https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

# AN EFFICIENT MARKETING POLICY RECOMMENDATION SYSTEM BASED ON USER CENTRIC SIMILARITY SEARCH USING REVERSE TOP-K QUERY

**Sindhu B Jigali[1], Nirmala C R[2]**

*[1]Department of Computer science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere*
*[2]Department of Computer science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere*

## Abstract

*Information administration can be characterized as advancement and implementation of strategies, practices and systems. The advancement procedure comprises of recognizing likeness among information products as major task. Similitude measurements like Euclidian separation and cosine comparability are utilized to decide resemblance between information products and is figured in light of their properties. An effective advertisement marketing approach can be built up by not just thinking about the credits to process similitude among information products yet additionally client inclinations and conclusion on information about products. The set up arrangement focuses on the particular group of onlookers with bunch of best Data products. The closeness checking framework as a rule utilizes top-k question that profits the k Data products with the best rank for a specific client. The displayed inverted top-K inquiry output in the arrangement of powerful information products where the information products have a place with their top-k group. Jaccard constant is utilized to perform comparability calculations among the subsequent arrangements of the turnaround top-k questions, where as it additionally figures the min and max bound on the client driven closeness of information products. θ-similitude and m-closest neighbor inquiries productively register likeness among information about products of the invert top-k result.*

***Keywords*** *– θ-similitude, m-closest nearest neighbors, top-k*

---------------------------------------------------------------------***---------------------------------------------------------------------

## 1. INTRODUCTION

**Similarity search** is a standout amongst the most utilized scope of systems it allocate the guideline of seeking (ordinarily, substantial) places of items somewhere the main accessible contrast is the closeness involving some combine of articles [1].It is ending up progressively imperative during a time of extensive data stores where the articles limited don't have any common request, for instance expansive accumulations of pictures, voice and other complex advanced items.

Closest neighbor inquiry and scope inquiries are critical subclasses of closeness look. Investigate in resemblance Search is overwhelmed through the inalienable issues of seeking in excess of tedious items. Such protests make most recognized systems lose footing over vast accumulations. Lamentably, as a rule the articles are naturally tedious and the closeness look is extremely vital [3]. The majority broad way to deal with likeness look depends ahead the numerical thought of key liberty, it permits the development of productive list arrangement keeping in mind the end goal to accomplish versatility in the inquiry area. Assessment of the comparability amid substance is single between the principal task in statistics administration. Euclidean separation method to find the distance between the specified points and the cosine closeness are a portion of the likeness measurements that are utilized as a part of assessing the similitude connecting two information things. As per these measurements the closeness is registered in view of their qualities, exclusive of taking into client's supposition for thought. Client sentiment and inclinations are wellspring of Big data .Bearing in mind just attributes of the items, a reasonable advertising approach cannot be constructed.[2].

By means of the thought of together the qualities and client inclinations a reasonable advertising strategy might be composed which brings about formation of bunch of items best for particular clients. Big data is a wide phrase for information groups consequently extensive or tedious that conventional statistics preparing applications are deficient. Difficulties incorporate examination, catch, information length, duration, look, distributing, stockpiling, exchange, perception and data security. The term regularly alludes basically to the utilization of prescient investigation or other certain propelled strategies to extricate an incentive from information, and only here and there to a specific amount of information group[1].

**Nearest Neighbor computation or** Closest Neighbor calculation strategy processes the separation or difference amid information things and utilized as a part of characterizing them by characterizing principle. It likewise has variations to be specific k-NN \, weighted k-NN strategy utilized for order and relapse, which utilizes the information comprises of the k nearest preparing cases in the component liberty.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

**Top-k Query** is related by means of preparing extensive information groups and bringing about best k statistics is things with most great score. It can acknowledge n tables to process on the double and deliver totaled outcomes.

**Reverse Top-k or Invert Top-k** inquiries distinguishes the best k the majority powerful items to clients, where impact is characterized as the count of the switch top-k result. This meaning of impact is helpful for market investigation, because it is specifically identified with the quantity of clients that esteem a specific item and therefore to its deceivability and effect to the advertises.

**Jaccard closeness coefficient** is a measurement utilized for contrasting the likeness and assorted variety of test information groups. Jaccard remove is registered as the proportion of the span of the symmetric distinction to the association.

**R-tree** information arrangement bunches adjacent protests and speak to them with their base bouncing square shape in the following more elevated amount of the tree; the "R" in R-tree is for square shape. Because all articles exist in this jumping square shape, a question it do not meet the bouncing square shape likewise can't cross any of the contained items. At children level, every square shape depicts a solitary question; at more elevated amounts the accumulation of an expanding number of items. This is likewise observed as an undeniably coarse guess of the dataset.

MongoDB is considered as a good example for the NoSQL means not only structured query language. It is a sort of document oriented database. It collects and stores the information as a documents or articles. It also permits clients to accumulate and merge any type of statistics and vigorously update the schema means structured format of planned database. Hence, information that is accumulated in MongoDB posses very flexible structured format. The groups of articles , properties or attributes in MongoDB both were similar the tables , tuples and attributes of very much flexible schema.

Client based community sifting system ascertains likeness among clients by looking at their appraisals on a similar thing, and it at that point figures the anticipated marking for a thing by the dynamic client as a biased normal of the evaluations of the thing by clients like the dynamic client anywhere weights are the similitude's of these clients with the objective thing. Thing based sifting strategies figure forecasts utilizing the likeness amongst things and not the similitude among clients. It assembles a dimension of thing similitude's by recovering all things evaluated by a dynamic client from the client thing lattice, it decides how comparable the recovered things are to the objective thing, at that point it chooses the k most comparative things and their relating likenesses are likewise decided. Expectation is finished by having a subjective normal of the dynamic clients rating on the comparative things k. A few kinds of closeness procedures are utilized to register comparability among thing/client.

## 2. LITERATURE SURVEY

The idea identified as client inclinations in light of the utilization of inquiries is developing quickly in some constant applications. Item inclinations based on strategy have to be adjusted and improved with additional database activities.

Items are bunched in light of the comparability of their highlights. The current examination suggests another strategy for item grouping. The novel strategy figures biased qualities as far as estimations of traits of items and their relating sentiment esteems indicated by clients for every item independently. Entirety of weighted qualities, estimations of properties of items increased by conclusion esteems, are registered by utilizing a straight biased capacity. In view of the biased totals items are bunched.

Top-k questions are quite a while ago examined point in the database and data recovery networks. Such questions restore the k majority encouraging information things, in view of accessible client inclinations. The suggested job tends to the issue of estimating the nature of best k outcome groups returned through a data recovery framework, similar to the instance of looking at web crawler comes about. The creators talk about a few elective measures and give quick guess calculations to the assessment of some of them. Then again, turn around top-k questions, presented restores the clients that locate an item (the inquiry point) in their best k result sets. An utilization for turnaround top-k inquiries is to recognize influential items, where influence is defined as the quantity of the invert top-k result group. The distinguishing proof of influence substance is valuable for advertisement investigation.

### Problem Statement

Top-k question restores the k most encouraging information things, in view of accessible client inclinations, yet it neglects to quantify the nature of the best k output groups and the upgraded similitude between the outcome group. So to beat this, Reverse Top-k (RTOP-k) question with Jaccard co-productive is utilized it brings about the k most encouraging information things counting persuasive information things additionally, by ordering and enhanced similitude is characterized between the k-data things of the outcome group, utilizing θ-closeness and m-closest neighbor inquiries by characterizing different limits on outcome group set liberty.

### Objectives

- To create a framework which makes use of user-centric approach for similarity computation, using Reverse Top-k query.
- To compute and use θ-similarity and m-nearest neighbor queries to validate the result.

- To design a recommendation system for efficient marketing policy.

## 3. METHODOLOGY

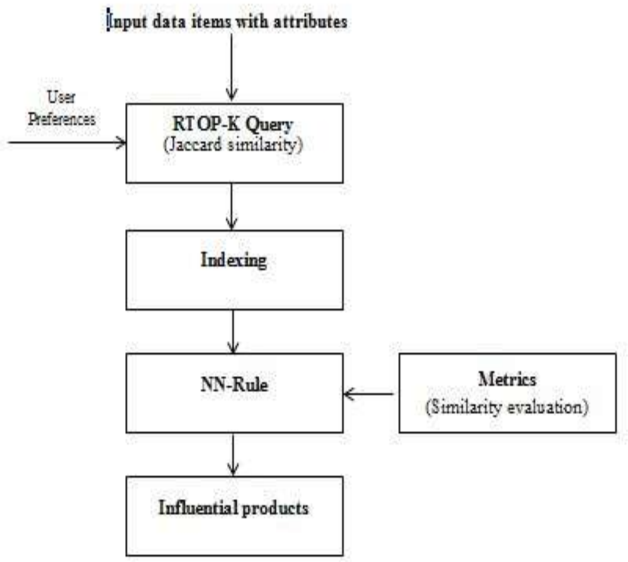


**Fig 1:** User Centric Similarity search Methodology

Breaking down the comparability between information things distinguishes the similitude in light of their properties alongside the inclinations provided by the client. The information things are handled by RTOP-k inquiry bringing about the best k set counting powerful items. To advance the things in the RTOP-k group the things are recorded and listed information group is additionally utilized with θ-likeness and m-closest neighbor inquiries. A few measurements are utilized to dissect the outcome group by changing k esteem in information thing liberty. The planned job guarantees that it proficiently recognizes the likeness between information things not withstanding when diverse similitude measurements are measured because client inclinations are measured and the closeness brings about a well fined way.

## 4. ARCHITECTURE

The suggested client driven resemblance seek application compose utilization of the above framework design. It utilizes MongoDB, a Hadoop database for storage purpose and preparing of Big data.

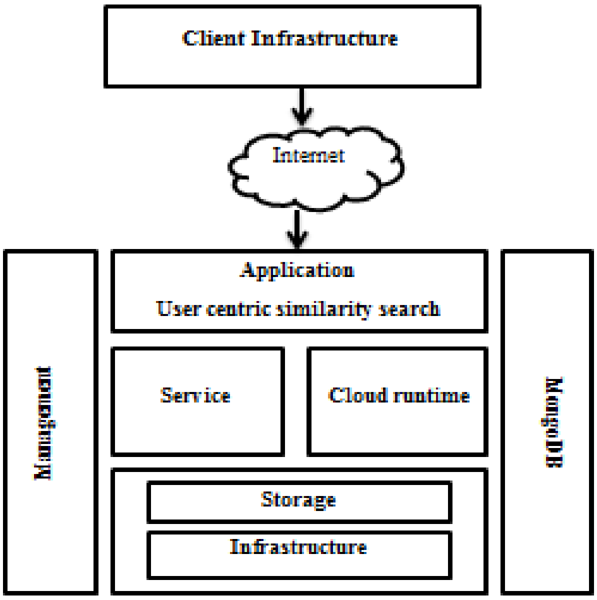


**Fig 2:** System Architecture

The user interface of the application is outlined utilizing servlets. This application is associated with gathering data on items and processes it utilizing reverse best K questions in light of client inputs, utilizing θ-similitude and closest neighbor calculation. The Products information groups are listed with previous to they are handled by θ-comparability and closest neighbor calculations. The clients and administrator are the two unique on-screen characters related with the application. Administrator has the approval to include new items and client can rate and remark the items. In light of the client see the application gives the market investigation to the administrator.

## 5. ALGORITHM

### 5.1 ThetaQuery (q, L, θ, RES)

**Input:** q is the query point
L is a priority queue
θ is the query parameter
RES is the result set

```
M = L. dequeue ()
if min_sim (M, q) ≥ θ then
    RES = RES ∪p, ∀p ∈ subtree (M)
end if
if M.type=LEAF then
    for all pi ∈ M do
        rpi = execute RTOPk(pi)
        if sim(pi,q) ≥ θ then
            RES = RES ∪pi
        end if
    end for
else
    for all Mj ∈ M do
        if max_sim (Mj, q) ≥ θ then
            L.enqueue (Mj)
        end if
        end for
end if
if L is not empty then
    ThetaQuery (q, L, θ, RES)
end if
```

### 5.2 Nearest Neighbor (q, L, m, nn)

**Input:** q is the query point
L is a priority queue
m is the number of Nearest Neighbors
nn is the list of Nearest Neighbors

```
M = L.dequeue()
if M.type=PRODUCT then
    nn.add(M)
end if
if nn.size== m then
    return nn
end if
if M.type=LEAF then
    for all pi ∈ M do
        rpi = execute RTOPk (pi)
        L.enqueue (pi)
    end for
else
    for all Mj ∈ M do
        L.enqueue (Mj)
    end for
end if
if L is not empty then
    Nearest Neighbor (q, L, m, nn)
end if
```

## 6. IMPLEMENTATION & PROTOTYPE

### Software Requirements

- Operating System: Windows 7 and Above
- Web Technology : Servlets, JSP
- IDE : Eclipse Mars
- Application server : Apache Tomcat 8.0
- Hadoop Database: MongoDB
- Database Connectivity: Robomongo-0.8.5-i386

### Hardware Requirements

- Processor: Intel I3 and Above
- RAM: 4 GB
- Hard Disk: 500 GB

The following are the procedures to be followed in designing the user centric similarity search application.

### 6.1 Pseudocode: To Upload New Product and View Products

```
Step 1: Upload the Novel Product
Construct a basic database object which accepts arguments or parameter as - product id, product type, product name, category, price, weight, manufactured data;

Step 2: Authorize the product id
Construct a database cursor object to find the specified query;
while(loop until cursor has next element)
assign product id to query
{
display the alert message as Product Id Already Exists;
}

Step 3: View Products Procedure
Construct a an object named collection to get product database
Construct a cursor object to search from the collection database;
while(until cursor has next element)
{
Construct a database object and assign its value to cursor has next;
}
```

### 6.2 Pseudocode: To look for Products

```
Enter a keyword to search an item label
Construct a basic database object which accept product name as parameter;
Construct a database object named clause2 which accept category as a parameter;
Construct a basic database list object;
Construct a database cursor object to search from collection database;
```

### 6.3 Pseudocode: To Generate Market Analysis.

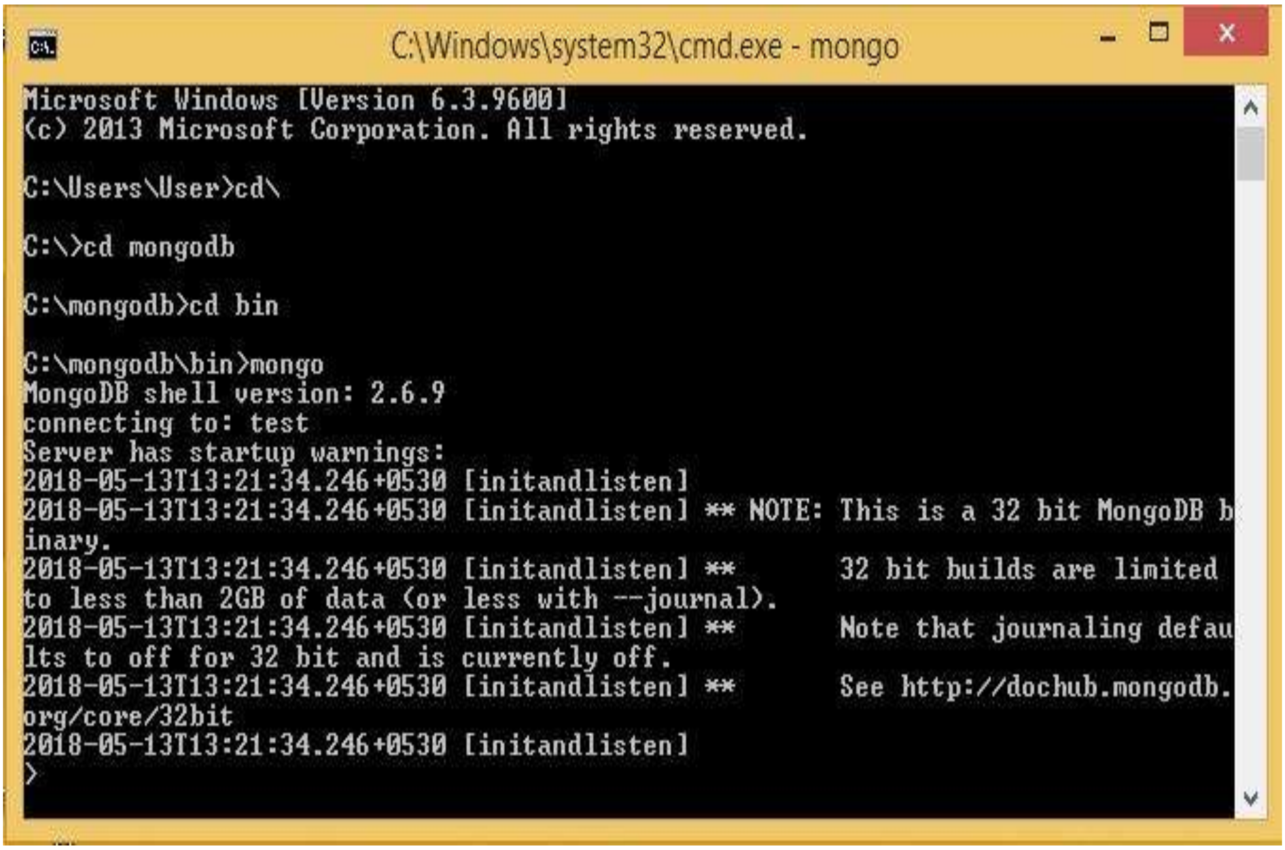
```
Step1: Client evaluates collection.
Search product based on unique product id;
Construct a database collection object to fetch the collection accepts parameter as "commands";
Construct a list of basic database object named as criteria;
Append criteria object must be appended by "count" value;
Update the collection database object by appending its name, category, new document;

Step 2: Producing the advertisement market investigation by using aggregate and group functions
Construct an array list of ranking named list;
Construct a database collection object named collection1 to fetch "commands" from database;
Construct a database object group fields by accepting arguments id and ranking;
```

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

```
Construct an object aggregation output named as output to fetch aggregation from collection database, its parameters are group and sort;
for (var doc in output results)
{
Set the group set count value as count;
insert group item to list;
}
for(var i in list.size())
{
Fetch a int i from list ;
Display product label, corresponding   product ranking, total count of the rankings        of a product;
}
```

```
Step 3: Submit or display advertisement Market analysis chart
function ()
{
construct a variable named chart is an object of canvas JavaScript;
{           data: arguments are
{// Change type to "bar", "area", "spline", "pie",
}
)chart.render();
```

## 7. RESULTS

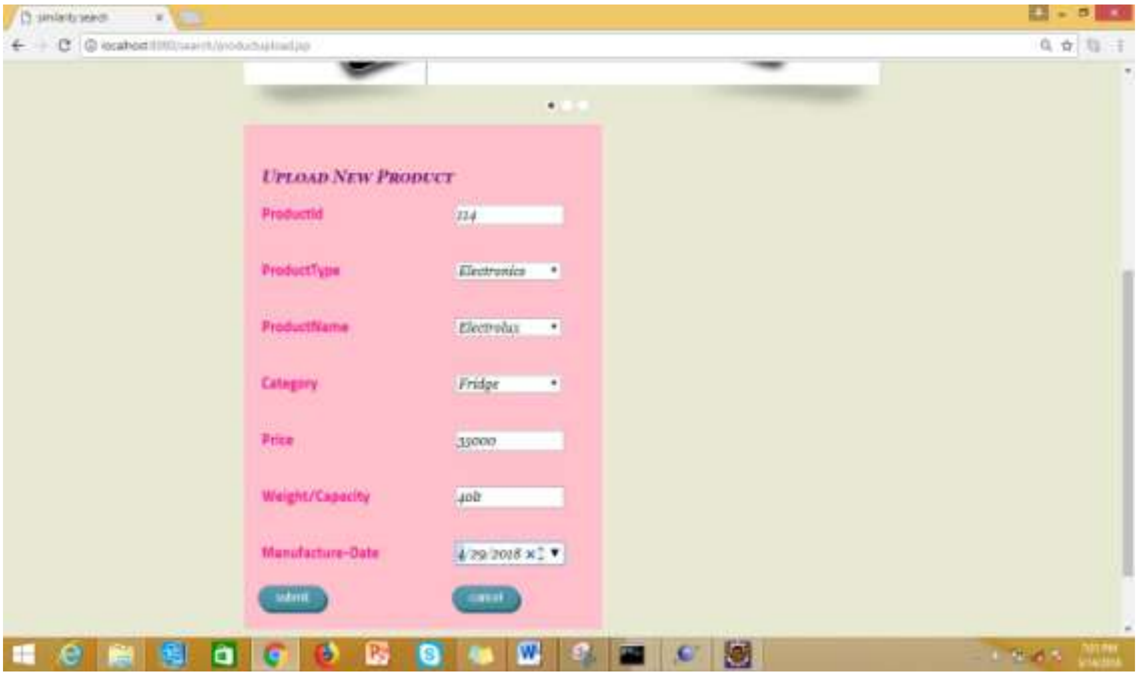


```
Microsoft Windows [Version 6.3.9600]
(c) 2013 Microsoft Corporation. All rights reserved.

C:\Users\User>cd\

C:\>cd mongodb

C:\mongodb>cd bin

C:\mongodb\bin>mongo
MongoDB shell version: 2.6.9
connecting to: test
Server has startup warnings:
2018-05-13T13:21:34.246+0530 [initandlisten]
2018-05-13T13:21:34.246+0530 [initandlisten] ** NOTE: This is a 32 bit MongoDB b
inary.
2018-05-13T13:21:34.246+0530 [initandlisten] **       32 bit builds are limited
to less than 2GB of data (or less with --journal).
2018-05-13T13:21:34.246+0530 [initandlisten] **       Note that journaling defau
lts to off for 32 bit and is currently off.
2018-05-13T13:21:34.246+0530 [initandlisten] **       See http://dochub.mongodb.
org/core/32bit
2018-05-13T13:21:34.246+0530 [initandlisten]
>
```

**Fig 3:** Launching MongoDB

MongoDB, a document based database classified as NoSQL database. It is used as database to store details of product and customers. It is launched using mongo command. It uses JSON like documents.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

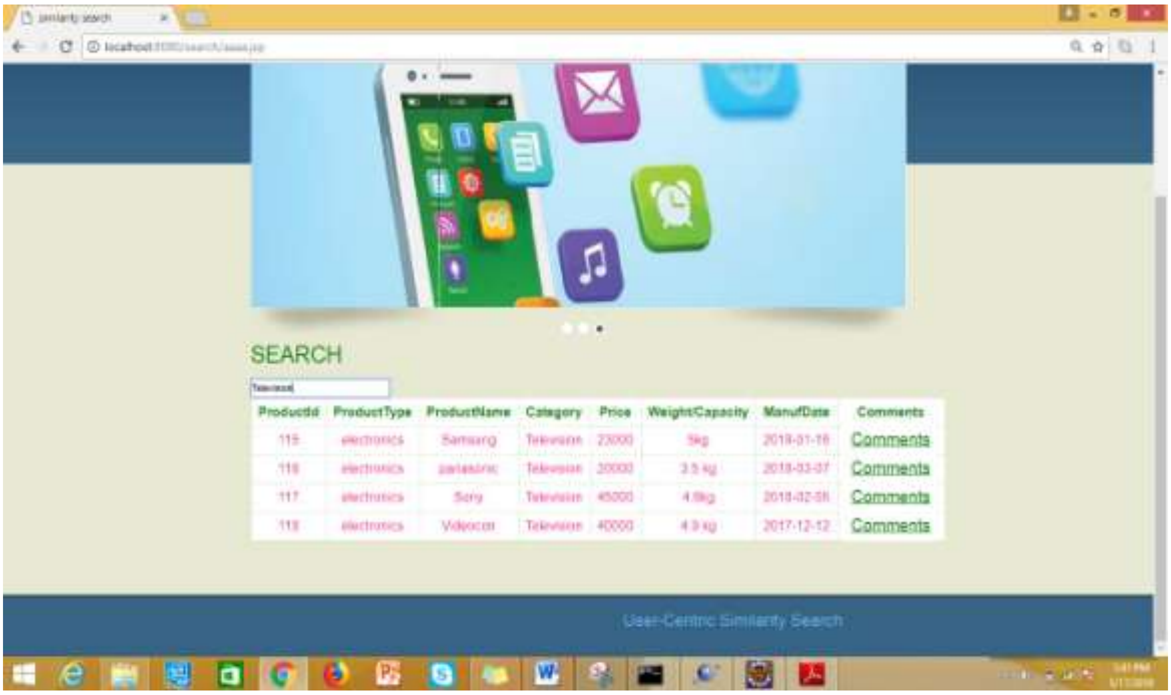


UPLOAD NEW PRODUCT

|  |  |
|---|---|
| ProductId | 114 |
| ProductType | Electronics |
| ProductName | Electrolux |
| Category | Fridge |
| Price | 35000 |
| Weight/Capacity | 40lt |
| Manufacture-Date | 4/29/2018 |

**Fig 4:** Upload New Product

Admin is allowed to upload new product by entering product features. Each ProdcutId is unique. Admin uploads a new product by entering ProductId, selecting product type, name, category, product price etc. Application validates for the unique ProductId i.e if a new product to be uploaded has the same ProductId as the existing ProductId it prompts the user saying 'ProductId already exists.



SEARCH

Television

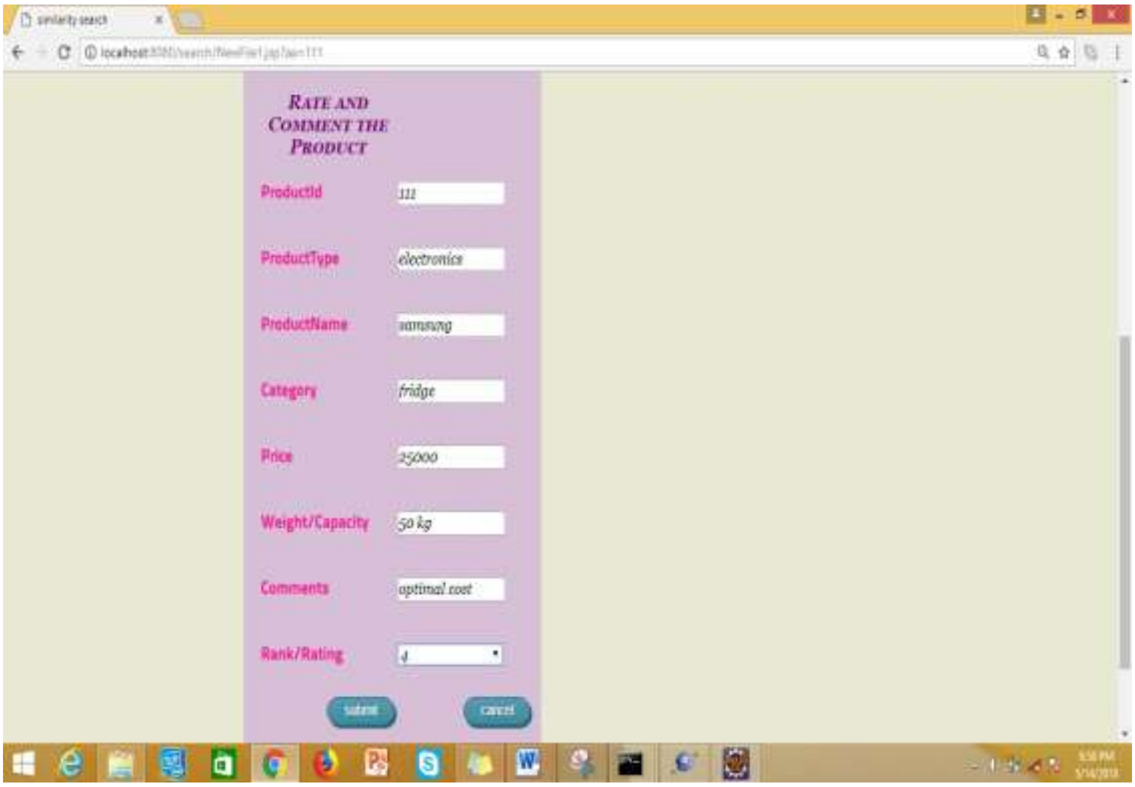
| ProductId | ProductType | ProductName | Category | Price | Weight/Capacity | ManufDate | Comments |
|---|---|---|---|---|---|---|---|
| 115 | electronics | Samsung | Television | 23000 | 5kg | 2018-01-18 | Comments |
| 116 | electronics | panasonic | Television | 20000 | 3.5 kg | 2018-03-07 | Comments |
| 117 | electronics | Sony | Television | 45000 | 4.8kg | 2018-02-08 | Comments |
| 118 | electronics | Videocon | Television | 40000 | 4.9 kg | 2017-12-12 | Comments |

similarity search
localhost:8080/search/search.jsp
User-Centric Similarity Search

**Fig 5:** Search a product by keyword

User has to enter the product keyword to search similar products based on category or product name. The products are processed by theta similitude query and nearest neighbor algorithm to display list of similar products. Once the products are displayed user is allowed to review them by clicking on comments link of particular product.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

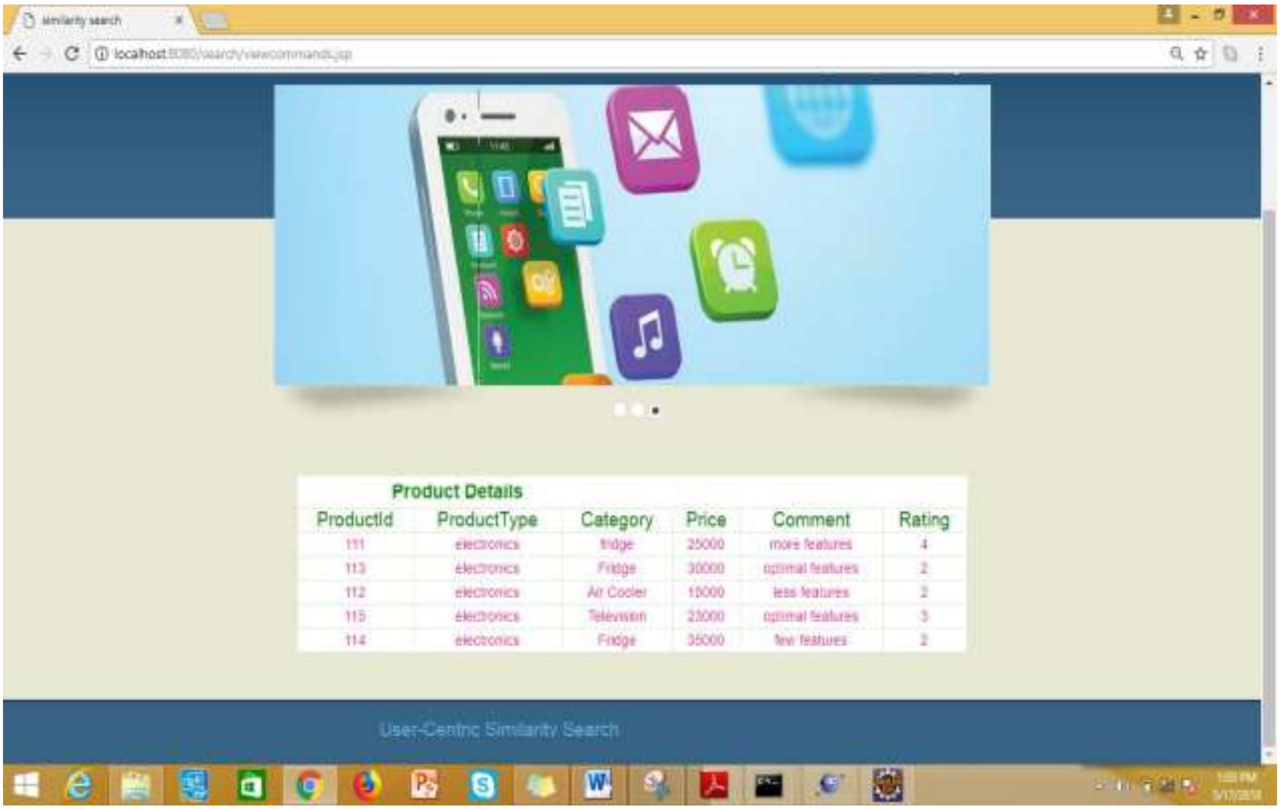


RATE AND COMMENT THE PRODUCT

| | |
|---|---|
| ProductId | 111 |
| ProductType | electronics |
| ProductName | samsung |
| Category | fridge |
| Price | 25000 |
| Weight/Capacity | 50 kg |
| Comments | optimal cost |
| Rank/Rating | 4 |

**Fig 6:** User review collection

Once the user search for similar products, he can rate and comment the product by considering the product features.

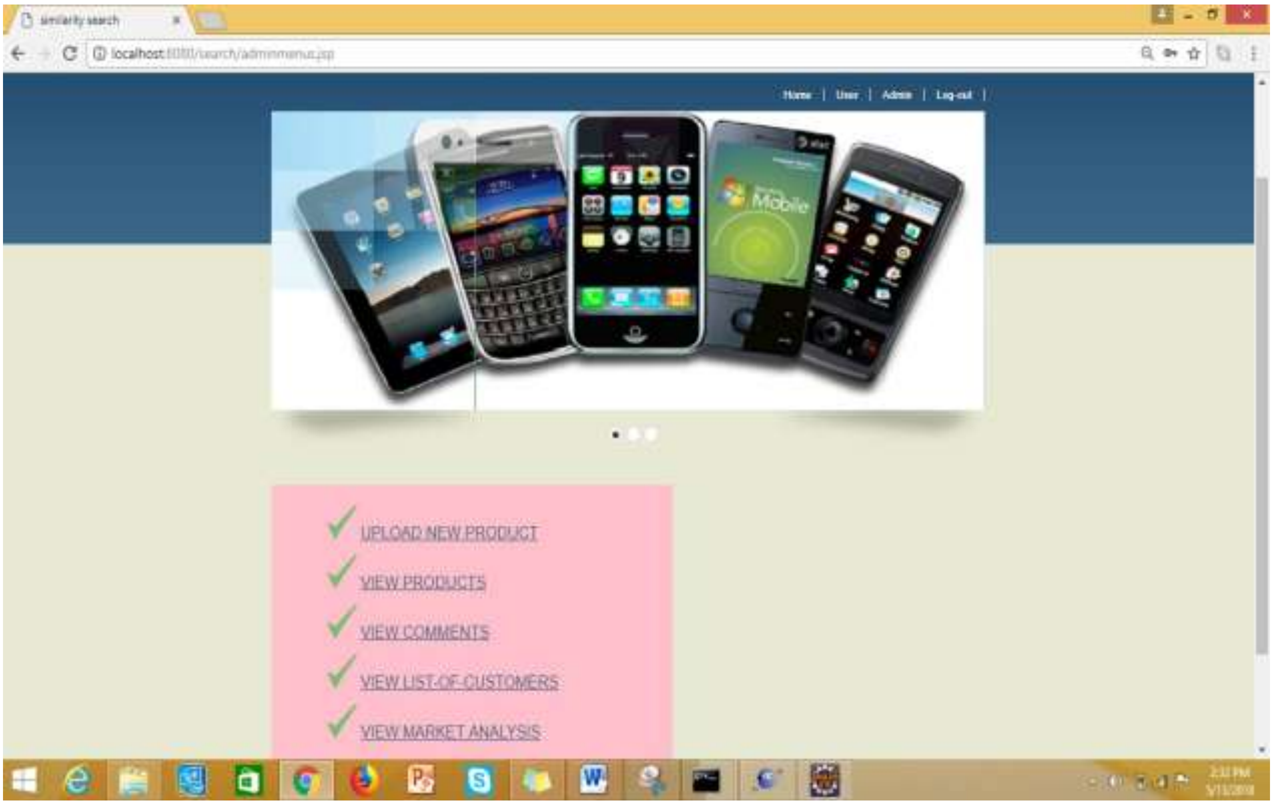


Product Details

| ProductId | ProductType | Category | Price | Comment | Rating |
|---|---|---|---|---|---|
| 111 | electronics | fridge | 25000 | more features | 4 |
| 113 | electronics | Fridge | 30000 | optimal features | 2 |
| 112 | electronics | Air Cooler | 15000 | less features | 2 |
| 115 | electronics | Television | 22000 | optimal features | 3 |
| 114 | electronics | Fridge | 35000 | few features | 2 |

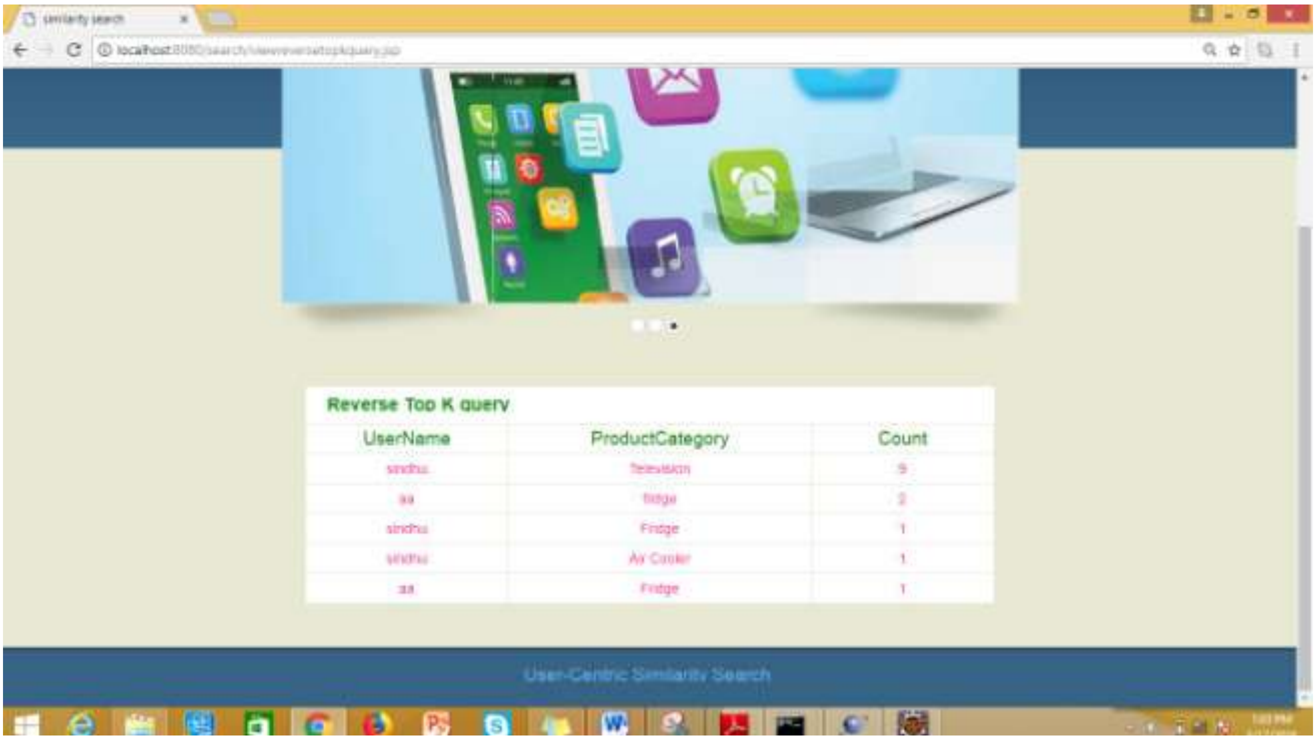
**Fig 7:** List of similar products reviewed by user

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018


similarity search
localhost:8080/search/adminmenu.jsp
Home | User | Admin | Log-out |
Mobile

- ✓ UPLOAD NEW PRODUCT
- ✓ VIEW PRODUCTS
- ✓ VIEW COMMENTS
- ✓ VIEW LIST-OF-CUSTOMERS
- ✓ VIEW MARKET ANALYSIS

**Fig 8:** Admin menus



Reverse Top K query

| UserName | ProductCategory | Count |
|---|---|---|
| sindhu | Television | 5 |
| aa | fridge | 2 |
| sindhu | Fridge | 1 |
| sindhu | Air Cooler | 1 |
| aa | Fridge | 1 |

User-Centric Similarity Search

**Fig 9:** Result of Reverse Top-K query

Reverse top k query is run on product and its related collections in MongoDB database. The result of reverse top-k query shows the user reviews of products based on product category.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

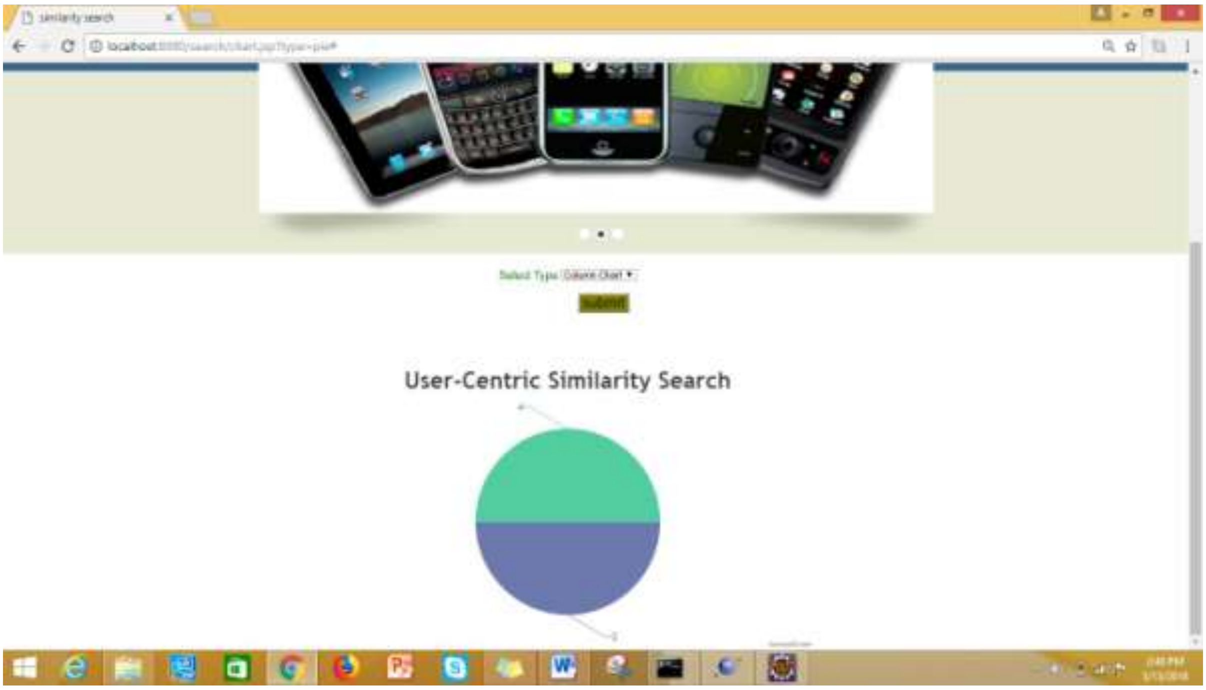


**Fig 10:** Market analysis-Pie chart(a)

The following screenshots shows market analysis in different charts and graphs like Pie chart, Column graph, Area chart, Spline chart.

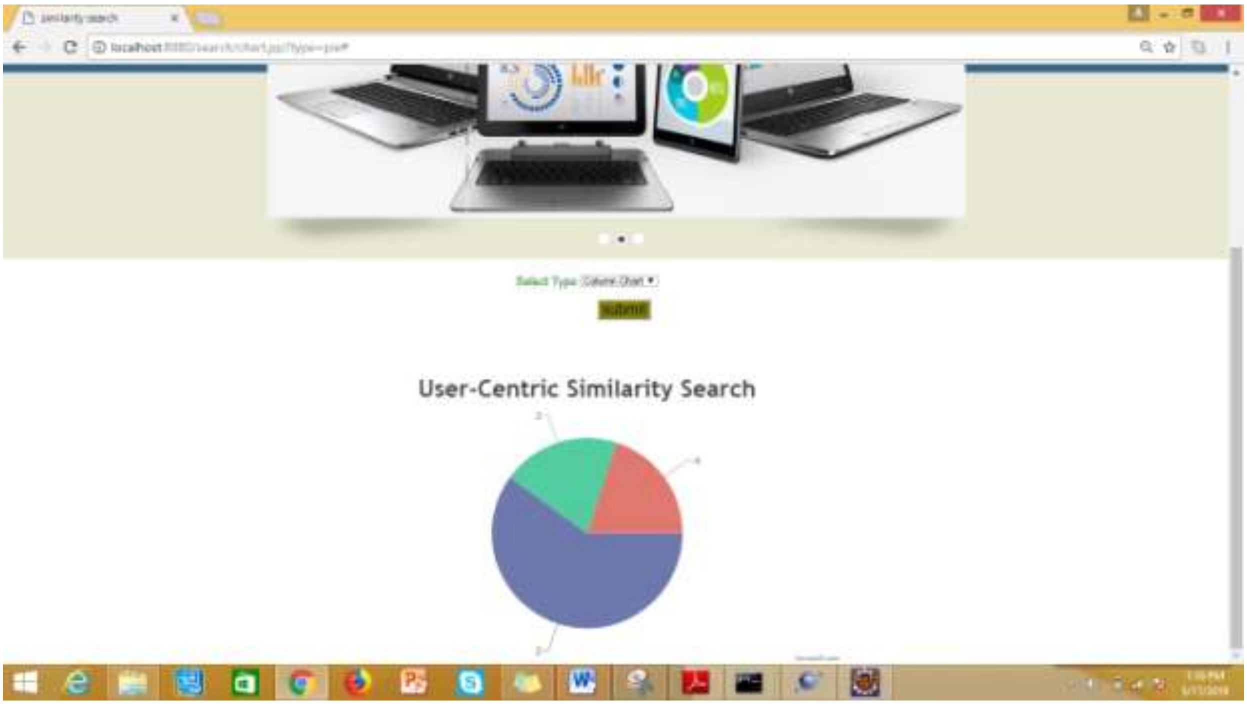


**Fig 11:** Market analysis Pie chart(b)

The above fig shows the pie chart based on user reviewed products. The different sizes of area in pie chart, represents the total number of products reviewed by users as a result of reverse top-k query.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

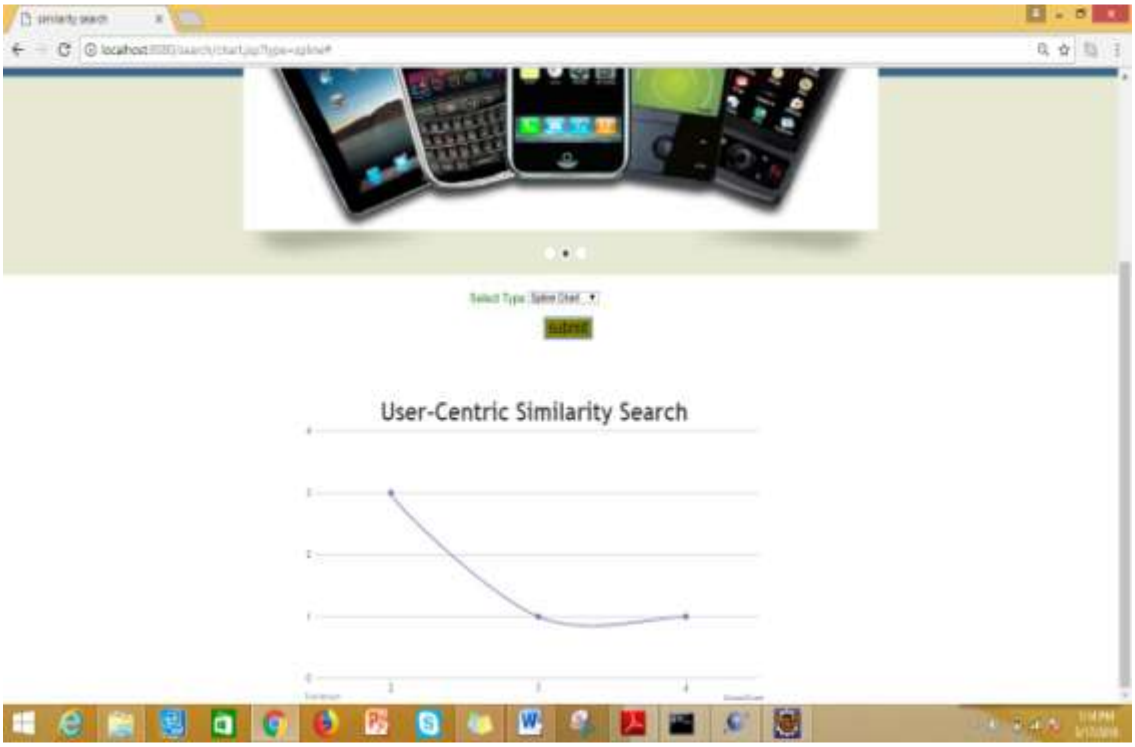

similarity search
localhost:8080/search/chart.jsp?type=spline#
Select Type Spline Chart
Submit
User-Centric Similarity Search
0
1
2
3
4
2
3
4

**Fig 12:** Market analysis-Spline chart

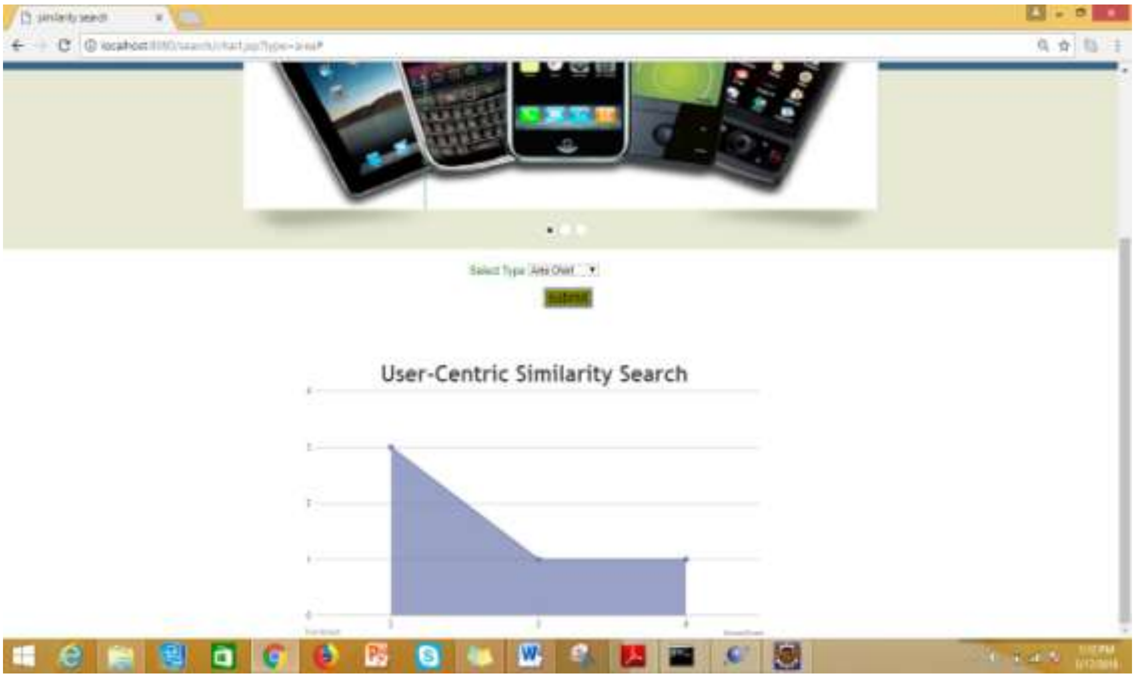
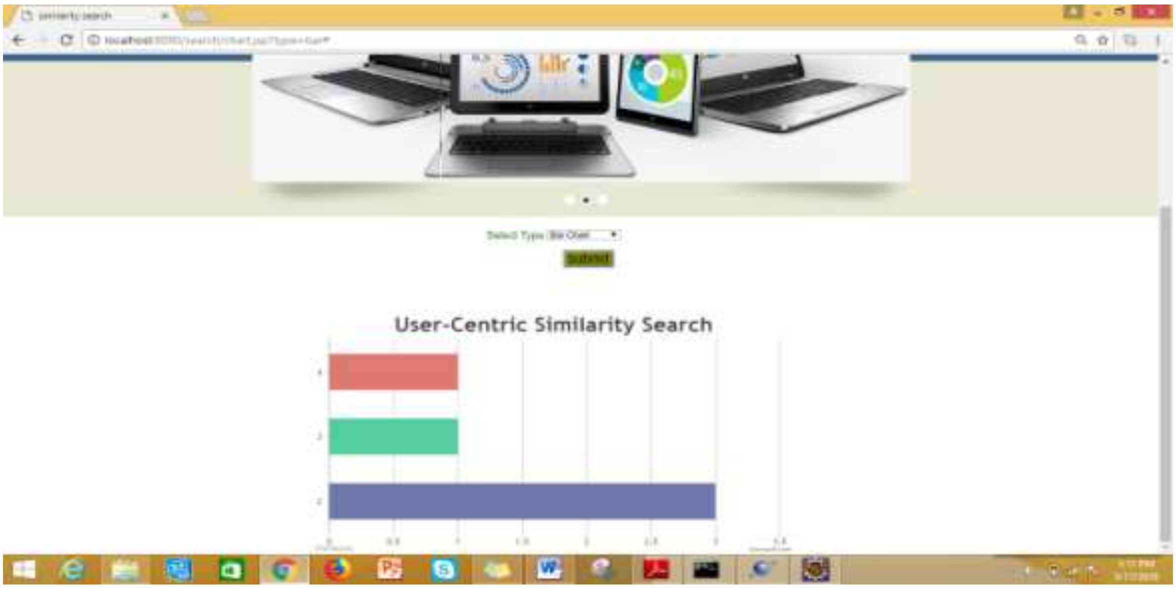

similarity search
localhost:8080/search/chart.jsp?type=area#
Select Type Area Chart
Submit
User-Centric Similarity Search
0
1
2
3
4
2
3
4

**Fig 13:** Market analysis –Area chart

The above fig shows the area chart based on user reviewed products. The different area covered for rating 2, and 4 shows the number of products rated the corresponding rating, represents the total number of products reviewed by users as a result of reverse top-k query. The figure shows that 3 products have been given rating 2.

The following column graph represents the market analysis which is the result of reverse top-k query. It shows the number of products belonging to different category, reviewed by different users. The three different bars represent the total number of products with different rating. This is computed by aggregate and grouping functions provided by MongoDB.

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

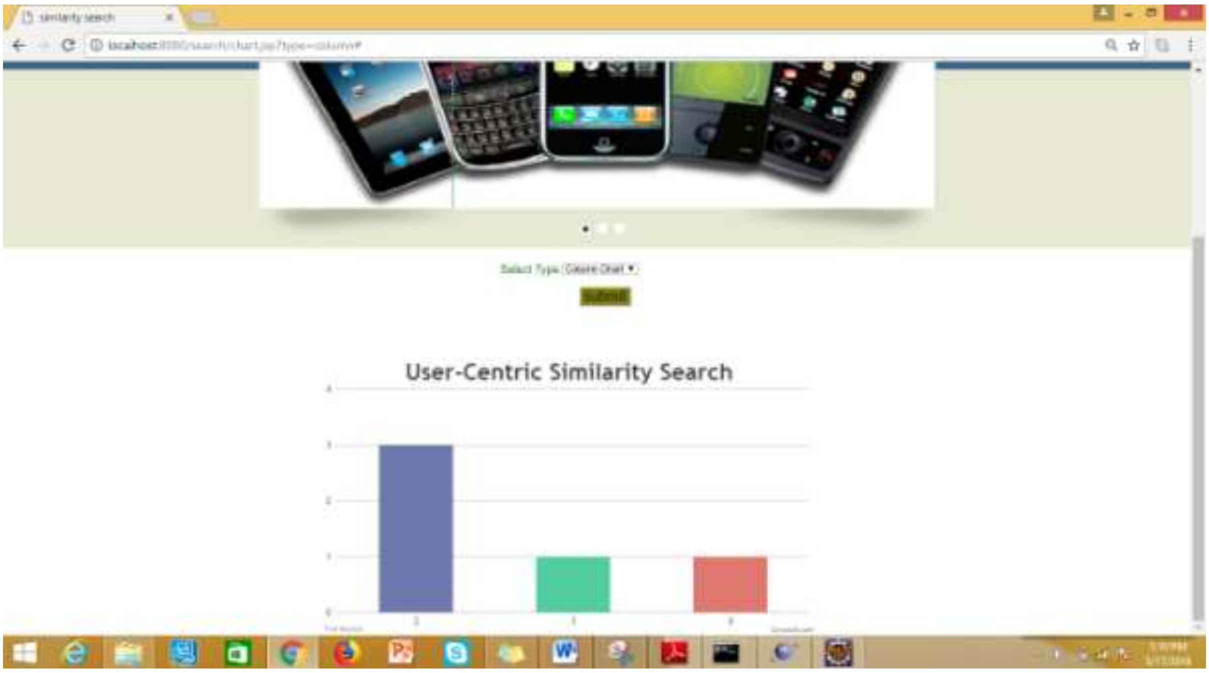


Select Type: Bar Chart

Submit

User-Centric Similarity Search

**Fig 14:** Market analysis- Bar Graph



Select Type: Column Chart

Submit

User-Centric Similarity Search

**Fig 15:** Market analysis -Column graph

## 8. CONCLUSION AND FUTURE SCOPE

Top-k inquiries are quite a while ago examined matter in the database and data recovery networks which restores the k majority encouraging items, in light of accessible client inclinations. The suggested job tends to the issue of estimating the nature of top k output set by utilizing reverse top-k inquiries that profits the clients who put an item in their top-k result outputs. This invert top-k questions are utilized to recognize persuasive items by searching the cardinality of the switch top-k result group. Consequently the invert top-k questions are successfully utilized to discover the impact of the novel item to the advertisement market which helps in market examination, while it is specifically identified with the quantity of clients that esteem a specific item.

The turnaround top-k is registered for every one of the item encased in M. The calculation price initiated is a generous supporter of the general handling expense of the calculations. In outlook, the suggested job would be upgraded by presenting improvements and other likeness measurements like expanded Jaccard closeness constant, keeping in mind the end goal to limit the quantity of turnaround top-k assessments.

## REFERENCES

https://doi.org/10.15623/ijret.2018.0707025 Received: 14-05-2018, Accepted: 25-06-2018, Published: 16-07-2018

[1] A. Rajaraman and J. D. Ullman, Mining of massive datasets.Cambridge:Cambridge Univ. Press, 2012.

[2] K. Georgoulas and Y. Kotidis, “Towards Enabling Outlier Detection in Large, High Dimensional Data Warehouses,” in Proceedings of SSDBM,2012, pp. 591–594.

[3] H. Becker, M. Naaman, and L. Gravano, “Learning similarity metrics for event identification in social media,” in Proceedings of WSDM, 2010.

[4] A. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Nørv°ag, “Reverse Top-k Queries,” in Proceedings of ICDE, 2010.

[5] A. Guttman, “R-Trees: A Dynamic Index Structure for Spatial Searching,” in Proceedings of SIGMOD, 1984, pp. 47–57.

[6] V. Hristidis, N. Koudas, and Y. Papakonstantinou, “PREFER: A System for the Efficient Execution of Multi-parametric Ranked Queries,” in Proceedings of SIGMOD, 2001, pp. 259–270.

[7] N. Roussopoulos, S. Kelley, and F. Vincent, “Nearest Neighbor queries,” in Proceedings of SIGMOD, 1995, pp. 71–79.

[8] R. Fagin, “Combining Fuzzy Information: an Overview,” SIGMOD Record, vol. 31, no. 2, pp. 109–118, 2002.

[9] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing Top-k Lists,” SIAMJ.Discrete Math., vol. 17, no. 1, pp. 134–160, 2003.

[10] A. Vlachou, C. Doulkeridis, K. Nørv°ag, and Y. Kotidis, “Identifying the Most Influential Data Objects with Reverse Top-k Queries,” PVLDB, vol. 3, no. 1, pp. 364–372, 2010.

[11] F. Korn and S. Muthukrishnan, “Influence Sets Based on Reverse Nearest Neighbor Queries,” in Proceedings of SIGMOD, 2000, pp. 201–212.

[12] E. Achtert, C. B¨ohm, P. Kr¨oger, P. Kunath, A. Pryakhin, and M. Renz, “Efficient Reverse k-Nearest Neighbor Search in Arbitrary Metric Spaces,” in Proceedings of SIGMOD, 2006, pp. 515–526.

[13] F. Korn, S. Muthukrishnan, and D. Srivastava, “Reverse Nearest Neighbor Aggregates Over Data Streams,” in Proceedings of VLDB, 2002, pp.814–825.

[14] K. C. K. Lee, B. Zheng, and W.-C. Lee, “Ranked Reverse Nearest Neighbor Search,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 7, pp. 894–910, 2008.

[15] A. Singh, H. Ferhatosmanoglu, and A. S. Tosun, “High Dimensional Reverse Nearest Neighbor Queries,” in Proceedings of CIKM, 2003, pp. 91–98.