

The randomized sampling algorithm considers the unweighted graph. Given a graph G, the node set and edge set are denoted by V_G and E_G of G respectively. We define $n = |V_G|$ as the number of nodes, and $m = |E_G|$ as the number of edges of G. The size of G is denoted by $|G|$ where $|G| = m + n$. The set of neighbor nodes of node v is denoted by N_v, that is, $N_v = \{u : (u, v) \in E_G\}$.

A triangle Δ_T is a cycle of length 3. Let $\{u, v, w\} \in V_G$ be the three nodes of the cycle. The set of triangles Δ_T is denoted by Δ_G, such that $\Delta_T \in \Delta_G$.

All nodes in graph G will be visited once for sampling monochromatic edges. The searching begins from the root node a in G where $a \in V_G$ as shown in the Fig. 1. The neighbor nodes $\{b,\ c,\ f,\ g\}$ of node a are explored and marked with 1 as visited nodes. The set of nodes $\{a,\ b,\ c,\ f,\ g\}$, however, no triangle anchors inside. Then, the searching continues to seek the neighbor nodes of node b, the set of nodes $\{a,\ h,\ i\}$. The triangle $\Delta_{T_{bhi}}$ is identified for the two nodes $\{h, i\} \in e_{hi}$. The searching strategy is accomplished

by enqueueing each level of the graph G sequentially as the breadth-first search. All neighbor nodes at the present breadth are explored prior to move onto other nodes at next breadth level. The searching stops till the last node in G is visited. Three triangles, $\Delta_{T_{bhi}}$, $\Delta_{T_{cde}}$ and $\Delta_{T_{def}}$ are identified in G. Figure 1 illustrates the path of searching triangles in G.

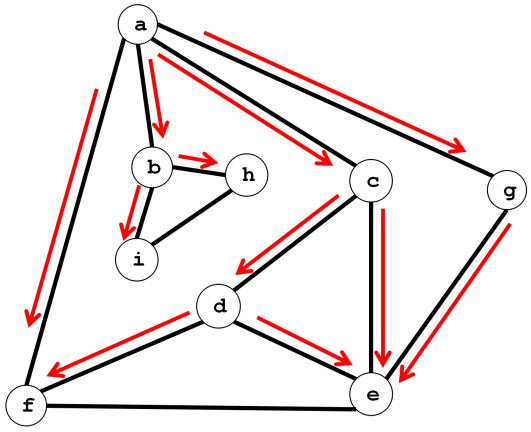

Figure 1: The path of triangle search based on Breadth-First Search

An adjcent list is mapped in Table I.

TABLE I: Adjacent List of nodes in G

node	neighbor nodes
a	b, c, f, g
b	a, h, i
c	a, d, e
d	c, e, f
e	c, d, g
f	a, d, e
g	a, e
h	b, i
i	b, h

Every node is colored once visited. Each node receives a random value denoted by R_v where $0 < R_v < |n|$.

Definition 1 Coloring: The coloring of a node $v \in V_G$, denoted by $cr(v, G)$, is defined as $\{cr(v, G) : R_v$ is uniformly given to each node v, where $0 < R_v < |n|\}$.

Definition 2 Monochromatic edge: $MONO_{e_{uv}}$ denotes the monochromatic edge e_{uv} if $R_u = R_v$. All monochromatic edges are sampled from E_G when all nodes in V_G are colored.

$$MONO_{e_{uv}} = \begin{cases} 1, & \{e_{uv} : R_u = R_v\} \\ 0, & Otherwise \end{cases} \quad (1)$$

Given the graph G=(V_G, E_G), where V_G=$\{a,\ b,\ c,\ d,\ e,\ f,\ g,\ h,\ i\}$. The set of neighbor nodes $\{b, c, f, g\}$ of node a are colored while they are explored. Node $\{a, c\}$ receive red color. Node $\{b, f, g\}$ recieve green color. The $MONO_{e_{ac}}$ can be sampled at the *Stage Two*. At the *Stage Three*, the neighbor nodes of node b, the node $\{h,\ i\}$ receive the green color, the same as node b. Three monochromatic edges, $MONO_{e_{bh}}$, $MONO_{e_{bi}}$ and $MONO_{e_{hi}}$ can be sampled. The iteration process stops till all monochromatic edges are sampled from E_G. Communities containing the $\Delta_{T_{bhi}}$ and $\Delta_{T_{cde}}$ can be extracted from G. Figure 2 illustrates the procedure of node coloring and monochromatic edge sampling.

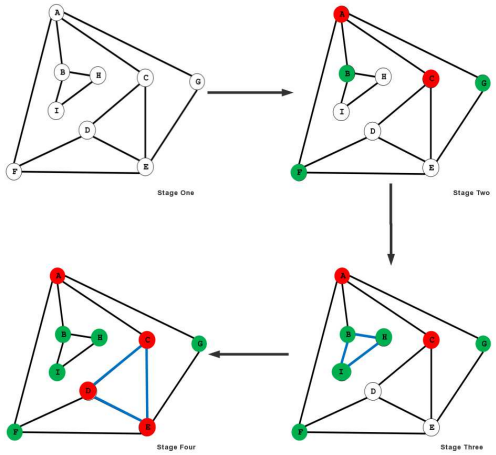

Figure 2: The Sampling of Monochromatic Edges in G

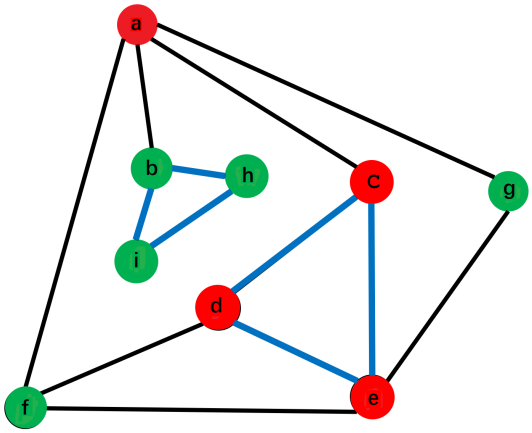

Figure 3: The Extraction of Communities in G

Figure 3 illustrates the extraction of frequent patterns from G. The $\Delta_{T_{def}}$ is not extracted from G as a pattern for e_{df} and e_{ef} are not monochromatic edges.

The randomized sampling algorithm is summarized in algorithm 1. Each node is colored with a random value R_v. Every edge is sampled with probability P_r where $0 \leq P_r \leq 1$.

A. Probability Analysis

1) Global Sampling
The probability of a triangle in G to be extracted as frequent pattern. An edge e_{jk} is monochromatic if its two endpoints j and k receive the same color where $R_j = R_k$. A tiangle that consists of three

Algorithm 1: *Randomized Sampling Algorithm*

- v := node $v \in V$
- e := edge $e \in E$
- R := random value for coloring node
- q := queue for Breadth-First traversal
- $MONO_e$:= the set of monochromatic edges $\in G$
- Δ_G := all triangles of graph G
- Δ_T := triangles contain the set of monochromatic edges

Input: $G = (V_G, E_G)$, R_V = 1,2,3,4....m
Output: a set of triangles Δ_T

```
begin
    init R_V, n, q.queue = ∅, Δ_T = ∅
    for i ∈ V_G do
        i.mark = 1
        q.enqueue(v_i)
        R_{V_i} is given to i
        if j is adjacent to i then
            v_i = q.dequeue()
            j.mark = 1
            q.enqueue(v_j)
            R_{V_j} is given to j
            if R_{V_i} = R_{V_j} then
                e_ij ∈ MONO_e
                e_ij is sampled
            end
        end
        if k is the common neighbor node of i and j
        then
            v_j = q.dequeue()
            k.mark = 1
            q.enqueue(v_k)
            R_{V_k} is given to k
        end
        Δ_{T_ijk} is identified
        for all triangles Δ_G in G do
            if R_{V_i} = R_{V_j} = R_{V_k} then
                e_ij, e_jk, e_ik ∈ MONO_e
            end
            Δ_{T_ijk} is extracted
        end
    end
    return a set of triangles Δ_T
end
```

monochromatic edges $\{e_{ij}, e_{jk}, e_{ik}\} \in MONO_e$ is extracted as a community of G. Then, the triangle $\Delta_{T_{ijk}} \in \Delta_T$. Thus, the probability of a triangle Δ_T to be extracted from G as a pattern is concluded as below.

$$P_r(\Delta_T) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

With the increasing numbers of colors which indicated by random values R_v, the probability P_r for every edge to be sampled as a monochromatic edge decreases.

$$P_r = \frac{1}{R_v}$$

2) Local Sampling

The probability of an edge $e \in E_G$ being sampled as a monochromatic edge. For every two nodes j and k that connected by an edge $e_{jk} \in E_G$ receive the same color. Such that, the edge e_{jk} is monochromatic. Then, the $P_r(jk) = P_r(j) \times P_r(k)$. Hence

$$e^2_{(V_G)} = \frac{V_G}{2(k-2)} = \frac{k(k-1)}{2}$$

Then, we obtain the probability P_r for any two nodes $\{j, k\} \in e_{jk}$ where

$$P_r(jk) = e^2_{(V_G)} \times \frac{1}{(V_G)^2} = \frac{(V_G)^2 - V_G}{2} \times \frac{1}{(V_G)^2}$$

B. Complexity Analysis

G is the given graph, let Ψ be the set of candidate frequent patterns contain at least one Δ_T with three monochromatic edges. For $0 \leq R_v \leq |n|$, every node in V is visited once with given a random value R_v. The algorithm initially require $O(|n|scan|G|)$ when giving random value R_v to each node. If the set of $\Delta_T \in \Psi$ of G is extracted, the process requires $O(|m|scan(|\Psi|)) = O(scan(|G|))$ I/O. A monochromatic edge will be removed from Ψ if it does not contained in any triangles. The worst case of the complexity, however, we simply employ an approach with lowest support and extract triangles one by one, which requires the worst case complexity of $O(|n| + |m|scan|G|)$. For the computation of the number of triangles with three monochromatic edges, we at most enumerate all triangles in the corresponding Ψ in which gives $O(\sum_{(\Delta_T \in \Psi)} |m|)$ complexity. Comparatively, the reservior sampling algorithm is a sampling approach that ensure the probability of each element being sampled is the same. The numbers of elements are unknown beforehand. The time complexity of reservior sampling algorithm is $O(n)$ if the total number of being sampled elements is relatively small. However, with increasingly large of the total number of both the elements in the reservior and the set of samples, the time complexity can be up to $O(n*(log(n)-log(n-m)))$. The sampling process uses constant space wiht $O(n*(1+log(m/n)))$ time complexity [42]. The *randomized Sampling algorithm* requires less time complexity than the reservior sampling algorithm when the given graph data is large.

C. Implementation

A $2 * n$ array list is employed for building the storage of all nodes in V_G and their corresponding random values R_v as Figure 4. A continuous storage space must be allocated statically or dynamically. Pointer is set to manage all elements. One is the header pointer in which pointing at the header element. Another is the end-of-list pointer in which pointing at the storage location of the next entry element. The number of elements in the array is constantly changing. Thus, the storage space that occupied by the array moves in the continuous space allocated for the node list. The set of neighbor nodes of node V_a is traversed, but without any anchored triangles. The node V_a and all of its neighbor nodes are removed out of the array for releasing memory space. The set of neighbor nodes of nodes V_b, V_c and V_d contains three triangles so that the set of neighbor nodes is saved to a cluster for storing triangles of G. Other neighbor nodes of nodes V_b, V_c and V_d are visited as

 ISBN: 978-1-61208-806-8

new beginning of traverse process. There is a cut of nodes if it connects to a null in order to prevent from self-loop exploring.

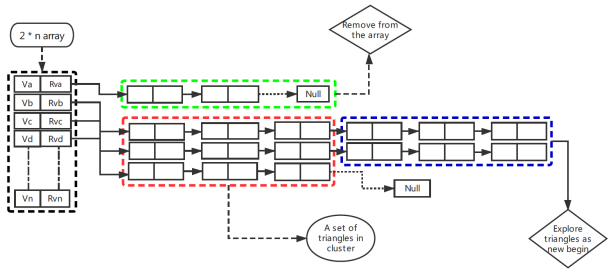

Figure 4: Implememtation of Randomized Sampling Algorithm

IV. Experiment and Evaluation

The performance test of the proposed algorithm is verified via a succession of experiments with the Intel i7 2.3GHz CPU, 16GB RAM, and the version 4.1.2 of C compiler in Mac OS 10.8.3 operating system as the experimential environment.

A. Characteristics of Datasets

Two real world datasets in different sizes are used. The *web-google* dataset is a directed graph, and the *com-LiveJournal* is an undirected graph. The proposed approach fits in processing both directed graph and undirected graph due to the characteristic of triangular structure. Table II shows the features of the two datasets.

TABLE II: FEATURES OF DATASETS

Dataset Statistics	web-Google	com-LiveJournal
$\|V\|$	875,713	3,997,962
$\|E\|$	5,105,039	34,681,189
Clustering Coefficient	0.5143	0.2843
Number of Triangles	13,391,903	177,820,130
Diameter	21	17
Type	directed	undirected

The *web-Google* released by Google is a part of *Google Programming Contest* source in 2002. The *com-LiveJournal* is a free online blogging community. The *com-LiveJournal* dataset provides friendship social network and ground-truth communities. *com-LiveJournal* can create groups when collecting community information according to different features, such as cultural backgroup, lifestyle, technology, entertainment preferences, etc. A community detection technique [18] is inspired by the matrix blocking problem. It is based on the connectivity occurence among all nodes in G. The similarities between a pair of columns in the adjacency matrix is exploited. Two columns in the same block should be more similar than two columns in different block if the patterns are non-zero. A cluster of nodes represents a dense subgraph in G.

B. The Performance of Randomized Sampling Algorithm

The *x-axis* of Figure 5 indicates the selected value of R_v, in which represents the numbers of color. We manully select

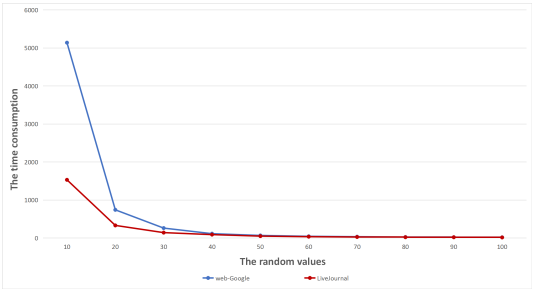

Figure 5: Time Cost of randomized Sampling Algorthm

the set of valuse $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ as random value for each round of communites extraction. The *y-axis* indicates the time consumption counted in second. Throughout the recorded experimental results by using both the datasets of *web-Google* and *com-LiveJournal* in Figure 5, the cost of computation time for communities extraction decreases with the increasing given value of R_v. The time consumption of communities extraction can be analyzed with following two aspects. All nodes are visited once whether each pair of nodes in G receives the same color or not. Therefore, the constant time consumption can be proved by using the Breadth-First Search algorithm for graph traversal. On the other hand, to determine the extracted numbers of communities in G, the process of triangle counting is required. The more counted triangles, the higher time cost, and vice versa.

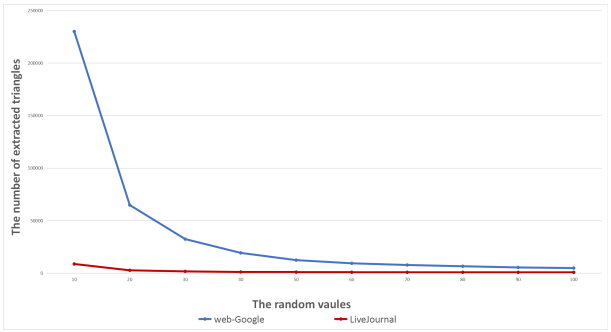

Figure 6: The Number of Extracted Triangles

The extracted triangles that contain the set of monochromatic edges are counted as shown in Figure 6. The *x-axis* indicates the value of R_v. The *y-axis* indicates the number of extracted triangles in G. The chosen values of R_v are the same as shown in Figure 5. Likewise, the numbers of extracted triangles in G decrease with the increasing value of R_v. For the probability of every edge $e \in E_G$ being sampled reduces. Therefore, the number of triangles with three monochromatic edges $e \in MONO_e$ decreases.

For verifying the effciency of the proposed algorithm, both the numbers and the sizes of communities within different values of R_v are examed. Communities in different sizes are extracted. The given graph G can be decomposed hierarchically. The experimental results recorded by Figure 7 proves the distribution results. The *x-axis* indicates the size of communities. The *y-axis* indicates the numbers of communities. The

 ISBN: 978-1-61208-806-8

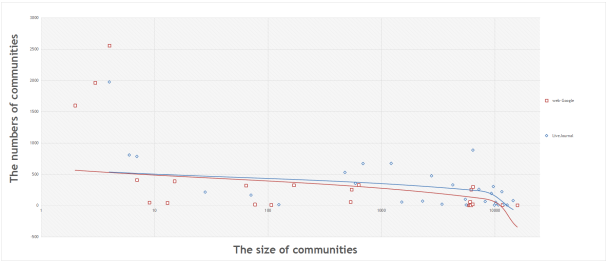

Figure 7: Distribution of Communities

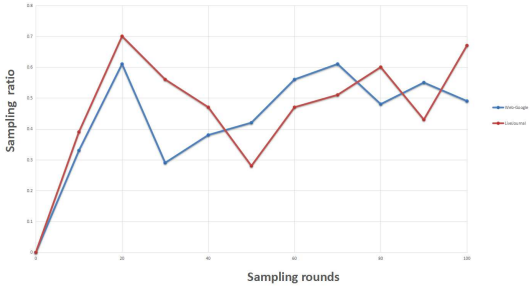

Figure 9: Statistics of Sampling Ratio

red graph shows the distribution of communities of the *web-Google*. The blue graph shows the distribution of communities of the *com-LiveJournal*. With relatively small value of R_v as the sampling threshold, less color for labelling all nodes $n \in N_G$. The probability of edge $e \in E_G$ to be monochromatic becomes higher. A few number of communities in large size are obtained. Contrarily, given relatively large value of R_v as the sampling threshold, more colors for labelling nodes $n \in N_G$. The probability of edge $e \in E_G$ to be monochromatic becomes smaller. Therefore, a large numbers of communities in small size are extracted from G.

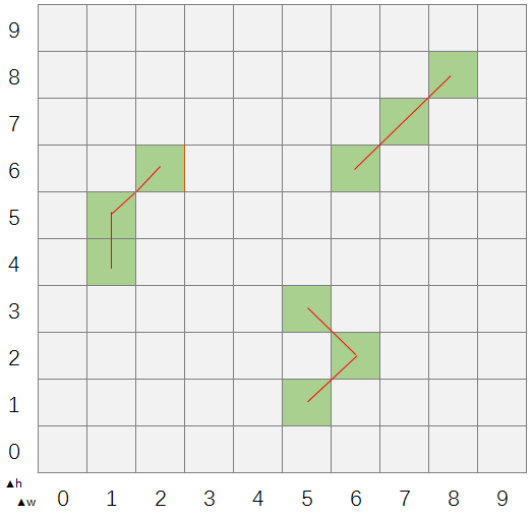

Figure 8: Location of Communities in a graph by using *2-Dimensional Grid*

To locate communities in G, a *2-Dimensional Grid* is proposed as shown in Figure 8. The red graph illustrates the path of community searching. The green blcoks indicate the location of communities in G. It can be computed as the formula below.

$$Grid[h_a][w_a] = \begin{pmatrix} \frac{y_i - y_0}{h_a} + 1 \\ \frac{x_i - x_0}{w_a} + 1 \end{pmatrix}$$

The notion of $\blacktriangle h_a$ and $\blacktriangle w_a$ indicate that community a locates at height of h and width of w respectively.

Figure 9 records the statistic of sampling ratio. The *x-axis* indicates the rounds of sampling process. The *y-axis* indicates the sampling ratio. The results of each sampling can be regarded as a random variable $Var_{(s)}$. Due to the unknown total number of communities in G, and the unpredictable number of extracted samples in the set of communities s, the n rounds sampling results $\{X_1, X_2, X_3, \ldots X_n\} \in s$ can be considered as a set of random variable $Var_{(s)}$. Let $\{X_1, X_2, X_3, \ldots X_n\} \in s$ are the samples selected from the population $G_{(fq)}$, then $\{G_{(fq)} : (X_1, X_2, X_3, \ldots X_n)\}$ is statistical quantity. The sample average can be computed as

$$\bar{Var}_{(s)} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Let the proportion of the amount of selected samples with certain attribute in the population $G_{(fq)}$ defined as the sampling ratio P_s. Then

$$P_s = \frac{s}{G_{(fg)}}$$

We employ the volume of the community for the computation of sampling ratio. Then

$$P_s = \frac{|V_s| + |E_s|}{|V_{G_{(fq)}}| + |E_{G_{(fq)}}|}$$

From the record of Figure 9, we gained remarkable results of sampling ratio both *web-Google* and *com-LiveJournal* at the 20^{th} round.

C. Comparison Experiments

In this experiment, we implement both the reservior sampling algorithm [42] and graph priority sampling algorithm [31] for comparing the performance of communities extraction.

TABLE III: MAXIMUM RUN TIME (IN SECOND)

Dataset	web-Google	com-LiveJournal
Randomized Sampling	5137.18	1529.02
Reservior Sampling	9255.6	6631.07
Graph Priority Sampling	2708.3	3889.032

Table III records the maximum time consumption of sampling. The *Randomized Sampling* is faster than the *Reservior Sampling* in processing both two datasets for the *Randomized Sampling* traverses entire graph G once. Contrarily, the *Reservior Sampling* needs to visit every node in G twice for the

 ISBN: 978-1-61208-806-8

in-degree and out-degree of each node are computed. However, the *Graph Priority Sampling* costs less computation time in extracting communities from web-Google. For *Graph Priority Sampling* (GPS for short) separates the function of edge sampling and sample estimation. The separation of estimation and sampling significantly save resource.

TABLE IV: MAXIMUM NUMBER OF COMMUNITIES

Dataset	Randomized	Reservior	GPS
web-Google	230018	13941	133925
com-LiveJournal	8632	7039	7780

Table IV records the experimental results of maximum numbers of extracted communities by *Randomized Sampling*, *Reservior Sampling* and *Graph Priority Sampling*. The *Randomized Sampling* extractes more communities than the *Reservior Sampling* and the *Graph Priority Sampling* (GPS for short). For a triangle is considered as the samllest community by the *Randomized Sampling*, but a node or an edge cannot be considered as a cohesive subgraph.

TABLE V: DENSITY

Dataset	Randomized	Reservior	GPS
web-Google	0.92	0.85	0.836
com-LiveJournal	0.87	0.69	0.776

Besides, Table V records experimental results of maximum density of communities. We employ the formular below for computating the density of both undirected and directed graphs.

$$dens = \frac{|E_s|}{|V_G| + |E_G|}$$

The results recorded in the table V show that the density of extracted communities by the *Randomized Sampling* are higher than both the *Reservior Sampling* and *Graph Priority Sampling* (GPS for short). The experimential results of the proposed algorithm are competitive and significantly improved.

V. CONCLUSION

We proposed randomized sampling algorithm for extracting communities in graphs. This approach combined the benefits of edge sampling and triangle count to offer high precision of communities extraction. The performance of the randomized sampling algorithm was evaluated based on four measurements, including time consumption, number of triangles, sampling ratio and density of community. Moreover, the superiority of the proposed method was proved by experimental results of comparing with the reservior sampling algorithm and graph priority sampling algorithm. Throughout the experimential results and theoretically analysis, the proposed method was highly confident estimations, and up to ten times sampling size reduction over the state-of-the-art alternatives when the sampling was low.

For the future work, we will prove the analysis of error bound of the randomized sampling algorithm. A modified version of randomized sampling algorithm based on hierachical pruning technique will be proposed.

ACKNOWLEDGMENT

This research is supported by the research fund of Shaoguan University under Grant No.SY2017KJ06 and the Scientific Technology Project of Shaoguan 2019 under Grant No.2019sn063.

REFERENCES

[1] A. Adriana, Gili, J. Elke, Noellemeyer and M. Balzarini, "Hierarchical linear mixed models in multi-stage sampling soil studies," The Environmental and Ecological Statistics, vol.20(2), 2013, pp.237-252.

[2] Ahmed, Nesreen, J. Neville and R. R. Kompella, "Network Sampling via Edge-based Node Selection with Graph Induction," The Computer Science Technical Reports, 2011, paper 1747.

[3] AI Mohammad and M. J. Zaki, "Output Space Sampling for Graph Patterns," Proceedings of VLDB Endowment, vol.2(1), 2009, pp.730-741.

[4] A. Montresor, F. D. Pellegrini and D. Miorandi, "Distributed k-Core Decomposition," IEEE Transactions on Parallel and Distributed Systems, vol.24(2), 2011, pp.288-300.

[5] A. Pavan, K. Tangwongsan, S. Tirthapura and K. L. Wu, "Counting and sampling triangles from a graph stream," The proceedings of the VLDB Endowment, 2013, pp.1870-1881.

[6] A. Zakrzewska and D. A. Bader, "Streaming graph sampling with size restrictions," The IEEE/ACM International Conference, 2017, pp. 282-290.

[7] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," The proceedings of the 10^{th} ACM SIGCOMM conference on Internet measurement ACM, 2010, pp.390-403.

[8] D. David, Chapman and Alexandria, "Selecting Unrestricted and Simple Random with Replacement Samples Using base SAS and PROC SURVEYSELECT," 2012, The SAS Global Forum 2012.

[9] D. Stutzbach, R. Rejaie and N. Duffield, "On unbiased sampling for unstructured peer-to-peer networks," IEEE/ACM Transactions on Networking, vol.17(2), 2009, pp.377-390.

[10] D. Türkoğlu and A. Turk, "Edge-Based Wedge Sampling to Estimate Triangle Counts in Very Large Graphs," 2017, The proceedings of ICDM.

[11] E. Satu and Schaeffer, "Scalable Uniform Graph Sampling by Local Computation," SIAM J. Computer Science, vol.32(5), 2010, pp.2937-2963.

[12] G. Didier, C. Brun, Baudot and Anaïs, "Identifying communities from multiplex biological networks," PeerJ, vol.3(7307), 2015, pp.1525-1545.

[13] G. W. Flake, S. Lawrence and C. L. Giles, "Efficient identification of web communities," The proceediings of the KDD conference, 2000, pp.150-160.

[14] H. Matsuda, T. Ishihara and A. Hashimoto, "Classifying molecular sequences using linkage graph with their pairwise similarities," Theoritical Computer Science, vol.210(2), 1999, pp.305-325.

[15] Ismahane Cheheb, Noor Al-Maadeed, Somaya Al-Madeed, Ahmed Bouridane and Richard Jiang, "Random sampling for patch-based face recognition," The 5_{th} International Workshop on Biometrics and Forensics (IWBF), 2017, pp.1-5.

[16] J. Abello, M. G. C. Resende and S. Sudarsky, "Massive Quasi-Clique Detection," Latin American Symposium on Theoretical Informatics, 2002.

[17] J. Cheng, Y. Ke, A. W. C. Fu, J. X. Yu and L. Zhu, "Finding maximal cliques in massive networks," ACM Transactions on Database Systems, 2011, vol.36(4).

[18] J. Chen and Y. Saad, "Dense subgraph extraction with application to community detection," IEEE Transactions on Knowledge and Data Engineering, vol.24(7), 2012, pp.1216-1230.

[19] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," The proceedings of the SODA conference, 1998, pp.668-677.

[20] J. Pei, D. Jiang and A. Zhang, "On mining cross-graph quasi-cliques," The proceedings of the 11^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.

 ISBN: 978-1-61208-806-8

[21] Jure Leskovec and Christos Faloutsos, "Sampling From Large Graphs," The proceedings of KDD, 2006, pp.631-636.

[22] J. Wang and J. Cheng, "Truss Decomposition in Massive Networks," The proceedings of the VLDB Endowment, 2012, vol.5(9), pp.812-823.

[23] J. Zhang, Y. Pei, G. Fletcher and M. Pechenizkiy, "Evaluation of the sample clustering process on graphs," IEEE Transactions on Knowledge and Data Engineering, 2019, pp.(99):1-1.

[24] L. C. Zhang and M. Patone, "Graph sampling," 2017, METRON, Springer, vol.75(3), pp.277-299.

[25] L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas, "Comparing community structure identification," The Journal of Statistical Mechanics: Theory and Experiment, 2005, vol.09.

[26] M. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, "On near-uniform URL sampling," The proceedings of the 9^{th} International Conference on World Wide Web, 2000, pp.295-308.

[27] Minne Li, Dongsheng Li, Siqi Shen, Zhaoning Zhang and Xicheng Lu, "DSS: A Scalable and Efficient Stratified Sampling Algorithm for Large-Scale Datasets," Network and Parallel Computing. Springer International Publishing, 2019, pp.133-146.

[28] M. Salehi, H. R. Rabiee and A. Rajabi, "Sampling from complex networks with high community structures," The Journal of Chaos, 2012, vol.22(2), pp.2202-2229.

[29] M. P. H. Stumpf, "Sampling properties of random graphs: the degree distribution," The Journal of Physical review. E, Statistical, nonlinear, and soft matter physics, 2005, vol.72(3).

[30] N. Alon and M. Krivelevich, "Testing k-colorability," SIAM J. Discrete Math., vol.15(2), 2002, pp.211-227.

[31] N. K. Ahmed, N. Duffield, T. L. Willke and R. A. Rossi, "On sampling from massive graph streams," Very Large Databases, 2017, vol.10(11), pp.1430-1441.

[32] P.Hu and W.C.Lau, "A survey and taxonomy of graph sampling," arXiv: Social and Information Networks, 2013.

[33] Q. Gao, X. Ding, F. Pan and W. X. Li, "An improved sampling method of complex network," International Journal of Modern Physics C, 2014, vol.25(05).

[34] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Extracting large-scale knowledge bases from the web," The proceedings of the VLDB conference, 1999, pp.639-650.

[35] Rong-Hua Li, Jeffrey Xu Yu, Lu qin, Rui Mao and Tan Jin, "On random walk based graph sampling," The proceedings of IEEE 31^{th} International conference on Data Engineering, 2015, pp.927-938.

[36] S. B. Seidman and B. L. Foster, "A graph-theoretic generalization of the clique concept," Journal of Mathematical Sociology, vol.6(1), 1978, pp.139-154.

[37] S. Chu and J. Cheng, "Triangle listing in massive networks and its applications," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol.6, 2011, pp.672-680.

[38] Stumpf, C. Wiuf and R. M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," The proceedings of the National Academy of Sciences of the United States of America, vol.102(12), 2005, pp.4221-4224.

[39] S. Wang, M. Dash and L. T. Chia, "Efficient Sampling: Application to Image Data," Knowledge discovery and data mining, 2005, pp.452-463.

[40] T. Wang, Y. Chen, Z. Zhang and T. Xu, "Understanding Graph Sampling Algorithms for Social Network Analysis," The proceedings of International Conference on Distributed Computing Systems Workshops, 2011, pp.123-128.

[41] V. Batagelj and M. Zaversnik, "An $O(m)$ algorithm for cores decomposition of networks," Advances in data analysis and classification, vol.5(2), 2003, pp.129-145.

[42] Vitter and S. Jeffrey, "Random sampling with a reservoir," ACM Transactions on Mathematical Software, vol.11(1), 1985, pp.37-57.

[43] Y.Bowen, G.Steve, and H.Frank, "Identifying communities and key vertices by reconstructing networks from samples," PLoS ONE, 2013, vol.8(4), e61006-.

[44] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," The proceedings of International Conference on World Wide Web, 2006, pp.367-376.

[45] Z. Bar-Yossef, A. Berg and S. Chien, "Approximating Aggregate Queries about Web Pages via Random Walks," The proceedings of International Conference on VLDB, 2000, pp.535-544.

 ISBN: 978-1-61208-806-8

Towards Inter-Rater-Agreement-Learning

Kai-Jannis Hanke, Andy Ludwig, Dirk Labudde and Michael Spranger
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*name.surname*}@hs-mittweida.de

***Abstract*—While technological advances and improved algorithms enhance most scientific fields, there remains a simple problem in many domains. If a decision has to be made we resort to simple majority votes or utilize agreement measures to determine how unanimous a decision is. Especially in text classification, a text is usually sorted into a specific category based on how many people agreed on it. However, the problem is that in these methods the individual that made the decision is neglected. Therefore, we propose a weighted approach that includes a flexible feature space and adjustments to the weights not only according to the individual's expertise but also to their performance on previous tasks. Preliminary experiments with a data set including short music related texts yield promising results with fewer cases for which no majority vote was achieved.**

Keywords–Agreement Measures; Weighted Majority Vote; Text Labeling.

I. INTRODUCTION

Scientific inquiries require sound and accurate measurements to ensure valid and reliable test results. While machine learning and artificial intelligence are becoming more prevalent in different research applications, the final judgment or the required previous labeling of data mostly remains a group decision of trained or untrained individuals. Human judgment is often considered gold standard. May it be text annotation for subsequent text classification, diagnoses by a collective team of medical professionals [1] or generally phrased: the correct labeling or interpretation of data in any other domain. This gold standard is frequently acquired by utilizing a majority decision.

However, ordinary majority votes neglect the unique strengths and weaknesses of the individuals participating in a given labeling experiment. In the case of a complicated medical case, a trained medical expert with decades of experience should have more influence on the final judgment than a student on their first internship. Obviously, the expertise and experience of an individual should play a decisive role, especially when it comes to human life. Still, in less lethal decisions the competence of individuals is often not taken into account. Instead, each individual receives equal importance in the final judgment, regardless of their predisposition to the topic at hand. Decisions of individuals with a keen interest in artificial intelligence and machine learning should have more weight when labeling texts about aforementioned topics than a nutritionist. The expertise on a topic should always be taken into consideration if the desired outcome is a gold standard. Furthermore, if an individual has the desire to perform well on a given task then this should induce a positive weight, whereas the goal of finishing as quickly as possible should lead to decreased influence in a majority vote. Additionally, internal consistency or the certainty with which individuals perform may also provide insight into the value of their contribution. If presented with the same task at different time points, then ideally the answers should be the same. If however there is a discrepancy, i. e,. if the individual makes a different judgment when repeatedly coming across the same question, then their overall value for the given assignment diminishes. In order to receive an ideal gold standard, these problems need to be addressed and implemented.

Hence, in this paper we propose a weighted majority vote to derive an increased amount of gold standard labels compared to an unweighted majority vote. When making majority decisions we inevitably come across items with no majority, but by implementing a flexible feature space we are able to push the boundary a little further and receive majorities, based on expertise, response time and internal consistency even when an unweighted approach did not meet the majority criterion.

This approach will be used to annotate a multilingual data set with music related texts in order to create a gold standard for the development of a cross-language classification algorithm. First steps were taken in testing the approach using an already labeled single-language data set from the music domain. However, for these preliminary experiments it was only possible to use those features that were already available when the data set was labeled.

The next section provides an overview of related work, whereas Section III explains the features we used for our weighted approach. Afterwards, the used approach is explained in Section IV. The results of the preliminary experiments are presented in V and the conclusion and future work in VI.

II. RELATED WORK

The quality of the data used for training text mining systems has significant influence on the final results. Thus, enhancing data quality is a requirement, if we want to receive improved outcomes. In case of processing data labeled by human beings, various quality altering approaches are known. One way of determining the integrity of a data set is to measure the extent to which humans agree on its content. The process of evaluating accordance is called Inter-Rater-Reliability or Inter-Agreement. Furthermore, the procedure for annotating texts and finding the agreement between annotators is also called Inter-Rater-Agreement-Annotator-Agreement.

An overview about methods measuring agreement among corpus annotators is given in a survey article by Artstein and Poesio [2].

The fundamental idea to assess the quality of ratings is to measure the agreement among individuals. A simple approach is the percentage of agreement as described by Scott in 1955 [3]. A combination of this observed agreement with an expected agreement is called chance-corrected agreement. The most common and widely used agreement approaches are

 ISBN: 978-1-61208-806-8

Cohen's kappa coefficient [4] for two raters and Fleiss' kappa coefficient [5] for more than two raters. An advanced approach using calculated weights is given by Cohen in 1968 [6].

A different method is to calculate the agreement by evaluating the disagreement, for example in Krippendorff's alpha coefficient [7].

Those measures are based purely on statistical analyses of the given data, they do not integrate additional features. Individual biases have to be considered to gain better results. For example the emotional state of a rater influences their judgment when dealing with emotional content [8]. Thus, individual stress levels or state of mind could be used as a feature to adjust individual weights in the final decision.

Another important point is the problem of global agreement, meaning the overall agreement of all raters. Usually, methods calculate global agreement. However, this procedure may mask the actual complexity and heterogeneity of the given data [9]. Hence, looking at a type of local agreement considering subgroups of raters with similar properties, for example, experts in a certain area, might be a starting point for accurately representing a data corpus. Moreover, Boguslav et. al indicated that the agreement between annotators might not actually be the upper bound for machine learning tasks [10]. A human labeled data set for which statistical agreement measures were calculated is not necessarily the gold standard but rather a heuristic, since there might be algorithms delivering better results than the human counterpart.

Lastly, there are major problems with basic agreement statistics. Inherently, agreement does not necessarily reflect facts. If an item belongs to topic x and two raters label it as y while only one expert rater judges it to be x, then the agreement is still in favor of y even though the actual topic is x. Furthermore, chance adjusted agreement has distinct problems in both directions. If we have low chance agreement, the influence after adjusting is marginal, thus chance adjusted agreement merely becomes ordinary agreement. On the other hand, high chance agreement can yield low chance adjusted agreement even when the individual raters have good accuracy. Hence, Passonneau and Carpenter [11] show the advantages of using the Dawid Skene model [12], which leads to avoiding problems of Cohens's kappa statistic.

III. The Ideal Feature Space

Usually, text labeling is based on a connection of individual decisions of a group of individuals. However, this connection reduces the whole decision making process to one single value whereas an important aspect is lost, the individual. Hence, we included additional features that can either be measured during or collected previous to the labeling process to receive a more holistic picture. Namely, these include the response time per rater and item, the internal consistency measured by Intra-Rater-Agreement, the conscientiousness per rater, the language proficiency and, finally, the topic expertise. For example, a native speaker should be given more value than a non-native speaker and in the same manner an expert on a given topic should have more weight than an individual barely having any knowledge in that subject area. In order to see how reliable a given individual is, we calculated their Intra-Rater-Agreement. It tests whether a rater expresses the same decision for the same item on different occasions. If decisions are frequently followed by uncertainty, then choices of this annotator cannot be valued as highly as the decisions from someone who is at least partially certain.

Furthermore, previous to a labeling process a questionnaire can be used to probe the conscientiousness of an individual rater to see how reliable and trusting this individual would judge themselves, even though self-judgment can be a two-edged sword as individuals might see themselves in a biased way, as John and Robins have shown [13]. Conscientiousness can be measured by using questions from a personality test using the "Big Five" personality dimensions as introduced by [14], which is widely accepted and has been used on many occasions in psychological studies. For example, a test based on the five personality dimensions was used to link personality traits to job performance as shown by Barrick and Mount [15] but also to overall career success [16]. As the multilingual data set for the planned gold standard is going to be labeled by German speaking annotators, it is planned to use the German personality test using the five personality dimensions as introduced by Satow [17]. Satow's questionnaire contains ten questions probing the conscientiousness of an individual and contrary to the classical "Big Five", Satow's scale goes from one to four and not to five. The reason behind this is to avoid an inherent tendency towards the middle. While we do not want to force a labeling decision on inconclusive texts in our data set, we do ideally push raters to take a more extreme stance on their self judged conscientiousness to ensure diversity in our data set and to avoid leaning towards the middle. Language proficency can be measured, for example, on a scale of one to six representing the different levels (A1-C2) of the "Common European Framework of Reference for Languages" (CEFR) [18] and later normalized to a scale of 1 to 5 to fit in with the other scales. Subsequently, the topic expertise needs to be judged by the individual raters themselves and can be measured on a scale from one to five. Finally, we also want to reward raters who perform well on multiple items. Thus, if a rater's decision for a specific item is in agreement with the majority of all raters for this item they shall be rewarded in subsequent decisions and, otherwise, be punished. As a result, experts for the labeling process might emerge that were not imminent by merely looking at the collected personal information.

IV. Weighted Learning Approach

Let J be a set of raters and I a set of items that are being labeled by these raters. In a classic majority vote, each rater $j \in J$ has the same weight $w = 1$ on every item $i \in I$, regardless of their item specific competence. The desired outcome is to adjust the original weight w for each individual rater j and for the individual items i, such that w_{ji} varies from item to item and from rater to rater depending upon the feature values that have been discussed in the previous section. The weight for rater j and an item i is written as w_{ji} and calculated in (1).

$$w_{ji} = R_j - f(t_{ji}) \tag{1}$$

R_j denotes the rater's competence and is a combination of Intra-Rater-Agreement, topic expertise and language proficiency. Further, $f(t_{ji})$ takes specific values depending upon the response time and conscientiousness for a given rater. In this case, t_{ji} refers to the time an individual j needed to label a given item i.

 ISBN: 978-1-61208-806-8

For the planned annotation of the multilingual data set we consider to include two measures for the topic expertise. Firstly, how frequently a person was in contact with music, e.g., listening to music and, secondly, how regularly they attended events with a high emphasis on music, such as festivals, parties or concerts.

As seen in (2) R_j utilizes n different features x. For the planned gold standard we will consider $x = 4$ features, namely the self-judged music event attendance, overall contact with music, language proficiency and Intra-Rater-Agreement. Additionally, each feature has a unique weighing parameter β which not only enables prioritizing certain features, but also optimizing the weights to receive ideal results.

$$R_j = \frac{\sum_{l=1}^{n} \beta_l x_{lj}}{\frac{1}{|J|} \sum_{j'=1}^{|J|} \sum_{l=1}^{n} \beta_l x_{lj'}} \tag{2}$$

The simplistic idea behind the rater competence R_j is to reward individuals that perform better than the average rater on the sum of features.

Furthermore, for the function $f(t_{j1})$ combining response time and conscientiousness-scores we need to bring the two in a similar format. We achieve this by $[0, 1]$ normalizing the conscientiousness-scores C for each rater j and the response times t for the individual items i, as normalizing over the entire set of items would mitigate the text length of the individual items. In the weighing process the relation of response time t and conscientiousness-score C will give each rater a unique interval. The baseline is the average response time $\bar{t_i}$ for a given item. If rater j has a bigger conscientiousness-score than the normalized average response time then their interval is $[\bar{t_i}, C_j]$. When the response time for the specific item t_{ji} falls within the specific interal the labeler will not face consequences, falling farther away from the given interval results in an increasing penalization. The same procedure applies to a conscientiousness-score lower than the average $[C_j, \bar{t_i}]$. The general function $f(t_{ji})$ is seen in (3) and it helps in pointing out individuals that may not be entirely focused on the task at hand. It can be assumed that individuals with low conscientiousness finish labeling tasks rather quickly while high conscientiousness individuals may need more time to come up with a decision. Contrary, if a low conscientiousness individual takes unusually long it could be due to distractions or lack of focus, which results in a slight penalization. Analogously, individuals with high conscientiousness just rushing through the labeling decisions are penalized, since we expect them to take more time before reaching a decision.

$$f(t_{ji}) = \begin{cases} 0, & C_j > \bar{t_i} \wedge t_{ji} \in [\bar{t_i}, C_j] \\ 0, & C_j < \bar{t_i} \wedge t_{ji} \in [C_j, \bar{t_i}] \\ t_{ji} - \bar{t_i}, & C_j > \bar{t_i} \wedge t_{ji} \notin [\bar{t_i}, C_j] \\ \bar{t_i} - t_{ji}, & C_j < \bar{t_i} \wedge t_{ji} \notin [C_j, \bar{t_i}] \end{cases} \tag{3}$$

For subsequent weights $w_{j(i+1)}$ how well rater j performed on the previous decisions is included. In this context performing well means being part of the majority for the previously weighted decision according to a pre-defined rule that defines majority. If rater j performed excellently, then part of their previous weight $w_{(i-1)j}$ will be transferred to the current decision w_{ij}, allowing them an increased error margin. At the same time, we do not want to over-penalize individuals that underperformed on previous items. Therefore, a parameter b is introduced to regulate how much of the previous weight can be used for the current item. In (4) the parameter b is defined. If an individual performed perfectly, e.g., is part of the majority for every single decision, then $b = 1$. Drastic underperformance yields a b approaching 0.

$$b_{ji} = \frac{1}{\lambda(i-1)} \sum \begin{cases} 1, \text{for j in majority for item i} \\ 0, \text{for j in minority for item i} \end{cases} \tag{4}$$

Since $b = 1$ results in an overweighting of the previous decision and completely neglects the current decision, we also utilize the parameter λ to declare an upper bound for b. Thus, $\lambda = 1$ would result in $b = 1$, whereas $\lambda = 2$ delivers a more reasonable $b = 0.5$. This has the advantage of giving experts some flexibility to bring in their previous expertise and turn the vote in their favor whereas underperforming raters can still utilize their estimated expertise for the current item without being too heavily penalized for their previous performance.

$$w_{ji} = (1 - b_{ji})[R_j - f(t_{j(i+1)})] + \frac{b_{ji}}{i} \sum w_{j(i-1)} \tag{5}$$

By extending (1) with the parameter b and the weight for previous decisions $w_{j(i-1)}$ we receive (5), which can be used for all subsequent decisions. Equation (5) generates unique weights for every single item and for each rater, while their performance on previous items is taken into account. Thus, not everyone is seen as equal and expertise can significantly increase ones impact on a vote, but at the same time drastic underperformance on previous items is considered and can in turn be used to penalize experts that may not perform well on this specific task. In the same manner, a novice may receive high scores due to excellent performance, hence receiving more weight than an expert. This approach allows flexibility within each vote, without being overly biased towards any specific group of individuals. Ideally, this yields an improved performance, which can be indicated by an increased amount of cases where a specific label can be derived from a majority vote. Items that have been seen as ambiguous can now receive a label because a smaller group of individuals with higher overall performance and expertise may swing the vote in their favor.

V. Preliminary Experiments and Results

Currently, it is planned to label a corpus consisting of music related texts from three different languages with 1000 texts per language. The labeling task will be done by 90 raters with different language levels. As the development of this corpus is still in progress, in this paper, preliminary results using an already existing data set including 3000 texts from the same domain labeled by a total number of 48 raters, whereas each text was rated by 20 to 27 raters, are presented. This dataset includes besides the text, the individual rating of each rater and the time stamp referring to the time at which the text was presented to the rater. Using these time stamps for successive texts, the response time was estimated. However, this data set does not include all features proposed in Section III.

For each text there were four possible cases: In the first case, the majority of raters decided that the text does not deal with music, thus giving it the label “not Music”. In the second case, the majority choose the middle ground or an uncertain outcome, meaning the item may contain music

elements yet it is not enough to strongly identify it as such. This case is referred to as “uncertain”. Analogous to the first case, it might be possible to identify a majority that voted for music content, giving a text the label “is Music”. Lastly, if aforementioned cases all fail due to a lack of significant agreement for the specific text, e.g., one third voted “is Music” one third voted “uncertain” and one third voted “not Music”, then we receive the label “no Majority”, leaving the text as ambiguous and hard to identify. Especially, the last case is of relevance as the number of “no Majority” occurrences determines the performance of a given parameter set.

A first baseline is received by conducting unweighted majority votes, utilizing the labels that have been provided by the data set. In order to prevent unrepresentative results, we repeated the calculation with different thresholds for the entire data set of 3.000 unique text items. The majority threshold is defined in the interval of $[0.5, 0.95]$ with steps of 0.05. This way, it is possible to determine the interval for which the best performance is achieved. Afterwards, majority votes utilizing the weighted approach were made using the same thresholds. A comparison of the results can be seen in Table I. It becomes evident, that there are two fundamental ways of reducing the “no Majority” count and thus increasing overall agreement. As expected, with a less restrictive majority threshold, the cases with no majority agreement are reduced. Furthermore, Table I shows that the weighted approach presented in this paper leads to a further decrease in “No Majority” cases if the threshold is kept stable.

TABLE I. COMPARISON OF MAJORITY OCCURRENCES, FOR DIFFERENT THRESHOLDS, WITH AN UNWEIGHTED (UW)AND WEIGHTED (W) APPROACH.

Threshold	is Music		Uncertain		not Music		No Majority	
	UW	W	UW	W	UW	W	UW	W
0.5	1534	1560	3	6	1344	1357	119	77
0.55	1456	1481	2	2	1298	1302	244	215
0.6	1373	1410	2	1	1240	1249	386	339
0.65	1278	1314	0	0	1175	1182	547	504
0.7	1128	1191	0	0	1115	1139	757	670
0.75	982	1072	0	0	1064	1081	954	847
0.8	825	918	0	0	998	1025	1177	1057
0.85	648	739	0	0	931	952	1421	1309
0.9 0	435	504	0	0	821	842	1744	1654
0.95	236	270	0	0	601	636	2163	2094

Figure 1 describes the discrepancy of “no Majority” cases for both, the original unweighted data and our weighted approach utilizing different thresholds. It becomes imminent that we receive fewer cases of “no Majority” with the new setup for all tested thresholds. Furthermore, by looking closely at the data we can see that our weighted approach performs especially well with majority thresholds in the interval of $[0.7, 0.85]$. Peak performance was acquired at a threshold of 0.8. Regardless of the seemingly ideal interval, we maintain a steady improvement of 10% or more for all tested thresholds. An increase beyond 10% is quite valuable, especially when taking into account that our data set did not include sufficient data for all the features that we implemented.

Since the data set did not include data for all required features, some aspects had to be artificially created. First of all, for the Intra-Rater-Agreement, the necessary values were drawn

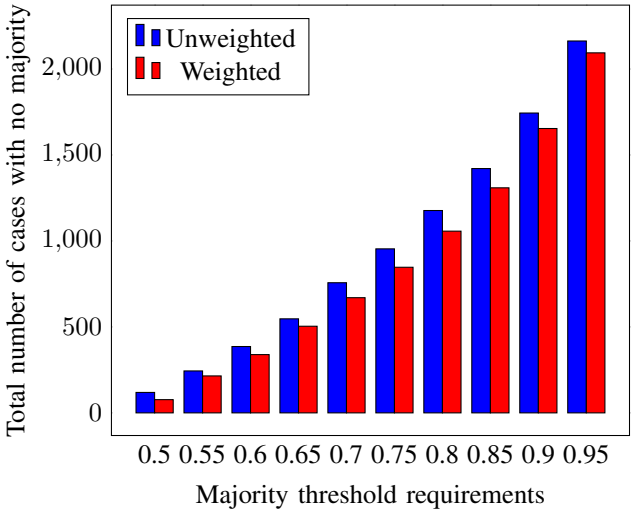

Figure 1. Comparison of “no Majority” cases for an unweighted and weighted majority vote with different majority thresholds.

from a normal distribution. The same procedure was used to get the values for the responses for the questionnaire, which included questions regarding the topic expertise. Furthermore, for the conscientiousness scores a normal distribution was used as well, yet a mean of 0.675 was assumed instead of 0.5 because of the wide availability of “Big Five Test” evaluations such as the one by van der Linden [19]. Finally, if a rater takes a break during the labeling process we may see a spike in the response time that could skew our data set in an unfortunate way. Thus, we shortened all response times longer than 5 minutes to a maximum of 5 minutes. In this experiment, all feature weightings $\vec{\beta}$ were set to 1.

Analogously to generating more majorities, this approach may also occasionally create disagreement. While the total “no Majority” count clearly decreases there could still exist individual cases in which the weighted approach derives the label “no Majority” whereas a classic unweighted majority vote may find a majority. It would be of great interest to see in which situations the weighted approach creates disagreement and if in some cases the majority in itself changes, meaning the unweighted majority labeled an item as, for example, “is Music” whereas the weighted majority labeled the same item as “not Music”.

VI. CONCLUSION AND FUTURE WORK

In this work, a flexible weighting approach for Inter-Rater-Agreement is proposed. Preliminary experiments have shown that using this approach may provide improved results for text labeling tasks even when missing data is artificially created. With actual available data the improvements might be even more significant since a correlation between response time and rater conscientiousness might exist.

We currently collect data to redo this experiment with non-generated feature data, which should provide more accurate insights. If the aforementioned or a similar correlation exists, the weighted majority vote may yield drastically improved results surpassing the current 10% mark. Additionally, in the next experiment we will evaluate the optimization potential

 ISBN: 978-1-61208-806-8

by utilizing different settings for the weighting parameters β_i which were set to 1 in this first research phase.

Furthermore, similar experiments should be conducted using varying feature spaces to not only ensure the validity of this approach but also to discover potential performance variations between individual features. Finally, while in this study we only resort to rater dependent features, there is also the option to include item dependent features such as the item specific competence level of the rater.

REFERENCES

[1] D. Yvonne, A. Eva, and G. Gunnar, "Inter-Rater Agreement Using the Instrumental Activity Measure," Scandinavian Journal of Occupational Therapy, vol. 7, 2000, pp. 33–38.

[2] R. Artstein and M. Poesio, "Inter-Coder Agreement for Computational Linguistics," Computational Linguistics, vol. 34, no. 4, Dec. 2008, pp. 555—596.

[3] W. A. Scott, "Reliability of Content Analysis: The Case of Nominal Scale Coding," Public Opinion Quarterly, vol. 19, 1955, pp. 321–325.

[4] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, 1960, pp. 37–46.

[5] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76, 1971, pp. 378–382.

[6] J. Cohen, "Weighted Kappa - Nominal Scale Agreement with Provision for Scaled Disagreement Or Partial Credit," Psychological Bulletin, vol. 70, Nov 1968, pp. 213–220.

[7] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Sage, 2004, ISBN: 978-07-6191-545-4.

[8] E. A. Kolog, C. S. Montero, and E. Sutinen, "Annotation Agreement of Emotions in Text: The Influence of Counsellors' Emotional State on their Emotion Perception," 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, pp. 357–359.

[9] R. Artstein, "Inter-Annotator Agreement," Handbook of Linguistic Annotation, 2017, pp. 297–313.

[10] M. Boguslav and K. Cohen, "Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing," Studies in Health Technology and Informatics, vol. 245, Jan 2017, pp. 298–302.

[11] R. Passonneau and B. Carpenter, "The Benefits of a Model of Annotation," Transactions of the Association for Computational Linguistics, vol. 2, Dec 2014, pp. 311–326.

[12] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of Observed Error-Rates Using the EM Algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 1, 1979, pp. 20–28.

[13] O. P. John and R. W. Robins, "Accuracy and Bias in Self-Perception: Individual Differences in Self-Enhancement and the Role of Narcissism," Journal of Personality and Social Psychology, vol. 66, Feb 1994, pp. 206–219.

[14] L. R. Goldberg, "The development of markers for the Big-Five factor structure." Psychological Assessment, vol. 4, no. 1, 1992, pp. 26–42.

[15] M. R. Barrick and M. K. Mount, "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis," Personnel Psychology, vol. 44, no. 1, 1991, pp. 1–26.

[16] T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick, "The Big Five Personality Traits, General Mental Ability, And Career Success across the Life Span," Personnel Psychology, vol. 52, Dec 2006, pp. 621–652.

[17] L. Satow, " B5T - Psychomeda Big-Five-Persönlichkeitstest," Personnel Psychology, 2011.

[18] C. of Europe, "Common European Framework of Reference for Languages: Learning, teaching, assessment," 2001.

[19] D. van der Linden, J. te Nijenhuis, and A. B. Bakker, "The General Factor of Personality: A Meta-Analysis of Big Five Intercorrelations and a Criterion-Related Validity Study," Journal of Research in Personality, vol. 44, March 2010, pp. 315–327.

 ISBN: 978-1-61208-806-8

Design of a Multimedia Data Management System that Uses Horizontal Fragmentation to Optimize Content-based Queries

Marcos Joaquín Rodríguez-Arauz, Lisbeth Rodríguez-Mazahua, Mario Leoncio Arrioja-Rodríguez, María Antonieta Abud-Figueroa, Silvestre Gustavo Peláez-Camarena, Luz del Carmen Martínez-Méndez
Division of Research and Postgraduate Studies
Tecnológico Nacional de México/Instituto Tecnológico de Orizaba
Orizaba, Ver., México
e-mail: marcos.joroar@gmail.com, lrodriguezm@ito-depi.edu.mx, marrioja@ito-depi.edu.mx, mabud@ito-depi.edu.mx, sgpelaez@yahoo.com.mx, lcmartinezmdz@gmail.com

***Abstract*— In the Technological Institute of Orizaba, a project was carried out to collect historical data of the Institute, where a large amount of multimedia data was retrieved. At the end of the project, it was detected that increasing multimedia content made it difficult to manage this data optimally. This work addresses the aforementioned problem since a system for the historical data management of the Institute is designed, which uses horizontal fragmentation to optimize content-based queries. Experiments using a cost model demonstrate that the system reduces the execution cost of content-based queries.**

Keywords-horizontal fragmentation; content-based retrieval; multimedia database.

I. INTRODUCTION

With the great growth of multimedia devices, large amounts of multimedia data are generated every day, the rapid access to these huge collections of data means increasingly greater challenges and the need for very efficient algorithms [1]. Therefore, multimedia database features, such as transactional updates, querying facilities, and indexing, become transcendent when the number of stored multimedia objects increases and such big challenges begin to appear [2].

Nowadays, efficient Content-based Image Retrieval (CBIR) techniques are a must for the optimal use of multimedia databases. The need for efficient image retrieval increases rapidly and therefore, to improve performance and reduce the margin between visual characteristics, the CBIR process is used [3]. Nevertheless, in CBIR systems, all the images of the multimedia database are processed to extract features and compare them to the features of the image query in order to retrieve the most similar images. Using fragmentation techniques to reduce the number of images processed to answer a query will greatly improve the performance of the system. The importance of fragmentation lies in the fact that it allows minimizing the number of access to irrelevant data, thus reducing the response time and the execution cost of the queries.

Therefore, this paper proposes the creation of a multimedia data management system that uses a horizontal fragmentation method to optimize content-based queries. The system will be responsible for storing and managing historical multimedia data of the Technological Institute of Orizaba. This paper is structured as follows: Section II describes recent work related to our proposal. The design of the proposed system is described in Section III. Section IV presents the results obtained with the cost model used to evaluate the efficiency of the horizontal fragmentation scheme proposed in this work and finally, in Section V, the conclusion and ongoing work are discussed.

II. RELATED WORK

Currently, several fragmentation methods for multimedia databases have been proposed, which include a cost model to evaluate their fragmentation schemas [4], [5], [6]. In [4], a horizontal partitioning method for multimedia databases was presented which is based on a hierarchical agglomerative clustering algorithm. The main advantage of the method is that it does not use affinity to create the horizontal partitioning scheme. The cost of a horizontal fragmentation schema is composed of irrelevant tuple access cost (ITAC) and transportation cost (TC). ITAC measures the amount of data from irrelevant tuples accessed during the queries. TC provides a measure for transporting between the sites of the network.

Most vertical fragmentation algorithms for MultiMedia Data Bases (MMDB) are static, which means that they optimize a Vertical Partitioning Scheme (VPS) according to a workload, but if it changes, the VPS could be degraded, this would result in long query response times. In [5], a system called DYnamic Multimedia ONline Distribution (DYMOND) was proposed, which uses an active rule set to execute dynamic vertical partitioning in multimedia databases. As a result, it was found that DYMOND improves query performance in multimedia databases.

Rodríguez-Mazahua et al. [6] proposed a hybrid fragmentation method for multimedia databases, called Multimedia Hybrid Partitioning Algorithm (MHYP), along with a cost model to evaluate hybrid fragmentation schemes. Experiments showed that MHYP outperforms both horizontal and vertical fragmentation in most cases.

One disadvantage of the above-discussed works is that they do not consider similarity-based predicates used in content-based queries. In contrast, in [7], it was introduced an approach to efficiently execute conjunctive queries on big

 ISBN: 978-1-61208-806-8

complex data together with their related conventional data. The basic idea is to horizontally fragment the database according to criteria frequently used both in traditional and similarity-based query predicates; nevertheless, it does not provide a cost model to evaluate its fragmentation schema.

On the other hand, there are several approaches to increase the efficiency of CBIR systems, but they do not report the use of a fragmentation method, e.g., to help the employees of a company that sells agricultural machinery and spare parts, called AGROMAQ, a CBIR system was proposed in [8] for the recognition of agricultural and spare parts, using two descriptors: Speeded Up Robust Features (SURF) and Scale Invariant Feature Transform (SIFT). The results showed that the SURF descriptor is faster and obtains a greater precision compared to the SIFT descriptor. In [3], a SURF-based CBIR system along with genetic algorithms for optimization was proposed. The results showed an increase in performance, with 98% of accuracy; this system was implemented using the MATLAB software image processing toolbox.

A CBIR system was introduced in [9], which proposed to use features derived from pre-trained network models from a deep learning convolution network trained for a large image classification problem. This approach appears to produce vastly superior results for a variety of databases, and it outperforms many contemporary CBIR systems. The retrieval time of the method was analyzed, and a pre-clustering of the database was proposed based on the above-mentioned features, yielding comparable results in a much shorter time in most of the cases. In [10], Prasomphan presented an algorithm for retrieving information from an image for telling a story about that image. The appearance of the shape inside an image can be used to distinguish characteristics of the image, for example, the era, architecture, and style of image. The architecture was created using machine learning and image processing. The experimental results for a cultural heritage information management system with a deep neural network were analyzed by using the classification results of the proposed algorithms to classify the era and architecture of the tested image.

In contrast, Ouhda et al. [1] applied a new CBIR method based on the K-Means clustering technique to provide accurate results with less computation time. For validation, the system was applied in two multimedia databases and the expected results were obtained. Also, a CBIR system was shown in [11] with two methods of extracting features: SIFT and Oriented Fast Rotated and BRIEF (ORB). ORB uses Feature Accelerated Segment Test (FAST), which is a key point detector, and Binary Robust Independent Elementary Features (BRIEF) as a descriptor of an image. The K-Means clustering method was used to analyze the data, which generates clusters using the descriptor vector. A precision of 88.9% was achieved as a result.

Today, the challenge of large-scale content-based image retrieval has had great approaches by numerous promising works, as, Skar et al. [12] which analyzed an efficient CBIR framework. Hadoop MapReduce was proposed to operate with great performance. Empirical tests proved to surpass in performance the techniques compared in the state of the art. Another approach is Kamel et al. [13] which proposed a new method that optimizes the search precision and runtime for large-scale content-based image retrieval. This was achieved by using local binary image descriptors, such as BRIEF, and binary hashing methods, such as Spherical Hashing (SH). As a result, the accuracy was improved.

Table I shows a comparison between the related works and this project. The table indicates if the work makes use of horizontal fragmentation, CBIR system, and a cost model to evaluate its fragmentation scheme.

TABLE I. COMPARISON OF RELATED WORKS

Article	Horizontal Fragmentation	CBIR	Cost Model
Ouhda et al.[1]	No	Yes	No
Prinka and Wasson [3]	No	Yes	No
Rodríguez-Mazahua et al. [4]	Yes	No	Yes
Rodríguez-Mazahua et al. [5]	No	No	Yes
Rodríguez-Mazahua et al. [6]	Yes	No	Yes
Fasolin et al. [7]	Yes	Yes	No
Rojas-Ruiz et al. [8]	No	Yes	No
Maji and Bose [9]	No	Yes	No
Prasomphan [10]	No	Yes	No
Kumar et al. [11]	No	Yes	No
Sakr et al. [12]	No	Yes	No
Kamel et al. [13]	No	Yes	No
This work	Yes	Yes	Yes

As can be seen in Table I, none of the other proposals applies a horizontal fragmentation method for the multimedia database to improve the CBIR system and includes a cost model to evaluate its fragmentation scheme.

III. DESIGN OF THE MULTIMEDIA DATA MANAGEMENT SYSTEM

This section describes the design of the multimedia data management system that also uses a horizontal fragmentation method to optimize content-based queries. The proposed solution will greatly improve the performance and the management of the historical data in the multimedia database for content-based queries.

A. Architecture Description

The system architecture was designed based on the Model View Controller architectural model (MVC) which distributes the system components in a way that facilitates its maintenance and is represented abstractly in Figure 1. This figure shows the components proposed for obtaining the solution, which are described below.

MongoDB [14] was selected due to its speed and simple system to query the content of the database. Java Server Faces (JSF) [15] was chosen due to the flexibility to create applications with the pure Java language, its ease of software development, and because is also free. NetBeans [16] was proposed as the Integrated Development Environment (IDE) due to the ease of developing applications with the selected framework. It was decided to use UML-based Web Engineering (UWE) [17] as the development methodology because it specializes in

developing Web and multimedia applications. Lastly, BoofCV [18] was selected because it is an open-source Java library for real-time computer vision that has ease of use and high performance.

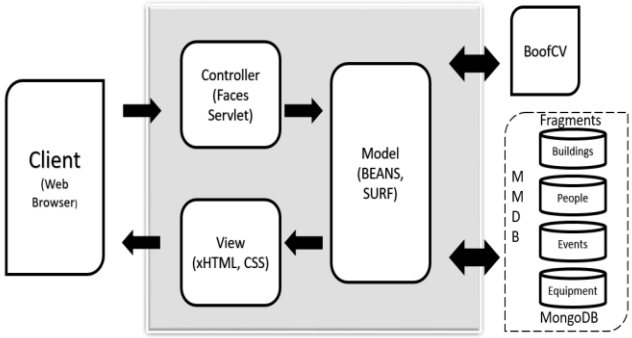

Figure 1. Application architecture.

1) *Model:* The system logic is represented in the managed beans that have access to the interface components and pass information to the model beans, the latter representing the important classes of the domain and using libraries, such as BoofCV to perform the content-based retrieval (images) and Java Database Connectivity (JDBC) to control access to the MongoDB database management system and manipulate information. The horizontal fragmentation method proposed in [7] was applied because it considers content-based queries, therefore, it produced four fragments which can be observed in Figure 1 as buildings, people, events, and equipment. This horizontal fragmentation schema is considered adequate because of the characteristics of the images in the database, and since the searches will be carried out considering these four types of images, in addition, the performance will not be affected by the imbalance between the fragments, since the queries will only retrieve images from the fragments corresponding to the image type.

2) *View:* Using eXtensible HyperText Markup Language files (XHTML), the model is represented and handles the interaction with the user; JSF's tags are used for this purpose, together with style sheets (CSS) to give the user a better visualization.

3) *Controller:* The JSF servlet is the link between the model and the view. It is responsible for managing the requests of the resources by accessing the model needed in the user's request and selecting the appropriate view to represent it.

As noted, the structure and organization proposed by the MVC pattern provide a good coupling between the components and the changes will only be noticeable to the parties directly involved.

B. Requirement Determination

Figure 2 shows the use case diagram of the Web application and announces the functional requirements. There are five actors: 1) The administrator has full control over the database and system and it is the only one that can perform deletions of documents; 2) The collaborator has access to the structure of the database but in a limited way; 3) The Professor is able to access the information and/or images where he or she appears as well as propose content; 4) The General user can view general information about the Institute and propose content, and 5) The anonymous user is only able to access general information of the Institute.

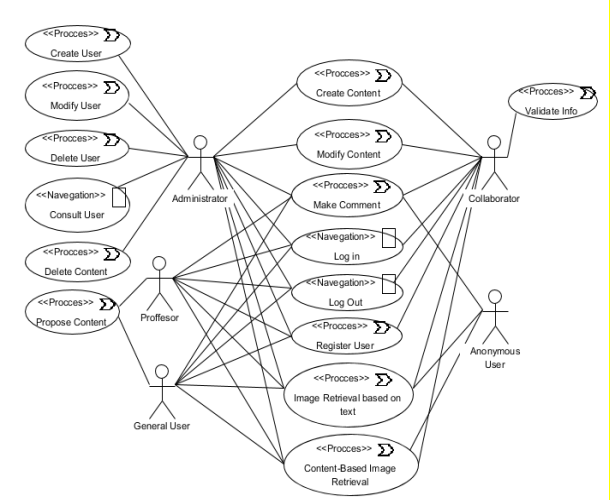

Figure 2. Use case diagram of the web application.

Figure 2 shows the use case "Content-based Image Retrieval", as suggested by the UWE methodology. This use case is described with the process diagram of the content-based image retrieval which can be observed in Figure 3. The Content-Based Image retrieval process diagram starts by displaying a form for uploading the image of the content to retrieve, the system then evaluates if there is a valid file in order to do the content-based retrieval. Once a valid file is uploaded, the system will analyze the image and retrieve the most similar content allocated in the specific fragment type of the image itself. Finally, a result page is shown where the user can appreciate the results of the content-based image retrieval.

C. System Design

The conceptual model in this work is represented by the physical diagram of the database of Figure 4. The images collection is fragmented considering the attribute *type* according to the Fasolin et al. [7] method, which gives as result the horizontal fragments named: buildings, events, people, and equipment.

Through the navigation model, all possible paths are known within the application for user navigation. Figure 5 shows the navigation model of the application. Figures 6 and 7 depict the content-based image retrieval and contents page, respectively.

 ISBN: 978-1-61208-806-8

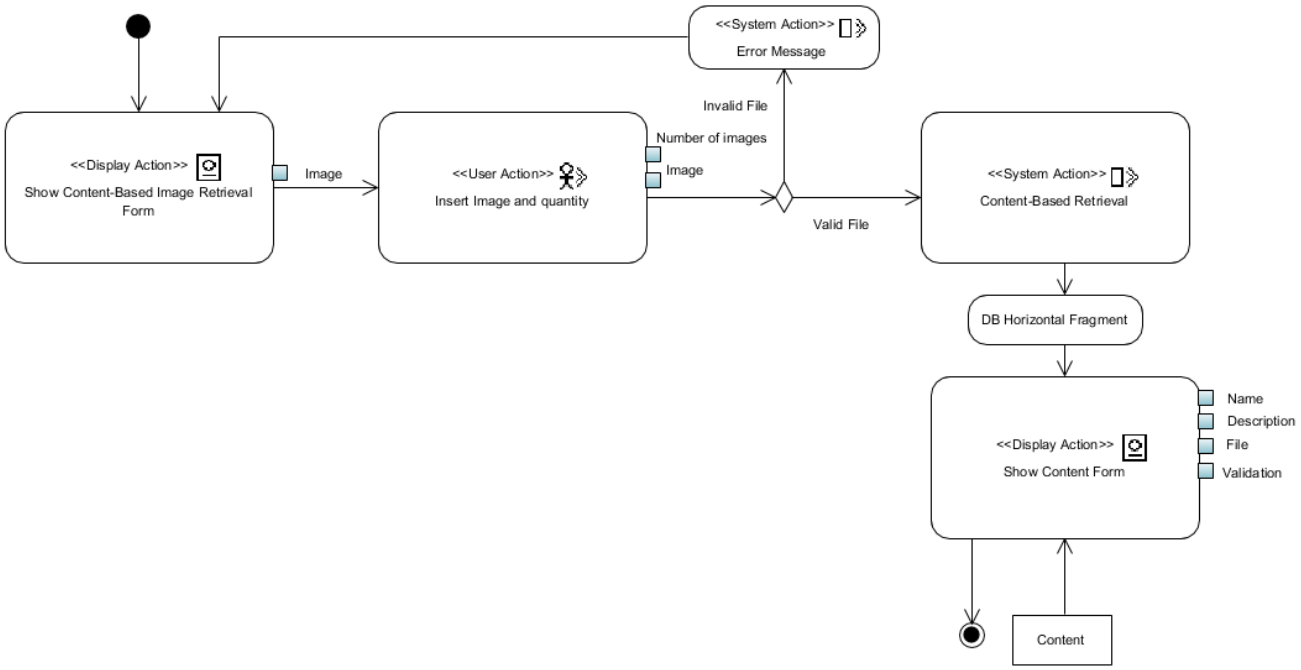

Figure 3. Content-Based Image Retrieval diagram of the process model.

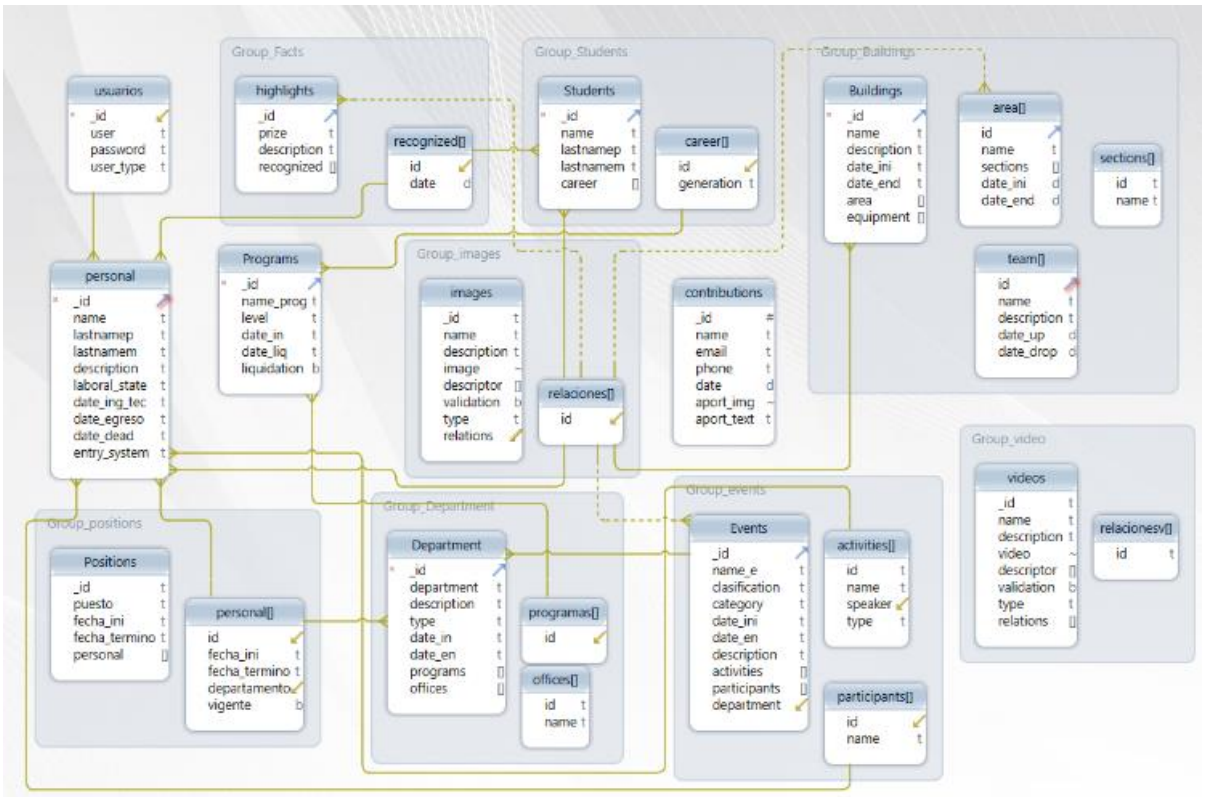

Figure 4. Physical Diagram of the database.

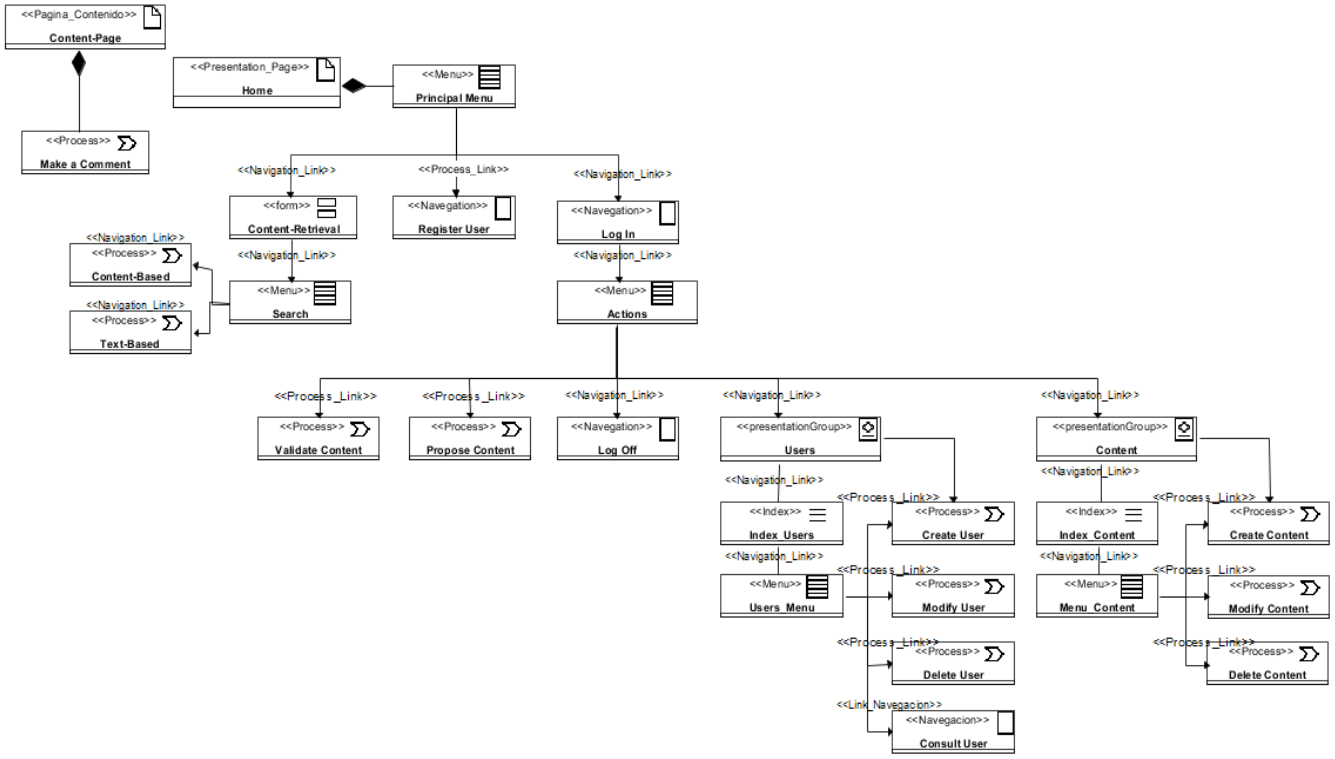

Figure 5. Web application navigation model.

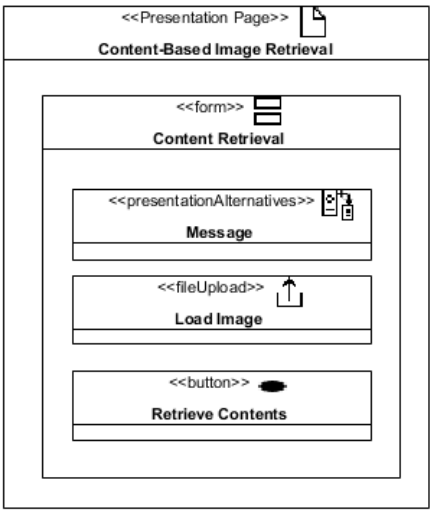

Figure 6. Content-Based Image Retrieval page of the presentation model.

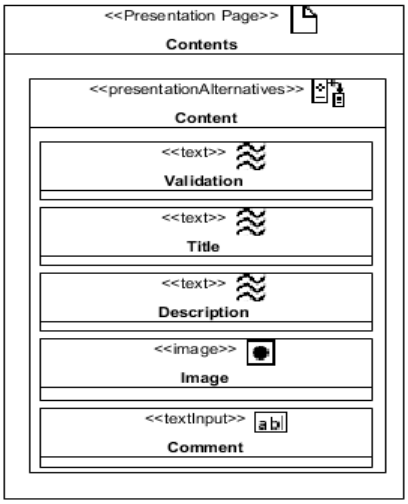

Figure 7. Contents page of the presentation model.

Figure 6 shows the content-based image retrieval page where the user can upload an image in order to retrieve the most similar contents allocated in the database. Figure 7 depicts the contents page where the user can appreciate the results of the content-based retrieval, showing the most similar images related to the one uploaded on the previous page.

IV. RESULTS

In this section, we compare the execution cost of four content-based queries in the multimedia database of the proposed system without fragmentation (NF) vs. horizontally fragmented (HF) using the cost model described in [4].

The first step is to determine the set of content-based queries Q={q_1, q_2, …, q_n}, in this experiment we consider nearest neighbor queries, which are described as:

q_1: Find the 5 images most similar to event_w.png
q_2: Find the 5 images most similar to equipment_x.png
q_3: Find the 5 images most similar to person_y.png
q_4: Find the 5 images most similar to building_z.png

The predicates for the queries *Pr* are described in Table II. As stated before in this paper, the results of the horizontal fragmentation method used [7] were four fragments, the number of images per fragment can be seen in Table III.

As the next step, it is needed the construction of a Predicate Usage Matrix (PUM), which is a matrix that contains queries as rows and predicates as columns and every cell indicates if the query q_i uses the predicate p_j. A PUM also includes the access frequency of each query and the selectivity of the predicates. Then a Fragment-Predicate Usage Matrix (FPUM) is obtained; this matrix has fragments as rows and predicates as columns, and every cell can have a value of 1 if the fragment uses a predicate or 0 otherwise. The FPUM also shows the cardinality of each fragment. With these matrices, it is possible to obtain the Irrelevant Tuple Access Cost (ITAC). Tables IV and V show the results. In the PUM of Table IV, f_i is the access frequency of q_i, and sel_j is the number of images that satisfy the predicate p_j, while $card_k$ in the FPUM of Table V is the number of images that each fragment contains.

TABLE II. PREDICATES USED BY QUERIES

Q	Pr
q_1	p1=5NN(event_w.png)
q_2	p2=5NN(equipment_x.png)
q_3	p3=5NN(person_y.png)
q_4	p4=5NN(building_z.png)

TABLE III. FRAGMENTS OF THE MMDB

Fragment	Number of Images
Events	756
Equipment	63
People	886
Buildings	158

TABLE IV. PREDICATE USAGE MATRIX

Q/Pr	p_1	p_2	p_3	p_4	f_i
q_1	1	0	0	0	15
q_2	0	1	0	0	5
q_3	0	0	1	0	20
q_4	0	0	0	1	10
sel_j	5	5	5	5	

TABLE V. FRAGMENT-PREDICATE USAGE MATRIX

Fragments	p_1	p_2	p_3	p_4	$card_k$
fr_1	1	0	0	0	756
fr_2	0	1	0	0	63
fr_3	0	0	1	0	886
fr_4	0	0	0	1	158

In order to obtain the Irrelevant Tuple Access cost of each query, the next equations are used, where Pr_j has the predicates used by a query q_j and located in the fragment fr_k, and n_p is the number of predicates in Pr_j. Figure 8 shows the comparison of the execution cost of the queries.

$$ITAC(q_j) = \begin{cases} f_j\left(card_k - \sum p_t \in Pr_j\ sel_t\right) if\ n_p \geq 1 \\ 0 \quad otherwise \end{cases} \quad (1)$$

$$Pr_j = \{p_t | PUM(q_j, p_t) = 1 \wedge \text{FPUM}(fr_k, p_t) = 1\} \quad (2)$$

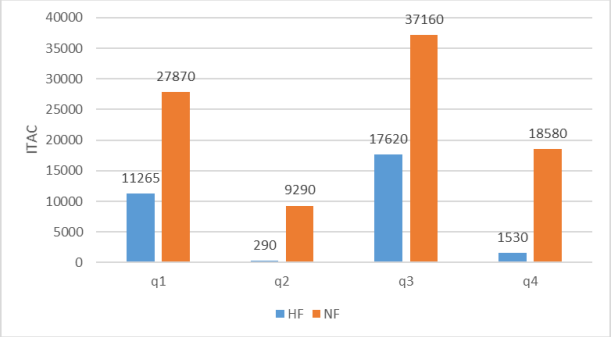

Figure 8. Execution cost of the content-based queries.

Figure 8 shows that the execution cost of each content-based query using horizontal fragmentation is significantly lower than when a fragmentation method is not used.

 ISBN: 978-1-61208-806-8

V. CONCLUSION

Data fragmentation is a very popular object of study today and is constantly used in the industry. A large number of fragmentation techniques are currently used, including horizontal fragmentation, which is the technique that is proposed to be used in the development of this work. This helps not only to improve the management of multimedia databases but also to optimize the efficiency and speed of access to the information that is needed.

This work presented the design of a system for multimedia data management. The system uses horizontal fragmentation to improve the execution cost of content-based queries. The users of the Technological Institute of Orizaba will benefit greatly, causing a social impact, which aims to raise awareness about the history of the institute. The system will also generate an economic impact, since by using a fragmentation method, advantages can be observed, such as lower execution costs or reduced retrieval times.

Future work includes the implementation of the system using the selected technologies and the incorporation of a vertical fragmentation method for even better optimization of the content-based queries. Also, conjunctive complex queries will be included for the evaluation of the horizontal fragmentation schema.

ACKNOWLEDGMENT

The authors are very grateful to the National Technological of Mexico (TecNM) for supporting this work. Also, this research paper was sponsored by the National Council of Science and Technology (CONACYT).

REFERENCES

[1] M. Ouhda, K. El Asnaoui, M. Ouanan, and B. Aksasse, “Content Based Image Retrieval Method Based on K-Means Clustering Technique,” Journal of Electronic Commerce in Organizations, vol. 16, pp. 82-96, Jan. 2018, doi: 10.4018/JECO.2018010107.

[2] A. Silberschatz, H. F. Korth, and S. Sudarshan, Database system concepts, 6th ed. New York: McGraw-Hill, 2011.

[3] Prinka and V. Wasson, “An efficient content based image retrieval based on speeded up robust features (SURF) with optimization technique,” en 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, may 2017, pp. 730-735, doi: 10.1109/RTEICT.2017.8256693.

[4] L. Rodríguez-Mazahua, G. Alor-Hernández, Ma. A. Abud-Figueroa, and S. G. Peláez-Camarena, “Horizontal Partitioning of Multimedia Databases Using Hierarchical Agglomerative Clustering,” en Nature-Inspired Computation and Machine Learning, vol. 8857, A. Gelbukh, F. C. Espinoza, y S. N. Galicia-Haro, Eds. Cham: Springer International Publishing, 2014, pp. 296-309.

[5] L. Rodríguez-Mazahua, G. Alor-Hernández, X. Li, J. Cervantes, and A. López-Chau, “Active rule base development for dynamic vertical partitioning of multimedia databases,” J Intell Inf Syst, vol. 48, pp. 421-451, Apr. 2017, doi: 10.1007/s10844-016-0420-9.

[6] L. Rodríguez-Mazahua, G. Alor-Hernández, J. Cervantes, A. López-Chau, and J. L. Sánchez-Cervantes, “A hybrid fragmentation method for multimedia databases,” DYNA, vol. 83, pp. 59-67, Sep. 2016, doi: 10.15446/dyna.v83n198.50507.

[7] K. Fasolin et al., “Efficient Execution of Conjunctive Complex Queries on Big Multimedia Databases,” en 2013 IEEE International Symposium on Multimedia, Anaheim, CA, USA, Dec. 2013, pp. 536-543, doi: 10.1109/ISM.2013.112.

[8] R. Rojas et al., “A CBIR System for the Recognition of Agricultural Machinery,” Journal of Research in Computer Science, vol. 3, pp. 9-16, 2018.

[9] S. Maji and S. Bose, “CBIR using features derived by Deep Learning,” arXiv:2002.07877 [cs, stat], Feb. 2020, Accessed: Aug. 08, 2020. [Online]. Available: http://arxiv.org/abs/2002.07877.

[10] S. Prasomphan, “Cultural Heritage Content Management System by Deep Learning,” in Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference, Nagoya Japan, May 2020, pp. 21–26, doi: 10.1145/3399871.3399894.

[11] M. Kumar, P. Chhabra, and N. K. Garg, “An efficient content based image retrieval system using BayesNet and K-NN,” Multimed Tools Appl, vol. 77, pp. 21557-21570, Aug. 2018, doi: 10.1007/s11042-017-5587-8.

[12] N. A. Sakr, Ali. I. ELdesouky, and H. Arafat, “An efficient fast-response content-based image retrieval framework for big data,” Computers & Electrical Engineering, vol. 54, pp. 522-538, Aug. 2016, doi: 10.1016/j.compeleceng.2016.04.015.

[13] A. Kamel, Y. B. Mahdy, and K. F. Hussain, “Multi-Bin search: improved large-scale content-based image retrieval,” Int J Multimed Info Retr, vol. 4, pp. 205-216, Sep. 2015, doi: 10.1007/s13735-014-0061-0.

[14] A. Boicea, F. Radulescu, and L. I. Agapin, “MongoDB vs Oracle -- Database Comparison,” en 2012 Third International Conference on Emerging Intelligent Data and Web Technologies, Bucharest, Romania, Sep. 2012, pp. 330-335, doi: 10.1109/EIDWT.2012.32.

[15] JavaServer Faces, “Introducción and JavaServer Faces”. [online] Available at: http://www.jtech.ua.es/, [Retrieved: August, 2020].

[16] NetBeans, "NetBeans Docs & support," 2016. [Online]. Available: https://netbeans.org/kb/index.html. [Retrieved: August, 2020].

[17] C. Nieves, J. Ucán, and V. Menéndez, “UWE in Learning Object Recommendation System,” Aplicando Ingeniería Web: Un Método en Caso de Estudio, Revista Latinoamericana de Ingeniería de Software, vol. 2, pp. 137-143, 2017.

[18] Boofcv.org, “BoofCV” [online] Available at: https://boofcv.org/index.php?title=Main_Page [Retrieved: August, 2020].

 ISBN: 978-1-61208-806-8