Navigational Pattern Based Relevance Feedback Using User Profile in CBIR

Syed Karim, Muhammad Harris, and Muhammad Arif
Department of Computer Science, City University of Science and Information Technology Peshawar, Pakistan

Abstract: *Content Based Image Retrieval (CBIR) is an application of computer vision and addresses the problem related to retrieval of digital images in large image databases. CBIR uses low level image features for retrieval task and tries to portray users intended results. Relevance Feedback (RF) is a technique for marking retrieved results as relevant or irrelevant by the user. People in the society have mutual interests and needs while searching for required data. Interesting and similar patterns can easily be found in the browsing behaviour of users pursuing required images from CBIR system. Recording users browsing behaviour and applying mining techniques to find frequent itemsets helps boost the retrieval performance of the CBIR system in terms of quality and processing time. User categorized into different groups on the basis of users age and gender specification helps fasten the mining process because of the similarity of thoughts in these users groups. This paper focuses on mining user browsing behaviour belonging to different user categories (user profiling) with FP-growth mining algorithm for revealing similar search patterns. The results show efficiency against the existing approach.*

Keywords: *CBIR, profile, apriori, FP-growth, support vector machine, navigational pattern relevance feedback.*

Received June 15, 2014; accepted October 26, 2014

1. Introduction

Retrieval system aims to retrieve results from databases and present them to the requesting users. Retrieval systems are a way to extract user perceptional data from large data reserves online and offline. Retrieval systems are broadly categorized into two groups commonly known as Text based and content based retrieval systems. The main differences rely on the nature of their working. Text based retrieval system retrieve images with the help of text descriptors given to define the images, while Content Based Image Retrieval (CBIR) processes visual features (low level features) of the images and after processing these visual features results are displayed [17, 18].

Labelling images manually to their exact representation is a tedious task in huge repositories and wrong annotations often results in irrelevant results to user perception [20]. As a result CBIR gained more attention by the research community as it portrays the actual user perception. Some areas of CBIR application are architecture design art and craft museums, archaeology, medical imaging and geographic information system, trademark databases weather forecast, image classification, image search over the Internet, remote sensing field for indexing biomedical images by contents and criminal investigations [14]. Some famous CBIR system are Query By Image Content (QBIC) [2], photo book [13], visual seek; [19], mars [6].

Figure 1 shows the general architecture of a CBIR system with Relevance Feedback (RF). RF performs query tuning and is iterative process. For retrieval task low level image features like color, texture, shape, etc., are extracted from images in the database and stored in feature database. Upon submitting the query image, image features from the query image are extracted and matched via a similarity measure with the features in feature database. The smaller the distance between feature vectors the higher is the degree of similarity [12].

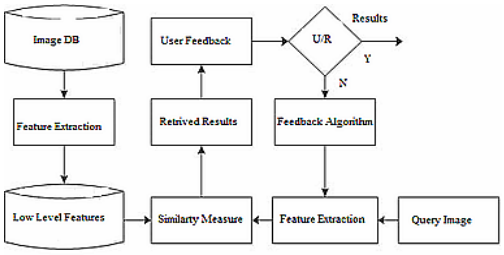

Figure 1. CBIR system architecture.

RF is a method that tries to capture user's perception utilizing different machine learning techniques. RF is considered to be learning and a classification problem by researches. Many classical machine learning schemes may be applied to the RF such as, decision tree learning [10], Bayesian learning [1, 21], Support Vector Machines (SVMs) [23], boosting [22] and so on. RF alone is not sufficient for the satisfactory performance of CBIR [20].

Navigational Pattern Relevance Feedback (NPRF) is a technique that mines patterns (frequent itemset) from useful user browsing logs for optimizing the search direction on RF [20]. NPRF gathers similar

thoughts of people and utilize them in searching process. The existing system [20] uses Apriori algorithm for mining user logs which is slower. Apriori algorithm has architectural flaws i.e., I/O and memory intensive [8].

To overcome the limitation of NPRF search using Apriori [20] we are going to use FP growth algorithm with NRPF profiling and SVM to provide better results compared to existing system with respect to time and retrieval quality.

The rest of the paper is organized as follows: A review of past studies is briefly described in section 2. In section 3, we describe the details of our proposed scheme. Empirical evaluations of the proposed method are expressed in section 4. Finally, we conclude the paper in section 5.

2. Related Works

The related work section is sub-sectioned as: Section 2.1 explains the RF related work, section 2.2 NPRF related work, section 2.3 FP-growth section 2.4 explains user profiling in CBIR.

2.1. RF Techniques

Different sort of machine learning techniques are used in RF to classify images as relevant and irrelevant. Some traditional RF methods used in CBIR are query refinement [16] and re-weighting [15], however these methods did not show performance due to the problems like is how to incorporate positive and negative examples to refine the query or to adjust the similarity measure [9].

SVM is a standard machine learning technique and has gained success in different empirical pattern recognition applications [5]. In the domain of CBIR SVM has been utilized to solve the problem of classification. SVM is a supervised classification technique which is trained using two different classes. In CBIR the classes are known as relevant and irrelevant classes based on the relevant and irrelevant image. SVM is trained using user feedback on each iteration to refine the search process. Each image in the database is passed through SVM regarded as relevant or irrelevant (+1, -1) by the user. It enables the system to reach maximum precision in less iterations [11].

2.2. Navigational Pattern RF

Utilizing user navigational patterns in browsing behaviour the user's intention can be precisely captured. The mining algorithm used generate useful browsing patterns is Apriori algorithm used in the current system [20].

2.2.1. Overview of NPRF

NPRF merges two sections, the three RF enhancement techniques and discovery of user navigational patterns for the effective and efficient exploration of relevant images. This job is divided into two major operations namely online image retrieval and offline knowledge discovery [20].

2.2.2. Online Image Retrieval

- *Initial Query Processing*: Without considering the feature weights, this phase extracts the features from query for matching purpose to the features in the reserves. The good examples selected by the user though RF is then further analysed. This initial phase is also known as iteration 0 [20].
- *Image Search Phase*: This phase is an integration of user navigational patters and proposed search algorithm which expands a single search to multi search point. A new query point is reached out by utilizing the previous positive examples [20].

2.2.3. Offline Knowledge Discovery

- *Knowledge Discovery Phase*: Learning from user browsing behaviour is a sort of knowledge discovery. This phase is primarily concerned with the formation of navigational models from the user browsing behaviour. This navigational model helps in prediction and enhanced retrieval results [20].
- *Data Storage Phase*: The database used stores time variant, integrated, non-volatile collection of valuable data including images, navigational patters, image features and log files. The knowledge ware house is valuable for the improvement of quality of image retrieval [7, 20].

2.3. FP-Growth Algorithm

FP-growth algorithm is an association rule mining algorithm and was proposed by Han *et al.* [3]. FP-growth algorithm is based on frequent pattern tree data structure and came into existence due to the bottleneck of Apriori-like methods i.e., candidate set generation and test as shown in [3]. FP-growth shows efficiency in mining in three ways: A large database is condensed into a small one with the reduction in passes of database scans, shrinks number of candidate sets and mining task is reduced to a set of smaller tasks in conditional databases which reduces search space [3].

2.4. User Profiling in CBIR

Customization of the user search process in CBIR is the area that has been broadly studies by the information retrieval community but still problems persists. User's interest can be modelled by the aggregation of user's interaction with the CBIR system [4].

Zhang *et al.* [24] proposed a personalized multimedia information retrieval by utilizing user profile mining. Active user historical data explains the

multimedia data which is stored and the association between semantic information and multimedia data is examined. User interest model in built on the basis of modified LDA model. The proposed scheme can retrieve the required multimedia results utilizing proposed personalized multimedia information ranking algorithm.

Pentalnd *et al.* [13] proposed architecture of personalization by implying the user by representing his interest and profile into an ontology of interest and using the ontology for re-indexing of the contents. A key part of personalization model is the user model.

In our proposed approach user profiles are built and specific user characteristics are collected that are user age and gender specifications and upon these user characteristics interesting patterns in user browsing behaviours are extracted and used in the retrieval process of CBIR.

3. Proposed Scheme

The proposed scheme NPRF profiling-FP-SVM is based on three critical modules i.e., NPRF profiling using FP-growth, RF using SVM and user profile. Rest of the architecture consists of CBIR based on color feature and image database.

3.1. User Profile

The proposed scheme is profiled based so user information is necessary part of the system. The users perform the search by providing the two parameters i.e., age and gender. Both of the user's characteristics are utilized in searching process of image retrieval, with NPRF. The two parameters of user profile are saved in log data with each new query process. This log data is then stored inside a database named as log database. Figure 2 shows the basic architecture of the log file with age and gender information.

```
1 <Root>
2 <Query>\image\set00001\190.jpg</Query>
3 <Relevant></Relevant>
4 <Irrelevant></Irrelevant>
5 <Age></Age>
6 <Gender></Gender>
7 </Root>
```

Figure 2. Log file architecture.

3.2. NPRF Profiling using FP Growth

Filtration of log files is done on the basis of user profile parameters and profile based log browsing behaviours are fetched by the association rule mining technique FP-growth to find the relevant and irrelevant frequent item sets, shown in Figure 3. These frequent item sets are then processed by the parser for parsing so that the SVM engine can do further processing on the parsed content.

3.3. Support Vector Machine Engine

As shown in Figure 3, NPRF profiling-FP feeds SVM with relevant and irrelevant classes.

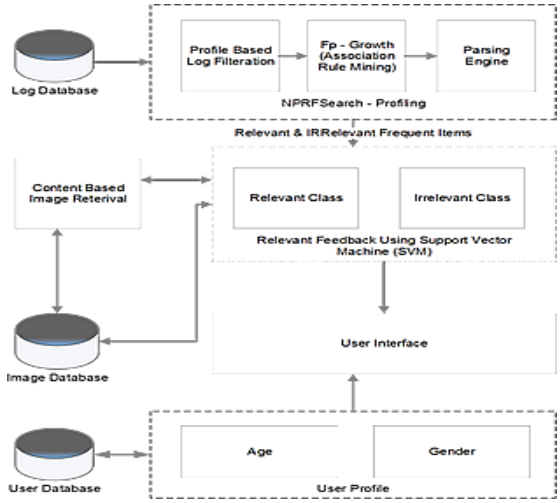

Figure 3. Proposed architecture.

This information is then used to train the SVM. After training the database is then searched for finding the relevant images. This step increases the efficiency of retrieval because SVM has already been trained using the majority user selected images. Below Algorithm 1 is the detailed for proposed approach.

Algorithm 1: NPRF profiling using FP growth and SVM.

// Main Variable and Constraints
L_i*=set of user log data in the log database*
U_i*=user profile database and user query*
R_i*=set of relevant images for user interface*
P_i*=relevant log data*
N_i*=Irrelevant log data*
$F_R=F_I=null$
$R_{IS}=IRR_{IS}=null$
// algorithm
$P_i=Filter(U_i, L_i, Relevant)$
$N_i=Filter(U_i, L_i, IRRelevant)$
//Extract Itemset with Frequency for Relevant Log Data
$F_R=FP\text{-}growth\ (P_i)$
//Extract Itemset with Frequency for IRRelevant Log Data
$F_I=FP\text{-}growth\ (N_i)$
// Find Most frequent Relevant Itemset String and Parse It
$R_{IS}=ParsingEngine\ (F_R)$
// Find Most frequent IRRelevant Itemset String and Parse It
$IRR_{IS}=ParsingEngine\ (F_I)$
$R_i=SVM(R_{IS}, IRR_{IS})$
Return R_i *to User Interface*

4. Implementation and Result Evaluation

Table 1 shows the computer system specification used for the implementation purpose for the proposed scheme.

Table 1. System description.

Language	C#.Net Framework 4 Client Profile
Libraries	Emgu C.v-SVM Open source Data mining Library SPMF-FP growth Matlab-cross correlation
R.A.M	1024MB
C.P.U	2.53GHZ
Environment	Virtual-VMware
OS	Windows XP

Image dataset from Google having dimension of 100*80 are taken as dataset for result evaluation shown in Table 2. Each dataset contains 200 images. Images are simulated utilizing custom developed software based on NPRF profiling-FP-SVM for log generation on the basis of the different image datasets. The log information is then stored in log database. Implementation tool for our custom software is based on visual C# with libraries and the simulation is processed using virtual environment on Windows XP platform.

Table 2. Datasets.

Datasets	Image Dimensions	No. of Items
Perfumes	100*80	200
Buildings		
Machine		
Animals		
Nature		

For measuring the quality of retrieved results we have utilized the precision and recall measures as measuring function.

$$Precision = \frac{|RelevantResults|*100\%}{|TotalRetrievedImages|} \quad (1)$$

$$Recall = \frac{|Relevant\ Results.|*100\%}{|Total\ Images\ in\ Database|} \quad (2)$$

4.1. Comparison w.r.t Time Factor

For time wise comparison between the existing and proposed approach, we have took dataset having different no. of log files all of these log data is generated using RF from SVM.

Figure 4 clearly depicts that FP growth profiling shows better performance both in worst and average cases. The worst and average case performs better due the information filtering on the basis of profiling. The profiled filtering helps in the reduction of log files due to selection of specific log files from the log database and improves the overall system efficiency with respect to time.

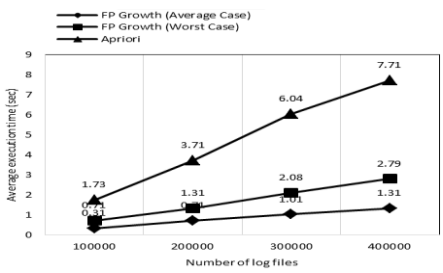

Figure 4. Comparison between Apriori and FP-growth.

4.2. Comparison w.r.t Quality Factor

For the purpose of testing retrieval quality, we have considered perfume dataset. As this dataset is clear and distinctive for both gender and can be overlapping in selection if the user does not have any idea about perfumes. Table 3 shows that the results based on profile information are more precise in terms of retrieval quality then existing approach. As the existing approach extracts frequent itemset from all the navigated data rather than age and gender specification, which leads to less precision value while retrieval of results.

Table 3. Comparison between different techniques with respect to precision.

Technique	Iteration # 1	Iteration # 2
CBIR using Cross Correlation	1/3	
NPRFSearch-Apriori	3/5	2/3
NPRFProfiling-FP Growth	2/3	

Figure 5 is graphical representation on the values based in Table 3.

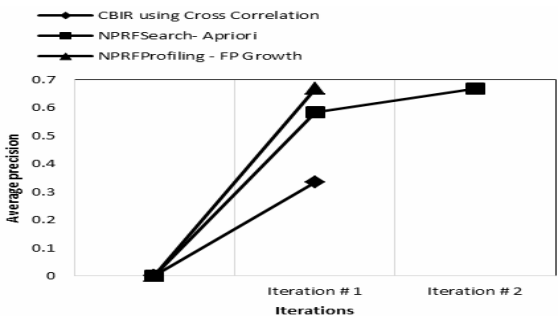

Figure 5. Graphical representation of precision values.

4.3. Practical Implementation.

Figures 6, 7, 8, 9 and 10 shows the snap shots of iteration achieved using different techniques. Results show that NPRFProfiling-FP growth has better efficiency both w.r.t time and quality against the existing approach. Profile information with FP-growth helps the NPRF search to achieve maximum precision and recall in less iteration and consume less amount of time while retrieving frequent itemset.

Figure 6. Test image.

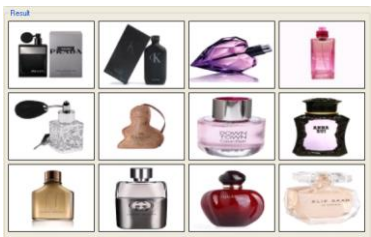

Figure 7. Iteration#1 using CBIR.

Figure 8. Iteration#1 using NPRF-Apriori.

Figure 9. Iteration#2 using NPRF-Apriori.

Figure 10. Iteration#1 using NPRF profiling-FP growth.

5. Conclusions

In this paper NPRF which was previously using Apriori algorithm for association rule mining on user log files is enhanced by implementing FP-growth algorithm with profiling and SVM. For system having large multimedia information database proposed system provides a much efficient approach both in terms of time and quality of retrieval results against the existing approach. Our proposed approach reduces the memory and IO consumption thus enhancing the overall functionality of the CBIR system.

Proposed approach can more be enhanced by implementing it over a distributed network.

References

[1] Cox I., Miller M., Minka T., Papathomas T., and Yianilos P., "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments," *IEEE Transaction on Image Processing*, vol. 9, no. 1, pp. 20-37, 2000.

[2] Flickner M., Sawhney H., Ashley J., Huang Q., Dom B., Gorkani M., Hafuer 1., Lee D., Petkovic D., Steele D., and Yanker P., "Query by Image and Video Content," *IEEE Computer*, vol. 28, no. 9, pp. 23-32, 1995.

[3] Han J., Pei J., and Yin Y. "Mining Frequent Patterns without Candidate Generation," *in Proceeding of the ACM SIGMOD International Conference on Management of Data*, Texas, pp.1-12, 2000.

[4] Har A., Ghen M., Ben H., and Sahbi S., "Architecture for Personalized and Semantic Information Retrieval: Approach Based on Content's Re-Indexing Using User's Profile," *in Proceeding of the 23rd International Conference on Software Engineering and Knowledge Engineering*, Miami Beach, pp. 7-9, 2011.

[5] Hoi S., Lyu M., and Rong J., "Integrating User Feedback Log into Relevance Feedback by Coupled SVM for Content-Based Image Retrieval," *in Proceeding of International Conference of Data Engineering Workshops*, Tokyo, pp. 1- 10, 2005.

[6] Huang T., Mehrotra S., and Ramchandran K., "Multimedia Analysis and Retrieval System (MARS) Project," *Graduate School of Library and Information Science University of Illinois*, pp. 100-117, 1997.

[7] Killikatt S., Kulkarni V., and Bijjal M., "Content Based Image Retrieval by Online and Offline," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 7, pp. 2761-2768, 2014.

[8] Li H., Wang Y., Zhang D., Zhang M., and Chang E., "pfp: Parallel FP-Growth for Query Recommendation," *in Proceeding of the ACM Conference on Recommender Systems*, Switzerland, pp.107-114. 2008

[9] Long F., Zhang H., and David D., *Multimedia Information Retrieval and Management-Technological Fundamentals and Applications, Chapter Fundamentals of Content-Based Image Retrieval*, Springer, 2003.

[10] MacArthur S., Brodley C., and Shyu C., "Relevance Feedback Decision Trees in Content-based Image Retrieval," *in Proceeding of IEEE Work-shop Content-based Access of Image and Video Libraries*, Hilton Head Island, pp. 68-72, 2000.

[11] Marakakis A., Galatsanos N., Likas A., and Stafylopatis A., *Relevance Feedback for Content-Based Image Retrieval Using Support Vector Machines and Feature Selection*, Springer, 2009.

[12] Patil P. and Kokare M., "Relevance Feedback in Content Based Image Retrieval: A Review" *Journal of Applied Computer Science and Mathematics*, vol. 5, no. 10, pp. 41-47, 2011.

[13] Pentalnd A., Picard W., and Sclaroff S., "Photobook: ContentBased Manipulation of Image Databases," *International Journal Computer Vision*, vol. 18, no. 3, pp. 233-254, 1996.

[14] Rehman M., Iqbal M., Sharif M., and Raza M., "Content Based Image Retrieval: Survey," *World Applied Sciences Journal*, vol. 19, no. 3, pp. 404-412, 2012.

[15] Rui Y. and Huang T., "A Novel Relevance Feedback Techniques in Image Retrieval," *in Proceeding of 7th ACM Conference on Multimedia*, Florida, pp. 3-10, 1999.
[16] Rui Y., Huang T., Ortega M., and Mehrotra S., "Relevance Feedback: a Power Tool for Interactive Content-Based Image Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644-655, 1998.
[17] Singh S. and Sontakke T., "An Effective Mechanism to Neutralize the Semantic Gap in Content Based Image Retrieval (CBIR)," *The International Arab Journal Of Information Technology*, vol. 11, no. 2, pp. 124-133, 2014.
[18] Sivakamasundari G. and Seenivasagam V., "Different Relevance Feedback Techniques in CBIR: A Survey and Comparative Study," *in Proceeding of International Conference* of *Computing, Electronics and Electrical Technologies*, India, pp. 1115-1121, 2012.
[19] Smith J. and Chang S., "Visual SEEK: A Fully Automated Content-Based Image Query System," *in Proceeding of ACM Multimedia Conferences*, Boston, pp. 87-98, 1996.
[20] Su J., Huang W., Yu P., and Tseng V., "Efficient Relevance Feedback for Content-Based Image Retrieval by Mining User Navigation Patterns," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 360-372, 2011.
[21] Su Z., Zhang H., Li S., and Ma S., "Relevance Feedback in Content-based Image Retrieval: Bayesian Framework, Feature Subspaces, and Progressive Learning," *IEEE Transaction Image Process*, vol. 12, no. 8, pp. 924-936, 2003.
[22] Tieu K. and Viola P., "Boosting Image Retrieval," *in Proceeding of IEEE Conferences Computer Vision Pattern Recognition*, USA, pp. 228-235, 2000.
[23] Tong S. and Chang E., "Support Vector Machines Active Learning for Image Retrieval," *in Proceeding of ACM Multimedia*, Australia, pp. 107-118, 2001.
[24] Zhang P., Xia J., and Li R., "Personalized Multimedia Information Retrieval based on User Profile Mining," *Journal of Networks*, vol. 8, no. 10, pp. 2340-2346, 2013.

Syed Karim is a MS Scholar at City University of Science and Information Technology Peshawar, Pakistan. His areas of interest are Multimedia Retrieval System, Data Mining, Computer Vision and Image Processing.

Muhammad Harris is a MS Scholar at City University of Science and Information Technology Peshawar, Pakistan. Pakistan. His area of interest are Multimedia Retrieval System and Multimedia Databases.

Muhammad Arif is a PHD Scholar at COMSATS Institute of information technology Islamabad, Pakistan. His areas of interest are Data warehousing, Multimedia Image Retrieval, Medical Images, Image Processing and Computer Vision.