THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

CONTENT BASED USER PREFERENCE MODELING FOR IMAGE RECOMMENDER SYSTEMS

Master Thesis

HARUN IŞIK

İSTANBUL, 2015

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

CONTENT BASED USER PREFERENCE MODELING FOR IMAGE RECOMMENDER SYSTEMS

Master Thesis

HARUN IŞIK

Supervisor: Assist. Prof. Dr. KEMAL EGEMEN ÖZDEN

İSTANBUL, 2015

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Name of the thesis:	Content Based User Preference Modeling for Image Recommender Systems
Name/Last Name of the Student:	Harun IŞIK
Date of the Defense of Thesis:	09 / 01 / 2015

The thesis has been approved by the Graduate School of Computer Engineering .

Graduate School Director:
Assoc. Prof. Dr. Nafiz ARICA
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Program Coordinator:
Assist Prof Dr. Tarkan AYDIN
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Committee Members	**Signature**
Thesis Supervisor: Assist. Prof. Dr. Kemal Egemen ÖZDEN	----------
Member: Assoc. Prof. Dr. Alper Tunga	----------
Member: Assist. Prof. Dr. Tevfik AYTEKİN	----------

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor Assist. Prof. Dr. Kemal Egemen ÖZDEN for his guidance, advice, criticism, encouragements and insight throughout the research.

I would also like to thank my parents, for their great support throughout her life. Without their understanding, and continuous support, I could have never been able to aspire for this level of education and complete this study.

İstanbul, 2015 Harun IŞIK

ABSTRACT

CONTENT BASED USER PREFERENCE MODELING FOR IMAGE RECOMMENDER SYSTEMS

HARUN IŞIK

Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Kemal Egemen Özden

January 2015, 46 pages

This thesis deals with evaluating image descriptors on whether they are useful to create a user preference model about user's taste on images and also whether these models can eventually be used in image recommender systems. Our aim is to address a simple user preference vector by using many visual descriptors of images. By means of image descriptors, we can reveal a correlation between user's taste and image features and easily build up a vector that models user's preferences. This content-based relationship may be used for image recommendation. Recommender systems can generally be considered as two headings such as content-based approaches and collaborative filtering approaches. Typical content-based methods computes content in user preference and compare it with other items. We want to use our image descriptor correlation as a content-based approach. But there are some natural challenges about this type content-based algorithm. For a very large image dataset, computing pairwise distances between vectors of image descriptors is very exhaustive process. To overcome this complexity, we have proposed a novel approach that we make cluster dataset through image feature vectors. This technique may be useful in different ways such that it speeds up image matching since you do not have to match each candidate against each image that a user likes. Also it can be able to group images very meaningfully in term of semantic according to your clustering algorithm success.

Keywords: Recommender Systems, Content Based Filtering, Image Recommendation, Image Features, Image Clustering

ÖZET

RESİM ÖNERİSİ SİSTEMLERİ İÇİN İÇERİK TABANLI KULLANICI TERCİHLERİ MODELLEME

HARUN IŞIK

Bilgisayar Mühendisliği

Tez Danışmanı: Yard. Doç. Dr. Kemal Egemen Özden

Ocak 2015, 46 sayfa

Günümüzde büyük önem kazanan öneri sistemleri üzerinde birçok araştırmacı çalışmaktadır ve bu sistemlerden biri de resim önerisi sistemleridir. Tezimizin başlıca araştırma konusu görsel resim özelliklerinin, kullanıcıların resim beğenilerinden yola çıkılarak bir kullanıcı beğeni modeli oluşturmadaki kullanılabilirliğini ortaya çıkarmaktır. Ayrıca amacımız üretilen bu modelin resim önerisi sistemlerinde kullanılması ve bu yöntemin başarı performansının ölçülmesi üzerinedir. Bilgisayar görüsü alanında daha önceden üzerinde çalışılmış ve üretilmiş birçok resim tanımlayıcısı bulunmaktadır. Bu resim tanımlayıcıların ve kullanıcıların beğendikleri resimler arasında mantıklı bir ilişki ortaya çıkartılabilir. Ortaya çıkartılan bu ilişki modellenerek sistemdeki diğer resimler üzerinden kullanıcılara öneri yapmak üzere içerik tabanlı resim önerisi sistemlerinde kullanılabilir. Bunlarla birlikte, çok fazla sayıda resim içeren veritabanlarında bu tür içerik tabanlı öneri algoritmalarını çalıştırmak fazla performanslı değildir. Birbirlerine görsel tanımlayıcılar üzerinden ikili yakınlıkları ölçmek tüm veritabanı için oldukça uzun süren bir işlemdir. Tezimizin son olarak önerdiği temel amaç ise resimlerin görsel tanımlayıcı verilerini önceden kümelemek etmek ve bu kümeleri yeni tasarladığımız resim önerisi yönteminde kullanmaktır. Bu kümeleme yöntemi resim önerisi mekanizmasının çalışma hızını artıracağı gibi öneri performansını da diğer klasik algoritmalara göre artırdığı yaptığımız deneylerle ispatlanmıştır.

Anahtar Kelimeler: Öneri Sistemleri, İçerik Tabanlı Filtreleme, Resim Önerme, Resim Özellikleri, Resim Kümeleme.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

API	:	Application Development Interface
CBIR	:	Content Based Image Retrieval
CEDD	:	Color and Edge Directivity Descriptor
CBF	:	Content Based Filtering
CF	:	Collaborative Filtering
RC	:	Recommender System

1. INTRODUCTION

In computer vision, visual descriptors describe the elementary characteristics of digital images such as the shape, the color, the texture and the motion. These descriptors provide a good information about images to search quickly and efficiently in the content of images. Nowadays, in social media, many image uploading sites have been rapidly increased and people have rated and commented on images in these website according to whether they like them. We want to evaluate image descriptors on whether these descriptors are useful in modelling user's preferences. It may be possible that we can determine user's likeness tendencies by using many image descriptors. For example, we select a user from social media and find images which is rated by this user. Then, for color descriptor of images, we try to find any correlation between rated images and any specific color histogram such as closer to red or yellow or blue frequency. Similarly, other image descriptors can be used to make decision what user's preferences look like.

Recommender systems (RCs) study for making-decision the best suggestions to the users about what items to buy, what movies to watch, what images to like and so on. Recommender systems simply offer a list of items which indicates whether users like it or not. We have proposed to use visual descriptors as we mentioned above for image recommender systems. As it is known, since given labels or metadata for images may not exactly define an image or may contain any incorrect information, image descriptors may be more productive about image recommendation. When user experience remains insufficient to decide to offer which images, using image features may be the best alternative way to the typical suggestion algorithms.

Nowadays image data and image data gathering have excessively increased. Whereas extracting useful and meaningful information from large dataset are challenging. So, there is no chance to search more effectively on image database without using data mining techniques. There are two main types of data mining tasks, that is, first one is predictive tasks and second one is descriptive tasks. We want to summarize the relationships in our image dataset over visual descriptors. Cluster analysis is the best choice to find groups of more similar items each other to which belong same cluster.

In our thesis, we aim that we want to create a simple user preference vector which models user's taste on images. We want to evaluate it against a basic content-based recommendation technique by combining clustering method. This vectorization technique can be useful in many different ways. In a typical content-based method, it try to find the closest item set based on content to which user will be given recommendation. Since we do not have to match each candidate against each image that a user likes due to min-min procedure have N (number of items) square complexity, making cluster before content-based procedure considerably speeds up time of producing any recommendation.

Finally, it should be required that we evaluate our novel approach with named as cluster based technique for image recommendation, after we implement our algorithm. General recommendation system techniques could be used to assess performance of our method. We also could recognize whether our algorithm is better by comparing other traditional algorithms. We compare our cluster based algorithm with three algorithm such as content based algorithm, collaborative filtering and random method. We expect that performance of our method should be higher than at least one of other methods. There are a lot of low level image features and we use some of them. Also we compare image features to understand which one is better because we make clusters by using image features. Moreover, we compare parameters of number of cluster to find the range which gives best recommendation results.

2. LITERATURE REVIEW

2.1 RECOMMENDER SYSTEMS

Recommender systems (RCs) make help us to search an item quickly. There has been supposed a lot of techniques for recommender systems: content based, collaborative filtering, knowledge-based, demographic techniques and hybrid approaches (Burke, 2007). There are some limitations about each of these techniques, such as the well-known cold-start problem for collaborative and content-based systems. This is due to fact that new users have with few ratings.

As these systems may specialize for each user, a common recommendation may also be made for all of them. A specialized recommender system may most likely increase diversity of suggested items and offer more funny experience to users. However non-specialized recommender systems make simply to generate top most selections regardless of any particular user preference.

2.1.1 Content Based Filtering

A basic content-based filtering algorithm make relationship item based correlation between the content of the items without any user dependency. It makes suggestions to user by computing the content in a user preference and comparing other items with this preference content. Content-based filtering does not interact a user profile with other users'. It only correlates with selected user profile and content of other items.

However, content-based approaches have some natural limitations about number of attributes or type of attributes. As it is known, since given labels or metadata for images may not exactly define an image or may contain any incorrect information, image descriptors may be more productive about image recommendation. To overcome this limitation we have proposed to use "pixel data" of our images. When we recommend an image to user, we use image features from image retrieval tools to discriminate the

images and to make decision what a user profile model. But one of challenges is there is very big image dataset and computing the pairwise distance between images as feature vector is quite difficulty process.

2.1.2 Collaborative Filtering

Collaborative filtering (CF) technique is considered to be the most applied recommendation technique due to its successful results. This technique is based on finding user-to-user correlation as opposed to content-based filtering method. Collaborative Filtering focuses on the similarity between users according to their ratings in the past. By comparing two users' taste on their history, we try to find the closest users each other. Then CF is used to recommend items to active user from other nearest user's items to him. There are several different implementation approaches deal with CF such as neighbor based approach, model based approach, memory based approach and item based approach.

There are several challenges about collaborative filtering. Data sparseness plays an important role on collaborative filtering. As CF is based on user's similarity in the past, if there is no sufficient rating information, CF cannot be able to generate any recommendation in this cases. Mostly new users may not be given any powerful suggestion, until they have enough number of ratings over time.

2.1.3 Hybrid Approaches

A hybrid recommender system combines more than one approaches to improve performance of recommendation system. For example, we can incorporate collaborative filtering and content based method so that collaborative filtering can include content information of items. Also since RCs techniques have some shortcomings, by combining multiple methods, we can compensate the limitations of one another.

2.2 DATA MINING AND RECOMMENDER SYSTEM

Data mining is a software tool that it helps to extract useful and meaningful information from large dataset. It is simply used to find patterns which are correlated among them over some attributes. And also data mining may be able to predict result of future dataset and to describe already existing relationship between objects in dataset. In the context of this thesis, data mining is used as the term of cluster analysis to build up our image recommendation models.

Image recommender system we have proposed cooperate with data mining concepts make its recommendation using cluster analysis. Our system is mainly based on building of user profile that is able to associate with visual descriptors from content based image retrieval.

2.2.1 Cluster Analysis

There are a lot of clustering algorithm in data mining. We used only k-means clustering that is the most prominent clustering method.

2.2.1.1 K-Means

The K-means clustering technique is very easy to understand how it works. K-means defines a centroid based group of objects and assign these objects to previously marked centroid. Firstly, it selects K initial centroids, where K is a user-specified parameter. K is the number of cluster how many you want to be as output. Each object is then assigned to the closest centroid, and each group of objects are formed a cluster. After all objects are iterated, centroids of clusters are recalculated and previous steps are run again. Until the centroids of clusters remain the same, we repeat the algorithm steps and eventually we make output the last result of clusters.

Pseudo-code of basic K-means algorithm:

1. Select K points as initial centroids.
2. **repeat**
3. Form K clusters by assigning each point to its closest centroid.
4. Re-compute the centroid of each cluster (mean of object in a cluster)
5. **Until** Centroids do not change.

2.2.2 Data Similarity Metrics

According to our data types, we should select the most correct similarity metric to calculate distance of image descriptor vector. For continues float data types in a specific range, we could use Euclidian similarity. For binary or discrete data types, we could use Jaccard similarity.

2.2.2.1 Euclidian Similarity

The Euclidian Distance can be applied on two or three or higher dimension of vector representation. Its familiar formula is:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

2.2.2.2 Jaccard Similarity (Tanimoto)

Suppose that x and y are the vector of two objects which refers to an image feature. If these vectors contain binary attributes correspond each pixel data in image, 1 indicates that the value exist in this pixel, while 0 indicates that the value is empty. Jaccard similarity says that these two vectors are close if and only if 1-1 attributes are matching at the same index. Its formula is:

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 mathces}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

This formula is only appropriate for data which has binary attributes. There is one other definition of Tanimoto similarity we used in calculation of distance of CEDD vectors.

$$f(A, B) = \frac{A \,.\, B}{|A|^2 + |B|^2 - A.B}$$

2.3 IMAGE FEATURES

There are a lot of low level and high level image features which are issued by many researchers before. We will mention about some low level image feature and use them in this thesis.

2.3.1 Simple Color Histogram

Simple color histogram specifies the color distribution of an image. A color histogram contains the number of pixels that have color values, for a digital image. These color values are in the ranges of the color space of image which is the set of all possible colors. The color histogram can be considered as a statistic that is an approximation of a continuous distribution of color values.

2.3.2 Basic Image Feature

This feature consist of eight basic image feature for image analysis and retrieval.

Brightness is an attribute of relative visual perception which is reflected by luminance of a light source. It is not a color property.

Clipping can be expressed as a result of processing an image where the intensity of a particular field of an image is maximum and minimum.

Contrast is the difference in the color and luminance of the objects that makes them separable from others.

Hue is one of the main properties of a color that is a more technical definition of our color perception which can be used to refer pure colors within a color space.

Saturation is the intensity in range from gray to pure color (hue) while lightness remains constant. A hue can be said to be a fully saturated.

In addition to above three more features exist in basic image feature, called as **complexity**, **skew** and **energy**

2.3.3 Color and Edge Directivity Descriptor (CEDD)

This feature is low level feature that is extracted from the images. Color and Edge Directivity Descriptor includes color and texture information in a histogram. CEDD size is limited to 54 bytes per image. If it needs to compare with the most MPEG-7 descriptor, CEDD has the less computational complexity during its extraction from an image. Performance of CEDD is evaluated some objective measure called ANMRR (Chatzichristofis & Boutalis, 2008).

2.3.4 Gist Descriptor

The GIST descriptor has recently taking into consideration in the context of scene recognition. It estimates shape of a real world scene such as a street, a landscape or a building. GIST represents low level dimensional properties of a scene which is some perceptual dimensions such as naturalness, openness, roughness, expansion, ruggedness (Olivia & Torralba, 2001).

3. METHODOLOGY

3.1 OUR MOTIVATION

We firstly want to address how user preferences are modelled according to their tastes on images. By using visual descriptors, we can reveal correlation between a user preference and image features. We have mentioned about several image features above such as simple color feature, basic image feature, CEDD and GIST. For example, if a user highly rates particular colored image like mostly red and blue, then we can easily build up this user's preference model by means of simple color feature. Whenever a new image which is close to this model in term of color comes to system, we can recommend to this user to see. Similarly, if a user's images that he liked are about similar theme like landscape, we can detect this tendency by using GIST feature and make a model for this user. We can eventually use these models extracted from user preferences in image recommender systems.

We have simply proposed to make cluster image dataset before applying content-based approach while making suggestion to users for a recommender system. We claim that this clustering method firstly provide more efficient way than computing distances to find any close item to a user preference. Also it secondly provides a cluster profile about a user and even though a user has too few number of rating that is having sparse data, our recommender system may give a suggestion more accurately through this user's cluster profile. "Clustering provides an abstraction from data objects to the clusters in which those data objects reside" (Tan & Steinbach & Kumar, 2006). We may think clustering as data preprocessing before applying content-based algorithm. Also, in the context of utility, we may find the most representative cluster model for a user rather than individual objects.

After clustering process, we have content-based clusters and we know which user ratings are in which clusters. Then we can draw a cluster histogram of a user for visualization. For example, we divide our image dataset into 30 clusters and sort them descending order for each user according to including number of ratings. My most

favorite cluster is fifth one. Our recommender system give a suggestion to me set of image that I have not seen yet before from fifth cluster. Consequently, we have experienced that by using clustering method we are finding items are to be suggested in a very efficient way and obtaining more accurate results.

3.2 IMPLEMENTATION STEPS

We firstly search on the internet for image website and start to gather image dataset. Flickr is one of the most popular image website and we downloaded numerous image and user rate information from this website. After we download images, we extract image feature by using some open source content based image retrieval (CBIR) tools and Matlab. Afterwards, we consider about cluster analysis and select a cluster method. Then we have clustered image dataset and user information so that we make recommendation to the users. Finally we recommend a list of image to users in system to like them.

3.2.1 Collecting Data

In order to perform our experiments about our proposition, we need to have image dataset which contains JPEG formatted images in RGB color space and user rating information. User data may simply consist of whether a user likes an image or not. If a user does not rate an image, it does not mean user not to dislike this image. He may have not seen this image yet. We have gathered our image dataset from “Flickr” website, which is well-known photo upload website, by using their API. Flickr offers an API to download their images and to get other user information related with images. We wrote down a java code by means of Flickr api and downloaded two dataset as small and big one. One of them contains about 300 images, while other contains about 1500 images. And we run our test on both dataset.

You can get more detailed information from Flickr api website:

https://www.flickr.com/services/api/

3.2.2 Extracting Image Features

We want to introduce concisely content based image retrieval (CBIR) system. CBIR is a field of computer vision and CBIR systems have been developed to search digital images in large databases. CBIR deals with content analysis of an image rather than the metadata such as labels, tags and other descriptions. There are many CBIR tools and research projects in order to get visual descriptors.

During our experiment, we have used open source LIRE application to extract image features from images. The LIRE library offers an easy Java API to retrieve images based on their texture and color characteristics. We extracted several image features by using LIRE, mentioned in chapter 2.3, which are simple color histogram, basic image feature and color and edge directivity descriptor. Also GIST descriptor is extracted by using Matlab.

In order to get more detailed information, see:

For LIRE, http://www.lire-project.net/

For GIST, http://people.csail.mit.edu/torralba/code/spatialenvelope/

3.2.3 Cluster Analysis

There are many clustering algorithms in data mining and to decide which one is to use is a sophisticated task. You should know your characteristics of your data and choose the most suitable method for your dataset. We have decided to use K-means clustering algorithm since we need number of groups of objects to make a specific model for each user for our recommender system. K-means is a partitioning clustering method. They construct as number of distance-based partitions as given k. As the result of good partitioning, items that are in the same cluster are close as much as possible, while items that are in other clusters are unlikely different. By building as such distinct clusters, we try to observe whether a user's taste on image is fit into particular model.

4. EVALUATION

4.1 DATASET

Firstly, we will describe our data structure on which we study. We downloaded some image dataset from Flickr website. Flickr is a widely used for uploading and rating photos from many people over the world. This image dataset also contains some user information such as userid and user rates. User rate means whether a user like an image or not. If a user rates an image, then this user saw it before and liked it (relevant to user). And also if a user do not rate an image, we do not know this user does not like the image or not to see yet. Sample user-image matrix is below:

Table 4.1: User – Image Matrix

	Image1	Image2	Image3	Image4	….	….	Image-N
User1	+	+		+			
User2						+	+
User3	+		+		+	+	
….							+
User-N	+		+		+		+

Source: This table is drawn by Harun Işık.

This dataset is named as structured representation because all attributes have been exactly matched user rows. Secondly, we describe our images as feature vectors consist of different size and types of numerical data. In chapter 2.4, we have mentioned about some image features we will use in our implementation. As we give detailed information about data types of these features:

Simple Color Histogram consists of size of 512 integer values:
Image#1 512 [1,1,0,0,1,0,1,0,0,0,,0,0,0,0…….……..0,0,28,0,0,0,0,0,0,0,8,48]

Basic Image Feature consists of size of 8 double values:

Image#1 8 [0.560, 6.5821, 0.536, 0.2, 0.1256, 0.1239, -0.0794, 0.00509

CEDD feature consist of size of 144 integer values:

Image#1 144 [1,1,0,0,1,0,1,0,0,0,0,0,0,0…….……0,0,2,0,0,0,0,0,0,0,8,4]

GIST consists of size of 512 double values:

Image#1 512 [0.1560, 0.2821, 0.536, 0.222…..……0.416, 0.9239, 0.454, 0.098]

4.2 PRECISION AND RECALL

In pattern recognition and recommender systems recall and precision measurement is commonly used as evaluation metric to understand performance of suggestions of system. Precision is defined as the ratio of retrieved items that are relevant, while recall is the ratio of relevant items that are retrieved. Suppose that you have size of 5 images that user liked before and you give size of 20 images to user as recommendation and only 3 of these are correct. Then your program's precision is 3 / 20 and recall is 3 / 5. As it is seen, recall may be maximum 5 that is number of relevant items. As size of list of suggestion is getting increased, recall value increases and eventually converges to 1. Whereas recall is so, the more you give list of items", the more precision is low and converge to 0.

4.3 DIVERSITY MEASURE

Diversity means how many items we distinctly recommend to users at overall. It is important that a particular user is not given suggestion same item set at different times. Mostly business departments want to see diverse item set as much as possible. Therefore, we understand large diversity number is a positive thing.

4.4 EVALUATION METHOD

Finally, we have clustered image dataset and user information about whether users like images. According to our simple proposition, we select a user from system and find his favorite cluster that contains his most rated image. And eventually we recommend a set of items from this favorite cluster. We have to evaluate this recommendation in term of accuracy rate. To do this aim, we set up an experiment set and divide image dataset into two partitions as test set and training set. We remove rating information in test set and run our algorithm on training set. We recommend increasingly number of image to thr users and try to get hit on test dataset. The bigger we get hit rate, the more successful our algorithm means. Also this is called as recall measurement.

Our pseudo code is:

```
1.  Input = 1500 images
2.
3.  Cluster (1500 images)
4.
5.  Split_Preference_Data() // output as test and training sets
6.
7.  recall = 0;
8.
9.  for all users in test set {
10.
11.     recList = Recommend_N_Image(topN, training set);
12.
13.     hit = 0;
14.     for all images in recList {
15.        if user_images() contains recList.image
16.            hit ++;
17.      }
18.     recall = hit / user_images_size();
19. }
20.
21. avgRecall = recall / test_users_size();
22.
23. return avgRecall;
```

Explanations of methods in pseudo-code:

Line 1: We downloaded about 1500 images and user rate information from Flickr website. Images are in jpg formatted and size of 640 pixels. We used this image dataset during all experiment.

Line 3: We cluster our dataset before we run recommendation section. We use R tool for making cluster. We select only one clustering algorithm that is K-means. Basic K-means algorithm takes parameter of cluster number. We try different cluster number from 30, 40, 50 and 60. We observe different results of hit rate about different parameter of cluster number.

Line5: Split preference data is an important step of the evaluation method. We should divide our dataset into two parts as test data and training data. It is preferred that test data is about 15 percentage while training data is about 85 percentage. Recommendation algorithm is run at training dataset not test dataset. User's rating information is removed from test dataset so that it can be used as unseen dataset whose user. And eventually test dataset is used in order to evaluate hit rates of recommendation algorithm.

Line 7: Recall value in this line indicate sum of values of recall which is belonged to each user in test data. After total recall is calculated, average of these recalls is returned.

Line 9: We iterate all users in test data and give a recommendation list for each user.

Line 11: Recommend N image method in this line is the heart of our algorithm. By using clusters we made before, we give a recommendation list as many number of given N parameter. It starts from the most favorite cluster of current user in test to the least rated cluster until number of N is reached. Finally number of N images are recommended and it is processed randomly.

Line from 13 to 18: Hit rate is counted in test items and recall value is calculated that hit count is dividing by total image number of current user in test. This recall value is only for one user.

Line 21: we add recall values are calculated in line 18 and we find average of these recall values are per user. This average recall value is the final result for given N parameter.

5. EXPERIMENTAL RESULTS

In order to evaluate our new recommendation method, we make several experiments by using recommender system techniques. We calculate precision and recall measurements to compare our algorithm with other traditional algorithms such as collaborative filtering, content based method and random recommendation. For our algorithm, some image features give more successful hit rates than others. Also we compare image features with each other through our algorithm and content based algorithm. Finally we compare cluster number parameters such as 30, 40, 50, and 60 clusters. We try to observe changes on precision-recall values while cluster number is increasing.

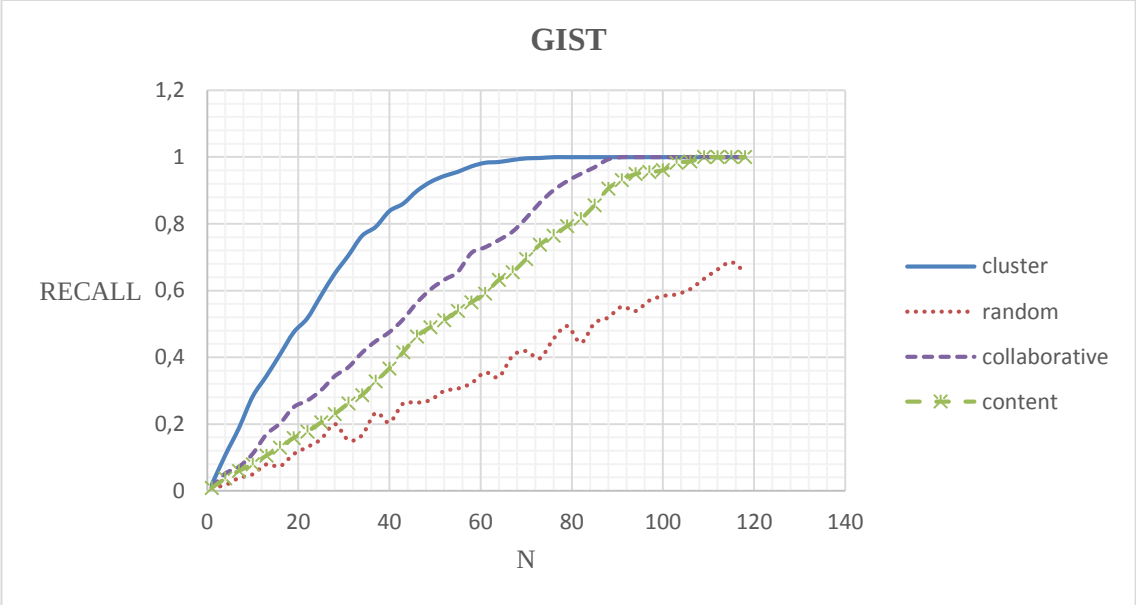
5.1 COMPARING ALGORITHMS OVER FEATURES

We used three recommendation algorithms except ours during our experiments. We separately compare these algorithms with our algorithms for each image feature we used.

5.1.1 Gist

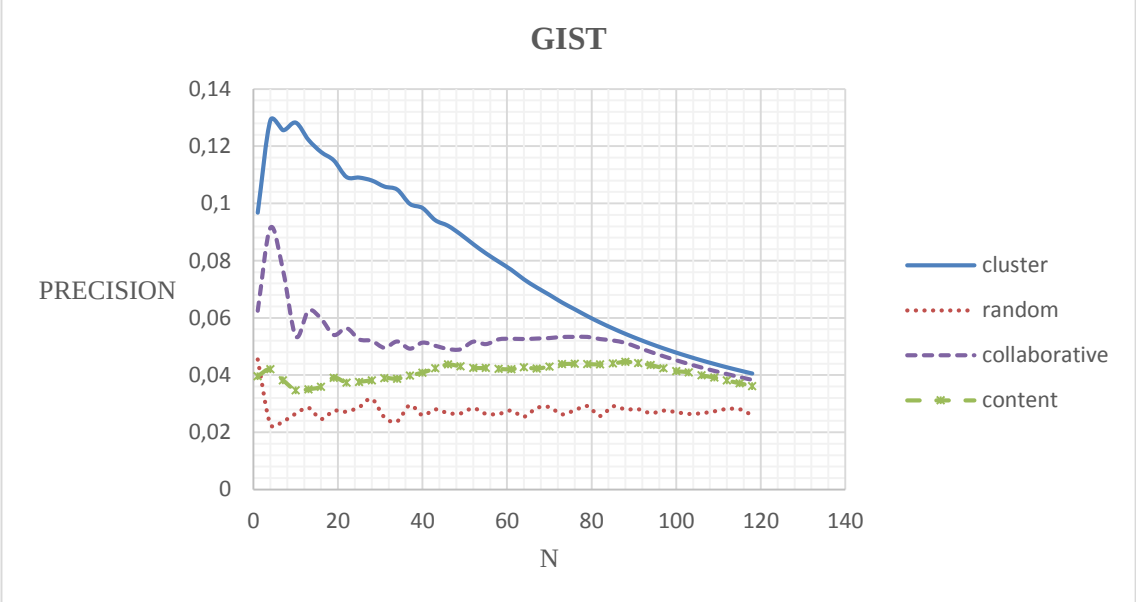
The GIST is a low level feature and it contains shape recognition information and semantic meaning information. So we expect that its hit rates are to be high when GIST is used.

Figure 5.1: Recall - N graphic of GIST feature



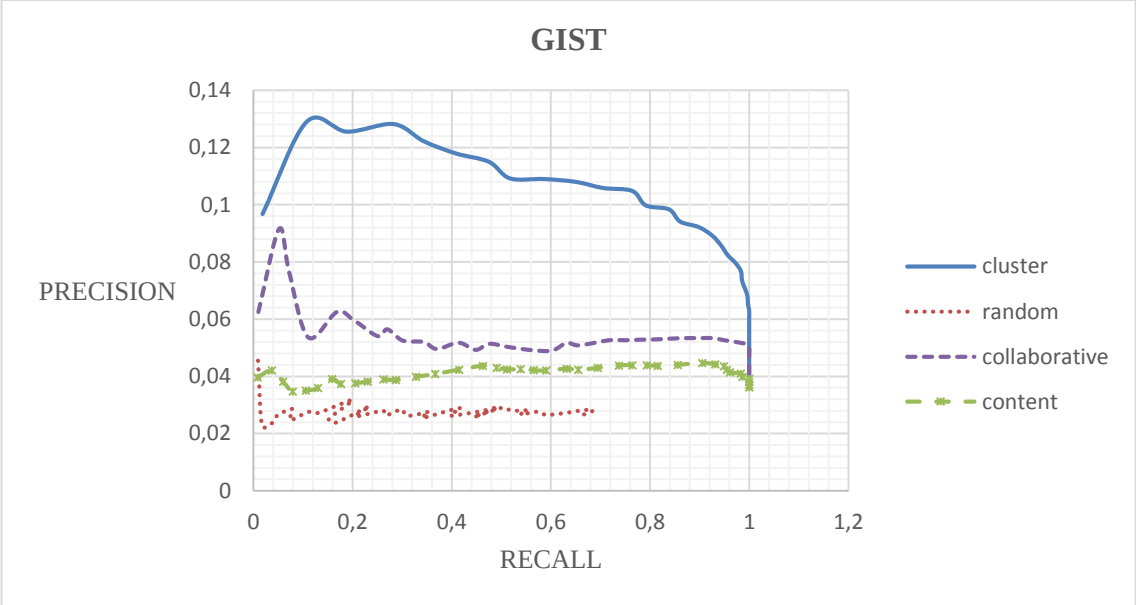
Source: Created by Harun Işık.

Figure 5.2: Precision - N Graphic of GIST feature



Source: Created by Harun Işık.

Figure 5.3: Precision - Recall Graphic of GIST feature



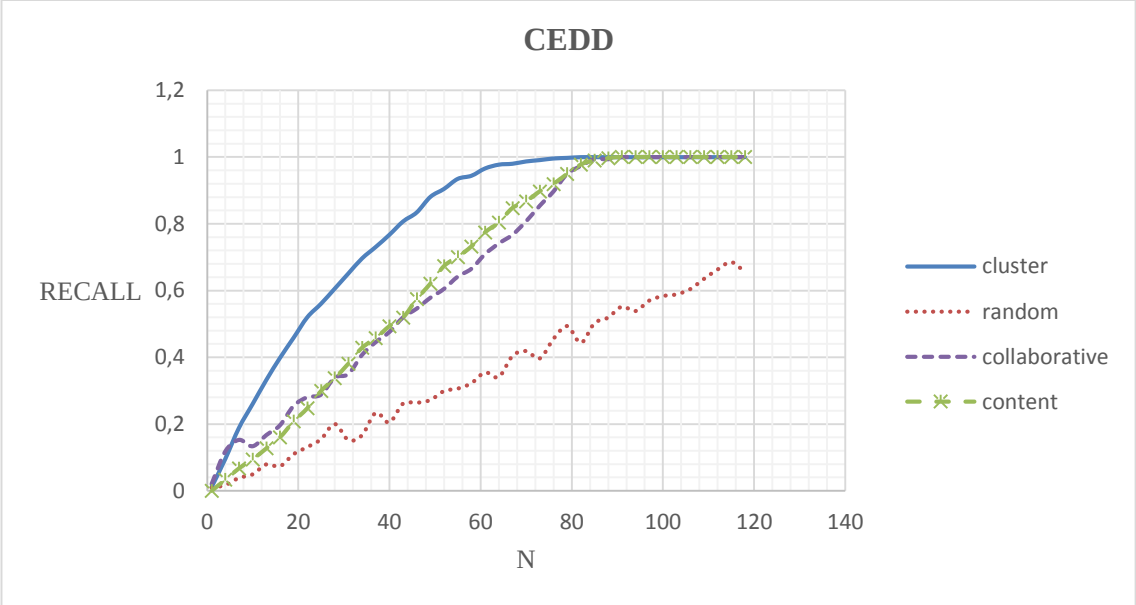
Source: Created by Harun Işık.

In figure 5.1, 5.2 and 5.3, we see that recall values of cluster based method exceeds values of other methods. GIST feature is perfectly able to make group image dataset in view of user's taste. Our cluster based algorithm is more successful than collaborative filtering and content based algorithm. When we divide image dataset into clusters by using GIST feature, we more accurately estimate images which are to be liked by users.

5.1.2 Cedd

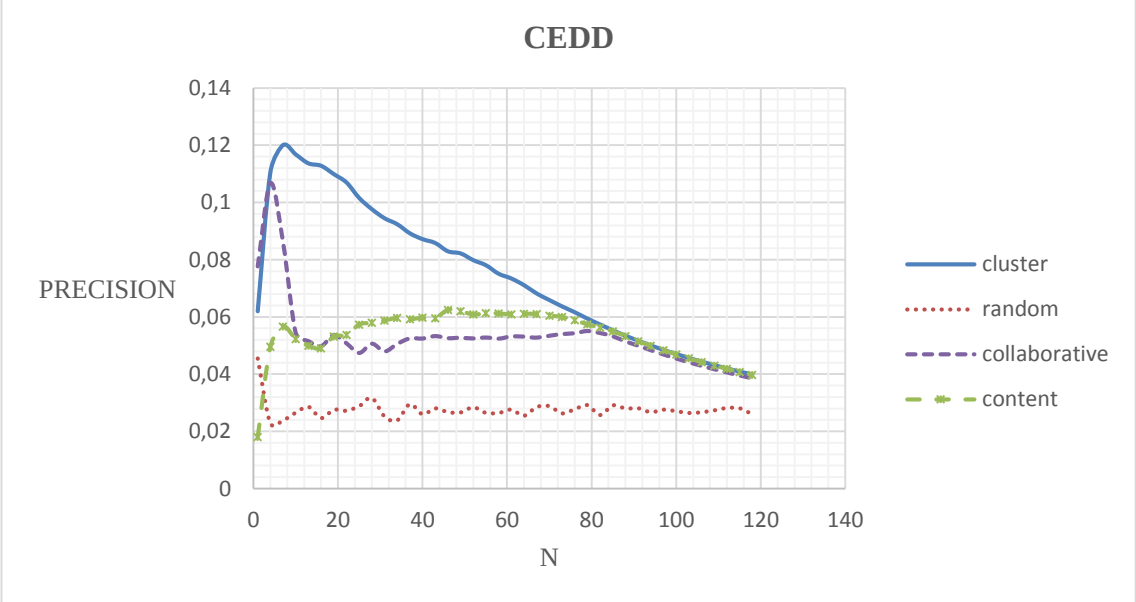
The CEDD feature stands for color and edge directivity descriptor. As it is mentioned in chapter 2.3.3, CEDD incorporates color and texture information in a histogram. This feature is more powerful than simple color and texture features.

Figure 5.4: Recall - N graphic of CEDD feature



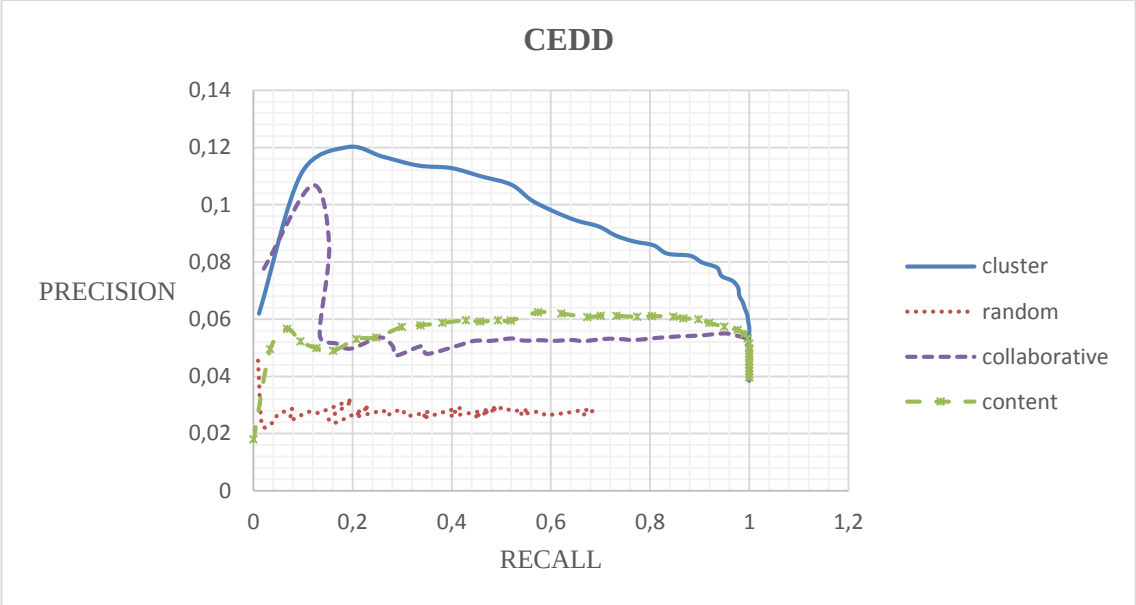
Source: Created by Harun Işık.

Figure 5.5: Precision - N graphic of CEDD feature



Source: Created by Harun Işık.

Figure 5.6: Precision - Recall graphic of CEDD feature



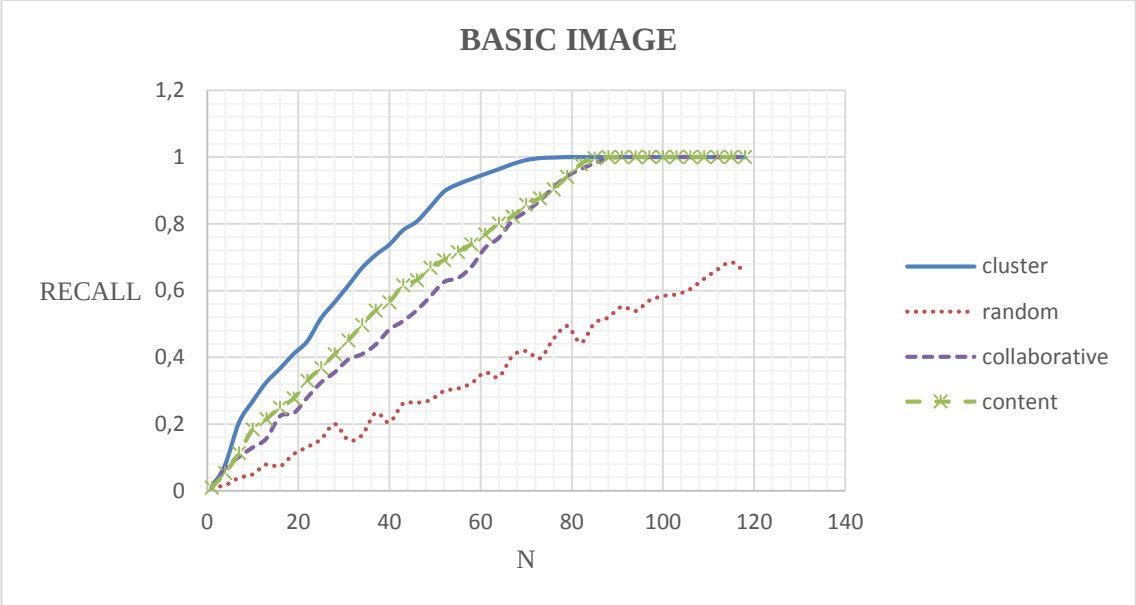
Source: Created by Harun Işık.

In figure 5.4, 5.5 and 5.6, we compare recommendation algorithms with each other by using CEDD features. Similarly in GIST feature, CEDD feature has also more successful hit rates than other algorithms. While content based algorithm is a little high in comparison with collaborative filtering method, cluster based method exceeds both of them.

5.1.3 Basic Image Features

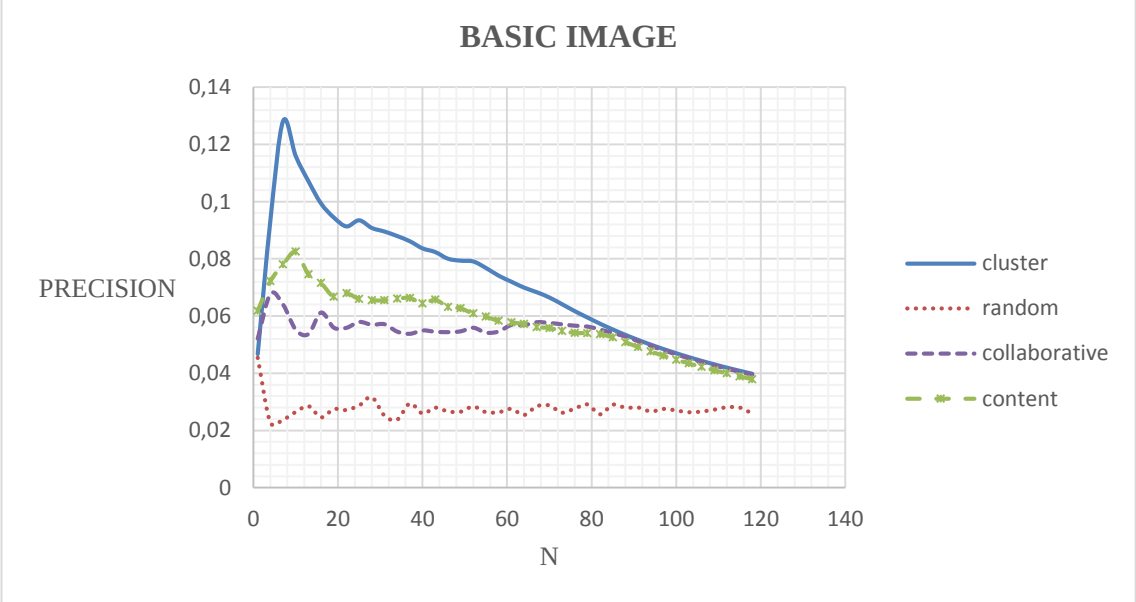
This feature contains basic image descriptors such as hue, contrast, saturation, brightness and some others. These information may be useful in making meaningful clusters.

Figure 5.7: Recall - N graphic of Basic Image feature



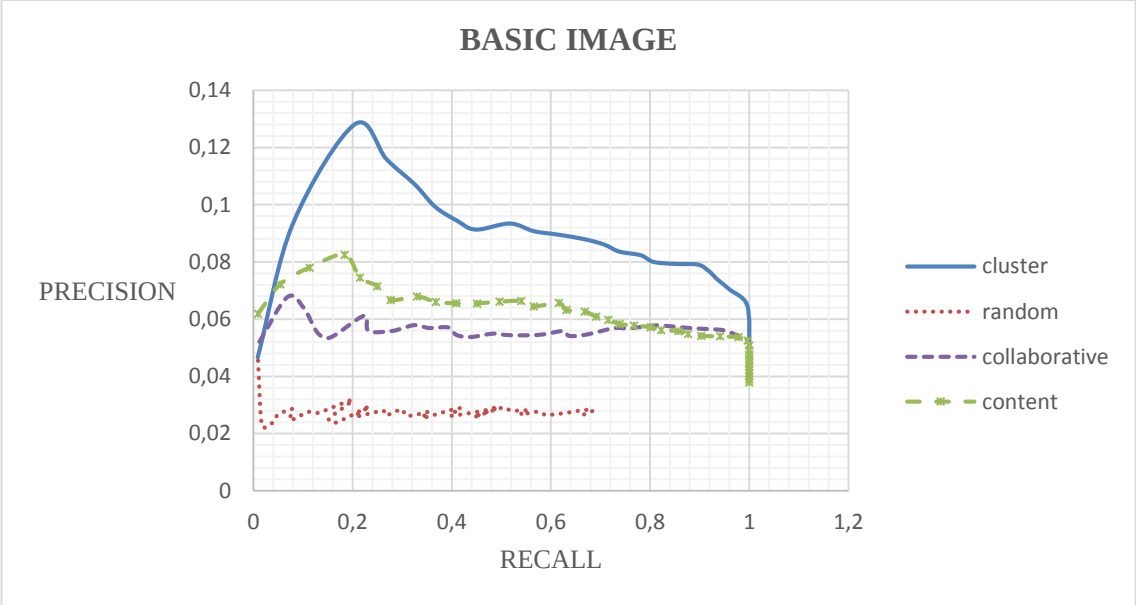
Source: Created by Harun Işık.

Figure 5.8: Precision - N graphic of Basic Image feature



Source: Created by Harun Işık.

Figure 5.9: Precision - Recall graphic of Basic Image feature

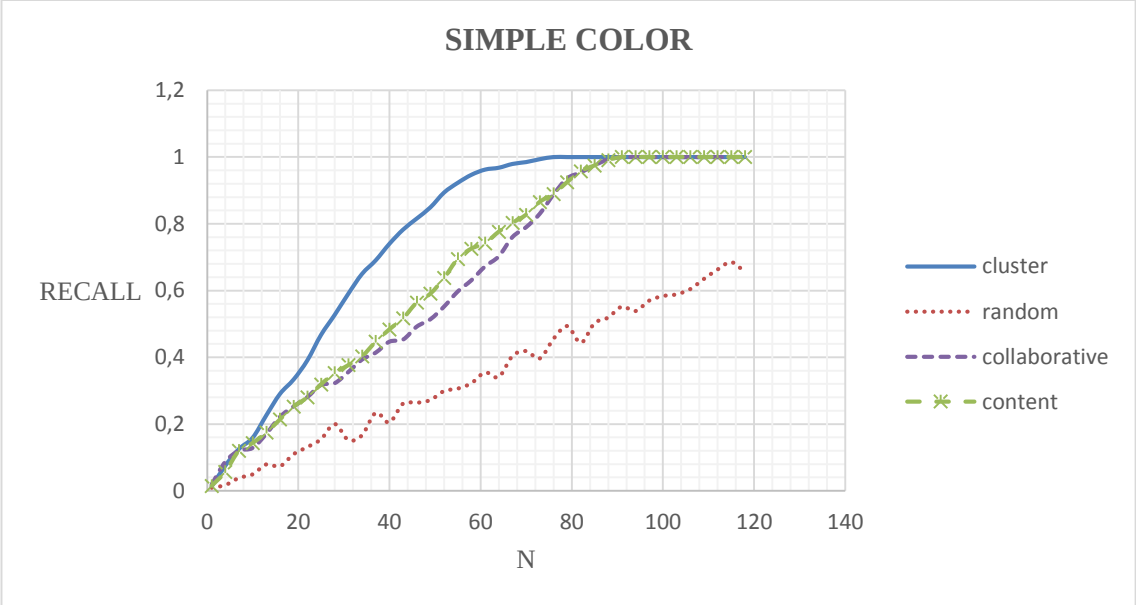


Source: Created by Harun Işık.

In figure 5.7, 5.8 and 5.9, recall values of cluster based algorithm is higher than recall values of other algorithms. But there is a little slight that its graph is getting close to cluster based graph. We can say that basic image feature is little much successful than GIST and CEDD.

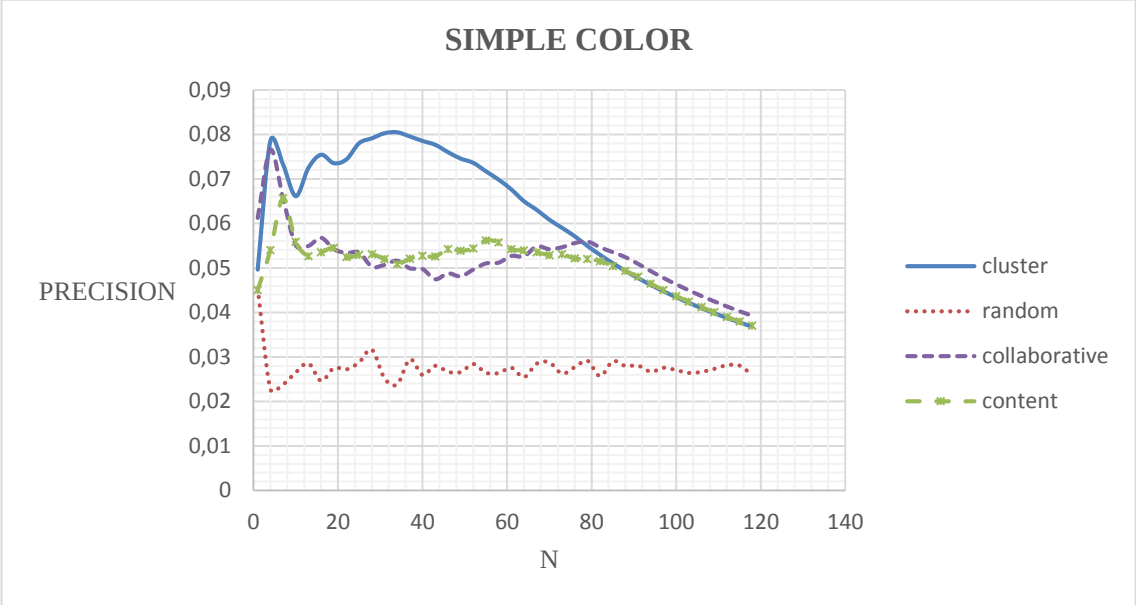
5.1.4 Simple Color Feature

Figure 5.10: Recall - N graphic of Simple Color feature



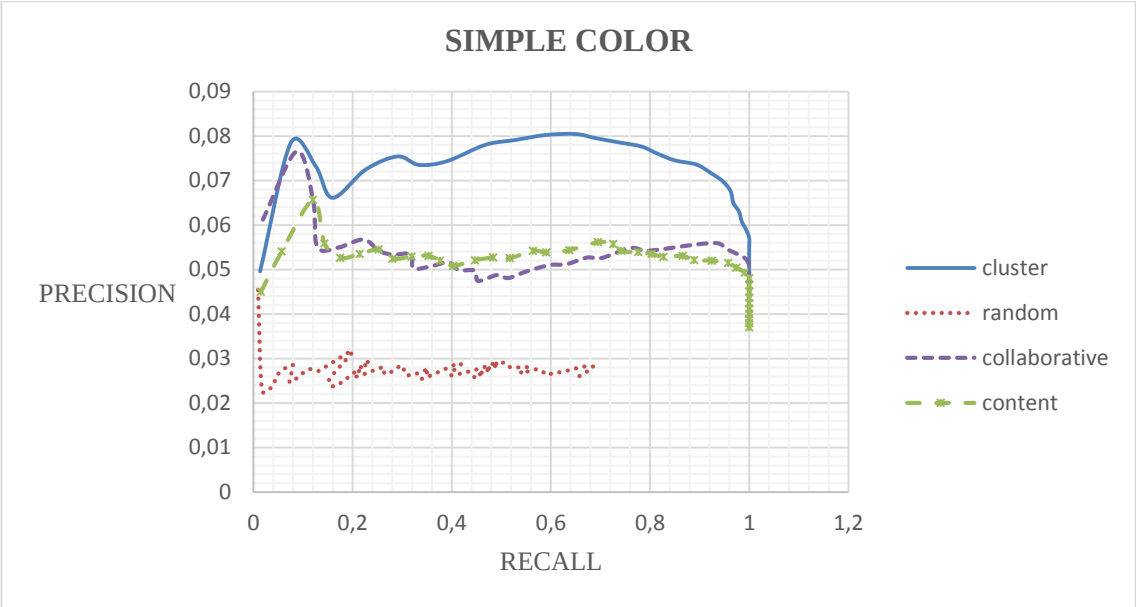
Source: Created by Harun Işık.

Figure 5.11: Precision - N graphic of Simple Color feature



Source: Created by Harun Işık.

Figure 5.12: Precision - Recall graphic of Simple Color feature



Source: Created by Harun Işık.

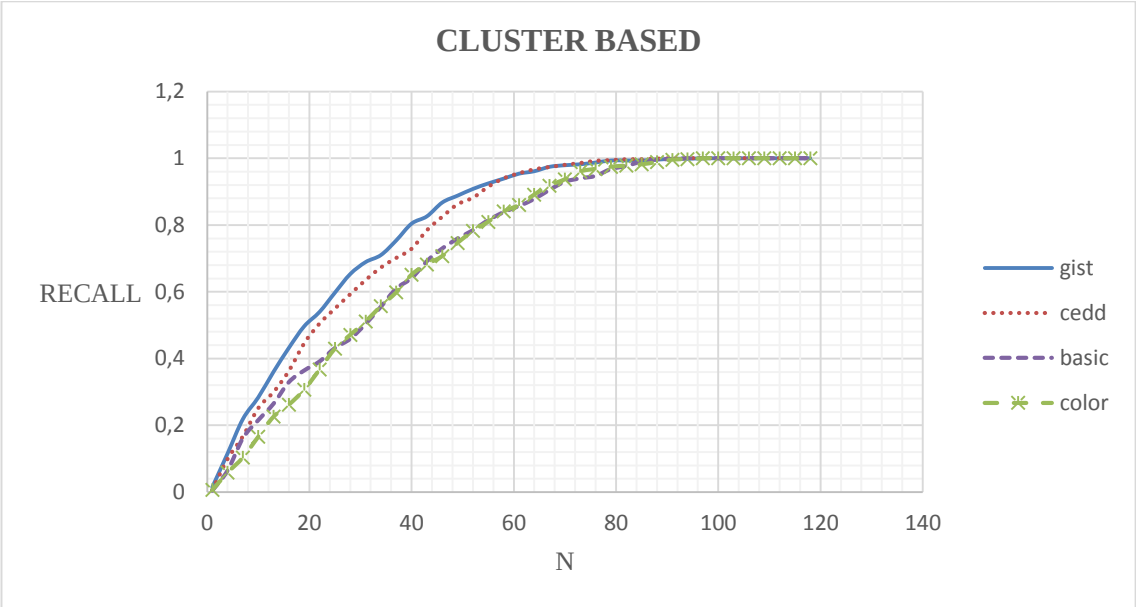
5.2 COMPARING FEATURES OVER ALGORITHMS

We used four image features which are basic image features, simple color feature, CEDD and GIST. We compare each feature with others for each algorithm we used.

5.2.1 Cluster Based Algorithm

Cluster based algorithm is our new approach to recommend images to users. We extract some mentioned image features as one dimensional vectors and make clusters by using these vectors. We observe hit rates of these features in cluster based algorithm and content based algorithm.

Figure 5.13: Recall - N graphic of Cluster Based Algorithm

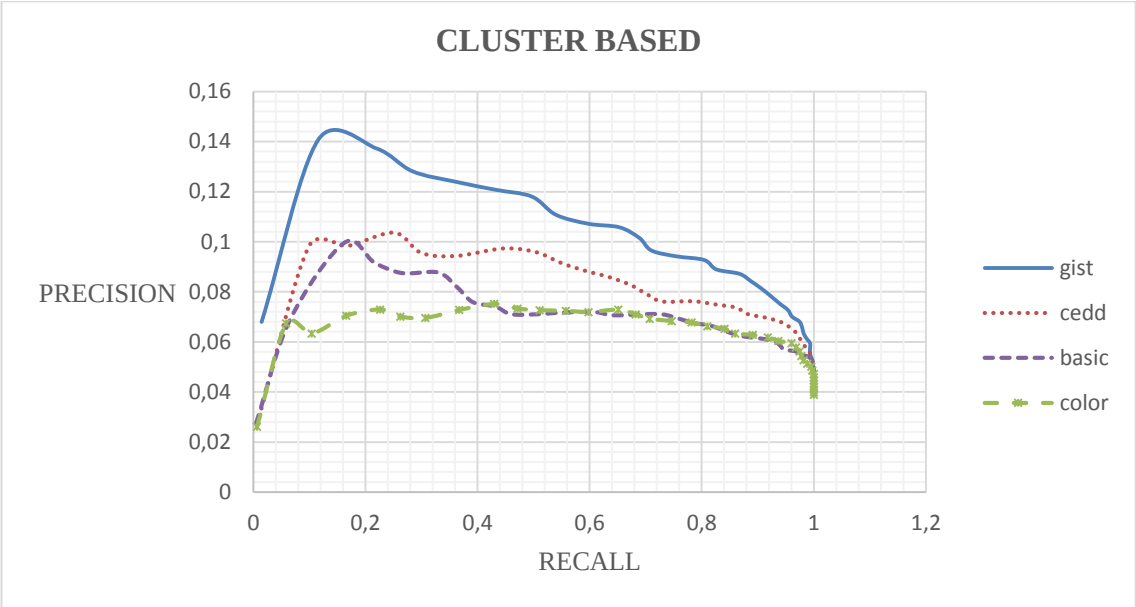


Source: Created by Harun Işık.

Figure 5.14: Precision - N graphic of Cluster Based Algorithm

Source: Created by Harun Işık.

Figure 5.15: Precision - Recall graphic of Cluster Based Algorithm



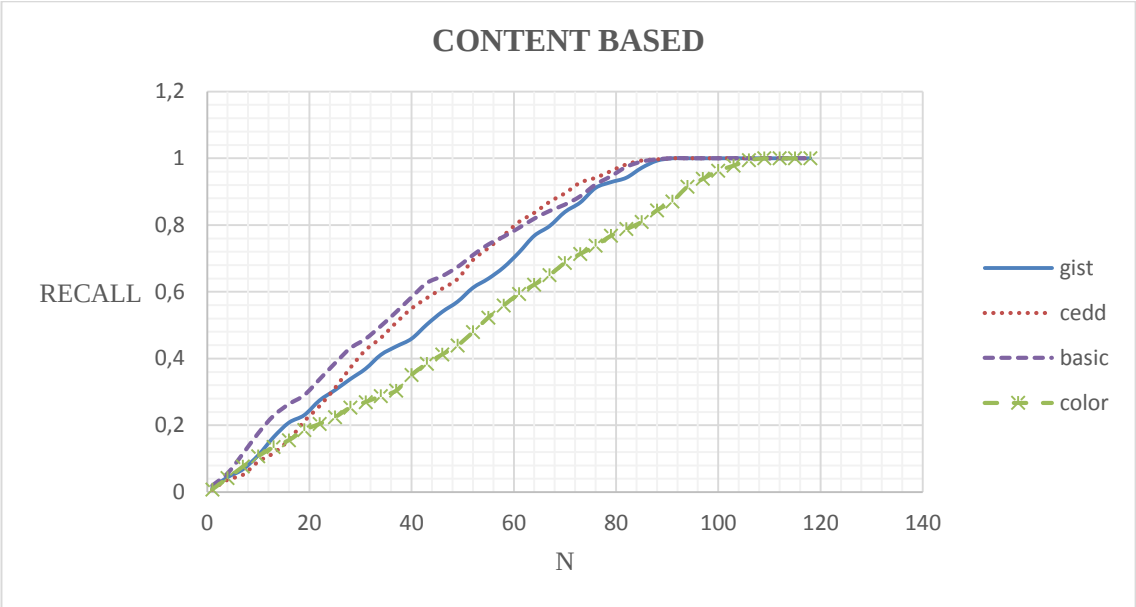
Source: Created by Harun Işık.

In figure 5.13, 5.14 and 5.15, GIST and CEDD features are seen as more successful than other two features. In order to get high hit rate in cluster based algorithm, we surely propose GIST and CEDD features. However these are still low level features and hit rates of recommendation algorithm can be increased by using more semantic features.

5.2.2 Content Based Algorithm

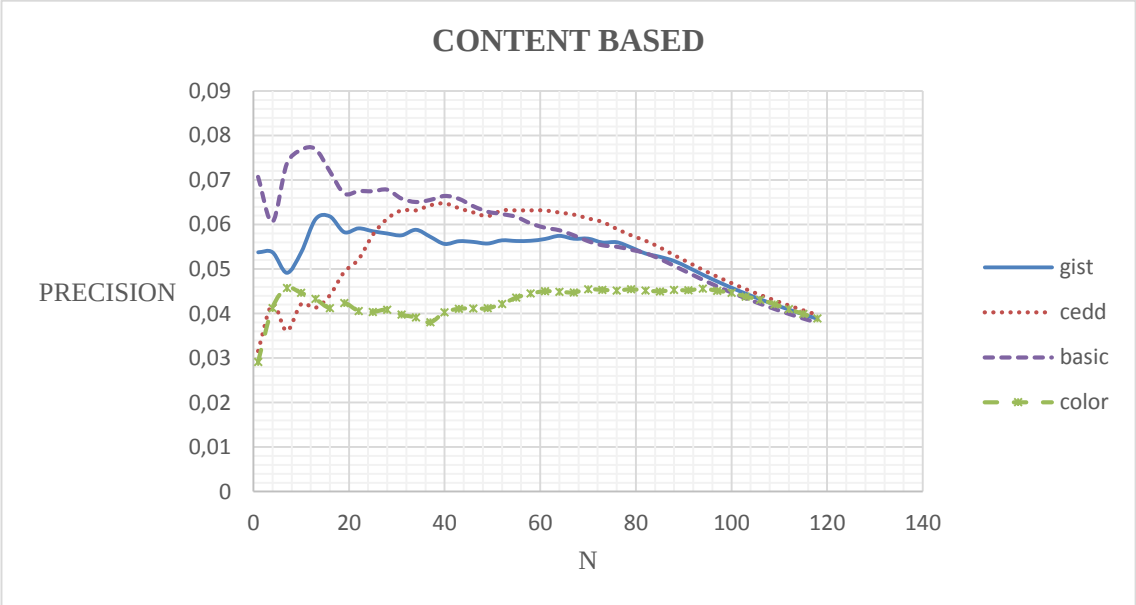
Content based filtering is a traditional method that it offers the closest images in ascending order to the selected user. Comparing image features in content based algorithm may give an opinion which one is more powerful feature than others.

Figure 5.16: Recall - N graphic of Content Based Filtering Algorithm



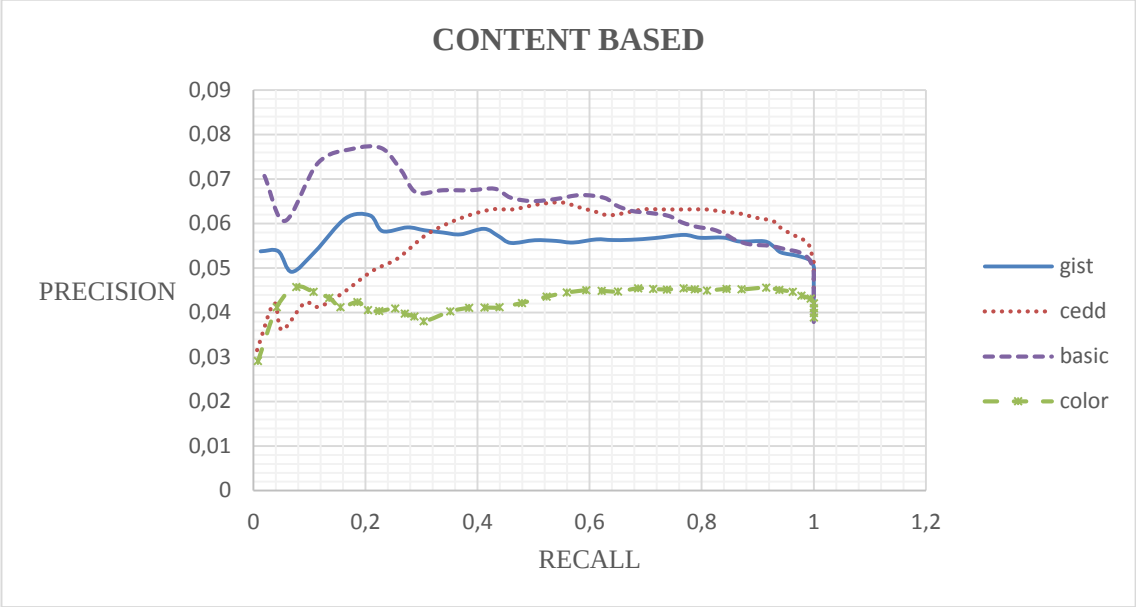
Source: Created by Harun Işık.

Figure 5.17: Precision - N graphic of Content Based Filtering Algorithm



Source: Created by Harun Işık.

Figure 5.18: Precision - Recall graphic of Content Based Filtering Algorithm



Source: Created by Harun Işık.

In figure 5.16, 5.17 and 5.18, as it is seen, simple color feature is weaker and basic image feature comes into the front of CEDD and GIST. Comparing features in content based algorithm is not directly related our cluster based algorithm but we understand that basic features and CEDD is better as content based.

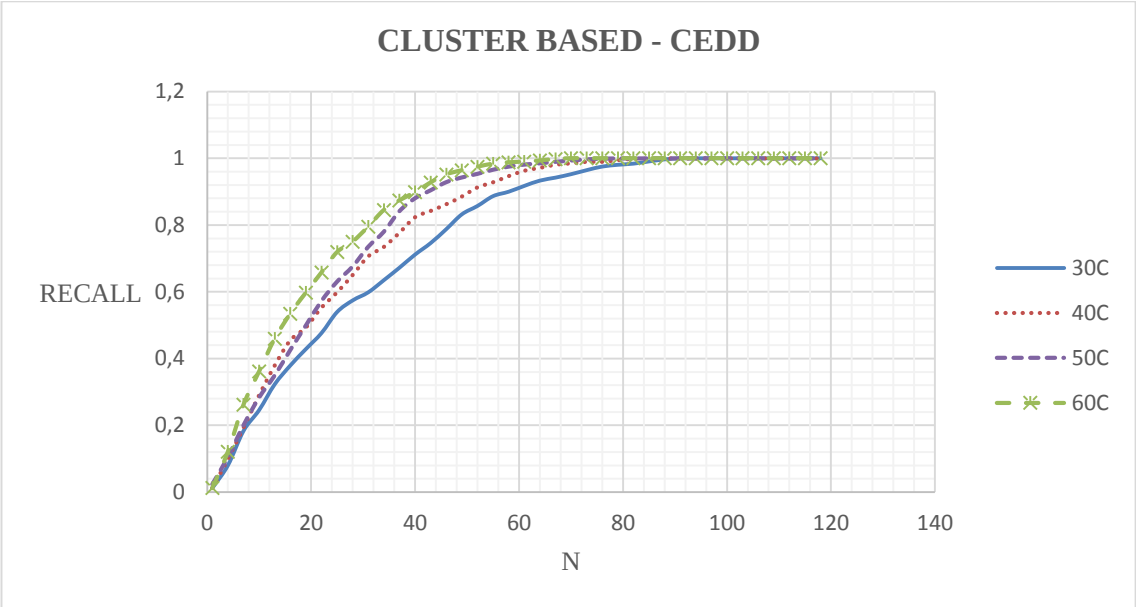
5.3 COMPARING CLUSTER NUMBER PARAMETERS

Cluster number parameter also is an important variable that effects the performance of our cluster based algorithm. As cluster number is increased, the number of images in each cluster decreases. So it could be easy to select remaining images after we eliminate images are already liked by the user in a cluster. As a result, we can get more correct results and we can have chance of offering images from more many clusters.

5.3.1 Cluster Based Algorithm - CEDD

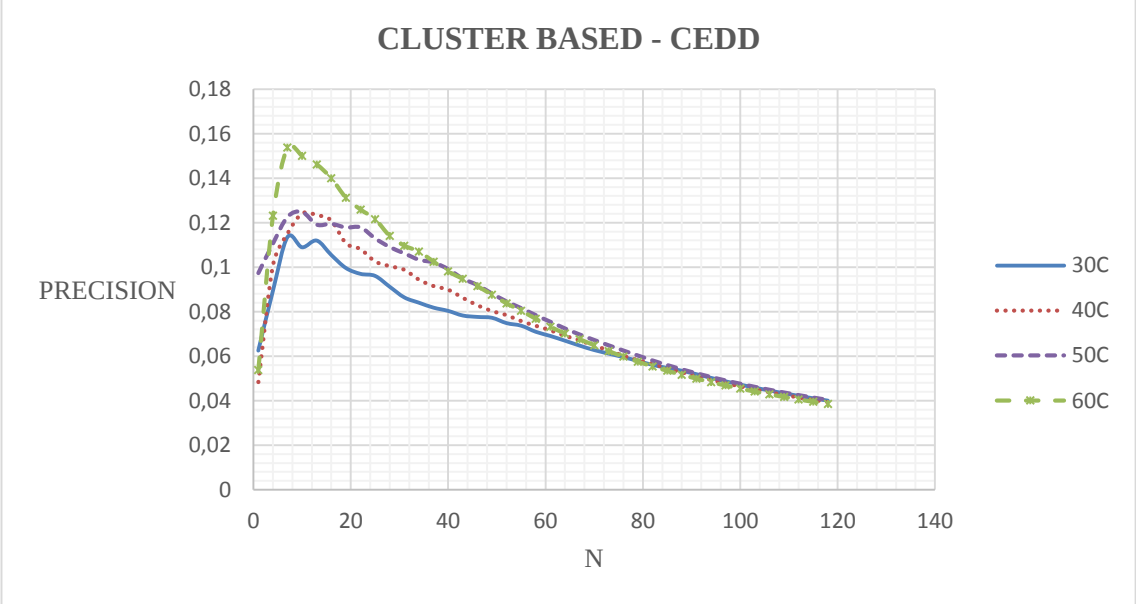
For the CEDD feature, we run our cluster based algorithm and compare number of cluster from 30 to 60.

Figure 5.19: Recall - N graphic of Comparing Number of Cluster (CEDD)



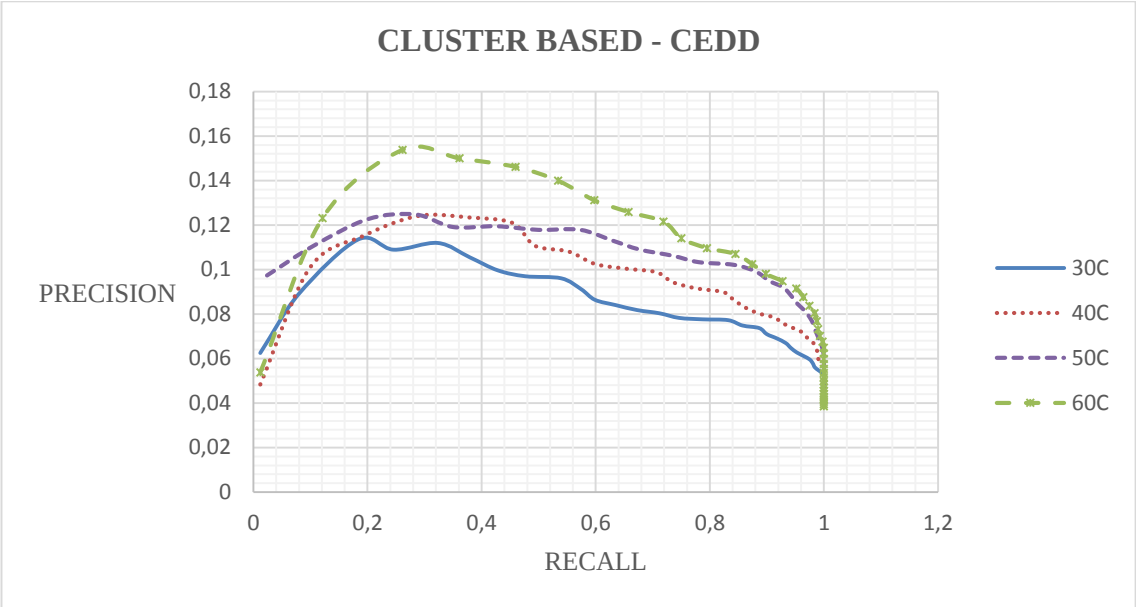
Source: Created by Harun Işık.

Figure 5.20: Precision - N graphic of Comparing Number of Cluster (CEDD)



Source: Created by Harun Işık.

Figure 5.21: Precision - Recall graphic of Cluster Based Algorithm



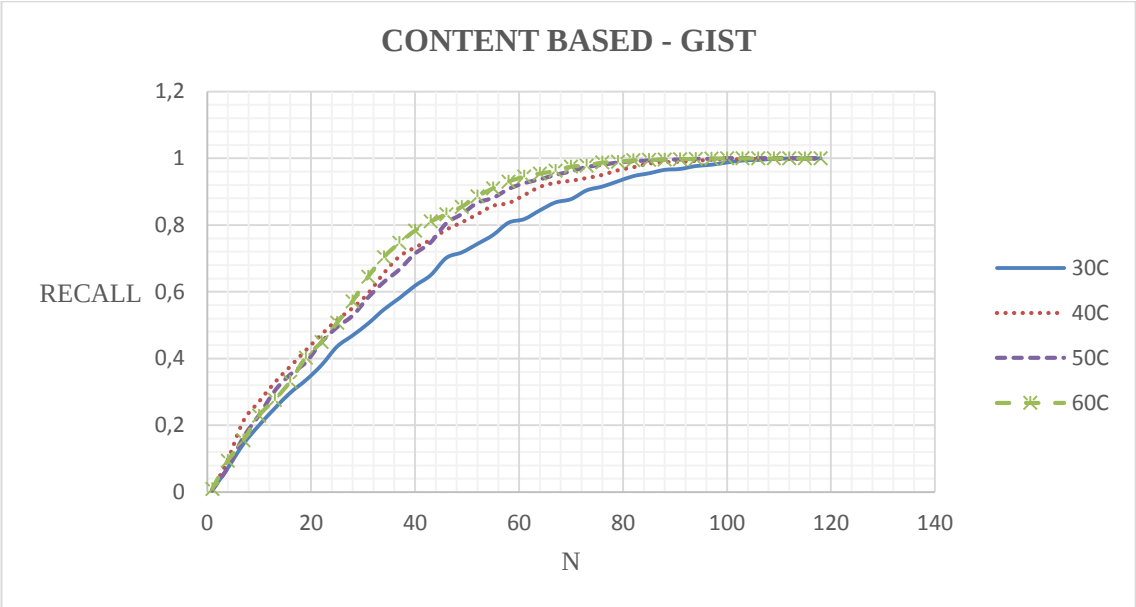
Source: Created by Harun Işık.

In figure 5.20, 5.21 and 5.22, we obviously see that as number of cluster is increased, the performance of our algorithm also increases. In case of 60 clusters, values of recall exceeds the other values of 30, 40 and 50 clusters.

5.3.2 Cluster Based Algorithm – GIST

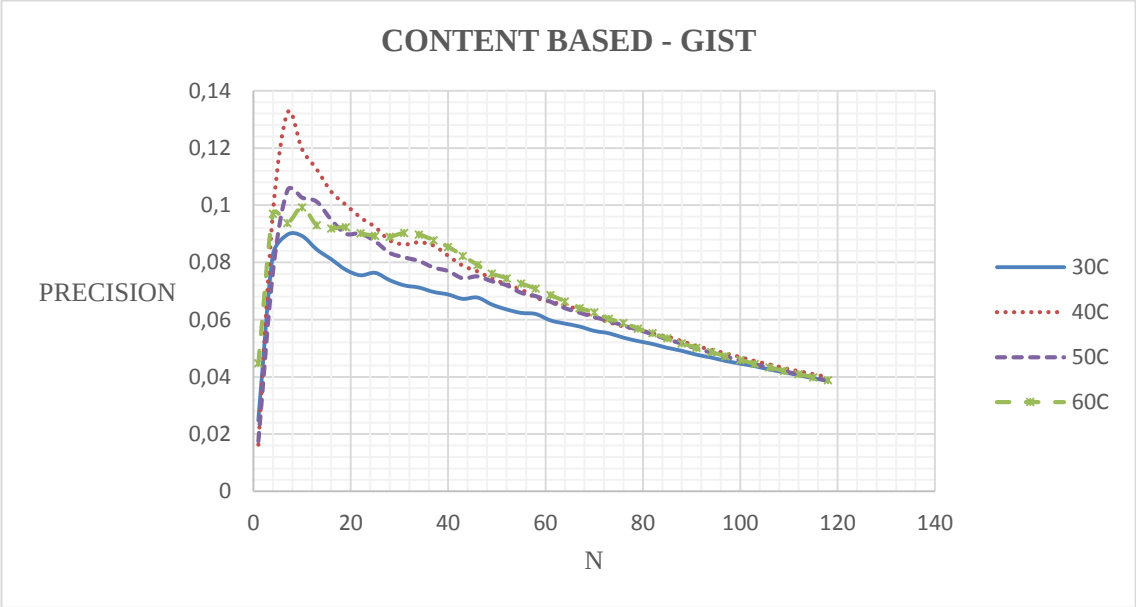
For the GIST feature, we run our cluster based algorithm and compare number of cluster from 30 to 60.

Figure 5.22: Recall - N graphic of Comparing Number of Cluster (GIST)



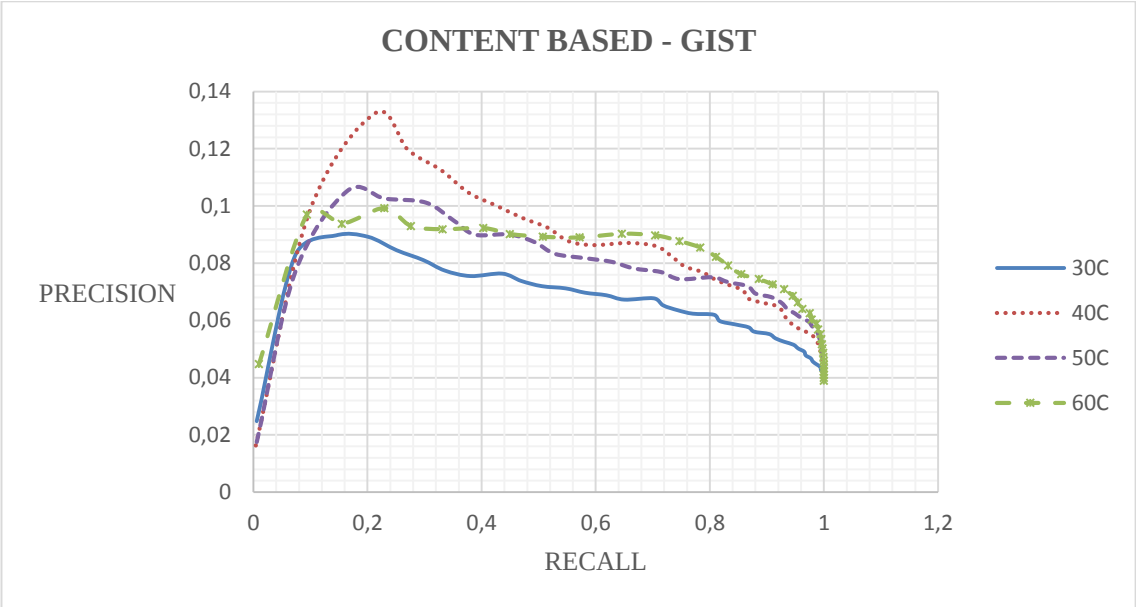
Source: Created by Harun Işık.

Figure 5.23: Precision - N graphic of Comparing Number of Cluster (GIST)



Source: Created by Harun Işık.

Figure 5.24: Precision - Recall graphic of Comparing Number of Cluster (GIST)



Source: Created by Harun Işık.

6. DISCUSSIONS AND CONCLUSIONS

Recommender systems have increasingly gained the popularity of many researchers. There are many application areas of RCs such as product recommendations, movie recommendations, news or article recommendations and others. Content based image recommendation term has been newly used. Although there are many image hosting sites and numerous photos have been uploading by people, there is no remarkable study about content-based recommendation which uses image features. Many recommended images to users in this websites have been offered according to most popular images or based on the people you follow. We want to improve accuracy of content-based recommendation methods and to contribute to image recommendation systems.

In most cases, people may want to search images they have not seen yet whereas labels or tags entered deals with images may not exactly define the images. These tags may contain sufficient or any incorrect information about images. So image search algorithms may not give much reliable results. We claim that image features model a user's preferences and help to search an image which he want to see. Unfortunately computers cannot "see" the image but they can only compare them with aspects of some features. Images can have color, texture, shape, size, orientation and more characteristics. While we determine a user's preference model to recommend him, we must consider all these characteristics. For example, a user usually likes photo of "red cars", we can create a preference vector for this user. By using color histogram and shape detection features, we can find all about "red car" photos and recommend them to user. Similarly, taste space can be modelled for many other users. So users will have reached photos without making tedious search and their models are able to use in different areas.

Clustering algorithms work in the way that gathers objects have similar properties and by distributing objects not like one another. Once the clusters are made, accordingly opinions of users in their favorite clusters, clusters can be used to make predictions for an individual. Dividing dataset into clusters may have some disadvantages. Clustering may lessen the accuracy of recommendations for users have flat distribution of items in

all clusters. For example, if a user has approximately same number of items in each cluster, this may not be useful. Clustering techniques mostly makes less personalized suggestions than other methods, and the clusters have worse accuracy than CF-based algorithms (Breese et al., 1998). However this technique is very useful in different ways for content-based algorithms.

In conclusion, we proposed a new approach about content based image recommendation. We firstly downloaded numerous images and user information from flickr website via flickr api. Secondly we studied on low-level image features and extracted several visual descriptors from image dataset. After we prepared descriptors, we determined appropriate clustering method to divide our dataset into groups. While we make these clusters, our aim is to create a user's preferences model about user's taste. We finally recommend to users by using our clusters, and we have done. It must be evaluated that our image recommendation algorithm is successful. In order to measure accuracy rate of our new method, we used precision-recall evaluation metrics on a recommendation system. We compare our algorithm with other approaches and compare image features on our algorithm. Finally we showed experimental results of algorithms via visual graphics.

6.1 FUTURE WORKS

During our experiments, we used some basic low level image features. There are many other low level image features from which are different one we used. We plan to include these features. We think if we use more feature, we are able to model user's preference more affluently. In addition to low level features, many researchers have been studying on extracting high-level image feature. High level image features try to describe an image semantically. These features are more intelligent than low level features to define an image. So we also plan to use next generation of our algorithm.

Recommender Systems can make common suggestion for all of users and also they can give more specialized recommendation for each user. In this thesis, we studied on giving common recommendations, and we did not use any specialized approaches for

each user. At next time, we will consider to run our algorithm per one user. We try to find image feature which gives the best result for single user. For example, GIST feature may more successfully result to model a user's preference but also it may not be more accurate for another user. Whenever we find more correct feature for modelling a user's preference, we increase our accuracy and scalability.

Finally, we want to talk about clustering method we used. Clustering makes groups of items which have similar characteristics. So cluster analysis is an important task in view of creating preference vector. In addition, we should proficiently know our data type of image features since we should use correct similarity metric. Recommendation result may change according to clustering algorithm you used. A good clustering algorithm provides more qualified recommendation results. So we also plan to study on other clustering algorithms such as hierarchical clustering and DBSCAN.

REFERENCES

Books

Breese, J., Heckerman, D., & Kadie, C., (1998)., Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pp 43-52.

Burke, R.: Hybrid web recommender systems. In: The Adaptive Web, pp. 377–408. Springer Berlin / Heidelberg (2007)

Han, J., Kamber, M., Pei, J., Data Mining Concepts and Techniques 3rd Edition

Ricci, F., Rokach, L., Shapira, B., Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 2011

Tan, P., Steinbach, M., Kumar, V., Introduction to Data Mining Book, p. 74, 2006

Other Publications

Adomavicius, G., Tuzhilin, A., Context Aware Recommender Systems

Chatzichristofis, S., A., Boutalis, Y., S., CEDD: Color and Edge Directivity Descriptor. A Compact Descriptor for Image Indexing and Retrieval, 2008

Fabrício D. A. Lemos, Rafael A. F Carmo, Windson Viana, Rossana M. C. Andrade, Towards a Context-Aware Photo Recommender System

Fan, J., Keim, D., A., Gao, Y., Luo, H., Li, Z., JustClick: Personalized Image Recommendation via Exploratory Search from Large-Scale Flickr Image Collections, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, No. 8, August 2008 1

Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer 42(8), 30–37 (2009)

Lops, P., Gemmis, M., D., Semeraro, G., Content Based Recommender Systems: State of the Art and Trends

Meteren, R., V., Someren M.V., Using Content-Based Filtering for Recommendation

Oliva, A., and Torralba, A., Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV, 42(3):145–175, 2001.

Pan, W., Xiang, E., W., Liu, N., N., Yang, Q., Transfer Learning in Collaborative Filtering for Sparsity Reduction

Pazzani, M., J., Billsus, D., Content Based Recommendation Systems

Potter, M. C. (1976). Short-term conceptual memory for pictures. Journal of Experimental Psychology: Human Learning and Memory 2, 509–522

Rae, A., Exploiting Social Networks for Recommendation in Online Image Sharing Systems, August, 2011

Sarwar, B., Karypis, G., Konstan, J., Riedl, J., Analysis of Recommendation Algorithm for E-Commerce, October, 2000