International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 4, Issue 3, May-June 2015 ISSN 2278-6856

Focusing User Modeling For Age Specific Differences

Prajakta Shitole[1], Mrs. M. A. Potey[2]

[1]Department of Computer Engineering
D. Y. Patil College of Engineering, Akurdi
Savitribai Pule Pune University

[2]Department of Computer Engineering
D. Y. Patil College of Engineering, Akurdi
Savitribai Pule Pune University

Abstract

When personalization and user satisfaction are the motives, demographic attributes play an important role. User's demographic attributes or demographics include traits such as age, gender, race, occupation, political and religious views etc. There are differences in search and browsing behavior of a user based on these attributes. We have focused on these differences based on user's age. The motivation to do so was to reduce the number of difficulties faced by users from different age groups, and to provide the most relevant result to the users query. In this paper, we offer a solution to this problem by showing how user age can be efficiently and accurately inferred based on their search query. Our approach uses Solr search engine in combination with the Bayesian classifier to predict the age of the user based on the query entered and the sequence of actions performed on the web.

Keywords:- User modeling, demographics, User Behavior, Bayesian Classifier, age prediction.

1. INTRODUCTION

The Internet is increasingly used by users with varied demographic attributes for all kinds of purposes. Nonetheless, there are not many resources especially designed for meeting the varied needs of all these users and most of the content on line is designed without taking there attributes into consideration. The demographic approach relies on a marketing approach which postulates that users with similar demographic back-grounds may have similar preferences. This approach uses an analysis of demographic data about users who rated a certain item, to learn and categorize the type of a person who likes the item. This information is stored and then used to provide for future recommendations. Although demographic data is used for marketing, its use in user modeling is relatively limited, since it is stereotypical by nature and requires a reference demographic data base that is usually not available. This paper focuses on age as a demographic attribute, and user modeling with respect to it. We consider the large differences between young users and adults using internet since their topic interests, computer skills, and language capabilities are completely the difficulties that different age groups encounter on the Internet when searching for information and browsing for content. Capturing age aims to reduce the difficulties faced on internet by different aged users. User Modeling is the process of building an user model. A user Model contains the systems assumptions about all aspects of the user which are deemed relevant for tailoring the dialog behavior of the system to the user [9]. The main goal of user modeling is customization and adaptation of systems to the user specific needs. User model is a set of information structures designed to represent one or more of the following elements: (1) representation of assumptions about the goals, plans preferences, tasks and abilities and the knowledge about one or more types of users; (2) representation of relevant common characteristics of users pertaining to specific user subgroups (stereotypes); (3) the classification of a user in one or more of these subgroups; (4) the recording of user behavior; (5) the formation of assumptions about the user based on the interaction history and/or (6) the generalization of the interaction histories of many users into stereotypes [10]. For personalized interaction with users, a system must have access to a wide variety of information about them ranging from relatively long-term facts such as areas of interest or expertise to quite short term facts such as the problem that a user is currently trying to solve. User modeling involves the tasks like classification, user profiling, and recommendation, and can be done implicitly or explicitly.

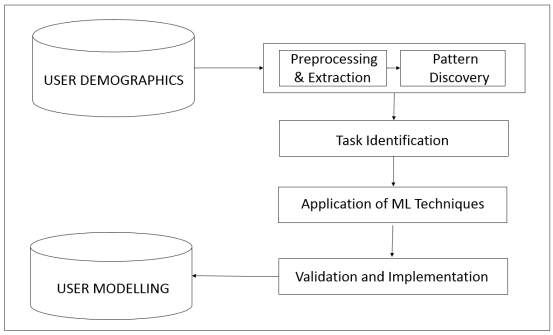

Figure 1: Implicit User Modeling

We are focusing on implicit creation of user model capturing the age of user. Figure 1 shows the process of implicit user modeling. In this paper we investigate the problem of predicting internet users? Age based on their

browsing behaviors, assuming that there are cognitive differences in the way of user interaction based on different age groups. Hence the type of query put is used to evaluate the demographic attribute: age.

2. RELATED WORK

Most of the previous research on user modeling is generally focused on patterns of Internet use, particular domains or services. The key concept of user modeling is to capture the online behavior of the user, and is one of the promising research works. Hongning Wang [15] studied the problem of user modeling in the search log data and proposes a generative model, DP Rank, within a nonparametric Bayesian framework. Weber and Castillo [17] presented a query logs study on how search differs in users with different demographics. They used demographic information that was derived from the US-census and user profile information to describe search patterns and behaviors for population segments with different demographic characteristics. They also employed an analogous methodology to show that the reading level of the urls clicked by children also varies across demographic features. Hidden Marokv models were also used extensively to find the online behavior [2], [18], [19]. The most relevant literature on user modeling, capturing search behavior and information retrieval based on demo- graphic attributes are described in the following paragraphs.

2.1Information seeking by different aged users

The first studies attempting to characterize the search behavior of children have been carried out using non-Internet systems, such as electronic libraries, CD-ROMs, and OPACs (Online Public Access Catalogs). Solomon [20] explored the search success of elementary school children when using an OPAC. The author found that children were able to use the system effectively when engaging in simple searches. However, they found that complex searches were hampered by the lack of mechanical skills of children. They pointed out that factors such as typing on the keyboard, spelling, limited vocabulary, and reading expertise are skills that are not developed enough in children in order to use the OPAC system studied [6]. Borgman et al.[5] found a similar behavior with high school children and a different OPAC. They also reported that these children had conceptual difficulties categorizing and browsing for searches that are more complex. Similarly, Neuman[18] found from a survey including 25 digital library administrators that the main problems children encountered during the search on digital libraries are the generation of keywords to construct the query and the lack of effective search strategies. Recent studies have explored the search behavior of children on the Internet with search engines. Nahl and Harada [17] carried out a study with 191 high school students to determine their search effectiveness after they have received special training to search the Internet. Users were asked to solve specific information tasks on the Internet. They were assessed based on the information they collected. Nahl and Harada [17] reported that most of the students had difficulties understanding how the search query is constructed with Boolean and default operators. In this study it was also observed that the lack of adequate vocabulary and content knowledge led to difficulties in the search process. Bilal and Watson [2] conducted a case study with children from a 7th grade science class (children between 11 and 13 years old) to determine how this group of users solve frequent school information tasks on the Web directory Yahooligans!. This Web service provides a directory structure in which users can browse from a large collection of Web sites. A search box is also provided to let users formulate search queries to find Web sites matching the query terms. Bilal and Watson [2] found that children tend to ignore the browsing categories and that they start their search directly using the search box utility.

3.Age Based Behavior

In the following paragraph we analyze each one of these types of factor. The focus is on finding metrics that give insight into the search difficulty that different aged users face. We will motivate in each section how each one of the metrics explored provides insight into the search difficulties of young users.

3.1Search Behavior of different Aged users

Search process in IR systems [12]. The correlation between query length and IR effectiveness has been explored before [2] [15]. On TREC ad hoc settings, it has been found that longer queries lead to better search performance and user satisfaction [2]. Nonetheless, recent studies show that this result does not always hold on the Web scale [12]. Recently, a strong association between the length of the query and the specificity of the users query intent has been found [22], in which longer queries lead to a more specific and less ambiguous set of results. Thus, the submission of longer queries (in respect to the average length) to the search engine is a strong indicator of the capacity of the user to construct queries that are more specific.

3.2Natural Language Usage in Queries

Natural Language Usage in Queries the aim of analyzing the usage of natural language in the queries is twofold: (1) As a mean to retrace child development: Children typically have a greater sense of curiosity, which we hypothesized is reflected in the searches they perform. For instance, we expect a greater amount of question queries for users below 10 years old and greater usage of superlative constructs. (2) Children have been observed to pose queries in natural language given their lack of familiarity with the keyword approach of search engines. Greater usage of this type of queries represents evidence of greater difficulty in expressing complex information needs through keywords which are better suited to

modern search engines. The following query types were created to quantify these phenomena. Query Preprocessing, we assumed the set that represents the query preprocessing steps as P.

1) P = where are the tasks stop – word removal, tokenization, stemming respectively.

2) Cosine Similarity D: D = $\{q_1, q_2, \ldots, q_n\}$ Where D is the set of all the cosine similarities measure between query terms.

$$\text{CosSim} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \qquad (2)$$

(1) Question queries. These are queries for which the first token is a question word (how, where, what, etc.), or the last character of the query is a question mark (e.g., what is the only immortal animal?).

(2) Modal queries. These are queries containing auxiliary verbs such as will, won't, don't, or modal verbs as shall, should, can, etc. (e.g., I don't want to go school).

(3) Knowledge questions. These are queries containing the words describe, about, explain, define, or interesting.

(4) Superlatives. These are queries containing superlative adjectives (e.g., the fastest dog).

(5) Kid-targeted queries. These are queries with the terms for kids or for children.

4.Different Age Groups and User Modeling

Given the small amount of content carefully designed for this audience and the lack of specialized search engines dedicated to help users find appropriate content on the Web. To characterize the search behavior of users and the browsing activities that lead to search. In terms of search behavior, we focus on identifying the difficulties that users encounter on the Internet when they search for information with a state-of-the-art search engine.

4.1Problem Definition

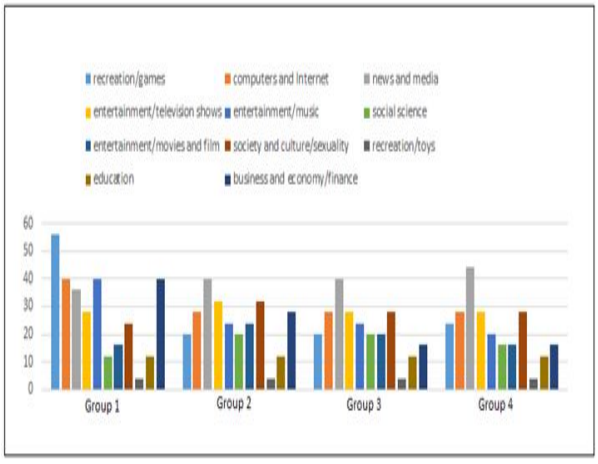

Figure 1: Type of activities different Age groups are involved in.

We formulize the problem in this section. The demographic attributes concerned in this paper include gender and age. We present a user's demographic attributes as two vectors gender and age. The age prediction is defined as classifying users into one of the following groups in Table 1.

Table 1: Age Groups

Group	Age-range
Teenage	14-17
Young	18-29
Mid-Age	30-64
Elder	65 and above

Given the webpage click-through log of some users with known demographic attributes, the problem is to find a general method to predict some users with unknown demographic attributes given their web-page click-through log.

4.2System Overview

The objective of our proposed system is to provide, effective Search by capturing the demographic attribute (age) of the user and hence user modeling Based on the current intention here we characterize the different user's age and search intents by analyzing the association patterns of their issued queries and corresponding clicks.

The demographic attribute concerned in this paper include age. We present a user's demographic attributes as a vector, of user's age. The age prediction is defined as classifying users into one of the following groups in: 14+, 18+, 30+ and 65+. Where 14+ is an abbreviation to a age group [14-18], 18+ is an abbreviation to a age group [19-30], 30+ is an abbreviation to a age group [31-65], 65+ is an abbreviation to a ages above 65 years. The demographic attribute concerned in this paper include age. We present a user's demographic attributes as a vector, of user's age. The age prediction is defined as classifying users into one of the following groups in: 14+, 18+, 30+ and 65+. Where 14+ is an abbreviation to a age group [14-18], 18+ is an abbreviation to a age group [19-30], 30+ is an abbreviation to a age group [31-65], 65+ is an abbreviation to a ages above 65 years.

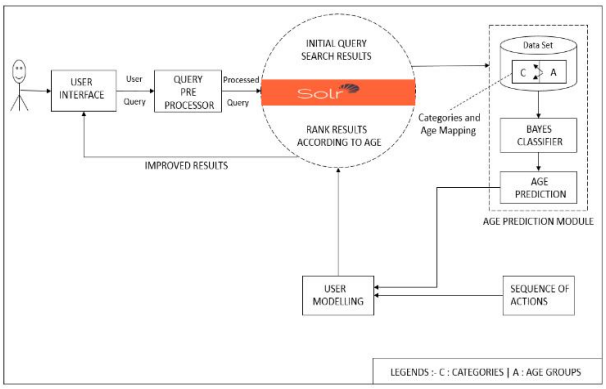

Figure 2: System Overview

4.3Demographic Prediction

Based on the age of user the queries entered will be classified in predefined groups, we use a Bayesian framework to predict the user's demographic attributes.

Similarly other classification techniques like Adaboost classifier, SVM Classifier are used for prediction of same. Users are classified in to predefined aged groups using the classifying techniques. Classification using Bayesian Approach is elaborated in this section. Table 2 shows few instances of the queries that resemble to the predefined age groups. Figure 4.3 depicts the types of activities users are involved on internet. Classification of users is done considered in all these aspects.

Table 2: Types of Queries Varying According to User Groups

14+	18+	30+	65+
School	College	Career	Knee Pain
Science	Graphics	Job	Doctor
History	Relationship	Bollywood	Pilgrimage
Election	Fashion	Music	World Tour
Elementary	Trends	Movies	Divine

4.4Classification

This task classifiers the users into predefined groups.

Bayesian Classifier with,

(a) Input: A input query q

A fixed set of age groups C = $\{g_1, g_2, \ldots, g_n\}$

A training set of m query categories,

$(q_1, c_1), \ldots, (q_m, g_m)$

Output:

(b) Bayes Classifier:

MAP is "maximum a posteriori" = most likely class.

$$C_{MAP} = \arg\max_{q \in A} p(g|q) \quad (1)$$

Bayes Theorem

$$C_{MAP} = \arg\max_{q \in A} p\frac{(g|q)p(q)}{p(q)} \quad (2)$$

Eliminating the denominator

$$C_{MAP} = \arg\max_{q \in A} p(g|q)p(g) \quad (3)$$

4 Implementation Details

The click-through data of the AOL search engine is used to. In total it spans 3 months of data and contains queries and clicked URL's pair along-with clicks instances. Click-through data is nothing but the activity performed by users and it reflects their interest along-with the semantic relationship between users and queries and between queries and clicked web documents. Click-through data quintuple representation is < u, q, l, r, t >, where u is the

user ID, q is a query issued by the user, l is an URL which user clicked on, r is the rank of the clicked URL and t is the time at which user submitted a query for search. But in this work as just query-URL bipartite graph is constructed using information of queries and URLs pair, < q, l> data is extracted from the quintuple of click-through data and used for constructing the required Query-URL bipartite graph. The other information is ignored. The data set used is being recorded by search engines, it contains a lot of noise as it is a raw data so the noise is cleaned, and only frequent and well formatted data was kept. Log data from users with ill-defined fields were also excluded (e.g., invalid zip codes). This filtering step is compulsory in order to be able to identify the age of the user. The resulting data was cleaned further by applying the following criteria:

i. queries containing only a single token and that contain exclusively non alphanumerical characters;
ii. queries that were issued by only a single user within a given age group;
iii. Queries containing personally identifiable information, such as credit card numbers or full street addresses.

The first criterion was carried out by using a rule-based approach. For instance, regular expressions were designed to detect non alphanumerical characters in the query string. For the second cleaning criterion we relied on the support of each query, which is obtained.

We use a dataset of KDD-Cup for the age prediction and hence the user modeling. The KDD-Cup 2005 Competition was held in conjunction with the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining and presented a real web search engine problem, categorizing search queries, to challenge the data mining community. The task of the KDDCup 2005 competition was to classify 800,000 internet user search queries into 67 predefined categories [11].

5.1 Evaluation Metrics

The performance of the presented methods was evaluated using the conventional precision (Prec), recall (Rec) and F1 measures [12]. Precision p is defined as the proportion of correctly predicted examples in the set of all examples assigned to the target class.

p = true positives / (true positive +false positive)

Recall r is defined as the proportion of the correctly predicted examples out of all the examples having the target class.

r = true positives / (true positive +false negative)

F1 is a combination of precision and recall defined as follows:

$$F_1 = \frac{2pr}{p+r} \quad (1)$$

5.2 Results

The first step of proposed work is Query preprocessing. After query preprocessing, the next job we have done is to find the similarities between consecutive queries. Cosine similarity serves the purpose. Table 3 gives clear idea of how similarity between queries is measured, between two consecutive queries sayQ_1, Q_2 where Q_1 is "User Modeling" and Q_2 is "User Modeling Approaches".

Table 3: Similarity between Queries

Quires	Query Vectors		
	Initial Weights	Weighted Calculated	Cos Sim
Q_1	[1,1,0,1,1]	[0.58,0.58,0.58,0.58,0.58]	[0.9]
Q_2	[1,1,1,1,1]	[0.58,0.58,1.58,0.58,0.58]	
	[2,2,1,2,2]	All words = 5	

The figures included in table are rounded off to two decimals. A separate query log is maintained for each query which will eventually help in classifying the user age and improve the ranking of results. Figure 4 depicts an instance of XML query log generated. Query Log is the XML data which contains the nodes: User ID, Search Word, Time Stamp, Related Query, and Cosine Similarity.

```
<input><user>U683367906878</user><id>420555280469</id>
<search_word>ipad air</search_word><time>2014-12-03
06:23:33.938</time></input><query_rel>420555280469</query_rel>
<query_rel_sim_index>0.0</query_rel_sim_index>

<input><user>U683367906878</user><id>381788046302</id>
<search_word>ipad air mac</search_word><time>2014-12-03
06:23:44.844</time></input><query_rel>420555280469</query_rel>
<query_rel_sim_index>0.5938759</query_rel_sim_index>
```

Figure 4: XML Query Log

A standard Dataset KDD Cup is used where the age groups are matched to the query categories on certain level of assumption. We reduce the scope to 3 categories as of now, viz. Entertainment, education, sports among all the 67 categories in the dataset. For the classification of the user in the predefined age group Bayes classifier is used and the user age is inferred. We evaluate the results retrieved based on the identified age group using Precision, Recall, and F1 measure.

Table 4: Precision, Recall and F-measure for the Bayesian Classifiers

Group	No. of Q	% Score	Accuracy	Precision	Recall
14+	50	70%	0.704	0.55	0.98
18+	50	76%	0.76	0.64	0.96
30+	50	81%	0.816	0.68	0.98
65+	50	86%	0.73	0.62	0.91

Hence we need to find the required attributes, True Positive, False Positive, and False Negative. True Positive is ranked results that are relevant to the user. False Positive is ranked results that are partially relevant to the user. False Negative ranked results that is irrelevant to he user. The above mentioned values are calculated on 50 queries for each age group. Figure 3 shows the evaluation metrics for relevance of retrieved by the system.

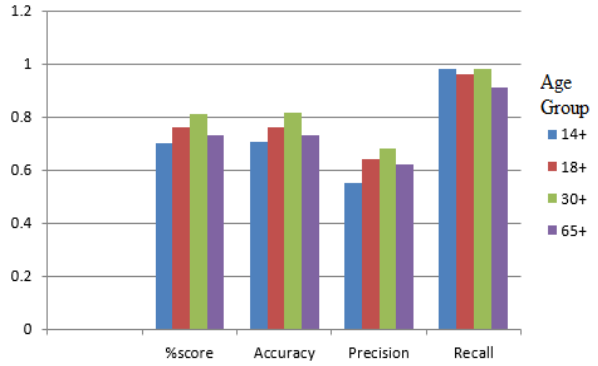

Figure 3: Evaluation Metrics for relevance of results

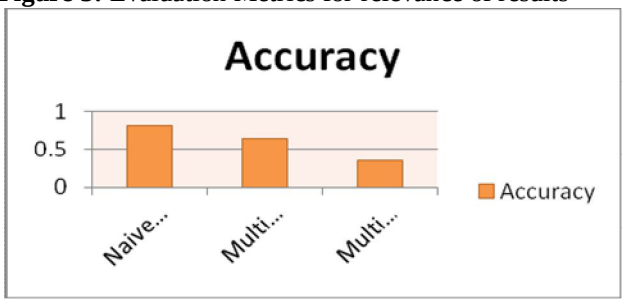

Figure 4: Comparison of Different Classifiers.

The user behavior is captured using two methods further (1) the classifier and (2) the user based method. Mainly the user based method is based on the queries entered in the respective session. Table 5 summarizes this scenario where in the respective groups are predicted using both the methods and the number of correct instances and wrong instances and accuracy for each instance is calculated accordingly.

Classification based on classifier is the process described in the previous sections. And classification based on user behavior is the maximum likelihood of the query trail to the query in the predefined group.

Table 5: Test case Results

Search query	Classification based on classifier	Classification based on UB	Accuracy	Correct	Incorrect
Games	Group 1	Group 1	1.001	1	
News	Group 2	Group 1	0.666		1
Education	Group 2	Group 1	0.5		1
Carrier	Group 2	Group 2	0.5	1	
Tv shows	Group 2	Group 2	0.57	1	
Football	Group 2	Group 2	0.625	1	
how to tie a tie?	Group 2	Group 2	0.666	1	
Salary report	Group 3	Group 2	0.6		1
hair styles for women	Group 2	Group 2	0.63	1	
funny college jokes	Group 2	Group 2	0.66	1	

As this approach is purely based on how well the system is trained, and the user groups with specific queries, interest,

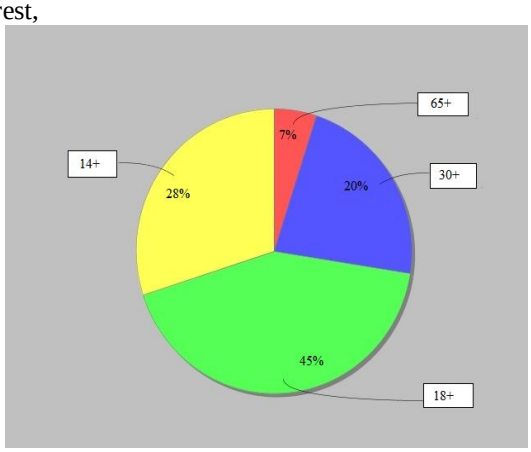

Figure 4: Testing response of different aged users

We tested our results with 20 participants from each age group and took an explicit feedback from each one of them for the relevance of results. Figure 4 shows pie representation of user's opinion, where the most relevant results were given to the age group 18+.

Table 6: Performance of Classifiers

Age Group	Bayes	Adaboost	Multiclass
14+	0.704	0.42	0.623
18+	0.76	0.53	0.71
30+	0.816	0.43	0.72
65+	0.73	0.32	0.51

In the above Table 5 we have tasted Bayes, Adaboost and Multiclass classifiers on different age groups whereas Bayesian has given better performance comparing to others.

Conclusion

User Models integrate ideas from machine learning, intelligent agents, and human-computer interaction. Demographic features such as age and gender require special attention to provide adequate tools aimed at improving the content readability. The proposed work focuses on demographic prediction (age prediction) based on the type of query put by the user. User intention is predicted by capturing the search sequences followed. When predicting user intentions, the accuracy is crucial; thus, increasing it is an important task. We have used a probabilistic method: Bayesian classifier for age prediction. The algorithm is used to classify the user in one of the predefined age groups based on the query category and is helpful to differentiate usage behavior of different users, particularly where there is a large number of training sequences and the goals in the system can be accomplished using more than one possible path. In future, this work can be extended by creating prediction model by using HMM and the technique which gives can be used to further process of user modeling. We also look forward to work on real time data. And also try capturing the gender of the user as there are noticeable differences in online behavior of male and female.

References

[1] W.-H. Au, K.C.C. Chan, and Xin Yao. A novel evolutionary data mining algorithm with applications to churn prediction. In Evolutionary Computation, IEEE Transactions on, volume 7, pages 532–545, Dec 2003.

[2] Nicholas J Belkin, Diane Kelly, G Kim, J-Y Kim, H-J Lee, Gheorghe Muresan, M-C Tang, X-J Yuan, and Colleen Cool. Query length in interactive

[3] information retrieval. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pages 205–212. ACM, 2003.

[4] D. Bilal and Watson. Children's paperless projects: Inspiring research via the web. In IFLA General Conference and Assembly. 1998.

[5] K.R. Bisset, A.M. Aji, E. Bohm, L.V. Kale, T. Kamal, M.V. Marathe, and Jae-Seung Yeom. Simulating the spread of infectious disease over large realistic social networks using charm++. In Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2012 IEEE 26th International, pages 507–518, May 2012.

[6] K.R. Bisset, A.M. Aji, M.V. Marathe, and Wu chun Feng. High-performance Biocomputing for simulating the spread of contagion over large contact networks. In Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on, pages 26–32, Feb 2011.

[7] Hirsh S. G. Walter V. A. Borgman, C. L. and A. L. Gallagher. Children's searching behavior on browsing and keyword online catalogs: The science library catalog project. J. Amer. Soc. Inf. Sci. 46, 9, 663684., 1995.

[8] E. Broch. Children's search engines from an information search process perspective. School Libr. MediaRes.3.http://www.ala.org/aasl/aaslpubsandjournals/slmrb/slmrc ntents/volume32000/2000.

[9] M.T. Cazzolato and M.X. Ribeiro. A statistical decision tree algorithm for medical data stream mining. In Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on, pages 389–392, June 2013.

[10] Fucai Chen, Xiaowei Li, and Lixiong Liu. Improved c4.5 decision tree algorithm based on sample selection. In Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on, pages 779–782, May 2013.

[11] Paul J. Dionne. Epidemiology. Biomedical Engineering, IEEE Transactions on, BME-19(2):126–128, 1972.

[12] M. Doerr and M. Papagelis. A method for estimating the precision of place name matching. Knowledge and Data Engineering, IEEE Transactions on, 19(8):1089–1101, Aug 2007.
[13] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. Understanding the relationship between searchers' queries and information goals. In Proceedings of the 17th ACM conference on Information and knowledge management, pages 449–458. ACM, 2008.
[14] C. Griffin and R. Brooks. A note on the spread of worms in scale-free networks. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36(1):198–202, 2006.
[15] F. Hashim, K.S. Munasinghe, and A. Jamalipour. Biologically inspired anomaly detection and security control frameworks for complex heterogeneous networks. Network and Service Management, IEEE Transactions on, 7(4):268–281, 2010.
[16] Bernard J Jansen. Search log analysis: What it is, what's been done, how to do it. Library & information science research, 28(3):407–432, 2006.
[17] Yanqing Ji, Hao Ying, J. Tran, P. Dews, A. Mansour, and R.M. Massanari. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. Knowledge and Data Engineering, IEEE Transactions on, 25(4):721–733, 2013.
[18] S.L. Kosakovsky Pond. Computational analysis of hiv-1 evolution and epidemiology. In Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, pages 60–63, Nov 2011.
[19] Dharminder Kumar and Suman. Performance analysis of various data mining algorithms: A review. In International Journal of Computer Applications (0975 8887) on, pages 389– 392, October 2011.
[20] A.R. Mikler, A. Bravo-Salgado, and C.D. Corley. Global stochastic contact modeling of infectious diseases. In Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS '09. International Joint Conference on, pages 327– 330, Aug 2009.
[21] D. Nahl and V. H. Harada. Composing boolean search statements: Self-confidence, concept analysis, search logic, and errors. School Library Media Quart. 24, 199207., 1996.
[22] D. Neuman. High school students use of databases: Results of a national delphi study. J. Amer. Soc. Inf. Sci. 46, 4, 284298., 1995.
[23] Nina Phan, Peter Bailey, and Ross Wilkinson. Understanding the relationship of information need specificity to search query length. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 709–710. ACM, 2007.
[24] Hong Qin, A. Shapiro, and Li Yang. Emerging infectious disease: A computational multi-agent model. Biomedical Computing (BioMedCom), 2012 ASE/IEEE International Conference on, pages 28–33, Dec 2012.
[25] P. Solomon. Children's information retrieval behavior: A case analysis of an opac. J. Amer. Soc. Inf. Sci. 44, 5, 245264., 1993.
[26] Wang Tao and Li Wei-hua. Naive Bayes software defect prediction model. In Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on, pages 1–4, Dec 2010.
[27] S. Volkova and W.H. Hsu. Computational knowledge and information management in veterinary epidemiology. In Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on, pages 120–125, May 2010.
[28] Juanying Xie and Shuai Jiang. A simple and fast algorithm for global k-means clustering. In Education Technology and Computer Science (ETCS), 2010 Second International Workshop on, volume 2, pages 36–40, 2010.
[29] Hui Xiong, G. Pandey, M. Steinbach, and V. Kumar. Enhancing data analysis with noise removal. Knowledge and Data Engineering, IEEE Transactions on, 18(3):304–319, March 2006.
[30] Xiwang Yang, Yang Guo, and Yong Liu. Bayesian-inference based recommendation in online social networks. Parallel and Distributed Systems, IEEE Transactions on, 24(4):642– 651, April 2013.

AUTHOR

Prajakta Shitole received the BE (2012) and pursuing M.E. degrees in Computer Engineering from D.Y. Patil COE Akurdi Pune under Savitribai Phule Pune University. Her area of interest is in Information retrieval.

Madhuri Potey is pursuing PhD at University of Pune in Government College of Engineering, Pune, India. She acquired BE(1993) and MBA(1996) from Amaravati University, India and ME (2006) from COEP, Pune, India. She is also a faculty at D Y Patil College of Engg., Akurdi, Pune, India. Her research interests are Information Retrieval, Advanced Databases, Cryptography and Software Engineering. She is ACM Professional member since 2009.