International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 3, Issue 4, July-August 2014 **ISSN 2278-6856**

Association Rules based Approach for Generating Informative ESP Game Labels

Hany Mahgoub[1], Mohamed Sakr[2]

[12]Menoufia University, Faculty of Computers and Information,
Menofia, Egypt

Abstract-*Games with a purpose (GWAP) and microtask crowdsourcing are considered two techniques of the human-computation. Image Retrieval (ImR) is an important field for every user. Every day hundreds of users go through the web sites and search for images matching given criteria. Many researchers focus on making the ImR systems more accurate. The using of GWAPs in ImR systems will make them more accurate and useful. They provide the ImR system's database with a rich of information by adding more descriptions and annotations to images. One of the systems of human-computation is ESP Game. ESP Game is a type of games with a purpose. In the ESP game, there has been a lot of work was proposed to solve many of the problems of it and makes the most benefit of the game. One of the problems of the ESP game is that it encourages players to assign "obvious" labels, which are most likely to lead to an agreement with the partner. This paper presents a new approach for generating informative ESP game labels with no need to extra un-useful game round between players using association rules mining. The results show that new informative labels can be generated automatically without any interference of extra game rounds or any human.*

Keywords: Human computation; Games with a purpose; ESP game; Association rules mining

1. INTRODUCTION

Games with a purpose (GWAP) are one of the human-computation and consider a way to make useful of the human desire to be entertained [1],[2]. Several GWAP systems have been proposed for image annotation and commonsense reasoning. Von Ahn and Dabbish [3] classified GWAP into three game-structure templates that generalize successful instances of human computation games: output-agreement games, inversion-problem games, and input-agreement games. Yuen et al. [4] added output-optimization game to these three templates. ESP game is one of the GWAP systems. ESP game was the first systems to clarify the advantages of using human computation and GWAP systems. It is example of output-agreement games and is a two player's game for labeling images [5]. Barnard et al. [6] reported that labeling images has proven to be a hard problem for computer vision, but it is something that humans can do easily. It has been shown that the image labels collected through the ESP game are usually of good quality. Moreover, the game results allow more accurate image retrieval, help users block inappropriate images (e.g., pornographic content), and improve web accessibility (e.g., the labels can help visually impaired people surf web pages [7]). In order to humans to label images, there must be some sort of motivation. One type of motivation is entertainment, which is achieved in the ESP game. In the ESP game, the players are chosen randomly and assigned the same image. Each player doesn't know the other player and the two players can't communicate with each other. The only thing they have in common is the image that they play with. Each player is asked to give a description to that image and has to guess what the other player is typing for each image to win the game and go to the next image. Once the two players have entered the same word, this word becomes the label for the image. The easiest way for both players to type the same string is by typing something related to the common image. The round lasts for 2.5 minutes. During the round the players try to describe as many images as they can. The players get number of points for each image they label. If the players agree on 15 images they get a large number of bonuses in points. Once there is a difficult image that the players can't agree on they both can press the Pass button. The game is attached with a scoreboard, with the names of players with the highest scores. Empirical studies of other peer-production systems have shown that points are a key feature in motivating users [8]. After analyzing the results of the ESP game Data set [9], we noticed that many of the labels that the player enters can be generated automatically with some ways of prediction or data mining techniques. One of the popular data mining techniques is association rules.In this paper, we address the informative labels problem and how to generate new labels with no need to extra un-useful game rounds between players. A new approach presented to generate new and more informative labels from ESP game labels using association rules as one of the data mining techniques. The resulted labels will add more useful information about the ESP game.The rest of the paper is organized as follows: A review of researchers and the work that has been done are presented in section 2. In section 3, the proposed approach for generating informative labels from ESP game is presented. In section 4 a case study for the proposed approach is presented. The results and simulation analysis of our approach are presented in section 5. Section 6 provides conclusions and future work.

2. RELATED WORK

ESP game is one of the successful applications of the games that harvest human intelligence and time to solve tasks, which is difficult by computer. Although the idea underlying the game is an extremely powerful one, more care needs to be taken in the design of the game. Moreover, the labels that are generated automatically by the players may be not very descriptive like colors. In the ESP game, there has been lot of works proposed to solve many of the previous problems and make the most benefit of the game. Games with a purpose have also been applied to obtain common sense facts as “Verbosity” [10], to locate objects within an image “Peekaboom” [11], to tag music “TagATune” [12], to elicit human-transcribed data for automated directory assistance “People Watcher” [13] and to get descriptions, rather than mere labels, for images “Phetch” [14]. Von Ahn and Dabbish [15] suggested a set of evaluation metrics, such as throughput, lifetime play, and expected contribution, to determine whether ESP-like GWAP systems are successful. Ho et al. [16] also noticed that the set of labels determined from the ESP game for an image, are not very diverse, and developed three-player version of the ESP game that involve the addition of a “blocker” to type in words that the other two players cannot use to match. Chen et al. [17] proposed a new metric called system gain, they used analysis to study the properties GWAP systems and implemented a new puzzle selection strategy to improve the GWAP systems. Jain and Parkes [18] presented game theoretic analysis for the ESP game, and they investigated the equilibrium behavior under different incentive mechanisms and provided guidelines to design incentive mechanisms. Weber et al. [19] noticed that the ESP game failed to collect informative labels so they proposed a language model to generate probabilities to the next labels to be added given the pre-added labels as training data. To the best of our knowledge, all the previous researches depend on various methods for generating more labels that require multiple game rounds. While we do not know of any research that exploit the using of data mining techniques for automatically generating informative labels.

3. THE PROPOSED APPROACH

The proposed approach for generating new informative labels from the ESP game dataset using association rules consists of three phases that are: the Preprocessing phase, Association rules mining phase and Visualization Phase, as shown in fig.1.

3.1 Preprocessing phase

The system begins with the preprocessing of the ESP game dataset. The ESP Game dataset consists of 100,000 image with their previous labels. Every image has a file named with the same name of the image and it contains all the labels that described that image as shown in fig.1 (A). All the images descriptions were obtained from the ESP game rounds. The images and all their labels must be inserted into database. It is not important to add the images in the database just their names. The table in the database is structured as follows: each row consists of image name and image label for example, if the image has five labels then there will be five records in the database. Each record contains image name and label number 1 and so on as shown in fig.1 (B). After finishing from inserting all the images names and their labels, the table of the database will contain 1444720 records. The data in the database will be represented using the binary representation model [0, 1] as shown in fig.1(C). If the label describes the image the value must be 1 if not it must be 0. At the end of this phase, the database will be exported as CSV file to be ready used in the mining phase.

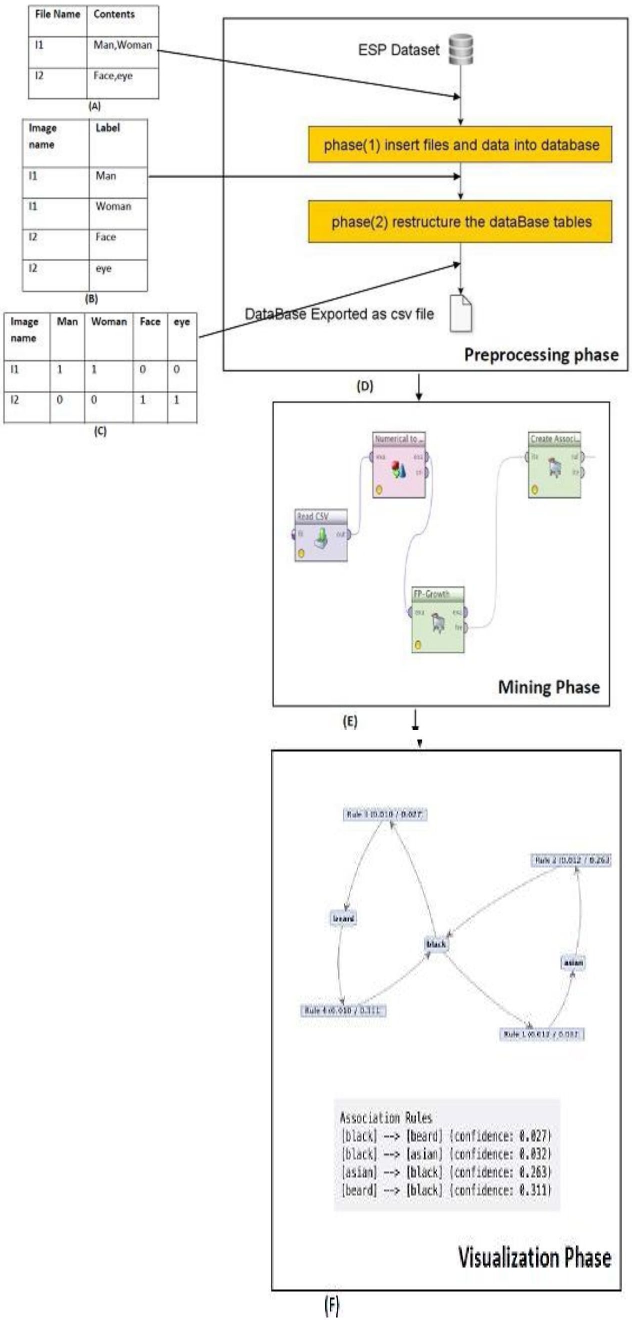
3.2 Association Rules Mining phase

Association rule mining is a popular data mining technique because of its wide application in marketing and retail communities as well as other more diverse fields. Association rule mining is a method of finding relationships of the form X→Y amongst Item-sets that occur together in a database where X and Y are disjoint Item-sets. Support and confidence measures serve as the basis for customary techniques in association rule mining. The support and confidence are predefined by users to drop the rules that are not so interesting or useful. The association rule indicates that the transactions that contain X tend to also contain Y. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. The problem of mining association rules can further be decomposed into two sub-problems:

1. Discovering the frequent Item-sets
2. Using the frequent Item-sets to generate the association rules for the database.

Many researches focus on mining frequent Item-sets as it is a hard problem when data is large.

FP-growth algorithm is usually considered as one of the fastest available algorithms to solve the problem of discovering the frequent Item-sets [20]. The major advantage of FP-growth is that it requires only two passes over the database; it is thus very I/O efficient. The two passes are used to build an FP-tree, which can be viewed as a compressed representation of the frequent items and their co-occurrence in the data. Based on the initial FP-tree, FP-growth recursively builds smaller FP-trees that are eventually used to obtain the actual frequent Item-sets.



(A) The structure of the ESP Game Dataset (consists of files)
(B) The output of phase(1) after inserting the ESP Game Dataset into the database
(C) The output of phase(2) modifying the database table to be ready to be used in RapidMiner
(D) Preprocessing phase to prepare the ESP Game Dataset to be used in RapidMiner
(E) Mining phase using RapidMiner to apply the FB-growth algorithm on the tables generated from preprocessing phase
(F) Visualization phase that shows the output rules of the RapidMiner

A frequent pattern tree (FP-tree) is a tree structure which consists of one root labeled, a set of item prefix sub-trees as the children of the root, and a frequent-item header table. Moreover, each node in the item prefix sub-tree consists of three fields: item-name, count, and node-link. Furthermore, each entry in the frequent-item header table consists of two fields: item-name and head of node-link, which points to the first node in the FP-tree carrying the item-name [21]. In the mining phase, the FP-growth algorithm is used for generating the association rules from the ESP game labels dataset. The FP-growth algorithm is applied using the RapidMiner software [22]. The RapidMiner done through four main blocks since each block accomplishes a specific task. The four blocks are Read CSV, Numerical to binomial, FP-growth and generate association rules, as shown in fig.1 (E). These blocks are illustrated as follow:

• Read CSV: it is responsible for reading the data that will be processed in the RapidMiner. From the pre-processing phase, after the data was prepared we save the database to CSV file. We input the path of the CSV file to enable RapidMiner to do mining process on the data. The output of this block delivers the CSV file in tabular form along with the Meta data.

• Numerical to binomial: it changes the type of the selected numeric attributes to a binominal type. It also maps all values of these attributes to corresponding binominal values. The data that was inserted in the CSV file was in Numerical type (0 or 1). To use these data in the RapidMiner to apply FP-growth algorithm the data must be in a binominal type. This block has one input which is example set. This input port takes an ExampleSet which it takes from the output of the Read CSV block. The output of this block is two ports ExampleSet and original. The ExampleSet is the input ExampleSet in numeric type converted to binominal type. The original is the input ExampleSet in numeric type with no changes. This is usually used to reuse the same ExampleSet in further operators or to view the ExampleSet in the Results Work.

• FP-growth: it efficiently calculates all frequent itemsets from the given ExampleSet using the FP-tree data structure. It is compulsory that all attributes of the input ExampleSet should be binominal. This block takes one input which is the ExampleSet that will be processed. The ExampleSet attributes must be in binominal type. This input is taken from the Numerical to binomial block output. There are two outputs for that block ExampleSet and frequent sets. The ExampleSet is the input that was passed to the block with no modifications. This is usually used to reuse the same ExampleSet in further operators or to view the ExampleSet in the Results Work. The frequent sets port gives the frequent itemsets with the specified min support. The min support is specified in the configuration of this block.

• Generating association rules: it generates a set of association rules from the given set of frequent itemsets. This block has one input itemsets which expects frequent itemsets. The frequent itemsets is given from the output of the FP-growth block. The output of this block is itemsets and association rules. The itemsets give the original itemsets that were given in the input with no changes in it. This is usually used to reuse the same itemsets in further operators or to view the itemsets in the Results Workspace. The association rules output port delivers the association rules with specified confidence. The confidence can be

International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 3, Issue 4, July-August 2014 ISSN 2278-6856

specified in the configuration of this block.All the resulted rules can be applied on the database to generate the new records.

3.3 Visualization Phase

In this phase, the generated association rules can be represented as graphical representation or textual representation as shown in fig.1 (F). After generating the rules from RapidMiner, these rules will be viewed as nodes which represent the label and arrows which represent the relations between them.

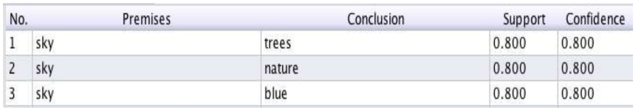
4. CASE STUDY

This case study illustrates the mining scenario and how the output of the mining process will be useful for the ImR system. This scenario is called association rule mining (ARM). Suppose the data that shown in table 1 to be a data collected from the ESP game as the first column identifies the image name and the second column identifies all the annotations associated to that image. For the images I1 and I2 the annotations associated with them are (Sky ,Blue ,Nature ,Trees ,Clouds) these annotations came all together about 40% of the total records in the table. In data mining it proves that all these annotations are related to each other and the occurrence of some of them guarantees the occurrence of the other. To prove that the labels have relations between them and that the occurrence of a given label will lead to conclude the occurrence of another label the dataset in table 1 is analyzed using RapidMiner software. The Results of the RapidMiner is shown in fig 2 with confidence 0.5 and support 0.5. In fig 2 the output of the RapidMiner is described as the premises and conclusion with support and confidence. The results shows that the label sky when occurred the label "trees" may be occurred by level of confidence 0.8 and support 0.8 the same for the labels "nature" and "blue". So after analyzing the output rules of the RapidMiner it was found that in the image named I3 that contains the annotations (Sky ,Blue ,Nature , Clouds) contain the label "Sky" and using the rules generated from RapidMiner it can be concluded that the labels "Trees","nature" and "blue" can all be added as a labels for that image I3. But the image I3 already contains the labels "Nature" and "blue" so the label "Trees" will only be added.

The same for I4 and I5 with labels "Blue" and "Nature" respectively. In fig 3 and fig 4 the output of the RapidMiner is shown using two different shapes Association Rules and Graph. So after the mining process the relations were generated and then apply the relations to the records of the database. The processes of mining and applying relations to the Database will improve the ImR system. By using the ARM if any one searches for the images that contains the word "Trees" the ImR system will retrieve the images I1 ,I2 ,I4 and I5 and will not retrieve I3 But after the ARM the ImR system will also retrieve the image I3 in addition to I1 ,I2 ,I4 and I5. This improves the ImR system as new images were retrieved from the system. This means that the using of ARM improves the ImR system by adding new annotations to the database which lead to retrieve new images.

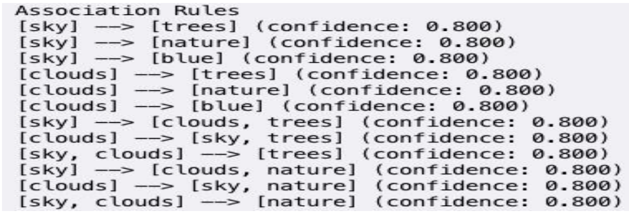
Table 1: Collection of dataset

Image Name	Image annotations
I1	Sky ,Blue ,Nature ,Trees ,Clouds
I2	Sky ,Blue ,Nature ,Trees ,Clouds
I3	Sky ,Blue ,Nature ,Clouds
I4	Sky , Nature ,Trees ,Clouds
I5	Sky ,Blue ,Trees ,Clouds



No.	Premises	Conclusion	Support	Confidence
1	sky	trees	0.800	0.800
2	sky	nature	0.800	0.800
3	sky	blue	0.800	0.800

Figure 1 The output of RapidMiner described as premises and conclusion



```
Association Rules
[sky] --> [trees] (confidence: 0.800)
[sky] --> [nature] (confidence: 0.800)
[sky] --> [blue] (confidence: 0.800)
[clouds] --> [trees] (confidence: 0.800)
[clouds] --> [nature] (confidence: 0.800)
[clouds] --> [blue] (confidence: 0.800)
[sky] --> [clouds, trees] (confidence: 0.800)
[clouds] --> [sky, trees] (confidence: 0.800)
[sky, clouds] --> [trees] (confidence: 0.800)
[sky] --> [clouds, nature] (confidence: 0.800)
[clouds] --> [sky, nature] (confidence: 0.800)
[sky, clouds] --> [nature] (confidence: 0.800)
```

Figure 2 The output of RapidMiner described as Association Rules

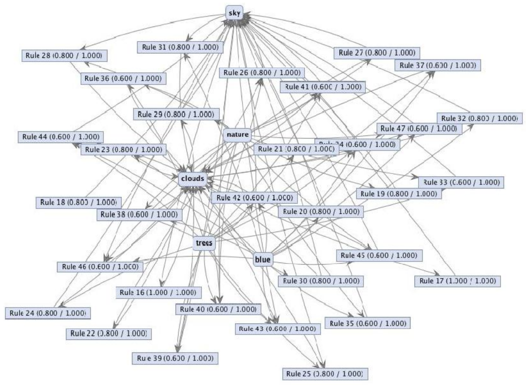


Figure 3 The output of RapidMiner described as Graph

4. RESULTS AND SIMULATION ANALYSIS

At the beginning, we start the experiments by applying the FP-growth algorithm on all the datasets once but we noticed that the execution time takes up to 15 hours and it didn't finish processing. Because of the delay of the execution time we divided the data into partitions and apply the FP-growth on each partition. The experiments were done on a Mac OS x machine with 2.5 GHz Intel core i5 processor with 8 GB 1600 MHz DDR3 memory and the database server was MySQL [20]. We prepared samples from the database with 53930, 55353, 91952, 87537 records. We applied the FP-growth algorithm on each data sample with Support 0.001 and 0.01 and Confidence 0.001. The results are shown in Table 2, moreover the diagrams for the results are shown in fig 5 and fig 3. The results shown that for the first sample of the dataset

(53930) when the support was 0.001 and confidence 0.001 the number of rules that were found are 206 rules. The support of 0.001 means that we want the rules that occur more than or equal (53930*0.001) about 53 time. For example if the labels man and face are usually come together give me a rule that containing both of them if the number of there occurrence together are more than or equal 53. For the confidence of 0.001, it means conditional probability that, given man present in a record, face will also be present with percentage and it is calculated using support (man, face)/support (man). Sample of the association rules that was generated from RapidMiner is shown in fig 7 and fig 8. We noticed that when we increase the minimum support, the number of rules is decreased. Also we noticed that the extracted association rules generate new labels for the images. This leads to improve the ImR system.

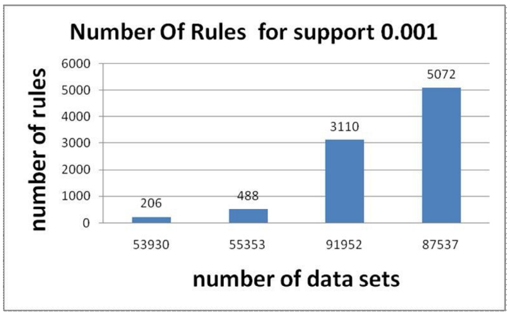


Figure 4 The number of rules at support 0.001

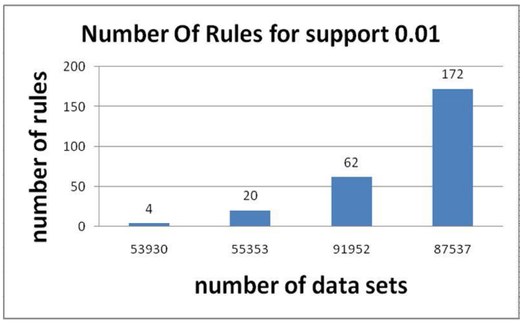


Figure 5 The number of rules at support 0.01

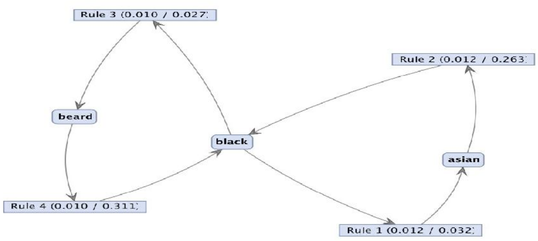


Figure 6 The graphical representation of rules

```
Association Rules
[black] --> [beard] (confidence: 0.027)
[black] --> [asian] (confidence: 0.032)
[asian] --> [black] (confidence: 0.263)
[beard] --> [black] (confidence: 0.311)
```

Figure 7 The textual representation of rules

Table 2: The number of association rules

Number of data sets	Support	Confidence	Number of Rules
53930	0.001	0.001	206
	0.01	0.001	4
55353	0.001	0.001	488
	0.01	0.001	20
91952	0.001	0.001	3110
	0.01	0.001	62
87537	0.001	0.001	5072
	0.01	0.001	172

6. CONCLUSION

This paper introduced an efficient approach for solving one of the human computation problems. The using of Association Rule Mining made an effective role in improving the ImR systems by generating more informative ESP game labels. The results show that new labels can be generated automatically without any interference of extra game rounds or any human. In the future, work we intend to take into account the execution time factor by using cloud computing and parallel processing.

References

[1] C. Harris and P. Srinivasan , “Comparing Crowd-Based, Game-Based, and Machine-Based Approaches in Initial Query and Query Refinement Tasks,” Lecture Notes in Computer Science, vol. 7814, pp. 495-506, 2013.

[2] S. Thaler, E. Simperl and S. Wolger , “An Experiment in Comparing Human-Computation Techniques,” Internet Computing, IEEE, vol. 16(5), pp. 52-58, October 2012.

[3] L. Ahn and L. Dabbish, “Designing games with a purpose,” Communications of the ACM, vol. 51(8), pp. 58-67, August 2008.

[4] M. Yuen, L. Chen and I. King, “A Survey of Human Computation Systems,” Proceeding CSE '09 Proceedings of the 2009 International Conference on Computational Science and Engineering, vol. 4, pp. 723-728, August 2009.

[5] L. Ahn and L. Dabbish, “Labeling images with a computer game,” Proceeding CHI '04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319-326, 2004.

[6] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei and M. Jordan, “Matching Words and Pictures,” Journal Of Machine Learning Research, vol.3, pp. 1107–1135, 2003.

[7] J. Bigham, R. Kaminsky, R. Ladner, O. Danielsson and G. Hempton, “WebInSight : Making Web Images Accessible,” Proceeding Assets '06 Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility, pp. 181 - 188, 2006.

[8] K. Nam, M. Ackerman and L. Adamic, “Questions in , Knowledge iN ? A Study of Naver ’ s Question Answering Community,” Conference on Human Factors in Computing Systems, CHI 09, ACM, April 2009.

[9] ESP Game dataset: http://server251.theory.cs.cmu.edu/ESPGame100k.tar.gz

[10] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In Conference on Human Factors in Computing Systems (CHI'06), pages 75–78, 2006.

[11] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In Conference on Human Factors in Computing Systems (CHI'06), pages 55–64, 2006.

[12] E. Law, L. von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. International Conference on Music Information Retrieval (ISMIR'07), pages 361–364, 2007.

[13] T. Paek, Y.-C. Ju, and C. Meek. People watcher: A game for eliciting human-transcribed data for automated directory assistance. In Annual Conference of the International Speech Communication Association (INTERSPEECH'07), pages 1322–1325, 2007.

[14] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In Conference on Human Factors in Computing Systems (CHI'06), pages 79–82, 2006.

[15] L. Ahn and L. Dabbish, "Designing games with a purpose," Communications of the ACM, vol. 51(8), pp. 58-67, August 2008.

[16] C. Ho, T. Chang, J. Lee, J. Hsu and K. Chen, "KissKissBan: a competitive human computation game for image annotation," Proceeding HCOMP '09 Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 11-14, 2009.

[17] L. Chen, B. Wang and K. Chen, "The Design of Puzzle Selection Strategies for GWAP Systems," Journal Concurrency and Computation: Practice & Experience, vol. 22(7), pp. 890-908, May 2010.

[18] S. Jain and D. Parkes, "A Game-Theoretic Analysis of Games with a Purpose," Proceeding WINE '08 Proceedings of the 4th International Workshop on Internet and Network Economics, pp. 342-350, 2008.

[19] S. Robertson, M. Vojnovic and I. Weber, "Rethinking the ESP Game," Proceeding CHI EA '09 CHI '09 Extended Abstracts on Human Factors in Computing Systems, pp. 3937-3942, September 2009.

[20] Jiawei H, Jian P, and Yiwen Y. Mining Frequent Patterns without Candidate Generation. Proceeding SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data. June 2000 29(2),1-12.

[21] A. Choubey, R. Patel and J.L. Rana, "A survey of efficient algorithms and new approach for fast discovery of frequent itemset for association rule mining (DFIARM)," International Journal of Soft Computing and Engineering (IJSCE) Volume-1, Issue-2, May 2011.

AUTHOR

Hany Mahgoub received the B.Sc. in Pure Mathematics and Computer Science and M.Sc. in Computer Science from Menoufia University, Faculty of science in 1993 and 1999, respectively. He received his PhD in Computer Science from Menoufia University, Faculty of Computers and Information, Egypt in 2010. His research interest includes data, text and web mining, Information Retrieval, software testing, distributed systems, intelligent systems, human computation and bioinformatics.

Mohamed S. Sakr received his B.Sc degrees in Computer Science from Menoufia University, Egypt in 2011. His research interests lies in the areas of human computation, data mining and cloud computing and currently he is demonstrator at Computer Science department in the Faculty of computers and information, at the University of Menoufia, Egypt.