IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

An Anatomy of Faceted Search on World Wide Web

Yogesh[1*], Shalu[2], Komal Kumar Bhatia[3], Neelam Duhan[4]

[1,2,3,4]Dept. of Computer Engineering. J.C. Bose University of Science and Technology, YMCA, Faridabad, Haryana. India

**Corresponding Author: yogeshymca7@gmail.com, Tel: 9582515851*

DOI: https://doi.org/10.26438/ijcse/v7i11.114120 | Available online at: www.ijcseonline.org

Accepted: 09/Nov/2019, Published: 30/Nov/2019

Abstract— With rapid development of online web shops and E-commerce data, it is evident that users get convenience in different fields such as lexical similarity, Vocabulary mismatch, information retrieval etc. Faceted search is becoming a popular method to allow the user to interactively search in online web shops and product comparison sites. Trying to figure out retrieval of information using facet search to reduce the number of search results quickly to improve the search results. There are many attributes, for example, filter, facet value, facet and facet count, which can also be used for information retrieval towards the user search query. Over the years, all kinds of improve search results techniques have tried to simplify this task such as WebPT, NextGen and Kareo. This paper gives a detailed survey of some recent algorithms of faceted search, the attributes handled by them and the methods used by them.

Keywords---Faceted Search, Semantic link**,** Data Mining, Probabilistic Model, Spatial Database, Navigation System, Information Retrieval

I. INTRODUCTION

Faceted search is a guided search system that provides relevant search results according to the user query. It permits storage of all search results/transactions into permanent records and each record conveyed crosswise over numerous data. Facet, facet value and facet count are the key components of this search; they are little arrangements of information retrieval that occurred inside the framework. Each new facet stores the reference of dynamic data by including single or multiple filters of the data. Users use facets to filter search results to browse items with multiple dimensions. Presenting narrowing options (facets) is easier for users because they do not have to know the syntax necessary to specify their search precisely. The main scopes of faceted search are semantic link and semantic web [1], [2], query facet search [4] and brute force system [5] and so on.

They are usually derived by analysis of the text of an item using entity extraction techniques or from pre-existing fields in a database such as author, descriptor, language, and format [2]. Thus, existing web-pages, product descriptions or online collections of articles can be developed with navigational facets. Researchers have shown that if the relationships between objects were known, the recall gain and reachability can be improved up to 266% for hard topics and 373% for very hard topics [6]. Since facet attributes are derived from the search result set, users are never left with an empty result set [7]. In addition, being able to see all available options, users can better understand how data are structured and alternatively use that information to specify better searches in the future.

Faceted search is dealing with the retrieval of information according to user search query that narrow down the search result. It plays an important role in our day to day life, for example, random forest system [7],[19] which is used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is mode of classification or means predictor (regression) of the individual trees. Specificity and accuracy can also be increased and helps in build better information retrieval system. Facet corresponds to properties of the information elements.

Customer's lookup for what they are expect to find as fast as possible. In fact, 80% of visitors search for a product once they hit a retailer's site in e-commerce. This classification system aims to increase product discoverability and conversions by denying any objects that don't meet the user's selected criteria. Faceted navigation refers to how Ecommerce websites allow visitors to filter and sort results based on product attributes. Today users are provided direct access to an unprecedented number of electronic information sources, yet most users have difficulty utilizing the full capacity of the substantial amount of information that information retrieval systems offer. Techniques such as relevance feedback, term suggestion, query expansion techniques, and query auto completion techniques are potentially effective ways to enrich the query and lead to improved performance in the information retrieval task. Query assistance also suggests that longer queries lead to higher satisfaction and less iteration in an interactive web searching environment.

Many researchers have worked on methods and algorithms to retrieve right information by narrow down the search. In this paper, different existing methods have been discussed. The methods have been compared with each other to have future aspects in the discussed area. Rest of the paper is organized as follows, Section I contains the introduction of facet search, Section II contains the detailed survey of some recent facet search methods and algorithms, Section III presents a comparison study of different techniques and algorithms with future directions and Section IV concludes the research work.

II. RELATED WORK

Ying Liu [1] proposed an information search and retrieval framework on the basis of semantically annotated multi-facet product family ontology. Firstly, framework develops a multi- facet product family ontology for a type of product of interest. Secondly, product entities with their associated properties are identified and extracted using concepts such as frequent terms identification and term weighting for the ranking of most relevant entities or concepts. Thirdly, the semantic annotation process is carried out with further processing of entities, concepts and their properties upon extraction, such as the selection of feasible tags and identification of their possible semantic relationships. This paper proposed a scalable document profile (DP) model that suggested good design analysis by using semantic tags. This paper used a case study of digital camera families that illustrated how the faceted search and reclamation of product information can be refined. Finally, all the relevant results are retrieved from the design and manufacturing repository according to the facets identified. This paper proposed a metric called averaged PMI (avg PMI) to measure the averaged adequacy of the semantic association features. avgPMI calculation is represented by

$$\text{Avg PMI} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} PMI(w_i,\ w_j)}{N},\ i \neq j \quad (1) \quad \text{where}$$

$$\text{PMI}(w_i, w_j) = \log\frac{p(w_i\ \&\ w_j}{p(w_i).p(w_j)},\ i \neq j \quad (2)$$

Where wi and wj are two different terms of Maximal Frequent Sequences (MFSs) discovered and N is the total number of features in the DP model. PMI between wi and wj is calculated based on the probability of co-occurrence of wi and wj in a window of words over the size of corpus. This paper worked on a small collection of 39 documents on the three D-SLR cameras. Corpus for tags generation contained about 1, 75,000 words and an average 17 words for each sentence.

Flavius Frasincar [2] proposed a forum for multifaceted product search with the help of Semantic Web technology. This paper used the XplorePrdoducts.com having two environments. First environment, Ping service, where the end- users can search and explore products on the Web and an environment where Web shops were able to ping our platform with their product information containing Web Pages. Second environment, searching on XploreProducts.com, platform uses multifaceted category navigation, integrated with keyword search. When the user enters a query, the SPARQL query generator translates the user's input to a SPARQL query in order to get relevant results from the RDF database. The XploreProducts.com evaluation incorporates in two steps. Firstly, they analyzed the results of the recognition of identical products. Secondly, this paper evaluated the category mapping task of XplorePrdoucts.com, and last, user interface is evaluated. This paper worked on collection of 700 product Web pages from both Amazon.com and Circuitcity.com. Notation used is represented by lv_{ij} which is normalized Levenshtein distance between two strings i and j and calculated as:

$$\text{lv}(x, y) = \frac{\text{alv}(x, y)}{\max(\text{length}(x), \text{length}(y))}.$$

Where normalized Levenshtein distance lv(x,y) is a function of absolute Levenstein distance. Function calcCosineSim (a,b) calculated the cosine similarity between two different products a and b. For evaluation of category mapping algorithm, author collected the product taxonomies of Amazon, ODP, and Overstocks contains total 2575, 44,181 and 1052 categories respectively. Results were computed by comparing the mappings obtained by the algorithm with the manual mappings. Optimal set of threshold were applied by using a hill-climbing procedure for each algorithm and each data sets. Tool were created which is used to automatically annotate Amazon.com and Cicuitcity.com Web pages. Product names identification, category mapping and multi-faceted search interface were main components for solution.

Yannis Tzitzikas [3] proposed a survey on facet exploration ofexploration of RDF/S datasets, with main spotlight on session-based interaction artifices for exploratory search. This study defined a precise and concise model that captured the essentials of RDF browsing approaches for recall-oriented information needs, allows accessing resources in group and is applicable to O-O information spaces (RDF/S). This model is suitable for comparing the existing approaches and can be used as a guide for designing, implementing and evaluating new systems, APIs or protocols. This paper described more than 30 systems in total (11 of them for single entity and 21 for multi entity type datasets like RDF/S datasets) according to the introduced aspects and the core model. Finally, the paper focused on HCI and IR methods for evaluating exploration approaches and discussed what kind of evaluation results are usually reported in the literature.

Anusree Radhakrishnan [4] proposed architecture of facet miner as shown in Fig 1. In this, similarity between the data

items is selected using the concept of cosine similarity. K means algorithm was used for clustering. Query was considered as XML data. An XML parsing mechanism is used for the type conversion. Similar data are grouped together for forming the clusters. In cosine similarity method, data are highly correlated if their cosine value variance is 0. Dissimilar if their cosine value is 1. Depending on our condition they go for the similar data. After clustering each of the clusters form the facets. Final results need to be arranged in a priority wise manner. The priority is assigned from the number of frequent access of the query. Two levels of scoring were needed, one is assigning priority for the data items within the facets and another is assigning priority in between the facets. Priority wise list was the final result. Utility mining was integrated to improve the searching. It adapts a cosine similarity method for finding the similarity the data items. Unique website model and context similarity models was the two proposed model in this study. Utility mining is a subset of frequent pattern mining to find the frequent pattern in the transaction database. FP growth algorithm can be used for mining the frequent items which incorporates concept of important scoring.

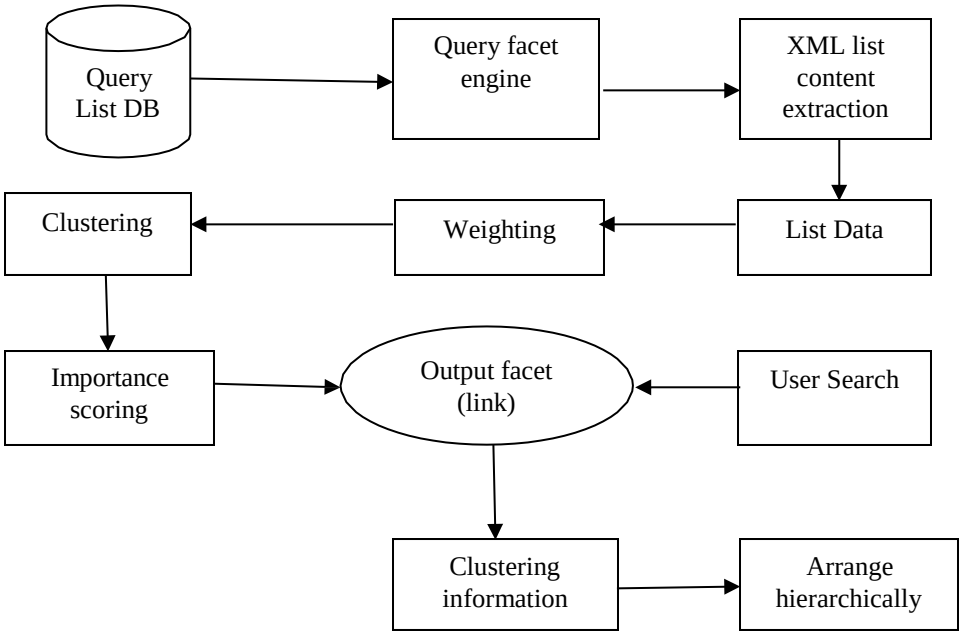

Fig 1: Architecture of facet miner algorithm

Anthony C. Robinson [5] has focused on multivariate faceted search through the application of a brute force computational process which has spatially-interesting results. General method for partitioning, combining and computing data to identify spatially-enhanced facet combinations comprises of two steps. Firstly, this paper presented their method to a point dataset evaluating the locations and engagement patterns across 35,000 students taking a Massive Open Online Course (MOOC). Secondly, they applied proposed method to a multivariate dataset describing restaurant food safety violations in New York City. This paper work contributes in several ways of the science of spatial analysis. Firstly, loading multivariate point data, classifying the point data and adding class information as a new field, aggregating point data using a grid or other boundary layer, counting how many points fall class information as a new field, aggregating point data using a grid or other boundary layer, counting how many points fall general workflow steps for identifying spatial facet combinations.

Secondly, they contributed a faceted search approach that builds upon an easy-to-understand evaluation workflow. Subsequently, method itself, which leans on a brute force approach, has been applicable to two case studies. Consequently, this paper presented how to easily integrate into a geovisual analytics system via a simple user interface.

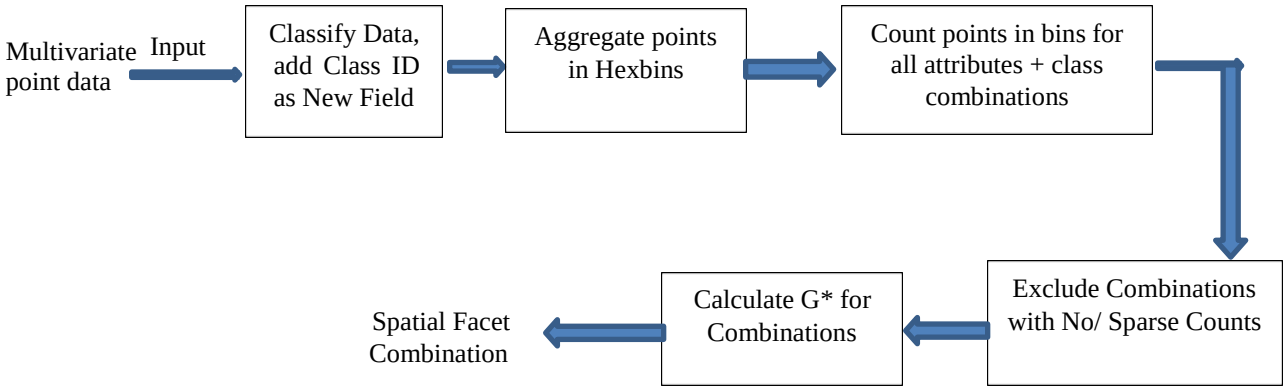

Fig.2 A general workflow for calculating spatial facet combinations

Serwah Sabetghadam [6] proposed a faceted approach to reachability analysis of multimodal graph modeled collections. This paper proposed Astera Information model to represent information objects from different modularity and their relationship to compute similarity. In this model, facet was characterized as an inherent criterion or property of an information object, otherwise depict as a resemblance of the information object. Each object in this graph may have a number of facets with the aim to influence cognitive and functional representations of information objects to make better IR results. It defined four categories of relations mediator objects in the graph named as Semantic, Part-of, Similarity, and Facet. It incorporated the following sequential parts. Firstly, it contains reachability analysis of relevant objects from different topics. Secondly, it shows the effect of enhancing semantic and similarity links in accelerating the reachability in graph. This paper followed an approach to decompose the query into a list of facets of distinct modality. For this, the function of relevance score value (RSV). The RSV value is represented by (1):

$$\text{RSV}(q, v) = \sum_{i=1}^{l} ((, vf_i)).wf_i \quad , (1)$$

Where sim is the similarity function between the two facets, norm is the normalization function, and w_{fi} is the weight of facet f_i for this query. The experiments were prosecuted on the Image CLEF 2011 Wikipedia collection having 400,000) documents and images with 50 topics. The topics were distributed into four distinct categories named as easy (17 topics), medium (10 topics), hard (16 topics), very hard (7 topics).

Xi Niu, [7] proposed a user real-time interactions by the help of facets search from both data science and human factor viewpoints. This paper used Random Forest Prediction model to predict facets using search dynamic variables which outperforms the K-Nearest Neighbors and regression models in terms of accuracy and specificity. Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outperforms the class which is the mode of the classes (classification) or means predictor (regression) of the distinctive individual trees. Random Forest model incorporated twofold study. First study was to adopt a data-mining approach to predict real-time facet use likelihood during the search dynamics from an action sequence view. The second study, they conducted a user study to better understand the "behind-the-scene" search context and added additional insights to the results from mining the logs. Random decision forest technique correct for decision trees habit of overfitting to their training set. Query formulation, facet operation, result manipulation and item examination are four types of level actions conducted by users. To capture the heuristics of query specificity, a measure called query entropy, developed to represent the specificity of the query. Calculation equation for query entropy is represented by (1):

$$\text{Query Entropy}(q) = -\sum_{i=1}^{k} p(c_i|q) log_2\, p(c_i|q) \;, \quad (1)$$

where c_i is a clicked result for the query q, and p (c_i|q) is the probability that c_i is clicked query q. This study chose an academic library from a research university for data collection where faceted search is powered by the Endeca search platform through several revisions. They have conducted the experiments with library Apache server logs data set of 1,556,707 useful records collected during six month. A detailed view of understanding human behavior perspectives were shown in this paper.

III.COMPARISON STUDY

This information system provides the right information efficiently to the users. Different methods and technique related to this have been studied in the previous section. In this section, a detailed comparative study of those methods has been presented and is tabulated in Table.

Table 1: Comparative study of Different Facet Search Algorithms

YYYY	Author	Methodologies	Pros	Cons	Outcome
2010	Ying Liu[1]	Semantically annotated multi-facet-product family ontology using document profile model.	Provided way how user can derive new variants of product on the basis of designer's query requirements via faceted search. Good decision support.	Performance degrades with complex texts e.g. test reports involving different branded products.	Scalable document profile (DP) model that suggested good design analysis by using semantic tags.
2012	Flavius Frasincar[2]	XplorePrdoducts. com platform for multifaceted product search using Semantic Web tech.	Processing of RDF annotated HTML Pages and Aggregation of product information coming from different stores is easy. Deals with the issue of Heterogeneous information.	Antonyms, Synonyms and product names of different lengths.	Provided solution for issues related to Web- wide information aggregation and parametric product search. Achieved an accuracy, recall, and specificity all are above 91% for a dataset with 603 Product names.
2016	Yannis Tzitzikas [3]	Session-based interaction schema model for RDF/S datasets.	Identified different kinds of information needs, information structures and configuration requirements for RDF/S data. Model was suitable for comparing the existing approaches as well as it can also be used as guide for designing, implementing and evaluating new systems, APIs or protocols.	Scalability issues arises when deal with exploration of big data. It cannot be specifically designed for exploratory search systems due to uncertainty, provenance issues and trust.	Efficient and effective framework. Generalized exploration / browsing approaches used Precise model comprising states and transitions.
2017	Anushree Radhakrishnn [4]	Query Facet Engine for easier search Results.	Facet eliminated multi linking and multi-page search issue. Avoided the usual web search problems.	Difficulty arises when to predict K-Value due to use of K-means Algorithm. Didn't work well with global cluster.	Efficient tool for fetching the facets based on the user query. Searching query performance improved.
2017	Anthony C.Robinson [5]	Brute Force Method for spatially- enhanced facet analysis.	Made possible to easily integrate into a geovisual analytics system via a simple user interface. Solved simple Problem of identifying whose synthesis hold observations to explore and which combination had potential pattern of spatial significance.	Must require to select a classification scheme to apply to each variable. Generalized level required in order to support combinatorial analysis.	Approached analytical utility to faceted search. Could play an important role on datasets that had been simplified by some other data mining approaches.
2018	Serwah Sabetghadam [6]	Reachability analysis of multimodal graph using Astera model.	Graph-based model for multimodal IR. Achieved better recall of Combination of Document and image textual facets as compare to document textual and image visual facets.	Weighted of different facets based on model analysis of their best combination. Recall for easy and medium topics were mainly reachable in initial steps only.	Recall increased around 266% for hard topics and around 373% for very hard topics. Achieved a recall increase around 10% by adding real semantic links.
2019	Xi Niu[7]	1.Random Forest 2.K-Nearest Neighbours(KNN) 3.Logistic Regression	Searched Dynamic Variables. Accuracy around 0.7804 achieved and specificity is more. More sensitivity & moderate specificity.	Less sensitivity. Moderate Accuracy. Less Accurate, Sensitive & specific.	Query entropy provided strong predictive power in facet addition. Mean and median are smaller than general search engine, such as Microsoft Bing.

A dedicate look at the convenient literature indicates the following issues which need to be inscribed in future research.

- The Navigational search system works well in a situation where the size of search result set is intractable.
- Search time to be minimized and enhanced accuracy is a concern.
- To build a faceted search system containing should be more robust and integrated with different facets and filters so as to retrieve relevant faster results in less time.
- The collection of data is a very challenging task.
- Researcher can work in this direction to build a novel facet search algorithm.

IV. CONCLUSION

Information Retrieval of a user search query in effective time is a task of great importance. Faceted search system has many applications in several domains like a commercial search engine, E-commerce websites and digital library catalogs etc. Some authors proposed a predictive model for search dynamic variable while some authors focused on reachability analysis for different level of topics but scalability and accuracy are the main concern by the authors. In this paper, a considerable survey on various faceted search methods and algorithms has been carried out such as RANDOM FOREST, ASTERA, RDF, LSTM etc. Each of these methods having its pros and cons has been tabulated which can be used to carry out information retrieval efficiently by narrowing down the search results using data mining techniques.

REFERENCES

[1] Ying Liu, Soon Chong Johnson Lim and Wing Bun Lee, "Multi-Facet product information search and retrieval using semantically annotated product family ontology", doi:10.1016/j.ipm.2009.09.001, pp 479-493, 2010.

[2] Flavius Frasincar, Damir Vandic and Jan-Willem van Dam, "Facet product search powered by Semantic web", doi:10.1016/j.dss.2012.02.010, pp 425-437, 2012.

[3] Yannis Tzitzikas, Nikos Manolis and Panagiotis Papadakos, "Faceted exploration of RDF/S datasets", doi.10.1007/s10844-016-0413-8, J Intell Inf Sys (2017), pp 157-171, 2017.

[4] Anusree Radhakrishnan, "Query facet Engine for easier search Results", International conference on circuits power and computing technologies (ICCPCT)

[5] Anthony C. Robinson and Sterling D. Quinn, "A brute force method for spatially-enhanced multivariate facet analysis",doi.org/10/1016/j.compenurbsys, 2017.

[6] Andreas Rauber and Serwah Sabetghadam, "A faceted approach to reachability analysis of graph modelled collections", International journal of Multimedia Information Retrieval (2018)

[7] Xiangyu Fan and Xi Niu, ACM Transactions on Information Systems, "Understanding Faceted search from data science and human factor Perspectives", Vol. 37 No.2, Article 14, January 19

[8] Siji Mol K Sijimol, International journal for scientific Research and Development, "A survey on Faceted Product Search Engines", Vol. 6 2321-0613.

[9] Lan Huang, "A distributed Multi-Facet search engine of microblogs based on SolrCloud", American journal of software engineering, vol.5 no.1 20-26, 2017.

[10] Daniel Sonntag, "Integrated Decision support by combining textual information, faceted search and information visualization", 2017 IEEE 30th International Symposium on Computer-Based Medical Systems.

[11] Hak-Jin Kim, Yongjun Zhu, Wooju Kim and Taimao Sun, "Dynamic faceted navigation in decision making using Semantic Web technology", http://dx.doi.org/10.1016/j/dss/2014/01.010.

[12] Ales Bosnjak and Vili Podgorelec, "Upgrade of a current research information system with ontologically supported semantic search engine", http://dx.doi.org/10/1016j/eswa.2016.09.01.

[13] Rajvardhan Patil, Zheng Xin Chen and Yong Shi, "A perspective from Optimization", International conferences on Web Intelligence and Intelligent Agent Technology, DOI 10.1109/WI-IAT.2012.188.

[14] Leo Breiman, "Using and Understanding Random Forests", Statistics Department, Vol. 3 No.1, 2002.

[15] Paul Hugh Cleverley and Simon Burnett, "A data driven information needs model for faceted search", Information Science 41, pp 97-113, 2015.

[16] Hak-Jin Kim, Youngjun Zhu, Wooju Kim and Taimao Sun, "Dynamic faceted navigation in decision making using semantic web technology", Decision Support System. 61 pp 59-68, 2014.

[17] Xi Niu, Tao Zhang and Hsin-liang Chen, "Study of user search activities with two discovery tools at an academic library", Internation Journal Humanities and Computer Interaction, pp 422-433, 2014.

AUTHOR'S PROFILE

Yogesh is a student of M.Tech. in Information Technology Engineering department of J.C. Bose University of Science and Technology, YMCA, Faridabad, Haryana, India. (Formerly YMCA University of Science & Technology, Faridabad), Haryana, India. He has completed his B.Tech (Information Tech. Engineering) in 2018 from YMCA University of Science YMCA, Faridabad, Haryana, India.
Email: yogeshymca7@gmail.com, Mobile No. 9582515851.

Shalu is a student of M.Tech. in Information Technology Engineering department of J.C. Bose University of Science and Technology, YMCA, Faridabad, Haryana, India. (Formerly YMCA University of Science & Technology,Faridabad), Haryana, India. She has completed her B.Tech (Information Technology Engineering) in 2017 from YMCA University of Science & Technology, Faridabad, Haryana, India.
Email: shalu4325@gmail.com, Mobile No. 9654881808

Dr. Komal Kumar Bhatia has a work experience of 16 year and currently working as a Professor & Chairman in Computer engineering department of J.C. Bose University of Science and Technology, YMCA, Faridabad (Formerly YMCA University of Science & Technology, Faridabad).
He has received the B.E., M.Tech and Ph.D. degrees in Computer Science Engineering with Hons. from Maharishi Dayanand University in 2001, 2004 and 2009 respectively. He has successfully guided 8 Ph.D. and is guiding 5 Ph.D. scholars. He has guided more than 25 M.Tech. Dissertations. He has published more than 100 research papers in reputed journals and conferences and his areas of interest are Information Retrieval and Web Mining. Currently he is handling additional charges of Dean (Faculty of Informatics and Computing) of J.C. Bose University of Science and Technology, YMCA, Faridabad (Formerly YMCA University of Science & Technology, Faridabad).
Email: komal_bhatia1@rediffmail.com, Mobile No. 9953537670

Dr. Neelam Duhan has a work experience of 15 years and currently working as a Professor in Computer Engineering Department of J.C. Bose University of Science and Technology, YMCA, Faridabad (Formerly YMCA University of Science & Technology, Faridabad).She has served as Associate Professor for three years at YMCA University of Science & Technology, Faridabad. She received her B.Tech in Computer Science and Engineering from Kurukshetra University, Kurukshetra and M.Tech in Computer Engineering from Maharishi Dayanand University, Rohtak. She completed her Ph.D. in Computer Engineering in 2011 from Maharishi Dayanand University, Rohtak. She has successfully guided one Ph.D. and is guiding four Ph.D. scholars. She has guided more than 25 M.Tech Dissertations. She has published more than 50 research papers in reputed journals and conferences and her areas of interest are database, information retrieval and data mining. Currently she is handling additional charges of TEQIP Nodal Officer (Academics) and Nodal Officer Digital India at University level. Email: neelam.duhan@gmail.com, Mobile No. 9818462006

