CHAPTER 2
REVIEW OF LITERATURE

2.1 INTRODUCTION

Page Ranking in Information Retrievals has a wide spectrum of applications and have been motivating researchers worldwide in this area. Boolean systems were developed 30 years back in minimal computing power and have syntactical restrictions on query. This limited retrieved document without ranks in relationship to a user's query is perplexed.

In spite of their powerful on-line search, they tend to provide poor services to end-users (Cleverdon 1983). The end-users lacked training to get consistently good results from the system mainly due to its complex query syntax. The ranking approach is oriented towards these end-users allowing them to input simple queries and retrieve a list of relevance ranked documents. Queries rank results based on the co-occurrence of query terms, as modified by statistical term-weighting. This chapter describes previous works on SNS research, content-based retrieval and several other aspects.

2.2 ONLINE SOCIAL NETWORK SITES

The internet, an information repository is a major part of people’s information flows and a useful tool for community building as they creation of new information avenues to mobilize group members [53]. Starting from 90’s, blogs, Social Network Sites and micro blogs have allowed people to interact with the intent of sharing common interests. The authors [20] found that in Social Network Sites 80% of internet users identified themselves as a group, thus making users accomplish goals. The percentage of adult Social Network Sites users has also increased considerably, where they only mail or search in additions to SNS. The authors [20] examined in Social Network Sites
amongst the adults found that majority of SNS users were on Facebook, supported by the findings of Madden and Zickhur.

2.3 TYPES OF SOCIAL MEDIA

Social Media can be classified into six types and are detailed below. Collaborative projects allow joint content creation by end-users and the most democratic demonstration of UGC. Users can change text content leading to a better outcome [17]. Second type is Blogs, the earliest Social Media forms. Blogs are usually managed by one person with interactions from others as comments. General Motors and Sun Microsystems have personal blogs. Blogs can be used by dissatisfied or disappointed customers to protest or complain through blogs [71], thus providing damaging information online for others to note. Third are Content communities, whose main objective is media content sharing. Cisco and Google, rely on content communities to share media content with their employees and investors. Fourth type is a Social networking site, used to for brand creation, marketing research. Next Virtual game worlds are platforms replicating a 3D environment imitating real life. These manifestations of Social Media provide the highest level of social presence and media richness. Virtual world’s applications have many users around the world. Users demonstrate behaviors resembling real life environments [43] offering marketing opportunities to companies or management of internal processes in human resources.

2.4 ISSUES IN SOCIAL MEDIA WEBSITES

The SNS have many issues in addition to users who are always a cause for concern. The generic issues of SNS are detailed below.

- **Privacy Issues:** SNSs are popular especially amongst students [15]. SNSs may be exposed to various risks, in addition to fake profiles, spam and fake links which lead to attacks.
- **Productivity Issues:** Employees might spend an entire day on SNSs. Soldiers are banned from accessing MySpace [18], Canadian government prohibits employees from Facebook and a U.S legislation bans youth from accessing SNSs in schools and libraries.

- **Resource Management:** Updates may ingest very little bandwidth, but the video links require higher bandwidths.

- **Viruses and Malware:** The threat from hackers who can commit fraud or attack SNS’s, is often overlooked by organizations.

- **Online Scams:** As more and more people follow SNS’s, the danger of online scams looms over applications.

- **Clickjacking:** It is a malicious technique of deceiving users when they click on Web pages, into revealing confidential information on the web.

### 2.5 EVOLUTION OF SOCIAL NETWORK SITES

SixDegrees.com was the first SNS launched in 1997. It allowed users to create profiles and list Friends, Classmates.com allowed people to connect to education friends, but users could not create profiles or list Friends. SixDegrees was the first to combine these features and was promoted as a tool to connect and send messages to people, but closed in 2000 as it failed to sustain its business. Between 1997 and 2001, many community tools supported articulation of public profiles like AsianAvenue, MiGente and BlackPlanet. LiveJournal launched in 1999 had buddy lists. Cyworld, 1999 complemented SNS features in 2001 with diary pages, Friends lists and guest books. Ryze.com (2001) promoted user based business networks. It allowed introduction to friends. Friendster, LinkedIn, Ryze and Tribe.net believed in supporting each other with less of competition. LinkedIn became a powerful business service. In 2003, witnessed the launch of many new Social Network Sites. Google's Orkut could not sustain a U.S. user base, but was the national SNS of Brazil and grew rapidly in India.
MySpace differentiated itself from others by allowing users to personalize their pages and updating user required features regularly. Teenagers joined MySpace in 2004, because they wanted to connect with their favorite bands. News Corporation bought MySpace in July 2005. The site was implicated for sexual content prompting legal action. Chinese QQ messaging became one of the world’s largest Social Network Sites after it made friends visible with profiles. Korean Cyworld introduced homepages and buddies. Facebook began in early 2004 as a Harvard-only Social Network Sites and registered users with edu email. Later it started supporting other schools and institutions. By late 2005, it expanded the network including anyone. Facebook users could personalize profiles and tasks. The social sites were mainly organized with people in a world of networks. SNS features have introduced vibrant research contexts.

2.6 SNS NETWORKS

SNSs provide rich sources of behavioral data by automated collections or datasets, enabling analysis on large-scale patterns of links or usage [24], Golder et. al. [19] examined a dataset of 362 million messages of Facebook users. Lampe et. al. in [42] explored the relationship between profile elements leading to a reduction in transaction costs. Network visualization was analyzed in [42], while Kumar et.al. investigated the growth of Flickr and Yahoo networks. Hsu et.al., studied Friendship classification scheme while Herring et.al analyzed the role of language in Friendship and research on geographical importance in Friendship. Study on people’s motivations in joining specific communities was done in [7]. Spertus et. al., in their study identified a topology of users from community memberships and their usage.

2.7 ADVANTAGES IN SOCIAL NETWORKING

Social Network Sites are interactive collaboration platforms exploiting different tools. They have many intuitive benefits and offer a means for self-mass communication capable of reaching global audiences. Users transform their personal social networks
with connections [36]. Social media is a multiple production and distribution source where users exercise some control over the information they provide on SNS sites [23]. Users understand the power and their intellectual property rights [36]. Social media introduces pervasive changes to communications. The SNS culture is on user’s thought on technology and its effect on their lives. The Figure 2.1 below shows the Social Media Landscape.

![Social Media Landscape](image)

**Figure 2.1 Social Media Landscape**

Individuals construct an online representation of self like dating profiles and SNSs constitute a part of self-profile management and friendship. Boyd’s article which
examined frienster, is one of the earliest academic articles on SNSs. It studied on how users connect and present themselves in an extended network. Marwick [45] in his study of three networks found that users had complex strategies in profiles and playfulness varied across sites. Skog [65] found that LunarStorm status feature strongly influenced users’ behavior. Impression management was one reason for Friendster users in choosing friends. Fono and Raynes-Goldie in their examination of LiveJournal described users' understandings as a catalyst for social drama. Walther studied friend’s links impacting impression formation. Online identity construction allows users to define themselves more by placing of labels.

Facebook offers sharing of interests and users online identity, thus producing a symbolic communication. Schau and Gilly explored Facebook reasons being prone to highlighting particular aspects of their identity. Any online representation creates users display their self-concept to others [25]. These underlying concepts lay the foundation for users to construct their identities. Hoyer and MacInnis social identity theory offers that individuals evaluate brands in terms of consistency [25]. Users may not make direct brand associations, but their behaviors show consistency with brands in groups. Zhang and Daughtery [35] state that Social Network Sites experience is a platform for user comparisons and enhancing self-identity. Education Social Network Sites can be used for survival and becoming an instructional leader [10]. Leaders in touch can give necessary information and be successful [10].

Hansford and Ehrich identified and reported beneficial outcomes for the participants in their analysis like personal ideas, problem-solving and professional development. Wenger’s term, informal units, consisted of people engaged in creating and communicating. SNS’s are an ever referable repository for educators and professionals. Those who use SNS do so because they view it as an encyclopedia of knowledge that is always available [4]. SNSs support a participatory context where participants create a media-rich information base.
2.8 FILTERED INFORMATION RETRIEVALS

People rely on recommendations for everything referring to books and Web. Tapestry, one of the first recommender systems, created the expression “filtering” which was widely adopted. Recommendations may advocate interesting items. Fundamental assumption in filtering is users’ ratings on items. Collaborative Filtering (CF) techniques use preferences on items to predict a new user like or dislike inferred through existing user behaviors. User ratings can be explicit or implicit in purchases [49]. Table 2.1 below is an example of user-item ratings matrix with likes and dislikes, while Table 2.2 below is a user recommendations table.

<table>
<thead>
<tr>
<th>User</th>
<th>Like</th>
<th>Dislike</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Serials, News</td>
<td>Superman</td>
</tr>
<tr>
<td>X2</td>
<td>News, Superman</td>
<td>Spiderman</td>
</tr>
<tr>
<td>X3</td>
<td>Spiderman</td>
<td>News</td>
</tr>
<tr>
<td>X4</td>
<td>Serials</td>
<td>Spiderman</td>
</tr>
</tbody>
</table>

Table 2.2 User recommendations

<table>
<thead>
<tr>
<th>User</th>
<th>Serial</th>
<th>News</th>
<th>Spiderman</th>
<th>Superman</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Like</td>
<td>Like</td>
<td>Dislike</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>Like</td>
<td>Dislike</td>
<td>Like</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>Dislike</td>
<td>Like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>Like</td>
<td>Dislike</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.9 CHARACTERISTICS OF COMBINED FILTERING

Though E-commerce Systems with predictions attract customers, their recommendation algorithms operate in a demanding environment. These systems depend on a few characteristics for providing fast and accurate recommendations.

Data Thinness

The data thinness appears on several occasions and it is difficult to find similar items. Any item is recommended only after it is rated. New users unlikely to give recommendations in lack of rating history. The effectiveness of a recommendation system is reduced, since it relies on comparing users in pairs for predictions. To reduce this problem, Dimensionality reduction techniques like Singular Value Decomposition (SVD) [9] have been proposed. They remove insignificant users or items for reduced dimensionality in user-item matrix. Latent Semantic Indexing (LSI) is based on SVD [9]. Information discard may result in losing useful recommendation information reducing the recommendation quality [58]. Hybrid CF algorithms address thinness in information with external information on new items. Kim and Li proposed a probabilistic model by classifying predictions into groups and using a Gaussian distribution of user ratings. Model-based CF algorithms like TAN-ELR provided accurate predictions for sparse data. A few model-based CF techniques addressed the thinness problem by applying an associative retrieval framework and explore transitive associations among users using their ratings and purchase history [33].

Volume

When existing information grows exorbitantly, traditional filtering algorithms suffer in terms of computational resources. It burdens the reaction time and recommendations [44]. Techniques produce recommendations quickly with dimensionality reduction Memory active filtering algorithms like item-based Pearson correlation CF algorithm achieved satisfactory results by calculating the similarity between the pair of co-rated items of a user [59]. Model-based filtering algorithms like
clustering CF algorithm seek user recommendation within similar clusters instead of the entire data set [57].

Similarity

Similar items with different names are treated separately and do not discover latent associations. For example, movie and a film, mean the same but are computed separately, thus decreasing system performance. Construction of a thesaurus can help solving such discrepancies. The LSI method used a matrix for document association and placed closely associated terms nearer. The LSI presentation in addressing similarity was impressive but the precision was low and was only a partial solution to the problem.

Sheeps

Gray sheep users inconsistency in does not benefit information filtering, while Black sheep users eccentric tastes fails recommendations, thus making black sheep an acceptable failure. Claypool et al. provided a hybrid approach used content-based and filtering basing their prediction on a weighted average of the content-based prediction and the Collaborative Filtering prediction. They determined the optimal mix for each user.

Attacks

Some users respond negatively to others and should be discouraged. Such models for collaborative filtering were identified studied for effectiveness. Lam and Riedl in their study observed that item-based Collaborative Filtering algorithm were affected marginally by attacks than the user-based Collaborative Filtering algorithm and suggested alternative ways to evaluate and detect attacks. O’Mahony et al. solved these attack problem by analyzing a recommender system’s flexibility to malicious alarms in the customer/product rating matrix Bell and Koren used a comprehensive data normalization approach to the attacks problem and working with residual of global effects for selecting neighbors thus achieving improved performance on the Netflix data.
2.10 MEMORY-BASED COMBINED FILTERING TECHNIQUES

Memory-based Combined Filtering algorithms generate a prediction by identifying every user as part of a group with similar interests and neighbors preferences on new items. The neighborhood-based Combined Filtering algorithm uses the following steps: calculate the similarity or weight, weight between two users or two items and produces a prediction by taking the weighted average of ratings of the user or item on a certain item or user, or using a simple weighted average [59].

Computing Similarity

A critical step in memory-based collaborative filtering algorithms is similarity. In item-based filtering identifying users who have rated the items similarly before the filter is applied [59]. In a user-based filtering algorithm, the similarity between users is computed before applying item ratings. Techniques to compute similarity is detailed below.

- **Correlation-Based Similarity**: Matches between users or items is measured by computing the Pearson correlation, which linearly relates two items by measuring two variables with each other. In user-based Pearson correlation, rated items of users are summed and averaged with co-rated items of the user. Table 2.3 below is a simple example of ratings matrix.

<table>
<thead>
<tr>
<th></th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>4</td>
<td>?</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>U2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td>3</td>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>U4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U5</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

**Table 2.3 User Ratings matrix**

User Ratings on Items and Blanks indicate they are not rated by any user
In an item-based Pearson Correlation is the average rating by the users. Variations of Pearson correlations can be found in [41]. Examples of other correlation-based similarities are Pearson’s correlation, Kendall’s and Spearman’s correlation. The number of users in similarity computation is the neighborhood size of the user.

- **Vector Cosine-Based Similarity**: Documents are visualized as a vector of frequencies and by computing the cosine angle formed by the frequency vectors are be used in collaborative filtering. A desired similarity matrix can be computed. Users with multiple rating fail the vector cosine method. To address this drawback, an average is subtracted from each co-rated pair, similar to Pearson correlation, but without negative values.

**Prediction and Recommendation for Computation**

Recommendations are the base step in collaborative filtering. Neighborhood-based CF algorithm identifies a subset of nearest neighbors of an active user, chosen based on similarity and a weighted total of their ratings is used to generate predictions for the active user [22].

- **Weighted Sum of Others’ Ratings**: The weighted average of all the ratings on an item can be used to predict for the active user [56].

- **Simple Weighted Average**: In item-based prediction, a simple weighted average can be used to predict the rating [59], where summations are done for all ratings and the weight between the user rating and overall weight on the item is computed.

**Top N Recommendations**

Top-ranked items can be generated for a user’s interest. Such techniques analyze and discover relationships between users for computing recommendations modeled on association rule mining techniques.
• **User-Based Top-Recommendation Algorithms:** They identify nearest neighbors of an active user and then apply Pearson correlation or vector-space models. Similarities in users are computed as vectors in the user-item matrix and aggregated to identify the Top-N items. The algorithms have limitations in real-time performance and scalability [38].

• **Item-Based Top-N-Recommendation Algorithms:** These algorithms were developed to address the scalability issues of top N-recommendations. The algorithms compute similar items and then identify top-N-set sorted in descending similarity order for item-based Top-lists [38]. This method can produce suboptimal recommendation which was solved by Deshpande and Karypis who overcame scalability issues.

### 2.11 MODEL-BASED FILTERING TECHNIQUES

Developed algorithms recognize patterns based on training data for predictions, which can be used further for filtering tasks. Bayesian or clustering models overcome of memory-based Collaborative Filtering algorithm shortcomings. CF models can be designed from classification algorithms for categorical user ratings, while regression and numerical ratings can be found using Singular Value Decomposition (SVD) methods.

• **Simple Bayesian Filtering Algorithm:** Simple Bayesian Filtering algorithm uses a Naive Bayes (NB) strategy for predictions in filtering and predict from the highest probability of a computed class. Observed data’s probability is smoothened by a Laplace Estimator and avoiding 0 (conditional probability). Since most real-world Collaborative Filtering data are multiclass, Su and Khoshgoftaar [67] applied the simple Bayesian CF algorithm to multiclass data for CF tasks. The predictive accuracy was bad but had better scalability than the Pearson correlation-based Collaborative Filtering.
- **Filtering by Clustering:** Cluster data objects are similar within the same cluster but dissimilar to other cluster objects. They can be hierarchical, partitioning or density-based methods. MacQueen partitioning method is easy to implement with a relative efficiency. Density-based clustering methods search for dense clusters separated by noisy sparse regions. DBSCAN and OPTICS are density-based clustering methods which are well-known. Hierarchical clustering methods like BIRCH, created a hierarchical decomposition of data using a criteria with clustering as an intermediate step. The resulting clusters are used for further analysis or processing. CF models with Clustering can be applied in different ways. Sarwaret al. and O'Connor and Herlocker used clustering to partition the data into clusters and used memory-based CF algorithm for CF tasks within each cluster to make predictions. Si and Rin extended existing clustering algorithms into a Flexible Mixture Model (FMM) for CF by clustering both users and items. The model modeled clusters separately but allowed each user in multiple clusters. Their experimental results showed better accuracy. Clustering models have better scalability because they make predictions within much smaller clusters. Optimal clustering over large data sets is near to impossible and dimensionality reduction is a necessity.

- **Regression-Based Filtering Algorithms:** The methods use an approximated rating based on a regression model for predictions. Regression matrix is sparse, if a random variable represents a user’s preferences. Canny proposed a sparse factor analysis using a regression model where missing elements are replaced with default values. Vucetic and Obradovic [70] proposed a regression-based approach to Collaborative Filtering addressing sparsity. Lemire and Maclachlan proposed slope one algorithms for faster CF based predictions.

- **MDP-Based Filtering Algorithms:** Shani et al. used a Markov decision processes (MDPs) model for recommender systems. After starting with an initial policy and then updating the policy based on the previous policymaking the iterations converge to an optimal policy like value function approximation, optimization and sampling.
2.12 HYBRID COMBINED OR COLLABORATIVE FILTERING TECHNIQUES

Hybrid Collaborative Filtering systems combine filtering with content-based systems in textual content information like messages, logs and user preferences. Textual contents have many important elements like words or similarity between items [73]. They use classification algorithms to make recommendations. Demographical recommender systems use user profile information. Content-boosted CF algorithms use Naïve Bayes as the content classifier. It then fills in the missing values of the rating matrix with the predictions of the content predictor. Content-boosted Collaborative Filtering recommenders have an improved predictor performance over pure content-based and memory-based Collaborative Filtering algorithms. Table 2.4 below illustrates Content-boosted Collaborative Filtering.

Table 2.4 Content-boosted CF and its variations

<table>
<thead>
<tr>
<th>Content information</th>
<th>Rating matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Sex</td>
</tr>
<tr>
<td>U_1</td>
<td>32</td>
</tr>
<tr>
<td>U_2</td>
<td>27</td>
</tr>
<tr>
<td>U_3</td>
<td>24</td>
</tr>
<tr>
<td>U_4</td>
<td>50</td>
</tr>
<tr>
<td>U_5</td>
<td>28</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Pseudo rating data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Pearson-CF prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

(c)
(a) Content / sparse data rating
(b) Pseudo rating by content predictor 333
(c) Predictions from (weighted) Pearson CF on the pseudo rated data.

Ansari et al. [3] proposed a Bayesian preference model to statistically integrate disparate information for recommendations.

- **CF combined Recommender Systems**: A mixed recommender combines results from different recommendations techniques like adjustable weights, majority weighted voting [16] and average weighted voting [66]. However hybrid recommender relies on external information and increase implementation complexity.

- **Combining CF Algorithms**: Hybrid approaches combining memory-based and model-based CF approaches generally perform better like Probabilistic Memory-based Collaborative Filtering (PMCF) which uses a mixture model on a set of stored user profiles, interior user ratings distribution to make predictions. CF or model based-correlations using Naïve Bayes have lesser accuracy than PMCF. Personality diagnosis (PD) combines memory-based and model-based CF retaining advantages of both algorithmic types [51]. The active user is generated by choosing users randomly and adding Gaussian noise to their ratings. It can also be regarded as a clustering method with a user for each cluster making better predictions.

2.13 INFORMATION FILTERING IN DATA MINING

Data mining is the process of extracting patterns from data and involves the following classes of tasks:

- **Regression** – Attempting to find a function which models the data with minimal error. It takes a numerical dataset and develops a mathematical formula that fits the data.
• **Classification** – Arranging data into predefined groups. Classification is one of the major data mining tasks. The goal of classification is to predict difference of objects. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification and neural networks.

• **Clustering** – It is a similar group classification without predefined groups. Clustering can be considered the most important unsupervised learning problem.

### 2.14 INFORMATION FILTERING APPLICATIONS ON SNS

Varying quality of information is disseminated on the internet, since anyone can easily publish information on the internet. Internet also has vast amounts of trash. Filtering this information can be removal of data from an incoming stream or extracting only non-trash information. Schools and Universities are selective in their online information disseminated to their students. Political organizations are selective in choosing and distributing information online. The advantages of such filtering that it is executed at the background and unwanted messages or downloads are hidden from the client. Social Networks [13] defined as a network of interactions, where the nodes consist of actors and edges contain relationships between these actors. Information retrieval is viewing relevant information after ignoring any irrelevant data. Information Filtering is a tool in getting the required information. Information retrieval has been characterized in a variety of ways. On the Internet, almost anyone can easily publish at a very low cost with a varying quality that is disseminated. It has interesting things combined with garbage information and retrieving valuable information by avoid trash can enhance a network’s value. People download millions of messages and web documents. None filter of information can lead to unwanted information and the gains of better filtering are enormous. Filters are also used to organize and structure information. Filtering implies removal of data not finding of data in a stream. This section discusses some of the applications of Information Filtering/Retrieval Algorithms in the field of Social Networks. The basic architecture of an Information Retrieval system is depicted in Figure 2.2 below.
AdaRank - A Boosting Algorithm for Information Retrieval

Xu and Li, in their method described a framework to rank retrieved documents. For a given query, the system returns a list of documents corresponding to relevance scores, calculated using a function. In learning (training), a number of queries and their corresponding retrieved documents are given. Further, the relevance levels of the documents with respect to the queries are also provided. The relevance levels are represented as ranks.

Suppose that $y = (r_1, r_2, \ldots, r_l)$ is a set of ranks, where $i$ denotes the number of ranks. There exists a total order between the ranks $r_i > r_{i+1} > \ldots > r_1$ where ‘$>$’ denotes a preference relationship. In a training set of queries $Q = [q_1, q_2, \ldots, q_m]$ where each query $q_i$ is associated with a list of retrieved documents $d_i = [d_{i1}, d_{i2}, \ldots, d_{in(q_i)}]$ and a list of labels $y_i = [y_{i1}, y_{i2}, \ldots, y_{im(q_i)}]$ where $n(q_i)$ denotes the sizes of lists $d_i$ and $y_i$, $d_{ij}$ denotes the $j^{th}$ document in $d_i$, and $y_{ij} \in Y$ denotes the rank of document $d_{ij}$.
Figure 2.3 Queries with Number of Document pairs

The objective of learning is to create a ranking function \( f: Q \rightarrow Y \) such that, for each query, the elements in its corresponding document are assigned relevance scores using the function and then ranked on these scores. Queries are clustered into different groups based on their associated documents pairs. Figure 2.3 above shows queries with number of document pairs.

A New Information Filtering Method for WebPages

Lopez and Josep [61] based their method on the use of Document Object Model (DOM) in representing the WebPages, assuming the document object as a tree and labeling every single web page element. The tree \( T = (V, E) \) is a tree where edges \( E \) are connected to vertices \( V \). Consider, \( \lambda(n) \rightarrow \) the label of a DOM node \( n \) then the \( \text{root}(t) \rightarrow \) the root of a DOM tree \( t \) and a webpage is a pair \( (u, t) \) where \( u \) is an URL and \( t \) is a DOM tree. It allowed specification of complex queries with several words and metadata to produce combinations of texts that force a particular order of words. An important
contribution was using tolerance as a measure of semantic relation. It uses syntax
distances to approximate semantic relations for information filtering. The technique
works online and extracts information from websites.

Application of Genetic Algorithm in Online Information Retrieval

- **Chromosome Representation:** GA based online information retrievals [8] is
  based on the vector space model, where both documents (vector of terms) and
  queries (query terms) are represented as vectors. A document vector (Doc) with a
  keywords and a query vector with m query terms can be represented as Doc =
  \{term_1, term_2, term_3, \ldots, term_n\} and Query = \{qterm_1, qterm_2, qterm_3,
  \ldots, qterm_m\}. Using binary term vector each term (or qterm) will be either 0
  or 1. Term is set to zero when term is not presented in a document and set to one
  when presented in document.

For example, user enters a query into the system that could retrieve 5 documents.

\[
\text{Doc}_1 = \{\text{RDBMS, Java, Web Technology, Cloud Computing, DBMS}\}
\]
\[
\text{Doc}_2 = \{\text{AI, Internet, Visual Basic, Natural Language Processing}\}
\]
\[
\text{Doc}_3 = \{\text{C Programming, Expert System, Dot Net, Multimedia}\}
\]
\[
\text{Doc}_4 = \{\text{Fuzzy Logic, Neural Network, Cloud Computing}\}
\]
\[
\text{Doc}_5 = \{\text{Object-Oriented, DBMS, Java, Visual Basic}\}
\]

All keywords can be arranged in the ascending order as AI, Cloud Computing, C
Programming, DBMS, Dot Net, Expert System, Fuzzy Logic, Internet, Java,
Multimedia, Natural Language Processing, Neural Network, Object-Oriented, RDBMS,
Visual Basic, Web Technology.

2.15 RANKING ALGORITHMS

Boolean systems were developed when computing power was at its least
compared with today and required restrictions syntactically in queries to retrieve lesser
number of documents. Thus the documents were not ranked in any relationship to the user's query.

**PageRanking**

The PageRank algorithms based on random surfer model namely FolkRank and SocialSimRank have been proposed by Larry Page et al [50]. It is mechanically using link structures of a page to rank it objectively. PageRank is a numerical value representing a web page’s importance. The link analysis algorithm measures each page by assigning a numerical weight to each page based on relative importance. It calculates quality ranking for each web page and computed offline for each page without depending on search queries. A page can have a high Page Rank if many pages point to it, or some pages having a high Page Rank point to it [5].

PageRank is usually a numeric value between 0 and 10. A page with PageRank 10 is considered to be the most important and a PageRank 0 is considered to be the least important.

**Weighted PageRanking**

WPR Algorithm [72] assigns larger rank values to important pages instead of dividing the rank value of a page evenly amongst outgoing linked pages. The importance assigned to out pages in terms of weight to both incoming and outgoing links. WPR performs better than the conventional Page Rank algorithm by returning more relevant pages for a query.

**HITS Algorithm**

Hyperlink Induced Topic Search (HITS) developed by the author in [39]. This algorithm uses hubs and authorities to define a recursive relationship between web pages. The idea behind the algorithm stemmed from a particular insight into the creation of web pages during Internet’s formative years i.e. certain web pages, known as hubs,
served as a large directory used in compilations of a broad catalog of information for leading users directly to authoritative information, where the hub was a page and an authority, the link to different hubs [14]. HITS algorithm, initially retrieves the most relevant pages a search called the root set, obtained by taking the top n pages returned in a text-based search algorithm.

A base set is generated by augmenting the root set with pages linked from the root. HITS calculation is performed on the focused subgraph formed from links. According to Kleinberg [39], the reason for constructing a base set is to ensure that most of the strong authorities are included. Authority and hub values are defined in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page, while a hub value is the sum of the scaled authority values of the pages it points to.

2.16 GROUPING OF SNS DATA

Many algorithms detect overlapping communities in complex networks like GaoCD [12], CONGA, CPM and GA-Net+. Cai et. al. algorithm is genetic and developed for detecting overlapping community using link clustering. It first finds the linked communities using an objective function and partitions based on density D [1]. It then maps the link communities to node communities based on a genotype representation method and the community determination is automatic. Other algorithms for overlapping community detection are Sequential Algorithm for Fast Clique Percolation and Lancichinetti’s algorithm which need prior information and suffer on efficiency. Information retrieval takes a users links into consideration [54]. Tags play an important role in constructing user’s profiles and for classifying likes and dislikes. Groupings can also be done using mutual awareness and comments or likes. Retrieval of the likeminded from textual posts and then group them is a better option, since many works investigate social relations between users from text [47]. The Latent Dirichlet Allocation and probabilistic Latent Semantic Analysis are implemented for regrouping tweets [64]. Similarly, Latent Allocation was applied to identify user interactions based
discussed subjects and then create communities. It built a model from tweets to discover users’ likes or dislikes, while Wikipedia was also used to discover topics of interest from user tweets. Social media networks in Twitter are formed around a wide range of terms like companies, technology, entertainment, and more. These connections are visible. A special characteristic of Facebook.com is that users can join networks that represent schools, institutions and geographical regions. Crawling regional networks allow researchers to cover a large fraction of a regional network’s users and social links among them (Wilson et al, 2009).

**GROUPING ALGORITHM**

Clustering is classification without predefined groups but groups similar. Analysis of social networking sites is dependent on clustering algorithms. Clustering, partitions data points into a small groups. Existing clustering algorithms are K-means, Fuzzy C-Means (FCM), CLERA, PAM and CLERANS. Among the existing clustering algorithms, K-mean is very fast algorithm.

**K-means Algorithm**

Lloyd's algorithm or K-means algorithm is used to solve K-means clustering problem. The K-means algorithm is an evolutionary algorithm. It aims to find the positions $\mu_i$, $i=1... k$ of the clusters that minimize the distance from data points to the cluster. This algorithm is used to classify or to group objects based on attributes/features into K number of group where K is a positive number and is provided as an input parameter. Each observation is based on the observation’s proximity to the cluster mean recomputed and repeated.

1. The algorithm arbitrarily selects k points initially.
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and cluster center.
3. Each cluster center is recomputed as the average of points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge based on the implementation.

The DBSCAN algorithm

The algorithm can identify clusters in spatial data sets by looking at the local density of elements with only one input parameter. User gets a suggestion on the suitable parameter value with limited knowledge of the domain. The DBSCAN can also determine what information can be classified as outliers. Its working process is quick and measures very well with the database size. Distribution density of nodes in a database is categorized by DBSCAN into separate clusters.

Hierarchical Clustering Algorithm

Hierarchical clustering involves a predetermined ordering while creating clusters. It is a method of cluster analysis which seeks to build two types of a hierarchies.

- **Agglomerative**: A bottom up methodology where each observation has its own cluster, and pairs of clusters are merged while moving up.
- **Divisive**: A top down methodology where all observations start in one cluster, and then split recursively while moving down.

The merges and splits are determined in a greedy manner and results presented in a dendrogram.