Chapter 1
Introduction

K. R. Venugopal, K. C. Srikantaiah and Sejal Santosh Nimbhorkar

1.1 World Wide Web

The Internet [1] is a communication network of computers forming a global system. The Internet has stimulated the adaption of website technology for publishing, blogging and Web feeds. With the advent of the Internet, new types of applications for interaction such as instant messaging and Social Networking are developed [2]. The power of the Internet has lead to taking business online with the development of e-commerce Web applications (i.e. business-to-customer, business-to-business and financial services) impacting across entire industries. Coming to the point of governance, it has no centralized technology or policies for access and usage. Each individual network connected to the Internet has its own policies and access controls [2].

The interconnected computing systems across the world form the Web [3, 4]. Web is also known as a collection of resources and information interconnected *via* the Internet. With the evolution of the Internet, the major challenges were: (i) how to establish a network of users such as researchers, who wish to share their research work with other researchers to get reviews and suggestion (ii) how to represent the data to be shared and finally how to access the shared data. The establishment of the network of researchers could happen with the development of Web applications. The challenge regarding the access of data was solved with the development of Hyper

K. R. Venugopal (✉)
Bangalore University, Jnana Bharathi, Bengaluru 560056, India
e-mail: venugopalkr@gmail.com

K. C. Srikantaiah
SJB Institute of Technology, BGS Health and Education City, Uttarahalli Main Road,
Kengeri, Bengaluru 560060, India
e-mail: srikantaiahkc@gmail.com

S. Santosh Nimbhorkar
BNM Institute of Technology, Banashankari, Bengaluru 560060, India
e-mail: sej_nim@yahoo.co.in

K. R. Venugopal et al., *Web Recommendations Systems*,
https://doi.org/10.1007/978-981-15-2513-1_1

Text Transfer Protocol (http) [5] and representation of information to be shared was solved with the advent of Hypertext Mark up Language (HTML).

It has become common in everyday life to refer to the term Internet or Web, when using a Web browser to see Web pages of interest. Often the term the Internet and Web are mistaken; Web is one of the services offered by the Internet. The World Wide Web (Web) [4, 6] is collection of heterogeneous content such as Web pages (interconnected documents) containing html, XML, JSON objects and other multimedia objects referenced by Uniform Resource Identifier (URI).

Search engine is a Web tool that accepts query keywords as inputs, searches the keywords in its database and provides the pages that contain these keywords as Search Engine Result Pages (SERPs); it operates in the order: *Web Crawling, Indexing, Searching and Ranking*. Web search engines are classified into the categories: *Crawler based, Topic Directories, Hybrid Engines, Meta Engines and Specialty Search Engines*.

Exponential growth of Web information had made it a challenging task for the search engines to meet the users' requirements. Handling Web information appropriately and organizing adequately is more demanding on the Web. To get any information from Web, the user issue queries, follows some links in Web snippets, clicks on advertisements and spends some time on pages. The user reformulates his query, if he is not convinced with the clicked page information. In order to enhance the user experience, the search engine provides various kind of recommendations like queries, Web pages and images.

1.2 Web Mining

The increase in the number of Web users and size of the data generated indirectly poses a challenge in satisfying the Web user needs in terms of their preferences, security, response time, etc. Hence, it is necessary to analyse the behaviour of the Web user from their browsing history to identify the user preferences, that help in improving the business process. There is a need for Data Mining to derive business rules from user transaction data.

Data Mining is the process of extracting useful information from a large repository of data. It is also referred to as Knowledge Discovery from Data (KDD). Alternatively, it can also be viewed as an essential step in the process of knowledge discovery which consists of an iterative sequence of: (i) Data Preprocessing, (ii) Data Mining and (iii) Data Post-processing. Data Mining is an integration of various fields such as Database Technology, Statistics, Machine Learning, Information Science, Visualization, etc.

The Data Mining functionalities determine the kind of patterns that can be mined from the given data and they consist of: *Characterization and Discrimination*, *Mining Frequent Patterns, Association and Correlations*, *Classification*, *Clustering*, *Outlier Analysis*, *Evolution Analysis*.

Data Mining can be used to perform trend analysis and predict future behaviour in various applications. Hence, it allows businesses to make proactive, helps in taking

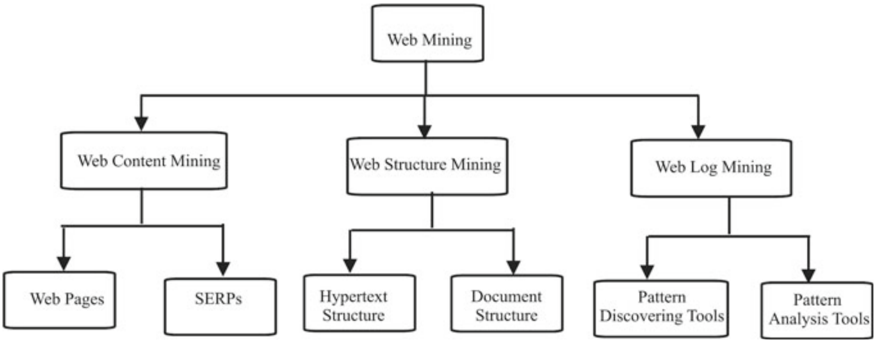

Fig. 1.1 Classification of Web mining

knowledge-driven decisions and can also answer business questions that traditionally were too time consuming to resolve, etc. It can be classified based on various criteria such as data models, types of data, applications involved, techniques utilized, etc. If the classification is based on the type of data, then we have Text Mining for text document database, Spatial Data Mining for spatial database, Time-series Mining for time-series database, Multimedia Data Mining for multimedia database, Web Mining for mining on Web data, etc.

Web mining intends to determine valuable knowledge and information from logs, Web page contents and URLs. Web mining can be broadly classified into three categories based on type of Web data: Web Content Mining, Web Usage(log) Mining and Web Structure Mining. Web Content Mining finds valuable knowledge and information from contents of Web pages. Web Usage Mining is the process of extracting information such as user behaviour from Web logs. Web Structure Mining determines information from links of Web pages.

The taxonomy of Web mining is shown in Fig. 1.1.

The main purpose of Web content mining is to analyse the Web pages in order to find information of interest or relevance embedded on the Web page. Web content mining can be broadly classified into two subdivisions. The first one is information retrieval and its purpose is to find useful information for locating relevant Web pages in a large collection (Web Searching). The refinement of the query is done by analysing the Web content [7]. The information embedded in the Web page is unstructured or semi-structured. Extracting information from the unstructured and semi-structured format poses greater challenges. The second subdivision is based on information extraction and its purpose is to find the structural information, which is saved in the database and to process it accordingly.

Web usage mining is also called as Web log mining, which is used to analyse the behaviour of online users [8]. This has led to two types of tracking; one is general access tracking and another, customize usage tracking [9]. The general access tracking is used to predict the customer behaviour on the Web. The Web log is located in three different locations such as Web server log, Web proxy server and

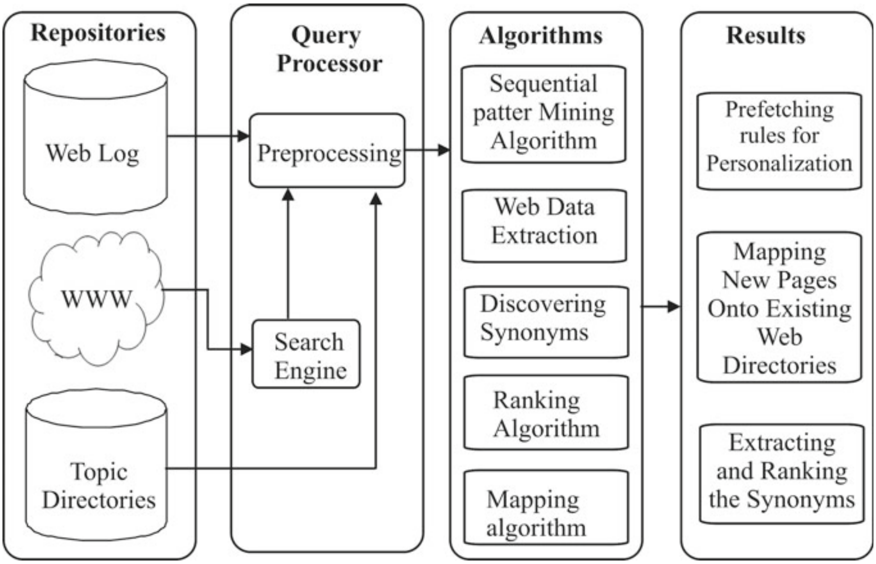

Fig. 1.2 Architecture of Web log mining and semi-structured data mining

client browser. Web log contains a large amount of irrelevant data such as noisy data, incomplete data, eroded and unnecessary information. Web server log files are used to identify the errors and failed requests, given by the webmaster or the system administrator.

Web Structure Mining is the process to analyse a Web graph, where the nodes represent Web pages and the edges represent the hyperlinks among them using graph theory. According to the type of Web structural data, Web structure mining can be divided into two categories: (i) extracting patterns from hyperlinks in the Web and (ii) mining the document structure, i.e. analysis of the DOM tree structure of the pages to describe HTML or XML tag usage.

Architecture of the proposed Web log mining and semi-structured data mining is shown in Fig. 1.2.

1.2.1 Issues in Web Mining

There are a number of research issues relevant to Web Mining and the most prominent among them are the improvement in the quality of keyword-based searches and ranking their results, effective extraction from deep Web, effective Web log mining and analysis tools, automatic construction of topic directories (such as Open Directory Project (ODP), Yahoo! directory), effectual automatic classifiers and clustering mechanisms.

Web Mining alleviates these problems by contributing solutions such as (i) mining SERPs, (ii) hyperlink analysis for ranking, (iii) automatic classification of Web documents, (iv) mining patterns from Web logs, (v) clustering of Web pages, (vi) construction of a multidimensional, multilevel Web and (vii) finding structure of the Web page. These Web Mining tasks can be used for applications like ranking, solving polysemy problem and finding user browsing patterns for customization, personalization and generation of prefetching rules.

1.3 Web Recommendations

Decades ago in a small town, everybody knew one another. Knowing each other, one can understand the interest of other and respond accordingly. The responses were usually matching the preferences of the other person. For example, when a woman visits a textile shop, the salesperson would recommend material according to the woman's preference. The salesperson understood the interests of most of the population in the town from their shopping history. So life in smalltown was all about social connections.

A few years later, the rate of interactions has gone down or the growth in the population might have made an impact. The salesperson in a book store usually can recommend new arrivals for a regular customer, but not for anybody from the town. Presently, the towns have become cities and small departmental stores have changed to supermarkets with a large number of choices. Hence, the problem arises in understanding customer preferences.

The trend changed with Amazon introducing the retail business online in the twenty-first century. If one wants to buy a book, Amazon has more than 2 million choices. Similarly, Netflix has more than 1,00,000 titles and *iTunes* has billions of tracks. With the diversity in choices, the problem is of finding the relevant item. The problem is getting worse as every minute terabytes of media are added on the Internet, hours of videos are getting uploaded on YouTube, and hundreds of books are published every hour. It gets more and more difficult to find the relevant stuff in the depth of possibilities. There is a necessity for some computational help to find items among the billions of choices. There is a need to explore methods combining the people's interest and disinterest to mine relevant items.

Recommendation technique which has branched out from data mining is not only about selling products and increasing profit; these methods of recommendation are also applied in politics to identify weaknesses to gain votes and police personnel to identify the trouble makers and terrorists.

Recommendation methods are focused on finding patterns in data. If the data size is small, anyone can become an expert to find the patterns in data using mental models. Recommendation enables the ability to handle a large quantity of information to make predictions.

1.4 Classification of Recommender System

1.4.1 Query Recommendations

Exponential gain of Web information is a challenging task for the search engines to meet the users' requirements. Handling Web information appropriately and organizing adequately is more demanding on the Web. To get any information from Web, the user issues queries, follows some links in the Web snippets, clicks on the advertisement and spends some time on pages. The user reformulates his query, if he is not convinced with the clicked page information. In order to enhance the user experience, the search engine provides various kinds of query recommendations.

Query recommendations is a technique to recommend related queries to users' input query by finding an association between queries from users' search log. It is an efficient way to enhance keyword-based search that is extensively useful to Web search systems. Users need to modify queries often because queries are informational. Users may seek discrete information on a distinct subject, hence they may check various query terms. Users may not have sufficient knowledge on a topic, and therefore adequate terms are not known to retrieve the required information.

Kato et al. [10] observed that query recommendations are frequently used when (1) an initial query is an exceptional query, (2) a single-term query is used as input query, (3) explicit queries are suggested, (4) suggestions are provided based on the modification of an input query, (5) various URLs have been clicked by users on the resulting search page. Query recommendations provided to the user efficiently can reduce the complexity of the search and help them to locate the required information more precisely. This method is extensively accepted by product, music, video search, retrieval of medical information and patient search information. Query suggestions techniques are implemented by commercial search, such as *Searches related* in Google, *Search Assist* in Yahoo!, and *Related Searches* in Bing Search.

Challenges in Query Recommendations: There are several challenges in designing query recommendations framework on the Web. First, usually users' submit short queries between one and three terms and are generally ambiguous. We observe from America On-Line (AOL) [11] data that 9.82% of Web queries contain one term, 27.31% of Web queries contain two terms and 26.99% of Web queries contain three terms. Second, in most of the cases, the users do not have sufficient knowledge of the topic that is searched, and they are not able to clearly phrase the query words. Then, users have to rephrase the query words and rephrase their queries frequently. Hence, it is necessary to solve query recommendation to satisfy users' information needs and to increase the search engine usability. Various kinds of data are used for suggestions and these data can be converted to graphs and can be used to solve many suggestions problems by designing a generic graph recommendation approach.

1.4.2 Web Page Recommendations

Internet usage has increased excessively as a result of evolution in *e*-commerce, research, *e*-banking, education, news, music, movies and electronic devices. Hence, a huge amount of information is archived and it keeps growing rapidly without any control. The decreasing costs in secondary storage have made it easy to store a huge volume of information. The valuable information is beneficial to determine interesting and useful patterns that are used by many researchers for guiding the users to visit the Web pages during their activity on the Web. This type of system is called Web Page Recommender System and helps to predict the user request.

The Web users generally spend most of the time on authoring and browsing than on search. Hence, search engines cannot efficiently predict users' search objective. Web prediction is performed on the navigation history as it is conducted in a passive manner. Hence, the prediction is performed only after the user submits his/her queries to search engines. The top *k* users are involved in a similar activity in Web prediction and it can be used in the recommendations system.

The likelihood of visiting a Web page by a user depended on the history of previously accessed Web pages is known as Web prediction. The Web users' prediction behaviour is trivial in Web mining to improve the search engine performance. The Web is configured as a graph, where each node represents a Web portal, and the edge represents the users' navigation. The user-visited Web pages distribution can be calculated and utilized in re-ranking and re-weighting results. The navigation path information is of prime significance than the user query. Storage of predicted Web pages in the cache can improve performance of search engine.

Behavioural targeting is a prime concern of predicting Web users' future behaviour. Behavioural targeting is an approach for efficiency improvement in advertising by online website advertisers and publishers by extracting knowledge of users' web-browsing practices. Behaviour targeting publishes advertisement through users' web-browsing actions. The analysis of user behaviour on the Web is the point of interest in on-line advertising and accurately targeted advertisements help in generating more consumer interest.

Web users' shopping style prediction has a crucial role in product recommendations, i.e. dynamic shopping recommendations across mobile, email and Web channels. It depends on each customer's current and past purchases practices. It also helps to improve conversions, website optimization and increase revenue by making related product recommendations to the customers.

Challenges in Web Page Recommendations: The World Wide Web (WWW) has generated immense opportunities to extract and gain massive online information. This inspires researchers to learn the navigation behaviour of Web site visitors from Web usage data to reduce access latency, Web page recommendations using efficient Web prediction technique and to improve the quality of service of that site. The Web log records the users' navigational behaviour. The preprocessing of the raw data is required before giving the data as input to prediction model. The preprocessing challenges include handling a huge amount of data, obtaining domain intelligence and session identification. Expensive training and low accuracy and are fundamental issues in prediction.

1.4.3 Image Recommendations

Billions of images are available on the Internet with the development of the World Wide Web. It is difficult for users to access and find image of their interest as the number of digital images has grown tremendously on the Web. Hence, additional processing is necessary to retrieve relevant images as per the user requirement. An image retrieval system provides an effective way to retrieve a set of images to meet the users' demand.

There are two basic image retrieval techniques: (i) Content-Based Image Retrieval (CBIR) and (ii) Annotation-Based Image Retrieval (ABIR). In the CBIR technique, images are retrieved based on texture, colour and shape features or by extracting knowledge of image rather than the metadata associated with the image such as tags, keywords or descriptions. The semantic meaning of the user query and low-level visual features of images do not match in CBIR. In CBIR techniques, search results are refined continuously by using the relevant feedback of user. This method is impractical for a very large dataset as it requires intensive computation.

Vertical search engine is used to perform domain-specific search and provides actual content (product) rather than links. Content-based and text-based search approaches are extensively used for these types of search engines to retrieve images. The text-based image search is dependent on the occurrence of input query terms either in surrounding text or metadata of images. This approach is widely used as it requires lower computation cost and provides faster response. Whereas, it fails to retrieve the images that are relevant but do not have the term in the surrounding text. Visual features of query-image are compared with images present in database in the content-based image search. It captures images that are relevant irrespective of the query-term as it performs content-based matching. This method requires higher time complexity and has a slower response time.

Sometimes, search engines fail to retrieve information as per the user wish because of various reasons: (i) improper input query (ii) lack of users' understanding about the input search query (iii) wrongly tagged images present in database. Hence, users are unable to obtain the desired output. This gap between users' search intention and understanding of the objects is called as semantic gap and is common in most of the image search engines.

Challenges in Image Recommendations: Web image search engines like Google and Yahoo! retrieve images with text-based queries. These text queries are matched with textual information such as comments, tags, surrounding text, URLs and titles along with Web images. Currently, only 10% of Web images have a meaningful description (annotation). Although the search engine retrieves images efficiently, they are able to maintain around only 42% precision and 12% recall [12]. Searches do not find relevant results on Google search for 52% of 20,000 queries [13]. This is an account of two main reasons: (i) generally, queries are short and ambiguous, e.g. the query *DM* has two different meaning *Data Mining* and *Data Mart*, and (ii) users may have different perspective for the same query, e.g. for query *apple*, users with apple product have different meaning than users who like apple fruit. Therefore, it is necessary to improve image recommendations results in order to satisfy users' needs and usability of the search engine.

References

1. https://en.wikipedia.org/wiki/Internet
2. http://oer.nios.ac.in/wiki/oer/ictapplication/internetanditsusage/internet_applications_and_services.html
3. http://www.internetsociety.org/internet/what-internet/history-internet/brief-history-internet
4. https://en.wikipedia.org/wiki/World_Wide_Web
5. https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol
6. http://www.slideshare.net/karthikanadar/world-wide-web-26195249
7. H. Han, T. Noro, T. Tokuda, An automatic web news article contents extraction system based on RSS feeds. J. Web Eng. **8**(3), 268284 (2009)
8. A. Ranade, A.R. Joshi, Techniques for understanding user usage behavior on the internet. Int. J. Comput. Appl. (09758887) **92**(7) (2014)
9. K. Bhalla, D. Prasad, Data Preparation and Pattern Discovery For Web Usage Mining
10. M.P. Kato, T. Sakai, K. Tanaka, When do people use query suggestion? A query suggestion log analysis. J. Inf. Retr. **16**(6), 725–746 (2013)
11. G. Pass, A. Chowdhury, C. Torgeson, A picture of search, in *Proceedings of First International Conference on Scalable Information Systems* (2006)
12. K. Stevenson, C. Leung, Comparative evaluation of web image search engines for multimedia applications, in *Proceedings of IEEE International Conference on Multimedia and Expo, ICME 2005* (2005), pp. 4–14
13. B. Smyth, A community-based approach to personalizing web search. IEEE J. Comput. **40**(8), 42–50 (2007)