Persian Multimedia Search Services' Users Propensities

Maryam Mahmoudi*, Masomeh Azimzade*, Mehdi Esnaashari**, Mojgan Farhoodi*, Reza Badie*
* Information Technology Research Group
Information and Telecommunication Research Center (ITRC)
Tehran, Iran
{Mahmoudy, Azim_ma, Farhoodi, Badie}@itrc.ac.ir
** Faculty of Computer Engineering
K. N. Toosi University of Technology
Tehran, Iran
esnaashari@kntu.ac.ir

***Abstract*— Nowadays, search engines are prominent tools, which are required by users, for finding information in web. Multimedia search engines are of special importance due to two different reasons; 1) attractiveness of multimedia contents and 2) growing rate of the creation and online dissemination of such contents. In this paper every effort is made to analyze and recognize the propensities of the users of Persian multimedia search services. For this purpose, behaviors of Iranian users of Parsijoo's image, voice and video search services has been studied by analyzing its usage log files. The analyses, which have been carried out by using users' queries for a time period of three months, can be categorized into two distinct types; holistic analyses and the ones based on using frequently used queries. The results of the analyses have shown that users are mostly after entertainments and amusement topics when they use multimedia search services.**

Keywords— Multimedia search services; Users' behavior analysis; Usage log files; Search engine;

I. INTRODUCTION

Enhancement of computer hardware capabilities on a daily basis and constant increase in the available bandwidth for accessing the internet motivate users to increase their usage of online multimedia services [5]. At the same time, the amount of available online multimedia contents and the number of websites which provide such contents have a noticeable growing rate. This phenomenon motivates search engine companies all around the world to provision multimedia search services.

Considering the huge volume of multimedia contents, indexing and making the contents searchable is a demanding task, especially in large scale scenarios, which asks for significantly high hardware and bandwidth requirements [6]. To conquer their hardware and bandwidth limitations, search engine providers must be able to recognize the propensities of users and their topics of interests in order to prioritize the indexing of different kinds of multimedia contents. This way, they will be able to provide high quality answers to users' queries.

Until now, user propensity recognition is verified in numerous papers [7-14]. Also for a Persian search engine, various researches have been carried out to analyze users' behaviors and recognize their propensities [1-4]. In all of these researches, one of the two approaches which are introduced below, were used for aforementioned purposes:

1- Using search engine's log file: In the papers of this category [2-4] [7-11], contents of the log file of a specific search engine for a particular time interval has been gathered and parsed. Then based on the achieved information, intended analyses have been carried out.

2- User's implicit behavior evaluation: Papers in this category [1] [12-14] have tried to track users actions during the time that they had interaction with the search service and in this way infer their tendencies and propensities. For tracking users, the researchers often use a plugin for a particular web browser. This plugin logs different activities of users, such as clicking on one of the URLs in the result set constructed by the search service, time elapsed in each search session, time spent reading a web page and etc. Then the acquired information have been sent to a central server where from the gathered information, users tendencies have been inferred.

The common aspect in these papers is that they all focus on textual search engines. There are a few number of researches devoted to multimedia search services for this same purpose. For instance, researchers in [15] have shown that more than 25 percent of the words used by Excite image search engine users were sexual and nudity–related ones and that there was a great amount of diversity in such words used. Also Spink and his colleagues in [16] concluded that users would often examine more web pages from the result set of sexual-related queries. In [17] researchers evaluated the behaviors of the Dogpile multimedia meta-search engine users in 2006. The results of these investigations have shown that among multimedia queries, 50% are devoted to image search service, 28% are of video search type and the remaining 22% belong to voice search service. All the frequently used video and image search queries were of sexual and nudity-related type. Frequently used voice queries contained song names or singer names. In [18] by leveraging a browser plugin and considering the users of the

978-1-5386-0420-5/17/$31.00 ©2017 IEEE

Yahoo! Search engine, the researchers have observed that adolescents within the age range of 10 to 19 use multimedia search services 2.4 times more than adults using the same services.

To the best of our knowledge, no research has been carried out in the context of the recognition of user propensities of Persian multimedia search services. In this paper, by investigating the log file of Parsijoo's multimedia search services, every effort has been made to provide an analysis on users' behavior of such services. Currently this search engine provides image, voice and video search services with the following characteristics:

- Image search service: About 100 million images have been indexed and this service is capable of handling 40,000 requests per day.
- Voice search service: This service makes it possible to search, download and play more than two million sounds and music files available in the Persian web. This service receives 4,000 requests per day.
- Video search service: Approximately, one million video files have been indexed and this service handles about 6,000 requests per day.

The analysis accomplished in this paper include: languages used in queries, typos in queries, whether queries sent from the internet or the intranet, and classification of queries. The results achieved from such analysis can identify the hot topics which are the most popular topics among the users of the multimedia search services. Hot topics identification can help the search service provider to prioritize well the crawling and indexing activities of the multimedia contents available on the internet.

The rest of this paper is organized as described below. In the second section, the steps taken for carrying out the analysis are described. The third section presents and verifies the results achieved from the accomplished analyses. The fourth section concludes this paper.

II. ANALYSIS STEPS

In order to be able to recognize users' propensities and behaviors in the usage of Persian multimedia search services, one must first gather required data, then carry out normal data cleaning and preparation and finally perform the various analysis tasks which are seen useful.

A. Data Acquisition

In this paper the data source that was used is the search engine's log file contents. A portion of this log file, which was for a period of 3 months was considered for the designated analysis tasks.

B. Data Cleaning and Preparation

In this step, firstly the log file contents were read and parsed based on special tags, in order to extract data about different multimedia search services (i.e. image, voice and video search services). Next, requests, which were made by other search engine bots, were identified and removed from the extracted data. Resulted data were then stored in a MySQL database. This database is then used as the input for data analysis process.

In addition, among the queries in the aforementioned time period, top-200 frequently used queries were also extracted separately for each multimedia service and stored in the same database for further analysis.

C. Data Analysis

Two types of analysis have been carried out for each of the image, voice and video search services. In the first type of analysis, which is called the holistic analysis task, all the queries from the considered time period were used during the investigations. The statistics gathered in the holistic analysis task are as below:

- The total number of queries received by each of the multimedia search services,
- The average length of queries received by each of the multimedia search services,
- Total number of queries sent from the internet or the intranet,
- Number of unique IP addresses that access each of the services
- Geographical distribution of the users of each multimedia search service

In the second type of analysis, the evaluation has been performed by considering only the most frequently used queries of each search service. This type of analysis is of special interest due to the fact that most frequently used queries are indicators of the topics which are most appealing to users. Analyses of this type are further classified into two different groups; general and specific. In the group of general analyses, we seek to provide analyses which are common among all multimedia search services. These include queries languages, problematic queries (those queries which have writing problems, contains sexual or nudity-related words, or like the query "((((((((((" has no meaning), whether queries were sent from the internet or the intranet.

On the other hand, in the group of specific analyses, each multimedia search service is considered separately for analysis. The aim was to identify the tendencies and inclinations of Persian users towards each multimedia search service. These analyses are dependent to the nature and characteristics of the queries used and the search service considered and as a result can be regarded as content based analyses. As an example, it is possible that identified classes of queries for each search service is different from that of other search services. In the voice search service, for instance, the classification is done based on names of singers, names of the albums, etc. which is completely irrelevant in other search services.

Yet, another example of analyses, within the group of specific ones, is that of "indicating whether a query is a specific or a general one". This analysis considers the degree of generality of queries used in each of the search services. For example the specific queries are those that are asking to get

information about a properly indicated entity which very intensively narrows down the domain of relevant web pages. For instance, a query that mentions the name of a person is a very specific query, whereas, rather ambiguous queries such as "film" or "clip" can be considered as general queries in the video search service.

III. Analyses Results

In this section, results of analyses, which were carried out on Persian multimedia search services, will be presented. The analyses performed on the usage log files within the time period of August 14th, 2015 up to November 14th, 2015.

A. holistic analyses

The first analysis was carried out to understand the average length of the queries sent to multimedia search services. By Fig. 1 you can observe that the average lengths of queries for all three services are greater than two words. This shows that Persian users mostly specify their information needs using two or three-word queries when using multimedia services. Another point obvious in Fig. 1**Fig. 1** is that the average length of queries sent to the video search service is lower than that of the image and voice services. This means that it is possible for Persian users to specify their information need for video contents by applying fewer keywords.

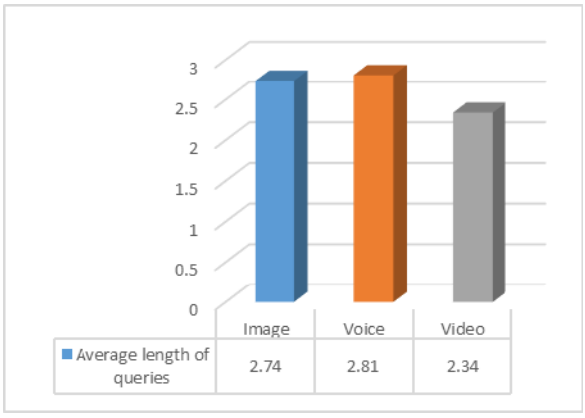

Fig. 1: The average lengths of queries for each multimedia service

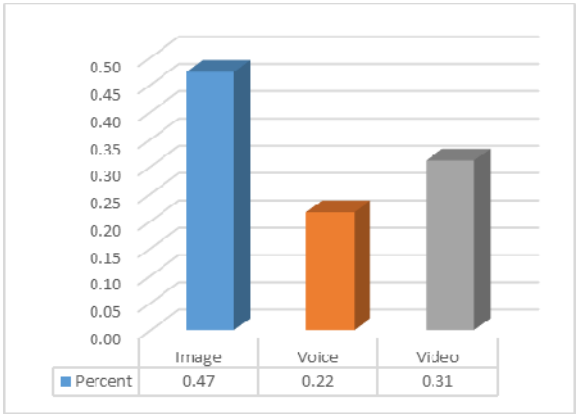

Fig. 2: Number of queries sent to each multimedia service

Another analysis was to find out how Persian people are engaged in different multimedia services. To this end, we have computed the number of queries sent to each of the services separately within the aforementioned time interval. Results, which are shown in Fig. 2, indicates that users were more engaged in the image search service rather than the other services. The number of queries sent to the voice search service was the least among all.

Fig. 3 depicts the number of queries received from the internet and the intranet users separately. For the image and video search services, the number of queries originating from the internet is higher than those originating from the intranet, whereas the opposite situation is present for the voice search service.

Fig. 4 shows the number of unique IP addresses that access each multimedia service during the aforementioned time interval. As it can be seen in this figure, the number of unique IP addresses for the voice service is the highest meanwhile this number is the lowest for the video search service. Perhaps the reason behind these statistics is the higher attractiveness of voice search service which draws more users toward itself.

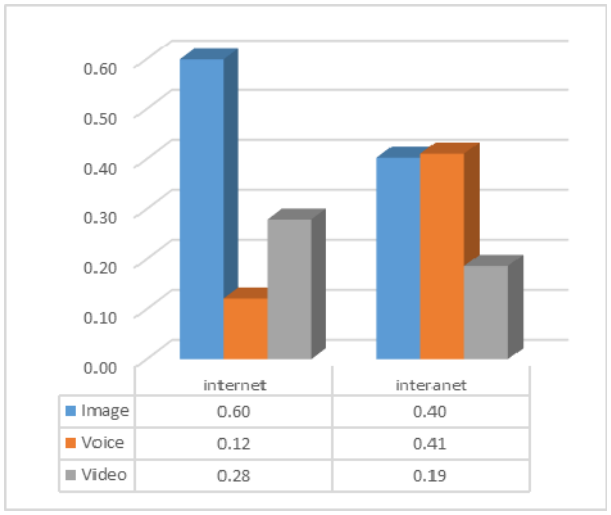

Fig. 3: Number of queries originated from internet or intranet

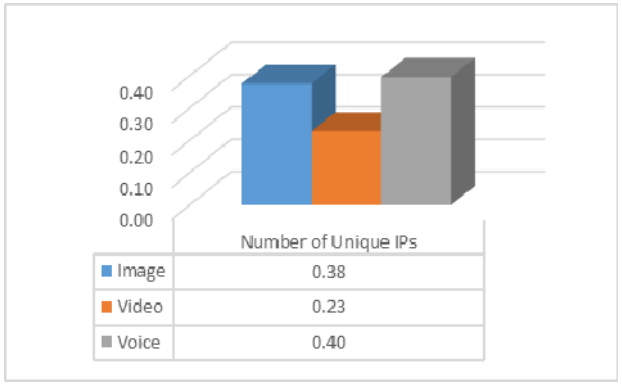

Fig. 4: Number of unique IP addresses

In order to be able to identify the geographical distribution of the users of multimedia search services, a mapping from IP addresses to Iran provinces is created. This mapping was extracted from the https://www.iplocation.net/ website. Before presenting the results of this measure, it is worthy to express that the geographical location for intranet IP addresses were unavailable so this measure is only calculated for queries sent over the internet. Fig. 5, Fig. 6, and Fig. 7 show the results for this measure separately for each of the multimedia search services. These figures show that the amount of inclinations toward multimedia services is different in various provinces of the country; Most of the usage of the image search service is originated from Ardebil, Isfahan and Yazd, that of voice search service is coming from Isfahan, Hamadan and Tehran and for the video search service, it is originated from Isfahan, Hamadan and Yazd. The point worthy to state is that Isfahan is among the three provinces with the highest quota of usage for all the services.

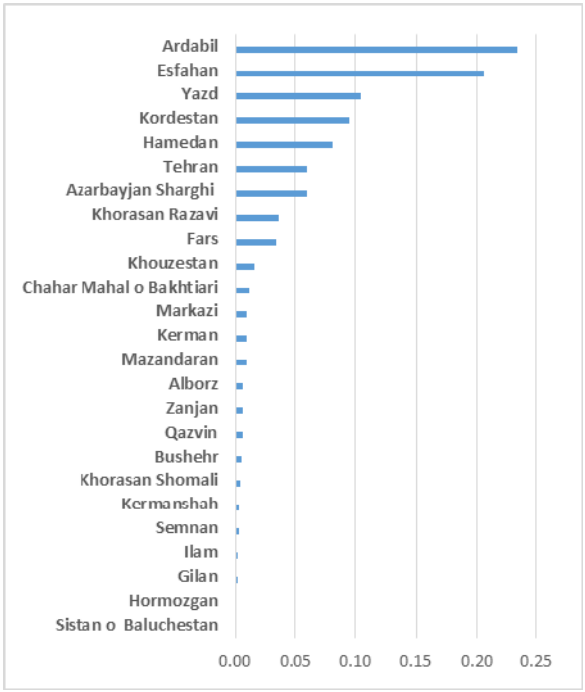

Fig. 5: Geographical distribution of the users of image search service

The other important point that should be noted here is that the highest amount of usage of image search service (approximately 40,000 requests) had happened in Ardebil. For voice search service the maximum usage were from Isfahan (about 20,000 requests) and for video search service the maximum usage was for Isfahan (roughly 12,000 requests). The interesting point here is that although the number of unique IP addresses were the highest for the voice search service but the number of queries sent to this service (in Isfahan province) was half of the number of queries sent to image search service in Isfahan. In other words, despite the higher amount of tendency within users for leveraging the voice search service, the number of queries sent by them to this service is less than the other services. From this observation we can infer that althoug a high number of users were eager to use voice search service, but because probably they weren't satisfied from the results this service provide, they abandon using the service. As a result the number of queries received by this search service is low. So we can conclude the maturity and effectiveness of the image search service of Parsijoo is higher than its voice search service.

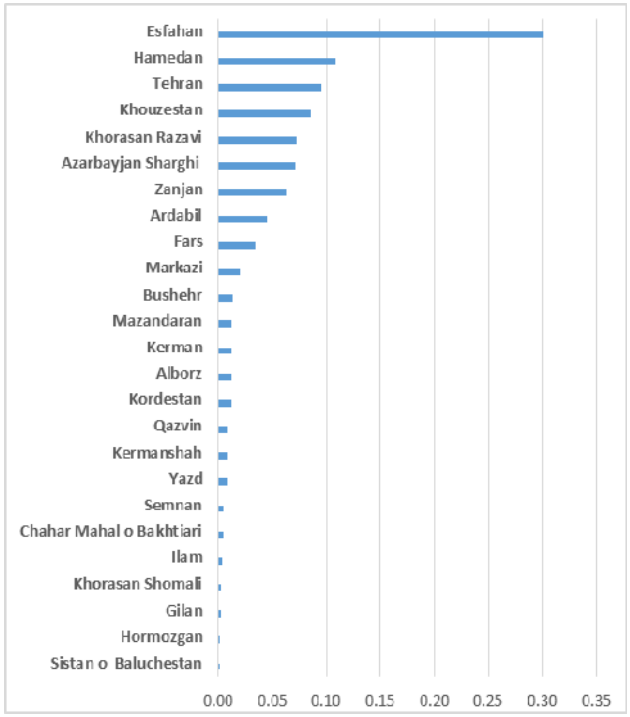

Fig. 6: Geographical distribution of the users of voice search service

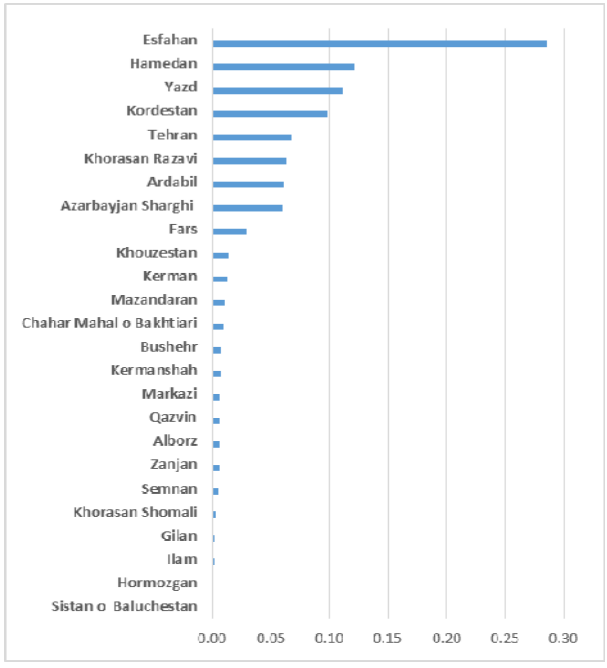

Fig. 7: Geographical distribution of the users of video search service

B. *Analysis based on frequently used queries*

1) *General Analyses*

In this section, the results of the analyses, carried out by using top-200 frequently used queries, are explained. As a sample, the top-5 frequently used queries for each multimedia search service is given in Table 1.

In Fig. 8 the percentage of queries written in English or Persian Language is shown for each of the services. As it is obvious from the figure, the percentage of queries written in English language is less than 10 %. This means that Iranian users primarily use Persian language for typing their queries. The interesting point here is that by considering the top-200 frequently used queries of the image, voice and video search services it is observed that the only service for which users use English and Persian languages together and compositely is the voice search service. Also the verifications have shown that such multilingual queries are used only when users are searching for non-Persian music files.

In this research the quality of queries written by users is also studied. In this study, the problematic queries, consists of ambiguous, erroneous, meaningless, or sexually and nudity-related queries, were counted and compared with the amount of non-problematic queries. As it is depicted in Fig. 9, in all three search services, lower than 20% of queries are problematic ones. Most problematic queries were ambiguous ones followed by meaningless queries. Meaningless queries can show that users either had no specific goal or information need or pursued other targets, such as planning an attack to the service, using such queries.

Another measure that was calculated is the percentage of queries originated from the internet or the intranet. As it is depicted in Fig. 10, the image and video search services receive more queries from the intranet than from the internet. This is completely in opposite for voice search service, which receives most of its queries from the internet. The reason behind this phenomenon could be the lack of the availability of any other similar voice search service in the internet in place of Parsijoo's voice search service.

TABLE I. TABLE 1: TOP-5 FREQUENTLY USED QUERIES FOR EACH MULTIMEDIA SEARCH SERVICE

Service	Persian Query	Query in English	Frequency of repetition
Image	داعش	ISIL	9418
	محرم	Moharram	7144
	صدف طاهریان	Sadaf Taherian	5532
	عکسهای بد حجاب بازیگران	Pictures of Badly Veiled Actresses	4689
	گل	Flowers	3792
Voice	کد پیشواز میثم مطیعی	Meysam Motiee Ring Back Music	976
	زمین	Earth	198
	علیرضا افتخاری	Alireza Eftekhari	127
	سزن اکسو	Sezen Aksu	109
	ناصر عبدالهی	Naser Abdollahi	98
Video	کریمی وقت جدایی	Karimi the time of separation	17816
	داعش	ISIL	11429
	کلیپ خنده دار	funny clips	4964
	اپارات	aparat	3394
	دوربین مخفی	hidden cameras	3336

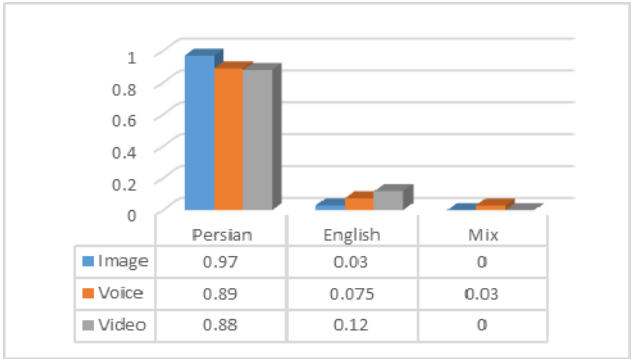

Fig. 8: Languages used by the users of multimedia services

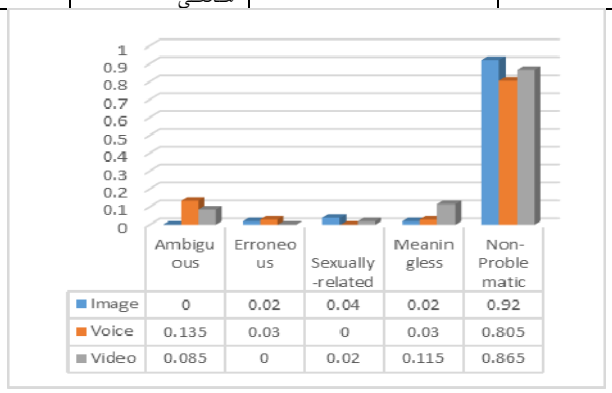

Fig. 9: Percentage of problematic and non-problematic queries

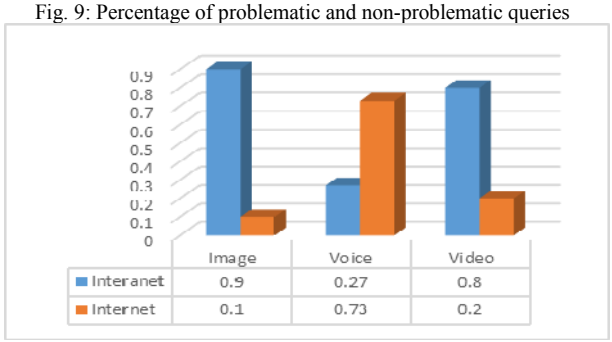

Fig. 10: The percentage of intranet and internet users of multimedia search services

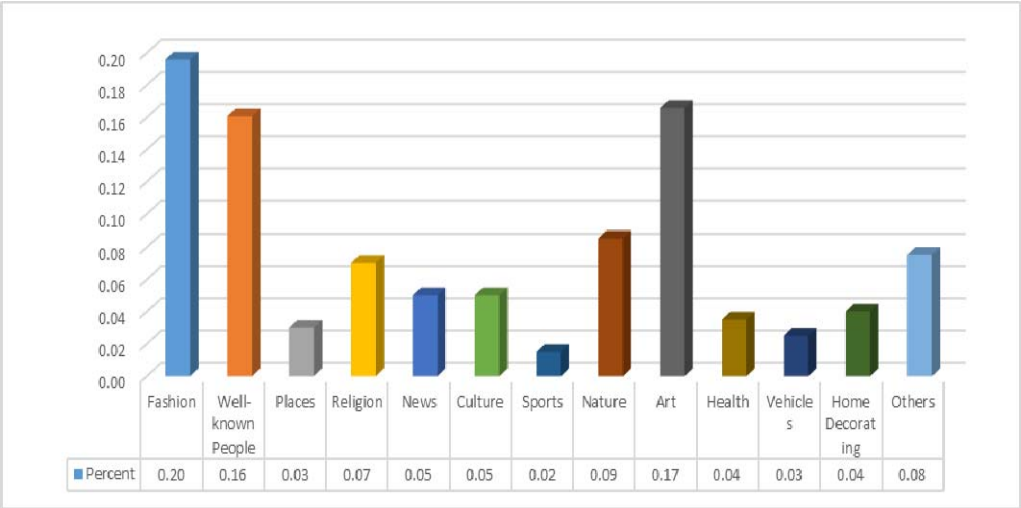

Fig. 11: Classification of image search service queries

Another aspect that was under consideration is the degree of specificity of the queries written by users for conveying their information needs. According to investigations carried out, the manner with which users interact with each of the multimedia search services is different; For using voice search service, users usually utilize specific queries whereas, for the other two search services, they mostly prefer to use general keywords. Fig. 12 presents the exact results of this analysis.

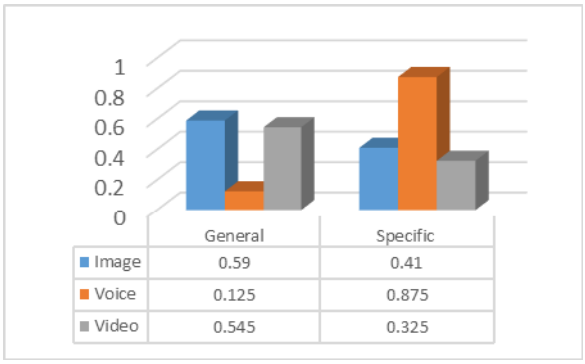

Fig. 12: Percentage of specific or general queries used in multimedia search services

2) Specific Analyses

a) Image Search Service

The image search service provides the opportunity for searching images that are crawled primarily from Persian web pages.

The results of the classification of queries used by image search service users are depicted in Fig. 11. It can be concluded from this figure that that the three most popular and attractive topics for the users of this search service are fashion, Art, and well-known people. Within the fashion class of queries, queries related to women, like Manto, dress or eveningwear, are dominant. It seems that Art-related queries are also more interested by women. Queries like weaving and knitwear learnings and tutorials, crystal flowers and etc. are belong to this class. In the group of well-known people, the majority of queries were typed with the aim of finding images of actresses, especially Iranian actresses.

On the other hand, the least frequently used queries were those about topics like sports, vehicles, and places. By considering the fact that topics of sports and vehicles are more interested by men rather than women, we can infer that probably a higher percentage of image search service users were women. We should mention that this is only a speculation that needs more investigations for proving. The low quantity of the queries under the topic of places, such as queries for finding historical and touristic places, is quite weird and surprising, since the majority of image queries worldwide are often of such class. The reason of this observation could be the higher quality and richness of results from search services like Google and Bing for this class of queries which inclines users to use them instead of using Parsijoo.

At the end, it is worth mentioning that a considerable portion of queries are classified as others. These queries were so general to be placed in a more concrete and specific class of queries. For instance, one may consider queries like “funny pictures”, “beautiful fancy pictures” and “appealing and attractive pictures”. From such queries it is inferable that the user itself was not sure what he wants exactly and only wanted to examine the outputs of the search services by using some broad and general words. The abundance of such queries among all the queries received by the search service may imply that the users did not trust this service and were only after verifying the quality of its result sets.

b) Voice search service

The voice search service provides the possibility for searching, downloading and playing sound and music files for users.

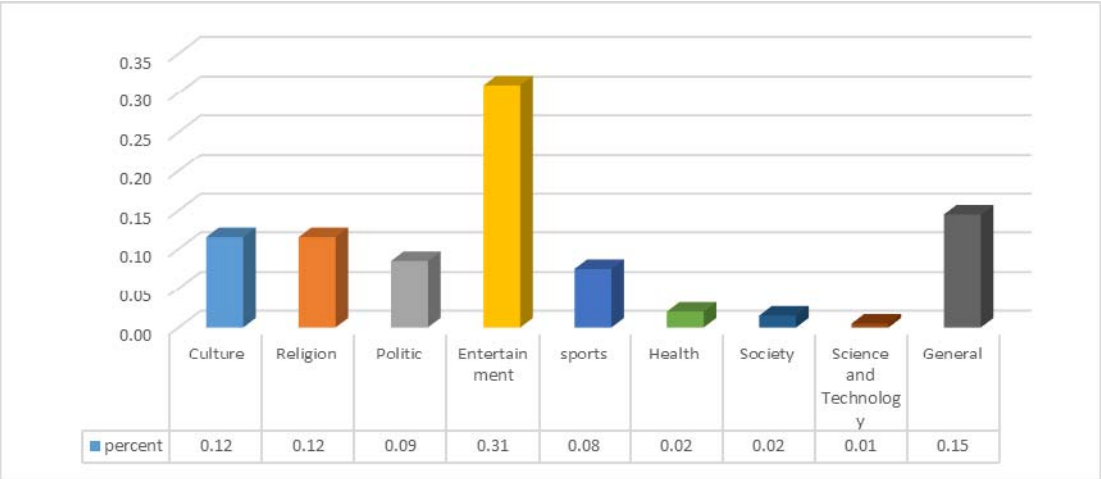

Fig. 13: Classification of video search service queries

One of the analyses made in this research activity was the verification of the way users express their information needs through queries while they are using the voice search service. As it is depicted in Fig. 14, the majority of the users have used names of artists in their queries. The next alternative is the song or the music name. Other ways of expressing information need in such a service are to provide album names, music styles and even words or phrases from the lyrics of a music file. Detailed results of this study and the percentage of utilizing any of the aforementioned ways of expressing information need can be obtained from Fig. 14.

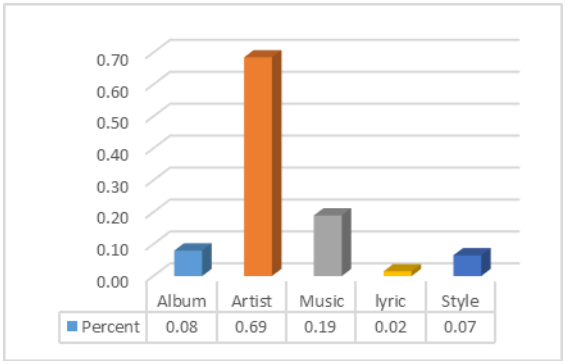

Fig. 14: Different ways of expressing information need when using voice search service

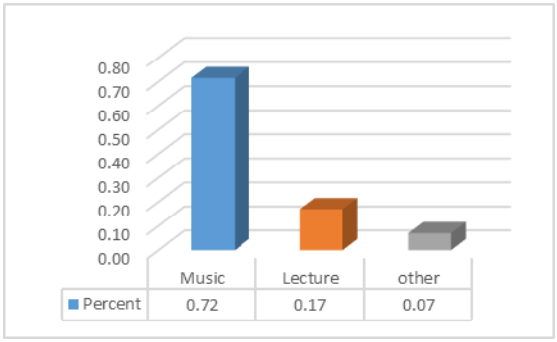

Fig. 15: Classification of voice search service queries

One another analysis which has been accomplished is the classification of voice search service queries. Considered classes for this search service are music, talk, and others. As it is shown in Fig. 15, it is perceivable that the music files are the target of the majority of the demands of users among the mentioned classes of queries.

c) Video Search Service

The video search service makes it feasible search for video files which are primarily extracted from the Persian web pages.

The results of the analysis performed on the usage of video search service are depicted in Fig. 13. It can be concluded from this figure that there is a considerable variety in the topics the users are after when they use this service, but the majority of their information needs (approximately 31%) is about entertainments and amusement topics like movies, concerts or live music Performances, comical and funny clips, etc. The next popular class of queries among users of this service is the class of general queries which constitutes 15% of all the queries. Queries like “movie” and “video” are among this class. This shows that for this service, like for the image service, users were more involved in examining the capabilities and the quality of results rather than pursuing their information needs. The third and fourth most popular classes of queries, sent to video search service are the cultural and religious ones which constitutes 12% of all the queries. Politic and health-related queries, are ranked next with roughly 9% and 8% respectively. The considerable point here is the lack of inclination of users for leveraging the video search service for searching about the science and technology topics.

IV. CONCLUSION

The attractiveness of multimedia contents for web users have attracted the attention of the search service provider companies. But it should be noted that the tremendous volume of such kind of contents is a serious obstacle for providing search services. As a result, one of the significant steps that search providers must do is the identification of information needs and recognition of propensities and inclination of users of such services. This way, search providers not only become

able to use the hardware and bandwidth resources optimally, but also gratify their users demands more successfully by prioritizing the process of the topics which are most popular among users. To this end, in this paper, we proposed a way of recognizing the propensities of the users of Parsijoo multimedia search services, by analyzing search engine log files. Analyses were accomplished in two distinct categories; holistic analyses and analyses based on frequently used queries.

The holistic analyses showed that the users are mostly eager to use the image search service among different multimedia search services. This indicates that the image search service has a greater level of maturity than the other two services and is able to provide higher quality results for users. On the other hand, there are lots of users who are inclined and eager to use the voice search service, but due to the poor quality of the service, they leave this service after a few trials. Another fact which could be concluded from the holistic analyses was that a considerable portion of queries were sent from intranet users. This means that the Persian search providers need to put more intense efforts for drawing the attention of more internet users to consume their search services.

Analyses based on the frequently used queries of the multimedia search service showed that Persian users leverage their own mother tongue for typing queries. In addition, it was revealed that the users are after finding entertaining and amusing topics when they use multimedia search services.

ACKNOWLEDGMENT

This research was supported by Persian native search engine program from Iran Telecommunication Research Center (ITRC). We also appreciate Webazma team members who have helped us most throughout our research.

REFERENCES

[1] M Azimzade, N Farhadi, M.M Esnaashari, “Analysis of search engine user based on implicit behavior”, International Conference in Web Research, April 2015.

[2] M.S Zahedi, M Azimzade, M Mahmoudi, R.badie, “Taste and behavior analysis of mobile user in native search engine based log file analysis”, 4th International Conference on Information Technilogy ManagementCommunication and Computer, June 2014.

[3] M Azimzade, M Mahmoudi, M.M Esnaashari, “Big data logs analysis of local search engine with purpose of deification Persian language behavior and taste”, 1th conference of BigData, 2014.

[4] M.S Zahedi, M Azimzade, N Farhadi, A.M ZareBidoki, “Behavior Analysis of Persian-language user in local search engine”, 19th National CSI Computer Conference, 2014.

[5] P. Daras, S. Manolopoulou, and A. Axenopoulos, “Search and Retrieval of Rich Media Objects Supporting Multiple Multimodal Queries,” *IEEE Trans. on Multimedia*, Vol. 14, No. 3, 2012, pp. 734-746.

[6] M. J. Swain, “Image and Video Searching on the World Wide Web,” *2nd UK Conf. on Image Retrieval,* 1999.

[7] K. Markey, “Twenty-five Years of End-User Searching, Part 1: Research Findings,” *American Society for Information Science and Technology*, Vol. 58, No. 8, 2007, pp. 1071-1081.

[8] S. Park, J. H. Lee, and H. J. Bae, “End User Searching: A Weblog Analysis of NAVER, A Korean Web Search Engine,” *Library and Information Science Research*, Vol. 27, No. 2, 2005, pp. 203-221.

[9] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, “Analysis of a Very Large Web Search Engine Query Log,” *ACM SIGIR Forum*, Vol. 33, No. 1, 1999, pp. 6-12.

[10] M. Costa and J. S. Mario, “A Search Log Analysis of a Portuguese Web Search Engine,” *2nd INForum-Simposio de Informatica*, pp. 525-536, 2010.

[11] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen, “U.S. vs European Web Searching Trends,” *ACM SIGIR Forum*, Vol. 36, No. 2, 2002, pp. 32-38.

[12] T. Joachims, “Optimizing Search Engines using Click-through Data,” *ACM Conf. on Knowledge Discovery and Data Mining*, 2002.

[13] T. Joachims, L. Granka, B. Pang, H. Hembrooke, F. Radlinski, and G. Gay, “Accurately Interpreting Click-through Data as Implicit Feedback,” *ACM Conf. on Research and Development on Information Retrieval*, 2005.

[14] T. Joachims, L. Granka, B. Pang, H. Hembrooke, F. Radlinski, and G. Gay, “Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search,” *ACM Trans. on Information Systems*, Vol. 25, No. 2, Article 7, 2007.

[15] B. J. Jansen, A. Spink, and T. Saracevic, “Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web,” *Information Processing and Management*, Vol. 36, No. 2, 2000, pp. 207-227.

[16] A. Spink, H. C. Ozmutlu, and D. P. Lorence, “Web Searching for Sexual Information: An Exploratory Study,” *Information Processing and Management*, Vol. 40, No. 1, 2004, pp. 113-124.

[17] D. Tjondronegoro, A. Spink, B. J. Jansen, “Multimedia Web Searching on a Meta-Search Engine,” *12th Australasian Document Computing Symposium*, Australia, 2007.

[18] S. D. Torres, I. Weber, and D. Hiemstra, “Analysis of Search and Browsing of Young Users on the Web,” *ACM Trans. on the Web*, Vol. 8, No. 2, Article 7, 2014.