IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 25-31
www.iosrjournals.org

Searching On the World Wide Web: A Survey

Najira Salam[1], Syamily K R[2]
[1]*Mtech Scholar, Department of Computer Science Engineering, Govt Engineering College, Mananthavady, Wayanad, Kerala, India- 670 644*
[2]*Assistant Professor (Adhoc), Department of Computer Science Engineering, Govt Engineering College, Mananthavady, Wayanad, Kerala, India- 670 644*

Abstract: *Searching for information is a crucial element in our lives. It is mostly the case that by issuing the same query, users search for various informations in a search engine. Thus it is very important to satisfy the needs of large number of users by providing a limited result set. Thus the idea of search result diversification was introduced to cover the different intents of user. Apart from providing diversified search results, it is also very important to increase the search experience of users. Query facets can be used to provide some useful knowledge about the query without browsing large number of pages. Thus it is very important to integrate the idea of providing search experience as well as diversified search results to users. This paper presents a survey on different types of search mechanisms and diversification algorithms.*

I. Introduction

A. Data Mining

An area that offers a great platform for data extraction is called as data mining. The important reason that attracted data mining everywhere comes from the fact that we have a large amount of data but does not have sufficient knowledge. Interesting relationships and regularities can be derived from the data stored in database. The retrieval of hidden knowledge from such huge data has a wider applications in online ebusi-ness [1]. Data Mining can be viewed as an algorithmic process that has data as the input and patterns such as association rules, itemsets, classification rules as the output. Data warehouses are designed to store the data from various industries. Data Mining can be also called as a knowledge discovery process. Thus data mining have applications in large number of industries like medical, aerospace etc. The large number of services provided by the Internet, Web etc are producing a vast amount of data. It is very difficult task to manage and visualize as well as to perform analysis on this large amount of data. Thus proper data mining techniques should be applied to acquire knowledge from this data and to make them useful for our lives. The five techniques of data mining are

Association: It is a simple technique to correlate two or more items.
Classification: Uses decision tree to determine the clas-sification.
Clustering: Clustering is a method of dividing data into groups of similar objects.
Prediction: Prediction method is the combination of trends, classification and relations.
Feature selection and extraction: It is an attribute reduc-tion process.

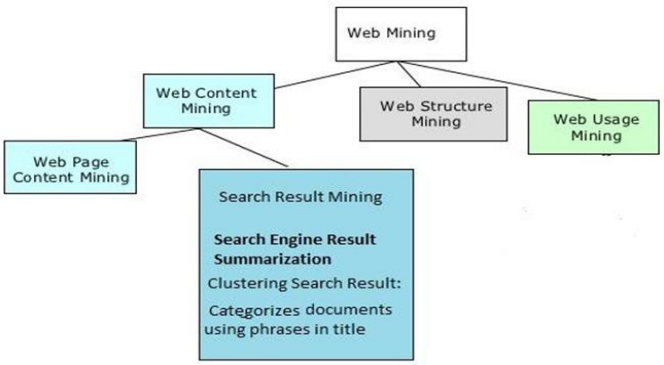

Fig. 1. Web Mining Process

B. Data Mining Technologies

Many tools have been used for data mining. Some of them are listed as below
Neural networks- The models that acquire knowledge through the process of training and looks like neural networks

Decision trees- Trees which are used to represent different decisions
Genetic Algorithms- Uses the techniques genetic algorithm, selection by nature etc.

C. Data Mining and Web Mining

Data Mining is the process of extracting useful knowledge from a vast amount of information. The data can be of different types like transactional data in applications of ecommerce. Thus data mining have a wide variety of applications. Nowa-days, data mining has been widely applied in the management of web data. The knowledge mined from different types of web data helps to create a relationship between various web object which effectively improves the web data management. Thus web mining is the use of data mining methods to extract useful knowledge from the web data. Based on which part of data to be mined, web mining can be classified into three. They are web structure mining, web content mining and web usage mining. Web content mining is the process of retrieving useful knowledge from web documents. Web content mining is also called as text mining which mines knowledge from web documents. Web usage mining extracts some interesting patterns about users search mechanism by using the information stored in web server logs. Understanding useful patterns help the researches to improve the design of the web. Web structure mining involves changing the model of web by extracting information. The different categories of web mining process is shown in Fig.1. Web mining has gained much attraction and contributes a lot to the practioners doing research in database management, information retrieval etc

II. Information Retrieval

The amount of information available in databases increases day by day. Extracting useful knowledge from such large collections to satisfy the needs of customer is called as information retrieval. This process of retrieving information is used in every activity. One of its major applications is for web users. The dependency of people on web has increased a lot nowadays. It helps the users in browsing as well as retrieving some useful documents. Thus the information retrieval community has gained much attention. Information Retrieval overtakes the traditional style of accessing the information through database style methods. Information Retrieval is mostly applied to unstructured form of data. The unstructured data mainly includes text which is usually stored in computers. Apart from unstructured data, information retrieval can be also used for semi-structured data. In semi-structured data, retrieval of information occurs by a query like finding a document that contains title as language and body as php. Information retrieval can be classified into three types based on the scale they operate. One important type is web search where the users have to gain information from a large collection of documents. The important issue in web search is how to index the documents which comes from a large set of collections. Another type of information retrieval is Personal information retrieval. This type of information retrieval is mostly used in consumer operating systems. Emails provide classification of text as well as search. Spam filters are used to block unnecessary mails and provide classification of mails into particular folders by manual or automatic process. There are several issues considering this type of search. It includes how to make the maintenance of system free and to handle different types of documents. The last type of information retrieval is domain specific search. Here the documents are collected from corporate internal database. The documents are stored in centralized systems and machines which are dedicated are used to perform the search.

A. Query reformulation and query suggestion

The needs of users vary diversely. To meet the growing needs of users in search engines, it is very essential to adopt some methods that help users to describe their needs. Query can be modified to represent exactly the needs of users. This is done through a process called query reformulation [2]. Queries can be further generated alternatively which has the same meaning of original query. This is done through query suggestion. Both these techniques help users to describe their needs in an appropriate manner.

B. Query based summarization

Algorithms which are used for summarizing a query can be classified into different types from different perspectives. They can be classified into single document or multiple documents based on the number of source of their summaries. Based on summary construction methods, algorithms can be classified into extractive or abstractive. Algorithms can be classified into generic or query based on the connection between summary and query and can be classified into indicative or informative based on the information type in the summaries [2]. Summarization techniques use sentences from documents to generate summaries. They also provide a list of sentences as the result.

C. Entity search

Nowadays, the problem of entity search has gained very much attraction. The structure of the webpage is mainly used to perform searching of entities in different searching for entities algorithms. It answers the

queries which are made for entities. The output of the search is entities, homepages and attributes associated with it.

D. Search Result Clustering

Clustering method uses the idea of query subtopics. In this method, search results are organized into clusters by subtopics. It provides an opposite approach to providing search results just in a flat manner. Different features regarding the text are gained from the input text and many clustering algorithms are applied on the text.

E. Query subtopic

The needs regarding the information which is more important to a query is called query subtopic. It can be also called as a single topic that describes a keyword or a group of words that describe the needs of a user. Query logs, anchor texts are used to mine the query subtopics.

III. Relevance Feedback

The interaction of user is very important to improve the search experience. In order to improve the result set, it is essential to make the user involvement. Based on the importance of documents, feedback is provided by the user. When a user issues a query, a set of documents are retrieved. Based on the retrieved documents, some are marked as important and unimportant by the user. The system computes the users feedback and a revised set of documents are produced as result. The process can be carried out through one or more iterations. The idea used in this approach is that by seeing certain documents, the user can filter their information needs. Image search is an example of this approach. By seeing certain images the user can identify whether it is relevant or not.

A. Relevance feedback on web

When a document satisfies a users information need, then it is marked as important by the user. Web Search engines rarely uses the method of relevance feedback. There is one exception for this which is Excite Web Search Engine. It works completely based on the relevance feedback. But due to the lack of use, the feature does not gained much attention. This is because the interface used in search engine is advanced and the users prefer to get their result as early as possible.

IV. Types Of Search

A. Keyword Search

The keyword provided by the user is accessed through an interface which contains only a single search box. The results are displayed in a manner which contains single list. Most of the currently used search engine like Google ad Bing uses this approach [4]. The keyword search method does not provide search result diversification method. It also suffers from the problem of vocabulary and does not provide navigation method.

B. Form-based Search

Compared to the keyword search method, form based search is easier and is flexible to use. It helps to do difficult searches by providing more advanced type of interface. This method is mainly used by websites which are hidden like HotelBook etc. This search approach uses the idea of keyword search in many perspectives. This method requires only a little approach of the data scheme.

C. Directory Search

A taxonomic method is used in directory search to perform navigation. But this method often leads to disorientation problem. In order to perform directory search, only a little prior knowledge of the data scheme is required. This method only supports ranking of search results.

D. Faceted Search

Faceted Search is an interactive form of search mechanism. It can be used to provide diversification of search results.

V. Important Terms In Faceted Search

A. Facet

The documents usually have different types of properties. In order to describe it, Rangaathan introduced the term Facet. Five different facets are introduced by him as Energy, space, time, personal and matter. The first classification approach was also proposed by Rangaathan. Later several approaches were proposed to modify the definition of facet. The idea of one aspect of a subject as a facet was introduced by Prietodas.

Table I Comparison Of Different Types Of Search

Items	Keyword Search	Form Search	Directory Search	Faceted Search
Search Interface	Have vocabulary problem	Provide multiple query options	Unable to adapt for different searchers	Multiple taxonomy
Prior knowledge	Required more	Required little	Required little	Requires little
Ranking function	Supports ranking	Supports ranking	Supports ranking	Supports facet And search result ranking
Navigation Function	No navigation Function	No navigation Function	Orientation problem	Supports navigation
Diversify search results	Not provide diversification	Not provide diversification	Provide diversification	Provide the best diversification

The set of terms that describe about a subject is called as a facet was introduced by Spiteer. The individual words in a facet are called as facet term or attribute. For example: the facet watch consists of its different brands, colors etc as facet terms.

B. Faceted taxonomy

A set of taxonomies that specify facet is called faceted taxonomy. Every faceted taxonomy is separate. There is no occurrence of a single word in multiple facets. Thus they are said to be orthogonal.

C. Faceted search

The searching mechanism that uses the interaction of user to perform search is called as faceted search. It provides successive refinement of search results. The user can select facets and make their search easier by narrowing the search space.

VI. Search Result Diversification

The importance of search engine has increased tremendously for the web users. Every piece of information is accessed through a searching mechanism. It is widely used but still suffers from many problems. It cannot correctly meet certain requirements of a user. One of the most important problems of search engines is vocabulary problem. It is the case that the query indexed by the search engine is different from what has been provided by the user which leads to vocabulary problem. Thus a mismatch occurs between users need and the search results displayed[3]. Another important problem is that most search engines uses one list only approach to display the search results. Search results of large number of topics are displayed as a single list. The result set displayed will be too long and causes dissatisfaction for users. Another problem is that a trial and error method is used by most search engine. This method does not have proper filtering mechanism. All of above mentioned problem causes overload of information for users. Thus a great deal of effort is needed by the user to select exactly what he/she needs. In order to address the problem of information overload, two methods are introduced. One is search result diversification and other is search result ranking. Search result ranking puts the most relevant result into the top of the result set. But the ranking approaches possess certain disadvantages. When the query is too general, too many results are produced and it is difficult to determine the most important result. The method of search result diversification makes each result varied thus avoiding redundancy. The search results are grouped into different types of categories like keywords, tags etc. Thus it helps the users to give more importance to the category to which they are most interested in. The users can ignore other categories of result. This help to avoid the problem of information overload.

Diversifying search result for a query is very important. This is because large number of duplicate results often annoys the user. Each search result for a query must be unique. When a user enters an ambiguous query or a query which is multifaceted, then comes the importance of diversifying the search results. Nowadays, the most important feature of web search engine is providing a ranked list of documents. Thus the ranked list is diversified to meet the different search intents of the user. But the method of search result diversification has a major drawback [5]. The user have no knowledge on how the results are reorganized which causes them to merely guess about it.A search engine is said to be ideal when it issues the most relevant results to a user who enters a simple query. Some queries are multifaceted and the needs of different users vary even if they issues same query. This situation is adversely affected when the query issued by the user is small or ambiguous making it difficult to understand the users intent. Consider for example a query Defender. This word has different interpretations like it can be Windows Defender or a newspaper Chicago Defender etc. When the user clearly specifies the query as Windows Defender, then also problem arises. The user may sometimes mean Defender homepage or its review etc. Thus it is very necessary to contain search results that satisfy different users intent at the top of the page to perform optimal search. To address the above issues, there are mainly two criterias; to identify the intent of the query and based on this identification, performing diversification of search results.

Many intent mining approaches have been introduced for this and diversification promotes diversity based on the intents. In order to perform diversification in the field of information retrieval many public tasks have been introduced. Their main aim is to return documents in a ranked manner by covering different intents of the query and reducing the redundancy of search results. In order to promote diversity, intentions of query can be implicitly and explicitly in the process of diversification. There are a large number of works carried out to promote diversification in an implicit manner. A Maximum Marginal Relevance algorithm was introduced by Goldsteen in [6]. The algorithm ranks the documents based on their importance and minimizes the redundancy of the result set. Based on the uniqueness of the content of the document, novelty of the re-sult is calculated. A risk reduction framework was introduced by Zhai in [7]. A negative feedback framework was proposed by Chen to increase the idea of acquiring one important document for a given query. In order to achieve diversity, the idea of ranking documents was proposed by Zheng. Agrawal introduced an explicit method to promote diversification. A greedy algorithm was proposed to increase the chance of getting at least one relevant document in the result set. The web search engines provide reformulations of the query. Based on this, search results are made diversified by Santos in [3]. An approach to diversification was proposed to understand about relevance and uniqueness. The subtopics from different data sources are aggregated by a framework proposed by Dou in [4]. The clicking behavior of web users is used to diversify the result set by Radiliski. Rafiee considers the clicks of users on a result as a sign of importance of that result. After then, many term level diversification approaches were proposed which uses subtopics to promote diversification. By the aggregation of different external sources, an algorithm was proposed in [8]. The problem of diversity was considered as a knapsack problem which has large number of subtopics and ranking of documents was made just as filling a knapsack by the researcher Ren. All of these proposed methods use different sources to develop different intents of query where each of them is a word or phrase.

A. Classification on search result diversification

Search result diversification can be classified into five different categories
Controlled vocabulary based diversification Thesauri based diversification
Clustering Method Taxonomy method
Faceted Classification

The search mechanism that uses the method of classification based on facets is called as faceted search. This mechanism is widely used in bibliographic databases, Amazon, Google Images etc. The search results are refined using facets. At earlier times, a single search button that provides ten links as search results when a query is entered is preferred by most organizations. But as the volume of data increases rapidly, the users need to get the correct information increases. The user gets frustrated when they do not correct results. Thus it is very necessary to narrow down the search results. One of the mainly used ecommerce website Amazon uses the method of faceted search to narrow the search results. It helps each user to choose their own path to reach the document they desire. The search results are classified based on the attributes of the content of the data presented to user. A large number of products are presented in ecommerce sites which makes it difficult for the user to choose the appropriate one. The faceted search helps the user to choose the product that they desire from huge collection of products. In enterprise world search classification is done based on types of contents, dates etc. These are the information that business users are looking for. Narrowing down of search results in enterprise world helps to reduce the frustration of business users because it helps to correctly sort out the information they need. They are also useful in medical sites where the users can learn about a particular disease properly. Care needs to be taken while choosing facets since search experience can be only improved through this process. Faceted search is also called as successive filtering.

VII. Search Result Diversification And Mining Facets

Most of the existing algorithms used for diversifying search results, considers different intents of a query as a word or a phrase. Thus there is a chance of misunderstanding documents. For example: when a user enters a query as ”Olympic Sports”, then the document that does not mention the word Sports will not be retrieved even if the documents says about all sports in Olympics. This has become a major limitation of existing diversification approaches. Another disadvantage is that the existing diversification algorithms cannot identify the better result set. For example: When two documents mentions about watches, then the existing algorithm cannot identify which one is better and which documents details more about watch. Thus, it is very essential to use query facets in the methods of diversification to increase the accuracy of search results. The items in different facets give us a detailed description of facets and thus they are very suitable for diversification approaches. The main attraction of the proposed diversification algorithm is that the document which covers the most faceted items is chosen with the condition as these items are not covered in documents which are selected before.

A. Advantages of Mining Facets

Different facets of a query are particularly used to improve the experience of users in a search engine. Mining the facets of query has typically many advantages. The most important is that the different aspects of the query can be displayed together with search results. Thus users can understand the important features of query without browsing large number of pages. Mining the facets of query has typically many advantages. The most important is that the different aspects of the query can be displayed together with search results. Thus users can understand the important features of query without browsing large number of pages. Mining of facets is most useful when the query entered by the user is ambiguous like apple. We can display the products related to Apple Company in one facet and those related to fruit apple in another facet. Thus query facets provide direct answers to the questions of users. Moreover, facets can be used to provide diversification of search results. Mining of facets can be also used in search using semantics of words. Query facets can be also used in reformulations and providing suggestions to the query.

B. QDMiner Approach to Automatically Mine Facets

Most existing facets mining systems works on a particular domain like product search. In order to extract facets that can be used in databases containing textual information, Dakkam proposed an unsupervised method. This method generates facets for whole databases, not for a particular query. In order to extract information from Wikipedia, a retrieval system was proposed by the researcher Li. The system extracts semantic information from databases. Compared to the previously mentioned approaches, QD-Miner extracts facets automatically based on a web search engine. The main feature of this system is that it can be applied on any domain. QD-Miner is not restricted to a particular domain and extracts facets automatically without any knowledge of the domain. QD-Miner develops facets by combining the lists which are frequent in the top results of a search engine. The basic idea behind this method is that the most important terms repeatedly occur in a sentence partitioned by commas or will preside in a table. Another idea in this method is that lists which are important occur many times in the top result set. Thus it is very easy to understand the lists which are good.

The important modules in the QD-Miner System are Extraction of lists and Context When a query is entered by a user, the top K results of the query are retrieved from Bing and all documents are combined to form a result set R. After forming the result set, each document in R is used to retrieve a set of lists based on patterns which are free text, HTML tags and patterns which have repeating regions. EX:Men's dresses, kids dresses etc is an example of list [2].

Weighing of Lists After performing extraction of lists, weight is assigned to each list. The lists which are not informative is given low weight. The list which is used in many websites or appears frequently are assigned high weights. The value of each list is calculated by document match weight and invert document frequency.

Clustering of List Each list may even contain noise. For example: dress type is a noise and cannot be used as facet. Thus the lists which are similar are grouped together to form facets. A Quality Threshold algorithm is used to perform grouping of lists.

Perform Facet and Item Ranking After performing the above three steps, ranking for facets and items is carried out. The ranking schema is based on the frequency of the occurrence of facets and the importance of corresponding documents. For example: while considering the watches of different genders, men and women are given high rank and unisex is given lower rank.

C. Search Result Diversification Based on Facets

The QD-Miner approach mines facets automatically from the top search results of a query. The mined facets cannot be used directly as subtopics to promote diveersification. The reason behind it is that the facets are much detailed than subtopics. So facets are grouped into clusters. IASelect is the famous algorithm used to promote diversity of search result. By using facets, a faceted IASelect algorithm is proposed which offers greater diversity than IASelect.

VIII. Conclusion

Every one searches for information. To satisfy the needs of any user, is an important challenge task of search engine. As the time of user is valuable, they often get depressed when their needs are not satisfied at the very moment. Improving search experience and providing diversified result set is an ultimate aim of any search engine. This paper describes about the idea of mining facets which can be used to improve the search experience. Various diversification algorithms are also discussed in this paper. The mined facets can be applied on diversification algorithms to improve the diversity of that algorithm. Thus the idea of incoprating automatically mined facets into diversification algorithm is a most thrilling task. The user can learn different aspects of the query along with acquiring the most diversified search result.

References

[1]. aHu S, Dou ZC, Wang XJ, "Search Result Diversification Based on Query Facets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA EN-GINEERING,2016.

[2]. Zhicheng Dou,"Automatically Mining Facets for Queries from Their Search Results", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2016.

[3]. Jianxlin Li, Chengfi Luii,"Context Based Diversification for Keyword Queries over XML Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2016.

[4]. Jansen B J, Spink A, Saracevic.T, "Real life, real users, and real needs: A study and analysis of user queries on the web", in Proc. of Int'l Conf. on Very Large Data Bases, pp. 487-499, 1994.

[5]. Carbonell J, Goldstein J,"The use of MMR, diversity-based reranking for reordering documents and producing summaries.", in Proc. of ACM SIGMOD Int'l Conf. on Management of Data, pp. 1-12, 2000.

[6]. Zhai C, Lafferty J., A risk minimization framework for information retrieval.", in Proc. of the Utility-Based Data Mining Workshop, pp. 90-99, 2005.

[7]. Zhang B, Li H, Liu Y, Ji L, Xi W, Fan W, Chen Z, Ma W Y., Improving web search results using affinity graph",IEEE Transactions on Knowledge and Data Engineering, Vol. 21(12), pp. 1708-1721, 2009.

[8]. Zhu Y, Lan Y, Guo J, Cheng X, Niu S.,"Learning for search result diversification."In Proc. the 37th SIGIR, July 2014, pp.293-302.