International Journal of Engineering

Journal Homepage: www.ije.ir

Discovering Popular Clicks' Pattern of Teen Users for Query Recommendation

H. Ghasemzadeh, M. Ghasemzadeh*, A. M. Zare Bidoki

Computer Engineering Department, Yazd University, Yazd, Iran

P A P E R I N F O

Paper history:
Received 11 November 2017
Received in revised form 13 December 2017
Accepted 17 Januray 2018

Keywords:
Search Engine
Query Log
Search Behavior
Teen User
Query Recommendation

A B S T R A C T

Search engines are still the most important gates for information search in internet. In this regard, providing the best response in the shortest time possible to the user's request is still desired. Normally, search engines are designed for adults and few policies have been employed considering teen users. Teen users are more biased in clicking the results list than are adult users. This leads to fewer clicks on the lowly-ranked search results. Such behavior reduces teen users' navigation and result extraction skills. With an increase in information load and in teen's demands, lack of efficient methods leads to inefficiency of search engines regarding teen users. For the purpose, this study discovers teen users' search behavior and its application in yielding an improved search is strongly recommended. In this way, the pattern of teen users' popular clicks is identified from a large search log through mining of users' search transactions based on the frequency and similarity of the clicks in the search log. Then, using binary classification, the closest query into the teen user's desired one is identified. To discover teen users' behavior, we took advantage of the AOL query log. System efficiency was examined on the AOL query search log. Results reveal that click pattern improves approaching the query to the one desired by teen users. Generally, this study can demonstrate that in data recovery, application of click behavior and its binary classification can result in improved access of teen users to their desired results.

***doi**: 10.5829/ije.2018.31.08b.07*

1. INTRODUCTION[1]

In recent years, improving user queries based on the content of web documents and the user's age is considered for providing results related to querying users [1]. By knowledge mining, the traditional Web search can be improved by completing queries and documents with additional latent structure. Knowledge representations have opened insider representations for a variety of search tasks [2].

With an increase in information load on the web in recent years, finding the desired information become even more difficult. Most new searching technologies on the web have been developed to deal with this issue. Despite the advances in this respect, search engines occasionally provide the users with inappropriate information. One reason for this can be the user's unawareness of entering an appropriate query. Search engine users often lack required skills for organizing appropriate queries. Furthermore, search engines often find difficulties showing short and precise results based on the user's informational request. Today, the expectations of search engines are beyond simple finding; a number of web pages upon the user's query terms. The problem here is that the search engine displays a large number of web pages relating to the user's query terms requiring considerable time to find one's desired information. This issue is known as surplus information overload. Search engines log users' interests, which are their queries, in query logs [3, 4].

Teen users have a higher click bias on a search engine's result list, compared to adult users, leading to fewer clicks on the results listed in lower ranks. This reduces teen users' navigation and data extraction skills. Certain groups of teenagers search for a limited number of websites or domains, such as games. Lack of appropriate queries and click bias on highly-listed results expose teen users to contents which are not their target, and can be harmful to them in some cases because current search engines provides all sorts of information [5]. On

*Corresponding Author Email: *m.ghasemzadeh@yazd.ac.ir* (M. Ghasemzadeh)

Please cite this article as: H. Ghasemzadeh, M. Ghasemzadeh, A. M. ZareBidoki, Discovering Popular Clicks' Pattern of Teen Users for Query Recommendation, International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications Vol. 31, No. 8, (August 2018) 1205-1214

the other hand, large data search and analysis are quickly progressing. Therefore, a new research field has emerged known as big data mining. Researchers have examined various methods and algorithms to extract knowledge from such data [6, 7].

Query log is a source of big data, which can be used to help improve user search; however, users apply short ambiguous queries to search the webs. The problem can be declined via query reformulation or recommendation. Query recommendation to users can be done through analyzing query logs using classification and clustering mechanisms. Normally, the query log of a search engine is massive and takes much time to analyze [8, 9].

The key issue in this study is to provide an appropriate query recommendation to teen user based on which, the user can access high-quality content through the search engine. Teen users, while entering a query, usually insert irrelevant keywords to their target. This leads to receiving irrelevant information, which makes them leave the search process. Query recommendation methods provide queries which are more comprehensive and relevant to primary one so that the users can modify their search upon the queries. Query recommendation method increases the user's chance to access relevant and appropriate information. In this regard, a query recommendation method is proposed based on teen users' search behavior with respect to AOL search log. The study applies analysis and mining of user's search behavior based on AOL Search Log. The teen users-related search transactions are extracted by matching the URLs on which users have clicked and already exist in AOL Search Log, with domains listed in "kids and teens" directory of DMOZ. On DMOZ, URLs that are related to teenagers suit the age of 13 to 15 years. These are used to improve query recommendation to teen users and to promote the search results. This is carried out through binary classification and the closest query to the one desired by teen user is determined and recommended to them. Based on the URLs clicked in transactions extracted from the AOL query log, teen users' search behavior pattern is discovered and, then, the closest query to the one desired by the teen users is determined via binary classification and is recommended to them.

2. RESEARCH BACKGROUND

A great deal of research has been carried out on providing the most appropriate response to the users' query in search engines. Certain studies have proposed various query mining methods based on the query logs. Clustering related queries based on the URLs in the query log of a search engine is a proposed method using four different distance calculations. The calculations are based on (1) keywords; (2) string matching of the keywords; (3) commonly clicked URLs; and (4) the distance between the clicked documents. The key issue, here, is to discover the most interesting queries submitted by various users [10].

Several studies have been conducted to analyze the query logs of commercial search engines in large scale. Silverstein et al. [11] carried out an analysis about the query log of AltaVista search engine including about one million entries. They also performed an analysis on query sessions and the correlation between query terms based on a set of qualitative measurements, including query length, query frequency, session length, and repetition of terms. Their results indicated that users tend to use short queries, including 2.3 words on average. Users' sessions are also short, including an average of two queries in each session. They also stated that most users do not change the queries, and 77.5% of the queries are individual, which calls for a considerable variety of users informational needs [11]. In another study, Spink et al. [12] reported similar results concerning query length and features based on the query log of Excite search engine.

Various aspects of an AOL query log, including query formulation patterns, search engine efficiency, demographic features of the user, and the user's interactions were examined by Pass et al. [13]. They demonstrated that 20% of the users submit almost 70% of the queries, also that less than 1% web domain accounts correspond to almost 50% users' clicks [13].

Other analyses have also been performed on the same query log regarding the classification of queries and sessions based on the popularity of the queries, studying various behaviors such as navigation coefficient, query length, and time. These studies have suggested various definitions for the user's session on query logs. In general, a session refers to a series of queries aimed at meeting an informational need [14, 15].

User's behavior based on toolbar data and yahoo based search was studied by Kumar et al. [16] and Cheng et al. [17]. This study was examined users' search sessions and the analyses have generally been performed upon age. Moreover, the search sessions' page views were classified according to content type (e.g. game, news, portal), type of communication (e.g. email, social networking), and search type (e.g. web, multimedia). As a result, almost half of the page views were in content category, one-third were in relation category, and one-sixth were in the search category [16, 17].

Torres et al. [18] conducted a study on query recommendation to children users. The issue was that children use limited vocabulary as query keywords. Furthermore, children have problems with choosing appropriate keywords. To recommend query, children-specific tags from social networks, and related keywords were used. Furthermore, a biased random walk was proposed based on a bipartite graph of web sources and tags. Their results showed that their proposed method outperformed the current search engines in the query

recommendation to children ages from 10 to 12 years. They also showed that social network serves as a valuable source of query recommendations. Their proposed method improved children's search [18, 19].

In a similar study conducted by Torres et al. [20], the researchers used topic tags of social networks related to children and appropriate keywords to create appropriate query recommendations. Again, they proposed a biased random walk based on bipartite graph of web sources and tags. In addition, they considered the quality of ranking the tags. They improved tags' ranking via a combination of topic-features and modeling the language used in children's query. Their results showed that their proposed method outperformed the current search engines in the query recommendation to children ages from 8 to 9 years [20].

Wang et al. [21] used the clicks entropy average to distinguish informational queries from ambiguous one. They also calculated the click entropy average over each user's click distribution.

Figure 1 depicts the bi-graph of clicks and queries. Based on the graph, one can calculate two queries with the same entropy average [21].

Duan et al. [22] introduced a click pattern for modeling the user's search behavior and showed that their proposed method improves query recommendation to users. However, our proposed pattern of user's search behavior concerns with a specific age range (teenagers) and is based on teen users' popular clicks, who are exposed to higher bias in click navigation than adult users.

3. PROPOSED METHOD

The present study applies teen user's search behavior pattern in order to recommend queries to teen users, which is the closest query to the primary one submitted by them. In this approach, the search transactions corresponding to teen are firstly extracted from the AOL search log. To do so and in accordance with the study by Torres et al. [20], search transactions corresponding to teen users are extracted from AOL search log using the clicks concerned with DMOZ teen directory and matching them with the clicks in the AOL search log.

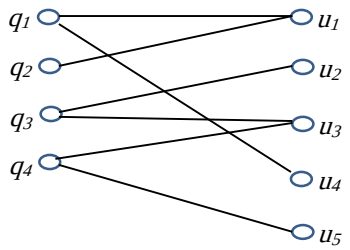

Figure 1. Bi-graph of clicks and queries

Then, popular clicks' pattern entropy and the patterns similarity were obtained from the extracted search transactions. To do so, first, popular clicks' patterns of the search transactions extracted from AOL Log is determined and, then, each pattern entropy and the patterns similarity are calculated. Finally, through binary classification via K-Nearest Neibours (KNN), the closest query to the one submitted by teen user is identified. The KNN model is the simplest and most intuitive for classification. It is the most popular for web usage classification, distance based text and recommended model. One of the advantages of this classifier is that it is conceptually very much related to the idea of collaborative filtering: Finding like-minded users (or similar items) is essentially equivalent to finding neighbors for a given user or an item (such as query recommendation). The KNN model, although simple and intuitive, has shown good accuracy results and is very amenable to improvements [23, 24]. While K-Nearest Neighbor is usually used for the collaborative filtering tasks, Support Vector Machine (SVM) is considered a state-of-the-art classification algorithm. Despite the KNN is dominant over SVM for collaborative filtering tasks, on the real-life corporate dataset with high level of sparsity, KNN fails as it is unable to form reliable neighborhoods. In this case SVM outperforms KNN [25].

On this base, a binary classification with higher accuracy is proposed. Generally two criteria can be used to find similar queries:

- **Using query content:** Two queries with the same or similar terms denote similar information needs. This method is more reliable in case of long queries. However, in most cases, users submit short queries to search engines. In such cases, they do not provide enough information to meet user's informational needs. Therefore, a second criterion is used to complement the first one. The second criterion is similar to the vision of clustering documents in information retrieval. It is believed that the closeness of the relationship between documents is in accordance with the assimilation of the queries. This vision is used as the reverse of the first vision.
- **Using user feedback:** Two queries are the same if they lead to the selection of the same content based on the user's click on the documents. Clicks on documents can be compared to the user feedback in traditional information retrieval environment.

The two criteria have different advantages. One can classify the queries with similar components using the first criterion. However, the second criterion takes advantage of the user's judgment. Furthermore, the second criterion is used to cluster the user's queries reported in literature [23]. Generally, content-based measurements tend to identify similar or the same terms. Measurements based on user's feedback, however, tend to identify queries related to similar topics [10].

Researchers have not extensively utilized the inherent dependency of documents and queries in a transactional database of a search engine. The distance between two documents can be evaluated without examining the content of the documents. This property is called content ignorance [26]. The issue concerning the second criterion is that individual users' behaviors are different even in case of the same search result. Therefore, it can be concluded that click behavior of a given user has some noise, but conforms to a set of common behavioral patterns. The proposed solution is to filter the user's behavior noises through identifying common patterns in their behavior to yield a more accurate modeling of the user. Then, it can be used in practical applications such as identifying similar queries as query recommendation.

Figure 2 presents the general outline of the method proposed for identifying pattern of the teenage user's behavior and its application in the offering query recommendation. In the first step, the search transactions corresponding to teen users are extracted from AOL Search Engine Log. By matching the clicks corresponding to the DMOZ teen directory with the clicks corresponding to AOL Search Log, search transactions related to teen users are extracted.

In the second step, using clicks extracted from AOL Search Log, the popular clicks' pattern of clicks related to a query is discovered.

In the next steps, through binary classification, the closest query to the one submitted by teen user is identified. To do this, corresponding to the clicks for each query from the search queries extracted from the AOL search log, the two features of the entropy of popular clicks' pattern and the similarity of popular clicks' patterns are calculated. Finally, binary classification with KNN method is carried out using the calculated features. On this base, a binary classification with higher accuracy is proposed.

3. 1. Popular Clicks' Pattern of Teen Users Each user with similar search results has different search behavior.

step 1: Extraction of search transactions related to teenage users

step 2: Discovery of the popular clicks' pattern related to each query

step 3: Calculation of features include entropy of popular clicks' pattern and Popular clicks' patterns similarity for each query

step 4: Binary classification

step 5: Analysis of classification accuracy and prediction nearest queries to the teen user query as query recommendation

Figure 2. Steps processing of the proposed method

It is assumed that noisy behaviors upon clicks of every search are based on a basic behavioral model according to which, users follow a set of common behavioral patterns. Through identifying the common patterns in user's clicking behavior, one can filter the noises within the user's behavior and achieve an accurate model of the user, which can later be used to identify similar queries and query recommendation. The proposed click pattern corresponds to a specific age range (teenagers) and the popular clicks of the user's search. Our proposed popular clicks' pattern is defined as follows:

- **Definition of popular clicks' pattern**: Given is a query, q, and a set of clicked documents, D_q. then, a popular clicks' pattern, $P_{q,d}$, which is a distribution of clicks' popularity on D_q, shows the degree of popularity to which a document is clicked based on the query q:

$$P_{q,d} = \{Pop(d_q) \mid d_q \in D_q, \sum_{d_q \in D_q} Pop(d_q) = 1\} \quad (1)$$

where, $Pop(d_q)$ is the popularity of the document d. Popularity of the document d is calculated according to the clicks made on the document through query q.

$$Pop(d_q) = C_d^{(q)}.(\sum_{d_q \in D_q} C_d^{(q)})^{-1} \quad (2)$$

Furthermore, $C_d^{(q)}$ is the number of clicks on the document d based on the query q. In fact, popular clicks' pattern of teen users is shown by three pairs indicating the clicks on the document with the highest popularity. Therefore, each popular clicks pattern is indicated by an ordered list of three pairs:

$$P_q = \{(u_1, Pop(u_1)), (u_2, Pop(u_2)), (u_3, Pop(u_3))\} \quad (3)$$

where, u_i denotes the clicked URL corresponding the i^{th} document, and we have $Pop(u_1) > Pop(u_2) > Pop(u_3)$.

3. 2. Similarity of Popular Clicks' Pattern To calculate the similarity between two queries, a bipartite graph of queries and URLs is used as in Figure 1. The graph $G(V,E)$ is a bipartite graph between the nodes Q and U such that $Q \cup U = V$, $Q \cap U = \varnothing$ and every edge in E is a node in Q and a node in U. A query vector is shown in Equation 4, where $rel(q_k,u_i)$ describes the relationship between the query k and the URL i [27].

$$\overrightarrow{q_k} = [rel(q_k, u_1), rel(q_k, u_2), ..., rel(q_k, u_m)] \quad (4)$$

Every query is shown as a vector, where the i^{th} element describes the relationship between the query k, and the URL i.

In the present study, the query vector (5) is shown by substituting the pattern of the popular clicks of Equation

(3) in Equation (4), and is used to calculate the similarity between the queries.

$$\overrightarrow{q_k} = [Pop(u_1), Pop(u_2), Pop(u_3)] \quad (5)$$

Figure 3 shows the vector depiction of queries. Every query is shown as a vector.
where, $Pop(u_i)$ is replaced with $rel(q,u_i)$ which shows popularity of u_i corresponding to the query, q.

Now, we introduce S_q as the set of common popular clicks' patterns corresponding to the query q as in Equation (6).

$$S_q = \{(P_{q'}, Sim(P_q, P_{q'}) \mid P_{q'} \text{ is Popular Click Pattern of } q'\} \quad (6)$$

S_q is the set of all popular clicks' patterns which is similar to popular clicks' pattern corresponding to query, q, and it shows the extent to which the popular clicks' pattern, $P_{q'}$, follows the given popular clicks' pattern, P_q, in the set, D_q. Generally, the similarity function is defined as Equation (7):

$$Sim(P_{q'}, P_q) = 1 - Dis(P_{q'}, P_q) \quad (7)$$

The function, *Dis*, determines the distance between two patterns, where any given distance function can be used. In the proposed method in this study, Equation (8) is used to calculate the similarity between two patterns:

$$Sim(P_{q'}, P_q) = \text{Cosine}(P_{q'}, P_q) = \frac{\overrightarrow{P_{q'}} \cdot \overrightarrow{P_q}}{|\overrightarrow{P_{q'}}| . |\overrightarrow{P_q}|} \quad (8)$$

3. 3. The Entropy of the Popular Clicks' Pattern

In previous studies, click entropy was calculated as the informational entropy of the distribution of the user clicks. Given the query, q, and the set of clicked documents, D_q, clicks entropy corresponding to query, q, is calculated through Equation (9):

$$ClickEntropy(q) = -\sum_{d \in D_q} p(d \mid q) \log p(d \mid q) \quad (9)$$

where, $p(d|q)$ is the likeliness of clicks on the document, d, among all clicks done upon q.

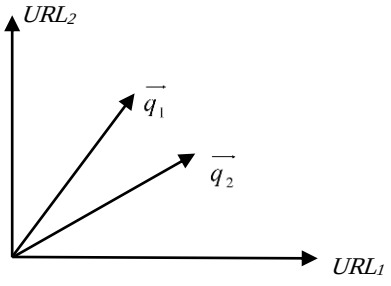

Figure 3. Query representation by URLs vector

In case of navigational queries where the intents are already clear, click entropy is assigned a low value. However, in case of informational queries, click entropy gives a high value. Here, the entropy of popular clicks' pattern is suggested as a feature for binary classification. In the present study, the entropy of popular clicks' pattern indicates the informational entropy of the distribution of the clicks' popularity in that pattern. Given the query, q, and popular pattern, P_q, the entropy of the popular clicks' patterns is calculated by Equation (10). Equation (10) is obtained by substituting the popular clicks' pattern corresponding to Equation (3) in Equation (9).

$$ClickPatternEntropy(q) = -\sum_{d \in D_q} Pop(d \mid q) \log Pop(d \mid q) \quad (10)$$

$p(d|q)$ is the same as $Pop(d|q)$ determining the popularity of the clicked document, d, upon the query, q. The informational entropy of the distribution of teen users' popular clicks pattern is significant in case of queries with ambiguous intentions. The main goal is to reduce the ambiguity of such queries, where a smaller value of the pattern entropy shows less ambiguity of the query.

In order for query recommendation, the closest query in the log to the user's query is used. For this purpose, the shortest distance between the user's query and the submitted one is calculated. The amount of supporting a query is determined by its popularity in the query log. Popularity pattern of the user clicks is used to measure the similarity between the queries, according to Equation (8), and the maximum similarity between the two patterns is considered.

4. IMPLEMENTATION

The proposed method was implemented on a microcomputer with a 4G RAM and an Intel processor of 1.8 G Hz, employing Alteryx tool to discover the popular teen users' clicks' pattern, and the WEKA tool for binary classification.
In order to identify the teen users' search behavior, the clicks on the AOL search engine log was used, containing over 20 million queries corresponding to 650000 users. Every record in this search log includes the following features.

- **AnonID:** an anonymous identifier assigned to each user,
- **Query:** query term supplied by the user,
- **QueryTime:** date and time on which the query is triggered by the user,
- **ItemRank:** rank assigned to each clicked URL,
- **ClickURL:** address of the clicked URL.

Figure 4 depicts a general approach of discovering popular clicks' pattern for each query on Alteryx tool. Since the AOL Search Log and the dataset of the teen directory corresponding DMOZ are large, Alteryx tool is

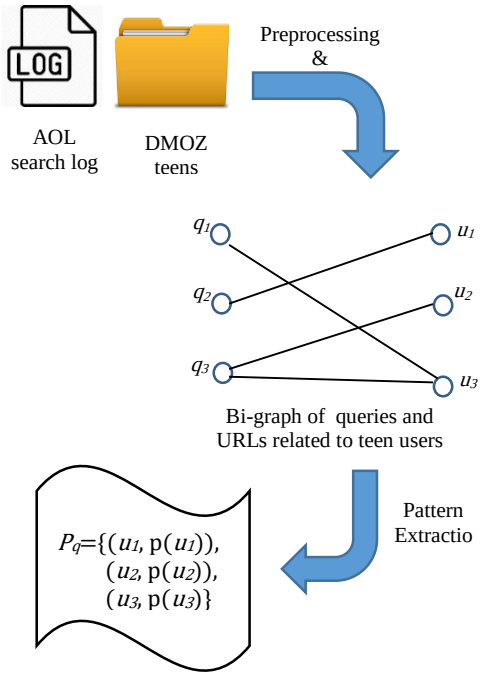

Figure 4. Discovering approach of popular clicks' pattern for each query on Alteryx tool

used to identify the popular clicks' pattern, with which big data can be processed. In this model, firstly we extracted teen users' search transactions from AOL Search Log, based on matching the URLs in teen directory of DMOZ with the URLs in AOL Search Log. After filtering the extracted search transactions, a bipartite graph of the teen users' queries and clicks is achieved. Then, according to the popularity of the clicks, grouping of the queries and clicks is carried out and the popular clicks' pattern is achieved for each query. Next the entropy of each popular clicks' pattern is calculated based on Equation (10) and the similarity between popular clicks' patterns is calculated using Equation (8) for each query extracted from the AOL log.

Equation (11) presents a popular clicks' pattern corresponding to the query "free coloring pages" including three clicked URLs with a higher popularity.

$$\begin{aligned} &PopularClickPattern = \{ \\ &\quad (\text{htttp://www.coloringcastle.com, 0.0045}), \\ &\quad (\text{htttp://www.actitvityvillage.co.uk, 0.0014}), \\ &\quad (\text{htttp://familycrafts.about.com, 0.0013})\} \end{aligned} \tag{11}$$

Figure 5 shows the entropy of popular clicks' patterns of teen users' clicks per each query extracted from AOL log. The lower entropy of the popular clicks' pattern means that the query is less ambiguous and also means lower ambiguity in navigation of clicks.

In the present research, query recommendation is considered as a classification task, where the addition of a term to the main query restricts the search space. The AOL query log is used for this purpose. To do the classification, two classes, namely "Yes" and "No", are considered for query recommendation and not allowing query recommendation, respectively.

In general, there exist a total of 812 queries as query recommendation alternatives, each labelled by an expert as either of the two classes "YES" or "NO". Then, binary classification is run for each alternative query. In this manner, the popular teen users' clicks' pattern was been identified from the search log using Alteryx tool, and the similarity between the patterns was calculated. Then, using a binary classification, the closest query to teen user's submitted query was determined.

The features used in the construction of binary classification for each query are presented in Table 1 At the second time, binary classification was based on the traditional feature of popularity as used in previous studies.

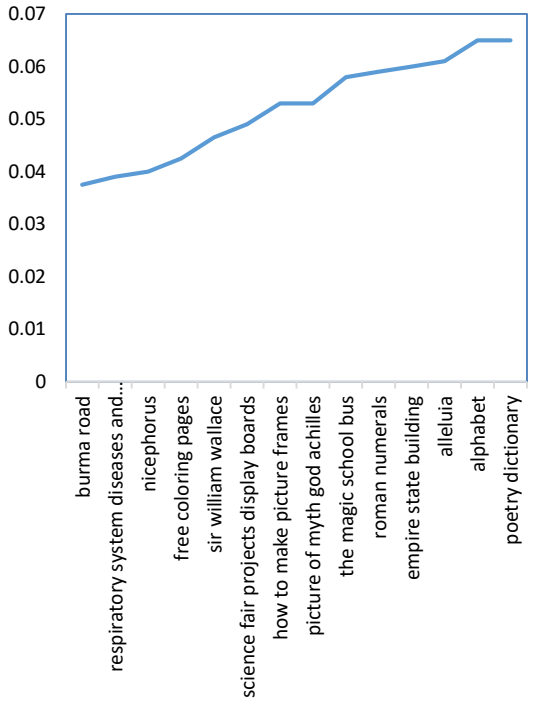

Figure 5. Entropy of popular clicks' patterns of teen users

TABLE 1. Features used in binary classification

	Features
1	Entropy of popular clicks' pattern
2	Popular clicks' patterns similarity
3	Average of clicks entropy
4	Popularity
5	Query Length

This was carried out using WEKA tool. After preprocessing, classification was performed twice using KNN method. The corresponding efficiency is presented in Table 2.

In fact, query recommendation varies depending upon the type of query. In case with information queries, the main goal is to reduce ambiguity. However, in case with navigational queries where the intention is clear cut, the main goal is to extract queries with highest possible similarity. Hence, click entropy and average entropy are not appropriate to informational queries since they contain a great value. Therefore, a recommendation system, which uses click entropy and average entropy, is biased to recommend navigational queries only. As a result, the use of entropy of popular clicks' pattern and similarity between those patterns for teen users' clicks is recommended to avoid such bias.

5. RESULTS

The system efficiency test was carried out on AOL Search Log queries. First, the search items related to teen users were extracted from AOL Search Log through matching the clicked URLs in AOL Search Log with the domains listed in "Kids and Teens" directory of DMOZ. Then, the popularity of the clicked URLs was calculated for each of the queries.

Figure 6 shows the popularity of the clicked URLs based on "dictionary" query. Among the clicked URLs, "www.wordcentral.com" had the highest popularity, followed by "www.wordsmyth.net".

Figure 7 shows the popularity of each teen user's query. On this base, the query "dictionary" has the highest popularity to the teen users, and the query "free coloring pages" considered as the main query in this test was the 13th popular query on AOL Search Log.

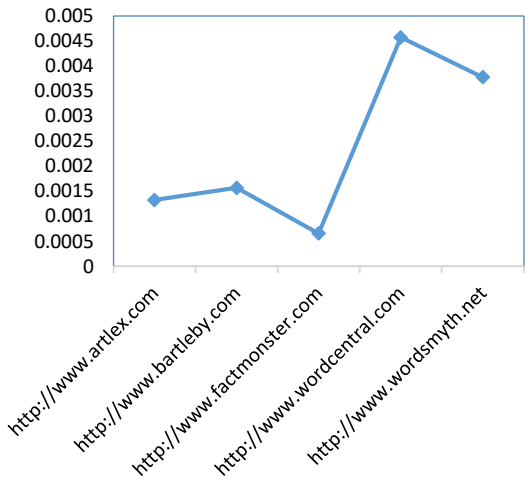

Figure 6. Popularity of the clicked URLs based on "dictionary" query

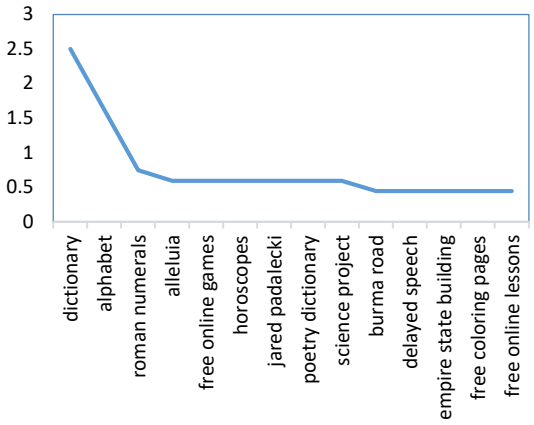

Figure 7. Popularity of each teen user's query

Table 2 presents the results of classification for query recommendation. Classification was carried out twice. The first row in the table gives classification results based on previous studies using the popularity of the queries. The accuracy of binary classification turned out to be 77.09%.

The second row in the table presents the results of classification using the features extracted in the present study including similarity and popular clicks' pattern entropy. Classification accuracy turned out to be 86.94%. Since there exists a higher noise in navigation of teen users in comparison with adults, the presented method provides a high-accuracy model for filtering the noise in navigation of teen user's clicks, and accordingly presents a binary classification with higher accuracy. As shown in Table 2, by adding similarity and popular clicks' pattern entropy of pattern sets, classification accuracy is increased.

Receiver Operator Characteristic (ROC) curve is depicted in Figure 8. The curve corresponding to classification with popular clicks' patten entropy and their similarity is higher than the one corresponding to the classification with the traditional criterion of queries popularity, showing that the classification with entropy and similarity classifies alternative query recommendations with higher accuracy.

Receiver Operator Characteristic (ROC) curves are commonly used to present results for binary decision problems in machine learning. An important difference between ROC space and PR space is the visual representation of the curves.

TABLE 2. Classification results for query recommendation

Features	Accuracy (%)	Precision	Recall
Popularity	77.0936	0.846	0.816
Entropy & similarity of patterns	86.9458	0.9	0.913

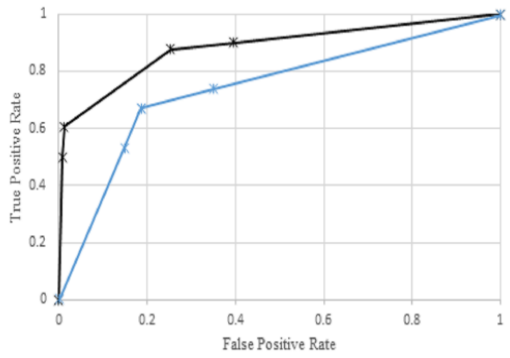

Figure 8. The ROC curve of binary classification

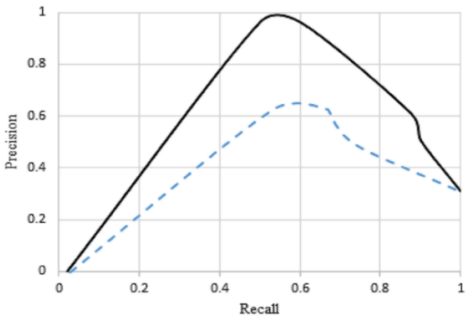

Figure 9. The PR curve of binary classification

Looking at PR curves can expose differences between algorithms that are not apparent in ROC space. The ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution. Precision-Recall (PR) curves, often used in Information Retrieval have been cited as an alternative to ROC curves for tasks with a large skew in the class distribution [28].

The goal in the ROC space is the upper left corner of the curve. The ROC curve in Figure 8 shows that improvement has been achieved. In PR space, the target is located in the upper right corner and the PR curve in Figure 9 shows that progress is relatively good.

Table 3 presents classification results for submitted query "free coloring pages". The two primary recommended queries which are predicted in "Yes" class have a low entropy of pattern, but their similarity with the main query is higher. However, the query "Free_online_lessons" which is predicted in "No" class has high entropy of pattern while its similarity with the main query is low. This shows the appropriateness of the test.

Generally, the test showed that using popular clicks' patterns is useful for query recommendation to teen users. Test results indicated that the popular clicks' pattern approaches the query recommendation to the query submitted by the user.

Figure 10 shows the ROC curves for three classifier models: Decision Stumpe Tree, KNN and Naive Bayes. The KNN model is better than the other two classifier models.

Table 4 presents classification results for the three classifier models. The KNN Classifier has a higher accuracy than the other two models.

The test shows that the KNN classifier model has a relatively suitable performance for binary classification of our data set.

6. CONCLUSION

Since there exists higher noise in navigation of teen users in comparison with adults, the proposed method of popular clicks' pattern yields a model with higher accuracy to filter navigation noise in teen users' clicks. Using the popular clicks' patterns of the clicks extracted from AOL search log can approach the query recommendation to the main query triggered by teen users.

The present study has used binary classification for query recommendation based on similarity and entropy of popular patterns corresponding to teen users' queries extracted from the AOL search log.

TABLE 3. Query recommendation for "free coloring pages"

Query Recommendation	Pattern Similarity	Pattern Entropy	Predicated Class
unicorn_coloring_pages	0.89	0.072	YES
free online games	0.87	0.075	YES
free_online_lessons	0.85	0.10	NO

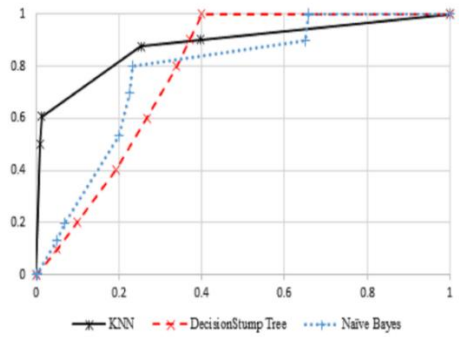

Figure 9. The ROC curves for the three classifiers

TABLE 4. Results of the three classifier models

Classifier model	Accuracy (%)	Precision	Recall
KNN	86.9458	0.9	0.913
DecisionStumpe Tree	72.4138	0.854	0.724
Naive Bayes	78.3251	0.817	0.783

It was concluded that popular clicks' pattern, as the user's search behavior, can approach the query recommendation to the query submitted by teen users. Moreover, it was concluded that the use of click entropy and entropy average for query recommendation is not appropriate in the case if the teen user's query is informational because they include a large deal of information. Therefore, a query recommendation system which is solely based on click entropy and entropy average is biased towards recommending navigational queries only. However, using the extracted characteristics of entropy and similarity of popular click pattern prevents such bias. The current study used popular clicks' pattern, as identified from the AOL search log, for the purpose of query recommendation for teen users.

As for future research, one can discover a clicking pattern based on popular topics such as "movie" to recommend queries with higher relevance to the users' submitted query. It is noteworthy that popular clicks' pattern has significant effects on the relevance of the information retrieval results for teen users.

7. REFERENCES

1. T. Raghunadha Reddy, B. Vishnu Vardhan and P. Vijayapal Reddy, "A Document Weighted Approach for Gender and Age Prediction Based on Term Weight Measure," ***International Journal of Engineering, Transactions B: Applications,*** Vol. 30, No. 5, (2017), 643-651.
2. M. Mitsui and C. Shah, "Query Generation as Result Aggregation for Knowledge Representation," ***Proceedings of the 50th Hawaii International Conference on System Sciences***, (2017), 4365-4374.
3. M. Caramia, G. Felici, and A. Pezzoli, "Improving search results with data mining in a thematic search engine," ***Computers & Operations Research***, Vol. 31, No. 14, (2004), 2387-2404.
4. R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," ***EDBT'04 Proceedings***, (2004), 588-596.
5. E. Foss *et al.*, "Children's search roles at home: Implications for designers, researchers, educators, and parents," ***Journal of the American Society for Information Science and Technology***, Vol. 63, No. 3, (2012) , 558-573.
6. M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: a review," ***SIGMOD***, Vol. 34, No. 2, (2005), 18-26.
7. L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "A new method for data stream mining based on the misclassification error," *IEEE Trans.* ***IEEE Transactions on Neural Networks and Learning Systems***, Vol. 26, No. 5, (2015), 1048-1059.
8. P. Domingos and G. Hulten, "Mining High-Speed Data Streams," ***Proceedings of The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***, (2000), 71-80.
9. Y. Liu, J. Miao, M. Zhang, S. Ma, and L. Ru, "How do users describe their information need: Query recommendation based on snippet click model," ***Expert Systems with Applications***, Vol. 38, No. 11, (2011), 13847-13856.
10. J. Wen, J. Nie, and H. Zhang, "Clustering user queries of a search engine," ***In Tenth International World Wide Web Conference (WWW)***, No. 49, (2001), 162-168.
11. 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), 6-12.
12. A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic, "Searching the Web: The Public and Their Queries," ***Journal of the American Society for Information Science and Technology***, Vol. 52, No. 3, (2001), 226-234.
13. G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," ***Proceedings of the 1st international conference on Scalable information systems InfoScale 06***, Vol. 152, (2006).
14. D. J. Brenes and D. Gayo-Avello, "Stratified analysis of AOL query log," ***Information Sciences***, Vol. 179, No. 12, (2009), 1844-1858.
15. R. Jones and K. L. Klinkner, "Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs," ***Proceedings of the 17th ACM conference on Information and knowledge management***, (2008), 699-708.
16. R. Kumar and A. Tomkins, "A Characterization of Online Browsing Behavior," ***Proceedings of the 19th International Conference on World Wide Web***, (2010), 561-570.
17. Z. Cheng, B. Gao, and T. Liu, "Actively predicting diverse search intent from user browsing behaviors," ***WWW '10: Proceedings of the 19th international conference on World wide web***, (2010), 221-230.
18. D. S. Torres, D. Hiemstra, I. Weber, and P. Serdyukov, "Query recommendation for children," ***Proceedings of the 21th ACM international conference on Information and knowledge management - CIKM '12***, (2012), 2010-2014.
19. S. D. Torres, D. Hiemstra, and T. Huibers, "Vertical selection in the information domain of children," ***JCDL '13 Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries***, (2013), 57-66.
20. S. D. Torres, D. Hiemstra, I. Weber, and P. Serdyukov, "Query recommendation in the information domain of children," ***Journal of the Association for Information Science and Technology***, Vol. 65, No. 7, (2014), 1368-1384.
21. Y. Wang and E. Agichtein, "Query Ambiguity Revisited : Clickthrough Measures for Distinguishing Informational and Ambiguous Queries," ***the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics***, (2010), 361-364.
22. H. Duan, E. Kiciman, and C. Zhai, "Click patterns: An Empirical Representation of Complex Query Intents," ***Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM' 12***, (2012),1035-1044.
23. Z. Markov and D. T. Larose, "Data mining the Web, Uncovering patterns in Web content,structure, and usage," ***John Wiley & sons Inc.***, (2007), 115-132.
24. X. Amatriain, A. Jaimes, N. Oliver, and J. Pujol, "Data mining methods for recommender systems," ***Recommender Systems Handbook***,***Springer***, (2011), 39–71.
25. M. Grcar, B. Fortuna, and D. Mladenić, "KNN Versus SVM in the Collaborative Filtering Framework," ***Learning***, (2005), 5-9.
26. D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," ***Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00***, (2000), 407-416.
27. M. Hosseini and H. Abolhassani, "Clustering Search Engine Log for Query Recommendation," ***Proceedings-Advances in Computer Science and Engineering***, (2008), 380-387.
28. J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," ***in Proceedings of the 23rd international conference on Machine learning-ICML'06***, (2006), 233-240.

Discovering Popular Clicks' Pattern of Teen Users for Query Recommendation

H. Ghasemzadeh, M. Ghasemzadeh, A. M. ZareBidoki

Computer Engineering Department, Yazd University, Yazd, Iran

P A P E R I N F O

Paper history:
Received 11 November 2017
Received in revised form 13 December 2017
Accepted 17 Januray 2018

Keywords:
Search Engine
Query Log
Search Behavior
Teen User
Query Recommendation

چکیده

موتورهای جستجو مهم‌ترین دروازه‌های جستجوی اطلاعات در اینترنت می‌باشند. در موتورهای جستجو تمهیداتی برای کاربران نوجوان پیش‌بینی‌نشده است. کاربران نوجوان انحراف بیشتری در کلیک کردن بر روی لیست نتایجیک جستجو نسبت به بزرگ‌سالان دارند. چنین رفتاری، مهارت‌های ناوبری و استخراج نتایج توسط کاربران نوجوان را مختل می‌سازد. در این پژوهش، کشف الگوی رفتاری کاربران نوجوانبرای بهبود جستجو توصیه می‌شود. در این روش، الگوی کلیک‌های محبوب کاربران نوجوان از لاگ جستجوو بر اساس محبوبیت و تشابه کلیک‌ها تشخیص داده می‌شوند. آزمایش کارایی سیستم بر روی بخش پرس‌وجوهای لاگ جستجوی AOL صورت گرفت.نتایج آزمایش حاکی از آن است که ویژگی الگوی کلیک‌ها موجب بهبود نزدیک شدن پیشنهاد پرس‌وجو به پرس‌وجوی مورد نظر کاربر نوجوان می‌گردد.

***doi**: 10.5829/ije.2018.31.08b.07*