IJCTA, 9(21), 2016, pp. 273-280

Cognitive Analytic Task Based on Search Query Logs for Semantic Identification

D. Suryanarayana* Prathyusha Kanakam* S Mahaboob Hussain* Sumit Gupta*

***Abstract* :** Understanding and analyzing the cognition of the user by the search engine to predict the subsequent actions to be performed and to provide better output results is a critical task. Technology is advancing ahead in such a way that, information is being provided to the user mostly through the web. Therefore, it is the major responsibility of a search engine to analyze user's cognitive actions to predict the consequent activity performed. Predictive analysis plays a significant role by utilizing the intent from the search history of the user. This paper furnishes a concise scenario about cognizance procedure in various phases to co-relate each and every activity that took place in antiquity to determine the consequent task performed by the user. Query logs are analyzed for recommending predictable activities (queries or visited URLs) based on their current access behaviours employing principal component analysis.

Keywords : Query log analysis, Query Recommendation, Search Engine, Cognition, Prediction, Crawler, Semantic Web

1. INTRODUCTION

A prediction is a form of personalization task that depends on the fact that individuals frequently utilize search engines to identify previously viewed resources. Predicting future activities can be employed for various purposes such as planning, resource allocation and identification of risks and opportunities. An individual past behavior is identified via query log analysis and used to estimate the future navigational behavior of the corresponding individual. Query log analysis is a key to understanding what users intend to do in their sessions and to predict their later activity. Query recommendation, Webpage re-ranking, advertisement arrangement are some of the applications associated with query log analysis and session search. Future activities of the users are predicted by utilizing current access behaviors and query logs. Inferences on users' information are utilized to measure the similarity between queries and how successful their new queries are anticipated to be.

To retrieve the information, session search will provide a path when a user issues aggregate queries continuously to a search engine in the quest if one or more information needs are satisfied. In this session search, a session is defined as the frequent number of communications or interactions with a particular search engine which is categorized in various approaches. To improve the better outcomes of a search engine and to satisfy the user with the precise information, one can enhance and interpret the essential procedure of interactions in a session search. Sessions are pursued by formulating the queries repetitively in various expressions. Queries will be reformulated and posed to search engines to retrieve novel outcomes.

In a session search, a user furnishes a sequence of queries involved with Uniform Resource Locator (URL) links. Subsequent to each query submission or URL click, forecasting of the queries will be done. Accurate predictions can facilitate the seek procedure of a user which orders the resulting Web pages or set up relevant advertisements. Each session is the combination of the queries posed in the past and the queries to be posed in the future.

$Q_L : S \rightarrow (P \Lambda F)$ is represented as, $S \rightarrow \{Q_1, Q_2, Q_3, Q_n\} \Lambda \{Q_{11}, Q_{12}, Q_{13}, Q_n\}$ where Q_L : Query Logs, S : Session, P : Past queries and F : Future queries.

* Department of Computer Science and Engineering Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India Email-{suryanarayanadasika, prathyusha.kanakam, mahaboobhussain.smh} @gmail.com and sumit108@hotmail.com

This paper gives a scenario about the state-of-the-art manoeuvre to figure out future activities of an individual by utilizing their session logs. The association between queries in the log is illustrated by analyzing the core part by applying principal component analysis. The interesting measures (support and confidence) for each query are computed and will help in estimate regression factor or co-relation factor of queries. The later section comprises related work and is followed by a neoteric procedure to predict future activities. Finally, the conclusion part is in the last session.

2. RELATED WORK

To grab their results, users always surf the internet using search engines. During their session, users may pose various informational, navigational, and transactional and connectivity type of queries into the query interface of the search engine. The input may be queries or URL links. It is the responsibility of the search engine to predict the future activity of its dependent user on working with their session logs.

Sessions consist of various combinations of queries containing core terms related to underlying information need and additional terms that reflect the user's cognitive changes. Over the course of a session, the core terms may change as well. At any point in a session, three possible term actions [1] are available to a user.

- **Term Detention :** The core terms are identified by following terms from one query to the next, for the current information required
- **Term Discard :** Amputating term from a query
- **Term Affixing :** Adding a new term not present in the preceding query to the query reformulation

The above three term-actions provide a path for query reformulations and they can be illustrated by considering the search log presented in Table 1. It represents the session of an explorative search of an individual on dental science.

Table 1. Explorative Search of an Individual

Activity	*Description*
Q1	Pedodontic Methodologies
URL1	http://www.pedodontics.com/surgery.html
Q2	Pedodontics surgery
Q3	Pedodontic experts
Q4	Role of pedodontics in dental science
URL 2	http://www.dentalscience.com/materials/pedodontics

The term 'pedodontic' is retained through three queries (Q1, Q2, Q3) with the user affixing and discarding terms 'methodologies', 'surgery', 'experts' in order to explore the topic. The focus shifts in Q4 changing to ' role of pedodontics in dental science', indicating a change in information need which is expanded upon in Q4.

Reformulation of the query, re-ranking of Web pages as well as the arrangement of advertisements is the potential applications to predict the future activities by understanding the users' behavior in current search sessions and query logs. For the most part of the research, the results are with the suggestion of the URLs of similarity, consists of proposals of queries and also context-attentive ranking.

The purpose of query proposition is to put forward a sequence of queries which are acknowledged by means of user current search intent. Fonseca et al. [2] work out the resemblance between users present queries with a set of candidate queries and recommend the candidate queries which comprise the uppermost similarity scores. In 2006, Zhang and Nasraoui [3] renovate each session in a query log dataset into a chart of query nodes and work out the relatedness of the queries based on the queries' path distance. In the same way, Boldi et al. [4] [5] create a bulky directed graph out of the queries in a query log dataset. Query suggestions are finished all the way through performing PageRank taking place on the diagram and queries are put forward by means of top PageRank values. In 2010, Cheng et al. [6] recommend queries associated with the Web page that a user is currently browsing.

The point of URL recommendation is to suggest a set of URLs which are associated with a users' in progress search objective. Wang et al. [7] present a structure to foresee the future clicked URLs of a user after the user poses a question. Their prediction technique aims at generating the URLs that a user will click during the rest of their investigating session.

In the revision of context-aware ranking, the documents in the search results are ranked by taking a user's precedent search activities into consideration. In 2005, Shen et al. [8] execute context-aware document ranking by promoting URLs that are additional and similar to a users precedent queries and clicked URLs. Agichtein et al. [9] unite users earlier period click actions by means of conventional information retrieval replica BM25 which is a function for ranking the documents by the search engines. Learn to rank is a research study of Xiang et al. [10] which is a unified algorithm, capable of reallocating the rank of the documents given by the search engines.

3. COGNITIVE PREDICTIVE TASK

This cognitive predictive task state of art technique will predict the user's future activities entirely depending on the computed measures (Support and Confidence) obtained by association query mining for different queries which are presented in the past query log of an individual and by applying principal component analysis, the core part of the sequence of questionnaire posed by the user to search engine is analyzed. Using the core part, the query is reformulated. Query reformulation of a user during a session includes cherishing terms from the unique query; evacuate others and toting up new terms. Simply, the query reformulation in a session search is typically closely related to the users cognitive shifting for the same previous query. Sometimes, searching amid session is an ambiguous task as the user may not have clear information at the starting and they are unsure of how to explicitly characterize their information need.

'Q' may be a query log of an individual which is the combination of both past and future activities (queries or URL clicks) Q = {P,F} where P represents the past activities of the user, P = {Q1,Q2,U1,Q3,...,Qn} where each activity either is a URL click or Query and F represents the future activities of the user, F = {U11, U12, U13, U14,...,U1n} which contains URLs in a specified table is represented as in Figure 1.

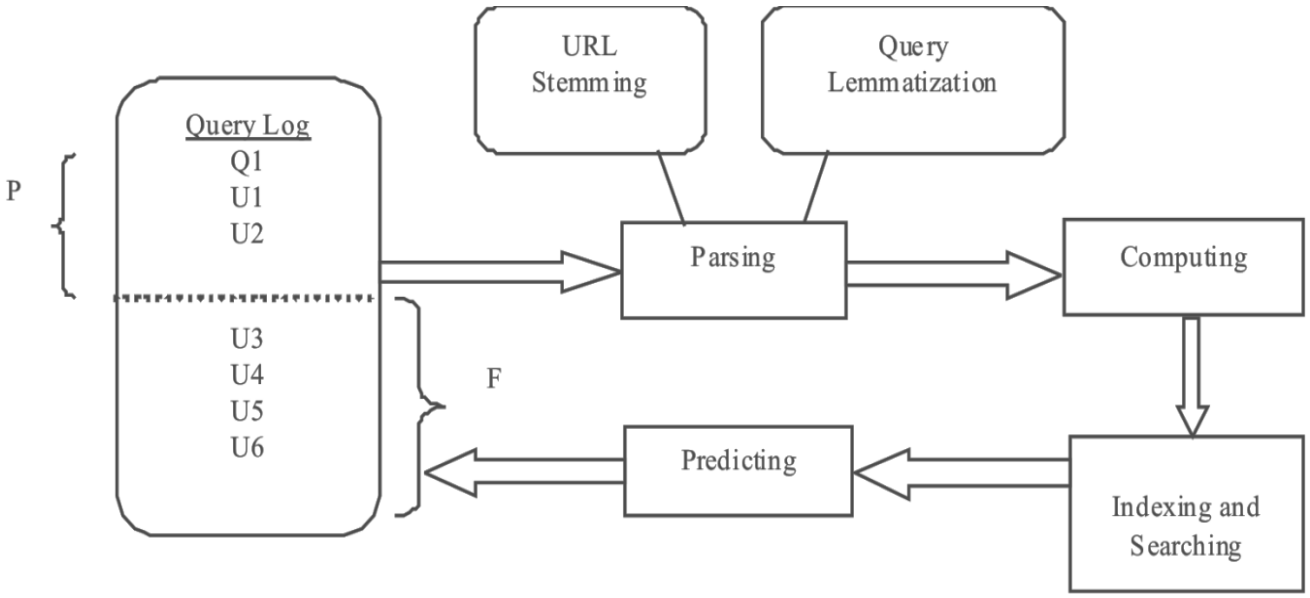

Fig. 1. Phases of cognitive predictive task

The cognizance applied manoeuvre predicts an individual's future activities through various stages. This approach constitutes four stages: Parsing; Computing; Indexing and Searching; Predicting; at every stage, the individual procedure takes place. The scheme begins with past query log, in which each activity is split in order to grab the unique component in the activity either it may be query or URL link.

3.1. Parsing phase

Depending on the activity performed by the user, parsing process varies. Each URL undergoes stemming[1] whereas query undergoes lemmatization[2] [11]. Thus, parsing in this state of art technology is of two different types - URL stemming and Query lemmatization.

URL Stemming : It is the process of chopping a URL into its parts- domain name, a path of the file, protocol etc. It is the reference to a resource available on the computer network, which exactly specifies the location of that resource. Stemming process is applied in such a way that a single URL click can be furcated as shown in Table 2 for URL clicks, for instance, U1 → http://www.pedodontics.com/surgery.html and U2 → http://www.dentalscience.com/materials/pedodontics.html.

Table 2. Separating the URL into elements

URL Click	*Protocol*	*Sub-Domain*	*Domain Name*	*Top-Level Domain*	*File Path*
U1	http	www	Pedodontics	Com	surgery.html
U2	http	www	Dentalscience	Com	materials/pedodontics.html
U3	http	www	gdc-uk	Org	dental professionals.html
U4	http	www	Isppd	Org	Publication
U5	http	www	Aapd	Org	Events
U6	http	www	Jisppd	Com	Articles

Query Lemmatization : It is the technique of acquisition of root word or lemma in the query. To understand the context and determining the grammatical structure of a statement in a Query, Lemmatization acts as a fundamental undertaking procedure. This can be illustrated as follows in the Table 3. The query is chopped into lemmas and the remaining stop words are removed from the query statement.

Table 3. Deriving lemmas from the query

Sl.No	*Query*	*Lemmas*
Q1	what are the pedodontics methodologies?	pedodontics, methodology
Q2	how to perform Pediodontics surgery?	pedodontic, surgery
Q3	who are the Pediodontric experts?	pediodontic, materials
Q4	Role of pedodontics in dental science	process, pediatric, dentistry

3.2. Computing phase

This is the primitive phase in the neoteric scheme. The evaluation takes place in two steps. At the first step, the support count for every term in the candidate queries as well as the domain names of the URL clicks is obtained using Palazzo Matrix Model [12] and at the second step, the principal variable is obtained from each activity, either a query or the URL click, through principal component analysis and correlation matrix which provides the correlation score among various queries of prior user's session.

Analysing principal component [13] is the procedure of decreasing a variable by obtaining a number of variables and measuring the correlation score within the variables where a variable represents a query. Each activity is furcated so that query is reduced in terms of its lemmas and URLs in terms of resource names and each activity is supplied a separate score calculated for all the individual terms in that particular activity or a variable. The correlation matrix for Table 3 of activities is mentioned in Table 4 as follows.

Values of the correlation matrix are given in such a way that all the activities specified in Table 1 fall under into two observed components say Q_1, U_1, Q_2, Q_3 related to various methodologies, surgeries, and experts in pedodontics and Q_4, U_2, related to the role of pedodontics in dental science. The typical row and column intersect to give the correlation between two corresponding variables or activities. For instance, the row for variable 2 intersects with the column for variable 1 gives correlation score 0.75 (the correlation between U_1 and Q_1 is 0.75).

Table 4. Correlation matrix table

Variable	Q_1	U_2	Q_2	Q_3	Q_4	U_2
Q_1	1					
U_2	0.75	1				
Q_2	0.35	0.83	1			
Q_3	0.28	0.65	0.62	1		
Q_4	0.03	0.04	0.07	0.06	1	
U_2	0.02	0.05	0.06	0.05	0.86	1

Now, principal component is calculated as the linear combination of two optimally weighted observed components. The selected observed components can be computed as,

$$c_1 = b_{11}(A_1) + b_{12}(A_2) + ... + b_{1p}(A_p)$$

Where C_1 is the score of an observed component, *b*1*p* is the weight for the activity Ap *i.e*., it gives the correlation score with the observed component. Let say,

$$c_1 = 0.44(A_1) + 0.40(A_2) + 0.47(A_3) + 0.32(A_4) + 0.02(A_5) + 0.01(A_6)$$

The activities $\{A_1 = Q_1, A_2 = Q_2, A_3 = Q_3, A_4 = Q_4\}$ are related to an observed component so the weight of the activities is between 0.32 and 0.44 and $\{A_5 = Q_4, A_6 = U_2\}$ are not related to component so their values are between 0.01 and 0.02. Similarly, the second component c2 computed as

$$c_2 = 0.01(A_1) + 0.04(A_2) + 0.02(A_3) + 0.03(A_4) + 0.48(A_5) + 0.31(A_6)$$

Equivalently, activities A_5, A_6 are related to component C_2 hence their values are between 0.31 to 0.48 and A_1, A_2, A_3, A_4 are not related so their values are between 0.01 to 0.04. Of the two components, which gives the highest score will be coined as the principal component.

3.3. Indexing and Searching Phase

After the principal component is computed, the activities are grouped independently according to scores and indexed in order to predict the future activities. Grouping the activities can be illustrated in Table 5. Here, group G1 furnishes the information associated to "Pedodontics" exclusively {surgeries, methodologies, experts} and group G2 provides knowledge about "Role of Pedodontics".

Table 5. Grouping of the associated activities

Activity	*Description*	*Group*
Q_1	Pediodontic Methodologies	G1
URL1	http://www.pedodontics.com/surgery.html	
Q_2	Pediodontics surgery	
Q_3	Pediodontric experts	
Q_4	Role of pedodontics in dental science	G2
URL 2	http://www.dentalscience.com/materials/pedodontics	

As a result, these groups indicate the observed components in order to derive the principal component. I.e., entire past session of activities of the user correlate to key elements of "pedodontics". Consequently, a separate ID is indexed to complete the searching process comfortable.

An inbuilt crawler indexes all the web pages related to a principal component is derived from the group of activities. Crawler generates a copy of visited pages by a search engine that provides downloaded pages for quick searches against a separate table for the next processing. The table is acknowledged as the crawler-related table that comprises both lists of collected pages including a list of URL's to be visited (for future activities).

3.4. Predicting phase

It is the ultimate phase to anticipate the next activities of the user using ID (which is indexed) for the principal component derived in the earlier phase. Using the ID of the principal component, next activities relevant to the past session is retrieved by using the crawler-related table which contains the description and ID of the principal component derived as shown in Table 6.

Table 6. Crawler-related table for the description and ID of the principal component

G1	Pedodontics	http://www.ncbi.nlm.nih.gov/pubmed/20516301
		http://wwrw.drcastro.com/non-surgical-periodontics/treatment-methods.html
		Pediodontics surgery
		Pediodontic experts
		http://www.drbretemoran.com/procedures/non-surgical/treatment-methods/
		http://www.pureperio.com/
		Essentials of periodontology and periodontics
G2	Role of pedodontics in dental science	Pediodontic Methodologies
		http://www.pedodontics.com/surgery.html
		http://soauniversity.ac.in/home/ids/pedodontics
		Role of pedodontics in dental science
		http://manipal.edu/mcods-manipal/department-faculty/faculty-list/n-sridhar.html
		http://www.dentalscience.com/materials/pedodontics
		http://manipal.edu/mcods-manipal/department-faculty/faculty-list/Rashmi-nayak.html

The brief procedure of this neoteric scheme presented in a pictorial form. Initially, it starts with storing of user sessions in the database which contains a description and IDs of activities performed.

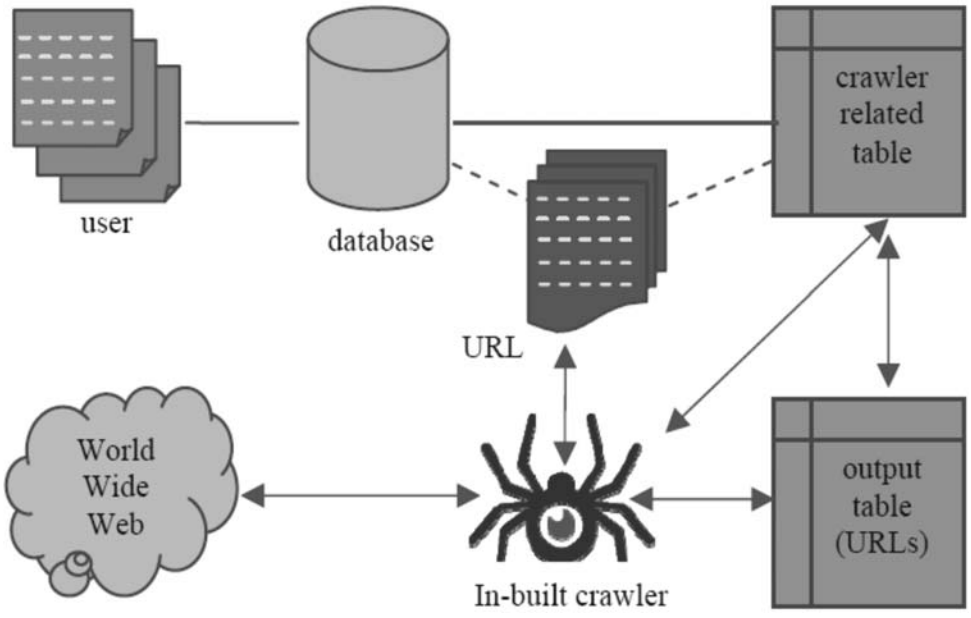

Fig. 2. Pictorial representation of cognitive predictive task for user queries

The database of the application stores two types of documents as, a list of collected pages and a queue of URLs to visit. *i.e.*, the database represents documents or web pages at two levels i.e., pages that already visited and pages to be visited. In-built spider as shown in the figure index all the web pages related to particular ID and description of the component derived and updates the web pages in the index table which is used for searching using the Internet.

Web crawlers or Web spiders do essentially employ for quick searches by a search engine by generating a copy of each visited pages and index the downloaded pages for later processing. The principal responsibility of the crawler is to look over two basic tables- Crawler-related tables and an output table, which stores the information to interpret future activities of the user.

As in Figure 2, a spider will regenerate all the URLs that are stored in crawler related table. These URLs are gained in an initial phase and which are correlated to an illustration of the group. This table is updated automatically by the spider when the user visits the URL which is not listed in the table. The spider acts as an intermediate between the Internet or the Web and crawler-related table for updating the URLs list[14]. Finally, it projects the output Table 7 which contains the URLs that the users wanted to visit (predictable) and is stored in terms of hypertext structure[3].

Table 7. Predictable URLS list

G1	Pedodontics	http://www.ncbi.nlm.nih.gov/pubmed/20516301
		http://wwrw.drcastro.com/non-surgical-periodontics/treatment-methods.html
		http://www.drbretemoran.com/procedures/non-surgical/treatment-methods/
		http://www.pureperio.com/
		http://soauniversity.ac.in/home/ids/faculties-ids/pedodontics
		http://www.gdc-uk.org/dentalprofessionals
		http://www.ada.org/en/education-careers/careers-in-dentistry/dental-specialties/specialty-definitions
		https://en.wikipedia.org/wiki/Pediatric_dentistry
		http://www.bue.edu.eg/index.php/pedodontics
G2	Role of pedodontics in dental science	http://soauniversity.ac.in/home/ids/pedodontics
		http://manipal.edu/mcods-manipal/department-faculty/faculty-list/n-sridhar.html
		http://manipal.edu/mcods-manipal/department-faculty/faculty-list/Rashmi-nayak.html
		http://www.isppd.org/%20publication
		http://www.aapd.org/events
		http://www.jisppd.com/articles
		http://www.ncbi.nlm.nih.gov/pubmed/21911945
		http://www.aapd.org/media/policies_guidelines/p_primaryspecialty.pdf
		http://www.wholelifedentist.com/procedures/pediatric-dentistry/

Table 7 is the subset of the crawler-related table as it only contains information about to be visited URLs. It is obtained after filtering crawler-related table with earlier user's activities performed in their sessions. As a result, this stratagem computes and anticipates the next activities to be performed by the user in easier, spontaneous and in an intelligent way. It evidently provides a track to identify users' behaviour by their logs.

4. CONCLUSION

This work clearly presents query log analysis which constitutes past and future activities performed by the user. The related work explains different methodologies to predict the future events of the user. It also gives a brief scenario about the neoteric scheme to anticipate the next activities using either queries or URL clicks through

various phases and eventually projects the output in the form of URLs to be visited by the user. This manoeuvre computes and envisages the subsequent activities to be carried out by the user in easier and intellectual mode. Certainly, it provides a path to identify user's behaviour by their logs.

5. ACKNOWLEDGEMENTS

This work has been funded and supported by the Department of Science and Technology (DST), Govt. of India, under the Grants No. SRC/CSI/153/2011.

6. REFERENCES

1. M. Sloan, H. Yang and J. Wang, "A term-based methodology for query reformulation understanding", Information Retrieval Journal, vol. 18, no. 2, pp. 145-165, 2015.
2. Fonseca, Bruno M., Paulo Braz Golgher, Edleno Silva de Moura, and Nivio Ziviani. "Using association rules to discover search engines related queries." In Web Congress, 2003. Proceedings. First Latin American, pp. 66-71. IEEE, 2003.
3. Z. Zhang and O. Nasraoui, "Mining search engine query logs for social filtering-based query recommendation", Applied Soft Computing, vol. 8, no. 4, pp. 1326-1334, 2008.
4. Boldi, Paolo, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. "The query-flow graph: model and applications." In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 609-618. ACM, 2008.
5. S. Momtazi and F. Lindenberg, "Generating query suggestions by exploiting latent semantics in query logs", Journal of Information Science, vol. 42, no. 4, pp. 437-448, 2015.
6. Cheng, Zhicong, Bin Gao, and Tie-Yan Liu. "Actively predicting diverse search intent from user browsing behaviors." In Proceedings of the 19th international conference on World wide web, pp. 221-230. ACM, 2010.
7. Wang, Xuanhui, Bin Tan, Azadeh Shakery, and ChengXiang Zhai. "Beyond hyperlinks: organizing information footprints in search logs to support effective browsing." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1237-1246. ACM, 2009.
8. Shen, Xuehua, Bin Tan, and ChengXiang Zhai. "Context-sensitive information retrieval using implicit feedback." In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 43-50. ACM, 2005.
9. Agichtein, Eugene, Eric Brill, and Susan Dumais. "Improving web search ranking by incorporating user behavior information." In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19-26. ACM, 2006.
10. Xiang, Biao, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen, and Hang Li. "Context-aware ranking in web search." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 451-458. ACM, 2010.
11. Hussain, S. Mahaboob, D. Surya Narayana, Prathyusha Kanakam, and Sumit Gupta. "Semantic Representation of Natural Language Query Using Combinatory Categorical Grammar and Lambda Calculus", in International Conference on Signal Processing, Control and Data Analytics, SAN DIEGO, USA, 2015.
12. Hussain, S. Mahaboob, D. Surya Narayana, Prathyusha Kanakam, and Sumit Gupta. "Palazzo Matrix Model: An approach to simulate the efficient semantic results in search engines." In Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on, pp. 1-6. IEEE, 2015.
13. I. Jolliffe, Principal component analysis. New York: Springer-Verlag, 1986.
14. Hussain, S. Mahaboob, D. Surya Narayana, Prathyusha Kanakam, and Sumit Gupta. " Stepping Towards A Semantic Web Search Engine for Accurate Outcomes in Favor of User Queries- Using RDF and Ontology Technologies." In International Conference on Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on, pp. 1-6. IEEE, 2015.