International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A New Approach for Information Retrieval in Multimodal Fusion using Association Rule Mining

Divya Pallavi[1], Naziya Pathan[2]

[1]Nagpur University, Nuva College of Engineering & Technology, Kalmeshwar Road, Nagpur, India

[2]Nuva College of Engineering & Technology, Nagpur University, Kalmeshwar Road, Nagpur, India

Abstract: ***The retrieving method proposed in this paper utilizes the fusion of the images' multimodal information (textual and visual) which is a recent trend in image retrieval researches. It combines two different data mining techniques to retrieve semantically related images: clustering and Frequent Pattern Tree Based Algorithm. This clustering technique is constructed at the offline phase where the frequent pattern rules are discovered between the text semantic clusters and the visual clusters of the images to use it later at the online phase. The experiment was conducted on images of Wikipedia collection.***

Keywords: feature extraction, frequent pattern mining, indexing.

1. Introduction

Today, Information Technology (IT) is presented in virtually all areas. This leads to raise the importance of computers for remote monitoring which is the basis for control systems. Large volumes of databases, diversity and heterogeneity of data sources require a new philosophy of treating them. In this case, data mining looks to discover implicit knowledge in a dataset based on different techniques that can be implemented independently or coupled. These techniques aim to explore data, to describe their contents and extract the information more meaningful. Because much of the information that exists in organizations is informal and unstructured, these techniques are not limited to digital data and evidence, but must also address the textual and multimedia. In the recent years the acquisition, generation, storage and processing of data in computers and transmission over networks had had a tremendous growth. This changed radically with the appearance of the Internet and the first web browser which revolutionized the distribution of information. The ease of information exchange incited millions of people to create their own web pages and to enrich it with images, video, audio and music. Due to this rapid development in the domain of digital and informational data people now live in a multimedia world. More and more multimedia information is generated and available in digital form from varieties of sources around the world, this expansion presents new challenges.

Section II discusses the background i.e. the background work for fusion. This helps to expand the idea of the proposed system in terms of using new techniques for the proposed methodology. Section III is the overview of the algorithm used. Section IV explains the module in which the project is divided. It discusses all the phases used in the proposed along with the approach used in it. Section V is the proposed system. Section VI is the experimental analysis where the result is analyzed. Section VII is the conclusion.

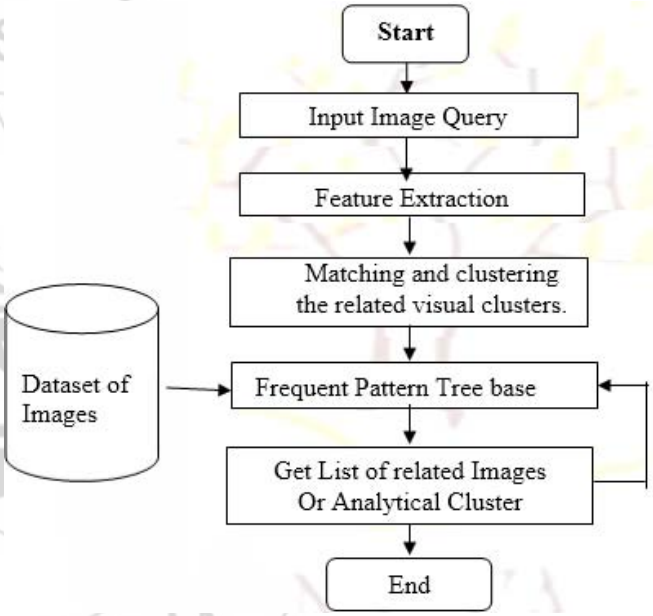

Figure 1: Architecture Diagram

2. Background

Raniah A. AlghamdiMouniraTaileb "A New Multimodal Fusion Method Based on Association Rules Mining for Image Retrieval" author in this paper deals with the combines two different data mining techniques to retrieve semantically related images: clustering and association rules mining algorithm. The semantic association rules mining is constructed at the offline phase where the association rules are discovered between the text semantic clusters and the visual clusters of the images to use it later at the online phase. The experiment was conducted on more than 54,500 images of Image CLEF 2011 Wikipedia collection. It was compared to an online image retrieving system called MM Retrieval and to the proposed system but without using association rules.

3. Overview

Image mining refers to set of tools and techniques to explore images in an automated approach to extract semantically meaningful information. The retrieval process represents a visual query to the system and extracts the images based on the user request such mechanism referred to as query-by-example and it requires the definition of an image representation a set of descriptive features and of some similarity metrics to compare query and target images. The additional mechanisms have been introduced to achieve better performance and relevance feedback proved to be a powerful tool to iteratively collect information from the user and transform it into a semantic bias in the retrieval process. RF increases the retrieval performance and it enables the system to learn what is relevant or irrelevant to the user across successive retrieval-feedback cycles. RF approaches critical issues yet unsolved. And user interaction is time consuming and tiring, and it is desirable to reduce as much as possible the number of iterations to convergence.

4. Module Division

The proposed work is divided into following four different modules:

4.1 Module 1

4.1.1 Collection of Dataset

We collect our dataset used for the proposed system from various sources. It is a data to be considered as real time police investigation reports. The documents we collected is in various formats like doc, docx, pdf, etc. Also, it is not necessary to use only pre maintained dataset rather we can use any dataset on runtime. For example: the dataset from external devices like pen drives and other. Preprocessing of text documents is necessary to clean data and to provide algorithms only the required data. The preprocessing techniques used in our system are described below:

(a) Removal of Stop Words

We maintained a stop word dictionary having all possible stop words. We scan our documents to find such stop words and remove it as well as we maintained the separate removed stop word list to keep the record for number of stop words found in particular document.

(b) Stemming

After stop word removal, we performed stemming of words. We maintained indexed stems. For first index position we kept the original stem, and then we scan the document to make the stems. For example: bail / bailed / bailing. So, if we found any word like bailed or bailing then we replace these words as bail.

(c) Synonyms

For better results, we maintained a synonym dictionary. If we don't gets accurate word matching then these synonyms could help us to create the related clusters. For example: bail, warranty, surety, bond, guarantee, and warrant. Our system finds any of word and considers it as similar word so that it places these words in same category. We put a text field to search any query by forensic analysts. There is no need to scan and manually check the cluster of interest. Instead, one can search for the interested clusters by entering any keyword or the query. We maintained the indexing of keywords and the files in which the keywords can be found. We retrieve all these files and then the above preprocessing steps are applied on these files. Thus, we get the keywords found in all files. We then input these result to three different algorithms i.e. K-means, K-medoid and K-representative. We find Jaccard coefficient as given below to calculate the similarity distance between two keywords. Thus, formed the clusters having similar group of words.

4.2 Module 2

4.2.1 Performing Indexing

The process of expressing the main subject or the theme of a text in document is called indexing. Text headings are often taken as indexes. There are primarily two categories of indexing:

- Classification
- Co-ordinate

With classification indexing, or classifying, the texts are included an appropriate class (one or several) depending on their content. All texts with basically the same semantic content are brought together. The index number of this class is assigned to each text within it and the number is then serves as its search specification.

In coordinate indexing, the basic semantic content of the text is expressed by a list of significant words selected either from the text itself or its headings or from a special normative dictionary. In the first instance, such lexical units are termed key words, and in the second descriptors. Each key word or descriptors designates a class that potentially includes all the texts that have the word in the basic semantic content.

4.3 Module 3

Performing Clustering using K-means algorithm:

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Clustering helps users to understand the natural grouping or structure in a data set. Clustering is unsupervised classification that means no predefined classes. It used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

A good clustering method will produce high quality clusters in which the intra-class (that is, intra-cluster) similarity is high. And the inter-class similarity is low. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. However, objective evaluation is problematic: usually done by human / expert inspection.

K-means clustering is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared

distances (Euclidean distances) between items and the corresponding centroid.

A centroid is "the center of mass of a geometric object of uniform density", though here, we'll consider mean vectors as centroids. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding

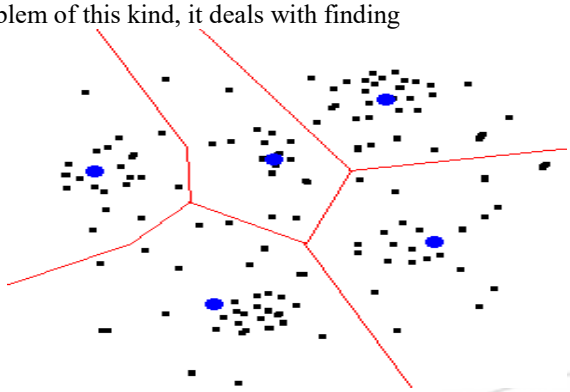

Figure 2: A clustered scatter plot. The black dots are data points. The red lines illustrate the partitions created by the k-means algorithm. The blue dots represent the centroids which define the partitions.

A *structure* in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A *cluster* is therefore a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to other clusters.

In this case we easily identify the 3 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called *distance-based clustering*, here I’m going to deal with is distance-based clustering. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

4.3.1 K-Means Algorithm

The k-means clustering algorithm is known to be efficient in clustering large data sets. This clustering algorithm was developed by MacQueen , and is one of the simplest and the best known unsupervised learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes/features, into *k* clusters, where *k* is a predefined or user-defined constant. The main idea is to define *k* centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related (in terms of similarity function; similarity can be measured by using different methods such as cosine similarity, Euclidean distance, Extended Jaccard) to all objects in that cluster.

4.4 Module 4

4.4.1 Implementation of frequent pattern tree algorithm

The FP Algorithm is an alternative way to find frequent item sets, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the item set association information.
In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity.

In large databases, it’s not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

The next subsections describe the FP-tree structure and FP-Growth Algorithm, finally an example is presented to make it easier to understand these concepts.

5. Proposed System

The proposed method is a Multimodal Fusion method based on Frequent Pattern Tree Mining. It is considered as a late fusion.

5.1 Methodologies

To achieve the objective, we have proposed following techniques:

- This method combines two different data mining techniques for retrieving: clustering and frequent pattern tree mining (FPTM) algorithm.
- It uses FPTM algorithm to explore the relations between text semantic clusters and image visual features clusters building a decision tree in the space of frequent patterns as an alternative for the two phases approach:

5.2 Online and Offline Phase

- In the offline phase, the relations among the clusters will be identified from different modalities to construct the frequent pattern rules.
- On the other hand, the online phase (retrieving phase) uses the generated Tree pattern, to retrieve the related images of the query.

6. Experimental Results

Table 1: Comparison of Frequent Tree and Apriori

Keyword (Image+Keyword)	*FP (w.r.t)*	*Aprior i(w.r.t)*
airrace.jpg+air	3.327	3.598
cake.jpg+car	3.245	5.234
wolf.jpg+water	5.234	6.098

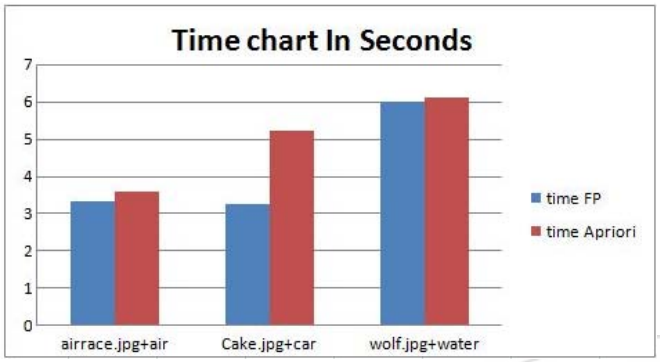

Figure 3: Comparison of Apriori and FPT in terms of time

For calculating Recall formula is,
Recall= A/ (A+B)
For Calculating precision, formula is
Precision=A/ (A+C)
Where
A - Number of relevant file retried.
B - Number of relevant files not retried.
C - Number of irrelevant file retried.

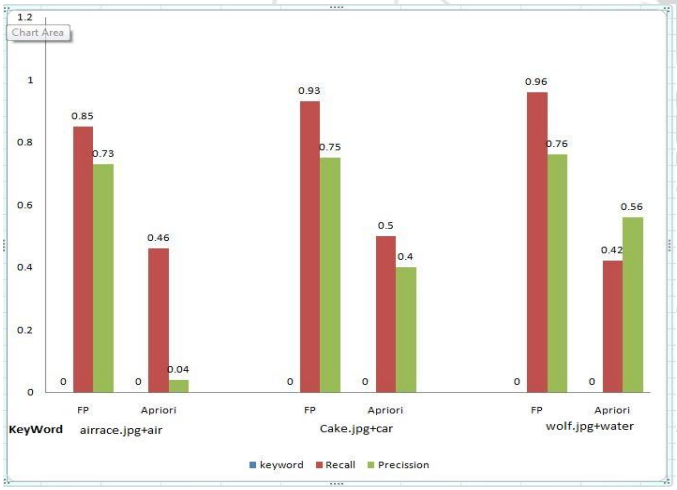

Figure 4: Comparison of Apriori and FPT in terms of recall and precision

7. Conclusion

In the proposed method, we used frequent pattern tree algorithm in our Web image retrieval system to construct a semantic relations between images clusters based on the visual features and the images clusters based on textual features for the same dataset. In this proposed system show the comparison between frequent pattern tree and apriori and found that frequent pattern tree get quickly result and show maximum search related to image and text.

In this project we propose a novel image retrieval approach which combines text, content and interactive based retrieval. The accuracy is higher in comparison to using the techniques separately. We designed a hybrid image retrieval system with the method proposed, which successfully achieves the demands with respect to the system requirements (i.e., allows the users to retrieve their desired images based on the text and/or sample image query). A new refining search algorithm has been provided, which optimizes the search results. The experiments on the sample data sets prove the effectiveness of the system.

References

[1] Raniah A. Alghamdi, MouniraTaileb, Mohammad Ameen, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia – Jeddah, "A New Multimodal Fusion Method Based on Association Rules Mining for Image Retrieval", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014

[2] GoastaGrahne, Jianfei Zhu Grahne, Member, IEEE, and Jianfei Zhu, Student Member, IEEE "Fast Algorithms for Frequent Item set Mining Using FP-Trees" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 10, OCTOBER 2005M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization," In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)

[3] S. Wei , Y. Zhao , Z. Zhu , N. Liu, "Multimodal Fusion for Video Search Reranking", IEEE Transactions on Knowledge and Data Engineering, 2010, v.22 n.8, p.1191-1199.K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, "A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II," KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[4] Adrien Depeursinge, Samuel Duc, Ivan Eggel and Henning M¨uller" Mobile Medical Visual Information Retrieval" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 17, NO. 11, OCTOBER 2011)

[5] M. Bastan, H. Cam, U. Gudukbay, and O. Ulusoy, "BilVideo-7: An MPEG-7 Compatible Video Indexing and Retrieval System", IEEE MultiMedia, 2010, vol. 17, no. 3, pp. 62-73.

[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACMSIGMOD Int'l Conf. Management of Data, pp. 207-216, May 993

[7] S. Wu and S. McClean, "Performance prediction of data fusion for information retrieval". Information Processing, Management, 2006. 42(4): p. 899-915.

[8] Ernst, R.: "Co design of embedded systems: status and trends", Proceedings of IEEE Design and Test, April–June 1998, pp.45–54

[9] H. Müller, P. Clough, Th. Deselaers, B. Caputo, "ImageCLEF" (ser. The Springer International Series

on Information Retrieval), vol. 32, pp.95 -114, 2010, Springer-Verlag.
[10] M. Ferecatu and H. Sahbi, "TELECOM ParisTech at ImageClefphoto 2008: Bi–modal text and image retrieval with diversity enhancement". In: Working Notes of CLEF 2008, Aarhus, Denmark.
[11] T. Gass, T. Weyand, T. Deselaers, and H. Ney. "FIRE in ImageCLEF 2007: Support vector machines and logistic models to fuse image descriptors for photo retrieval". In: CLEF 2007 Proceedings. Lecture Notes in Computer Science (LNCS), 2007, vol 5152. Springer, pp 492–499.
[12] R. Bellman, "Adaptive Control Process: A Guided Tour". Princeton University Press, 1961.
[13] P. K. Atrey , M. A. Hossain , A. E. Saddik and M. S. Kankanhall "Multimodal fusion for multimedia analysis: A survey", Multimedia Syst., vol. 16, no. 3, pp.1432 -1882, 2010.
[14] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu, "Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images" 2007, p. 345-357.
[15] H. Frigui, J. Caudill, and A. Ben Abdallah. "Fusion of multimodal features for efficient content-based image retrieval.", IEEE World Congress on Computational Intelligence, pp. 1992-1998, 2008
[16] Y. Liu, D. Zhanga, G. Lua, and W-Y. Ma , "A survey of content-based image retrieval with high-level semantics", Pattern Recognition, Vol. 40, No. 1. (2007), pp. 262-282.
[17] R. Datta, D. Joshi, J. LI, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age", ACM Computing Surveys (CSUR), April 2008, 40(2):1-60.
[18] T. Deselaers, T. Weyand, and H. Ney, "Image retrieval and annotation using maximum entropy", in Evaluation of Multilingual and Multi modal Information Retrieval, 2007, pp. 725-734.
[19] I. Bartolini and P. Ciaccia., "Scenique: a multimodal image retrieval interface", in Proceedings of the working conference on Advanced visual interfaces, 2008, ACM, Italy. pp. 476-477.
[20] X. Zhou, A. Depeursinge, and H. Müller, "Information fusion for combining visual and textual image retrieval," in Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos, 2010, pp. 1590-3.
[21] R. He, N. Xiong, L. Yang, J. Park, "Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval". In: International conference on information fusion. 2011.
[22] T. Tsikrika, A. Popescu, J. Kludas, "Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011". In: Working Notes of CLEF 2011, Amsterdam, the Netherlands. 2011.
[23] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Int'l Conf. Very Large Data Bases, pp. 487-499, Sept. 1994.
[24] R.J. Bayardo, "Efficiently Mining Long Patterns from Databases," Proc. ACM-SIGMODInt'l Conf. Management of Data, pp. 85-93, 1998.
[25] D. Burdick, M. Calimlim, and J. Gehrke, "MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases," Proc. Int'l Conf. Data Eng., pp. 443-452, Apr. 2001.

Author Profile

Divya Pallavi, M Tech student of Nuva college of Engineering & Technology, Nagpur, Maharashtra, India. Her area of Interest is Data Mining.

Prof. **Naziya Pathan,** received M.Tech degree from RTMNU, Maharashtra, India. She is working as a faculty of Computer Science at Nuva college of Engineering & Technology, Nagpur, Maharashtra. Her area of interest is Data Mining.