Middle-East Journal of Scientific Research 19 (8): 1047-1052, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.19.8.1506

A Survey of Information Retrieval in Web Mining

T. Nalini and G. Sangeetha

Department of Computer Science and Engineering,
Bharath University, Chennai, India

Abstract: There has been lot of research in recent years for efficient web searching. Several papers have proposed algorithm for user feedback sessions, to evaluate the performance of inferring user search goals. When the information is retrieved, the user clicks on a particular URL. Based on the click rate, ranking will be done automatically. In this paper, we generate an algorithm called "A Fuzzy Self, Constructing Algorithm" for clustering the feedback sessions. Mostly fuzzy logic is used for clustering the data sets. The proposed algorithm significantly reduces the computation time required to partition the dataset. It will reduce the original data set into simplified data set. It simplifies the data set and find relevant documents based on user feedback sessions. This will automatically iterate every time and reduce the number of iterations while speeding up the calculations and improve the run time performance.

Key words: Feed Back Sessions · Click Rate Ranking and Fuzzy Self Constructing Algorithm

INTRODUCTION

The information retrieval goal is to find the documents that are most relevant to a certain Query. The problem of information retrieval is to find the documents that are relevant to an information need for a large document. It deals with notions of Collection of documents, Query (User's information need), Notion of Relevancy. The types of information's are text, audio, video, xml structured and documents, source code, application and web services. The types of information needs are Retrospective, Prospective (Filtering). Retrospective means "searching the past". The different queries are posed against a static collection. Prospective means "Searching the future". The static queries are posted against a dynamic collection. It is time dependent. The components in information retrieval are user, process and collected. User- What computer cares about? Process and collection tend to what we care about. The information retrieval cycle consists of five phases. Source selection, query formulation, search, selection and result. The search process consists of the Index and document collection. The indexing is a Black box function; its process is not visible.

The main tasks of information retrieval are indexing the documents, process the query, evaluate similarity and find ranking and display the results. The documents are searching that are most closely matching the query. The index consists of stop word removal and stemming and inverted index. The removal of stop word usually improves the effectiveness of information retrieval. The lists of stop words are about, afterwards, according, almost, above its [1].

The stemming is based on suffix stripping. The reason for stemming is that the words that have similar meaning to each other. The stemming removes the some ending of words. E.g.: include, including, includes, included. A porter algorithm is used for suffix stripping. The results of indexing are based on some set of weighted keywords. The results of indexing are in the form of [2]:

$$D1= \{(t1, w1,), (t2, w2).....\} \quad (1)$$

Inverted file is used for retrieving the information for higher frequency.

Problems in Information Retrieval:

- How we represent the documents with selected keywords?
- How document and query representations are compared to calculate the weight?
- Mismatching of vocabularies.
- Ambiguous query.

Corresponding Author: T. Nalini, Department of Computer Science and Engineering, Bharath University, Chennai, India.

- Depicting of content may be incomplete and inadequate.

The effectiveness of information retrieval can be improved based on keywords. The keywords cover only the part of the contents. [3] User can identify the relevant/irrelevant documents based on the weight of the words. We need to be interacting with the user and getting the user feedback. The evaluation is based on recall and precision. The more information retrieval process available is open source IR tool kits.

Web Mining: The World Wide Web has been dramatically increased due to the usage of internet. The web acts as a medium where large amount of information can be obtained at low cost. The information available on the web is not only useful for individual user and also helpful to all business organizations, hospitals and some research areas. The information available in the online is unstructured data because of development technologies. Web mining can be defined as the discovery and analysis of useful information from the World Wide Web data [4]. It is one of the data mining techniques to automatically extract the information from web documents. The three issues in the WWW are web content mining, web structured mining, web usage mining. Web structure mining involves web structure documents and links. Web content mining involves text and document and structures. Web usage mining includes data from user registration and user transaction. WWW provides a rich set of data for data mining. The web is dynamic and very high dimensionality. It is very helpful to generate a new page, lot of pages are added, removed and updated anytime. Data sets available on the web can be very large and occupy ten to hundreds of terabytes, need a large farm of servers. A web page contains three forms of data, structured, unstructured and semi structured data. A number of algorithms are available to make a structured data, one such algorithm is a fuzzy self constructing. An unstructured data can be analyzed using term frequency, document frequency, document length, text proximity.

We have to improve searching on the web by adding structured documents. Using clustering techniques we have to restructure the web information. We provide a hierarchical classification of documents using web directories Eg: Google. While increasing the annual band width at ten times its average is increasing three times, because of that the traffic management is important in web mining.

Literature Review

Google Keyword Tool: Google's keyword tool is a free online research tool to help the user to find the appropriate keywords. The tool based on ranking is shaped by three participants [5]: Search engine company programmers, Webmasters and SEO practitioners, Search engine users. The Google ranking algorithm is based on "click through rate" and "bounce rate".

Click Through Rate: In a given query, how many percentage of time, the user clicks on a particular URL in a web page.

Bounce Rate: The bounce rate has a reverse effect; it calculates that how many searchers clicked on a particular webpage.

Webcap Browser Side Tool: Web cap is a browser side tool. It does not maintain log information. It can collect relevant information from different user based on the different user search goals. Web cap uses the different inference algorithm to collect the relevant documents. It involves two steps: i) collecting the relevant information based on the user's interests ii) apply inferring algorithm and different techniques to learn about the information. Use implicit indicators associate with each user behaviour's action to compute the user's degree of interest [6]. Implicit indicators such as reading actions such as reading time, mouse clicks, mouse movements and scroll bar movements and saving action and printing action.

Page Ranking Algorithm: The page rank algorithm determined the rank for single page not for the whole website. Based on the user clicking, the page rank can be determined. The original page rank algorithm is [7]

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + + PR(Tn)/C(Tn)) \quad (2)$$

Where:

PR (A) is the Page Rank of page A,
PR (Ti) is the Page Rank of pages Ti which link to page A,

C (Ti) is the number of outbound links on page Ti d is a damping factor which can be set between 0 and 1. How to increase the page rank value? The sum of the page rank is same; we can add a new page to the web site.

Increase the number of links with other inbound links. A one off website is considered as a small web, if we want to raise a page rank, add a new page to it. When you add a new page, be sure that link connected to the front page is correct.

Lexical Pattern Extraction Algorithm: In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. We need to identify the numerous semantic relations that exist between two given words. Efficient retrieval is based on words. The information retrieval is problematic when the words are same [8]. On the other hand, the snippets returned by a search engine for the conjunctive query of two words provide useful clues related to the semantic relations that exist between two words. A snippet contains a window of text selected from a document that includes the queried words. Snippets are useful for user while searching, users can learn the snippets in the without opening the URL. Using snippets is efficient it is helpful to download the documents from the web; it is time consuming when the document is large.

Construction of Virtual Documents: The virtual document terms are extracted from its detailed graph [9]. The virtual document of a concept consists of terms extracted from its description graph: For each entity identified by a URI in this graph, extract its local name and label; for each literal in this graph, extract its lexical form. For example, in Fig. 1, the virtual document of swrc:Student consists of terms "Student," "subclass Of," "Person," "type," "Restriction," "on Property," "studiesAt," "allValuesFrom," and "University." It is worthy to property names are also included so that the system can support more diverse keyword queries, e.g. swrc: Student can be retrieved by the keyword query "subclass of person."

User and Query Similarity Model: When the multiple users in the organization, the two users asking the same set of queries. The ranking functions have been derived on browser choices. The user similarity between the two users can be expressed as the average similarity between the ranking functions. The goal of the similarity is to determine the ranking functions derived from the similar query from the similar user.

When the user in the system enters the query in the browser, the results are obtained. Let be the ranking functions derived from the each individual query. The same query having different ranking functions.

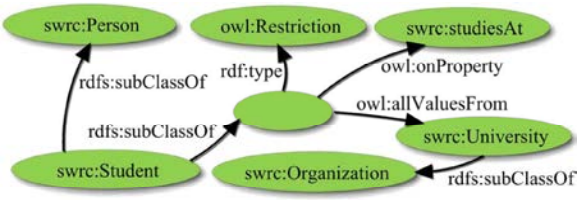

Fig. 1: RDF graph [9].

We cannot choose ranking functions randomly. In order to ascertain the correct ranking function, we use the concept called query similarity. It has two distinct approaches, query condition similarity and query result similarity [10].

Virtual Feature Updating Algorithm: VFs capture the high level semantics imposed by multiple users [11]. The semantics concepts are learned information retrieval experiences in two ways: In breadth and in depth. The virtual feature technique uses the both types of learning, to produce the desired result in real-world applications. In breadth is learned from one session in a single query. In depth is learned from multiple sessions in a multiple query. For in-breadth learning, the user can utilize the system to increase the weight of the conception in a single query session [12]. For in-depth learning, is based on previous information retrieval with multiple users, the system automatically utilizes the long-term knowledge and it is used to search the relevant images in the future [13].

Index Based Threshold Algorithm: The algorithm is based on two accesses: Sorted and random access. Sorted access retrieves the tuples information based on decreasing order of their attribute value. It maintains a buffer and allows to use a stopping condition due to the detection of final Top K tuples before processing the all tuples [2]. Random access efficiently retrieves the tuples in any form. If the similarity of this hypothetical tuple to the query is no more than the tuples in the Top-*K* buffer with the lowest similarity, the algorithm successfully terminates [13].

Filtering Algorithm: The algorithm stores the data in multidimensional points in a Kd - Tree. [2] Kd-tree is a binary tree having nodes and leaf. Each node of a kd-Tree is called cell, which contains closed box. The more than one point in a cell is called a leaf. The points in the cell are partitioned into one side. The remaining cells are children of the original cell. There are a number of ways are already described to partition the cell, one of the best way is to

Table 1.1: Shows that various techniques that are used for retrieving the information from web.

Title	Author	Year	Concept	Techniques used	Advantage	Disadvantage	Future Enhancement
How to Use Search Engine optimization Techniques to Increase Website Visibility	John B. Killoran	MARCH 2013	Web content audiences analyze the keywords and insert the key words into web text, based on that result pages will appear.	Google's keyword tool. Returns a millions of keywords. It follows click through rate and bounce rate	Frequently changing the any ranking algorithm for improves the efficiency of the result page.	-	-
WEB CAP: Inferring the user interest based on a Real time implicit feedback.	Nesrine Zemirli	2012	Web cap is a smart web browser and it can collective relative implicit documents during user search.	Web cap- a browser side tool	It does not require any log file.	Approach not effectivefor explicit documents.	Future is based on web application mobile indicators
A Collaborative Decentralized	Approach to Web Search Athanasios Papagelis and Christos Zaroliagis, *Member, IEEE*	SEP2012	Using a bottom up approach, data can be browsed, collected, tagged and studied.	Page ranking algorithm, bottom up approach	It improves the ranking and accuracy.	Based on the capital, it should build, it is an asynchronous connection and ad-hocmannr	To evaluate its effect on current search engine.
A Web Search Engine-Based Approach to Measure Semantic Similarity between Words	Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, Member, IEEE	JULY 2011	Semantic similarity between the words based on the page count and snippets retrieved from the web	Lexical pattern extraction algorithm	Improves the accuracy in a community mining task	-	-
Falcons Concept Search: A Practical Search Engine for Web Ontologies	Yuzhong Qu and Gong Cheng	JULY 2011	How the interaction helps the user to find the ontologies.	Constructing virtual documents for Keyword-based concept search	One main advantage is ranking concept.	-	We have to improve the snippet generation to better ontology structure
One Size Does Not Fit All: Towards User and Query Dependent Ranking For Web Databases	Aditya Telang and Chengkai Li and Sharma Chakravarthy	2009	Based on investigating the user browser choice, we can infer the workload of ranking functions	Similarity model: user and query similarity	Reduce the workload of the ranking function based on learning method	Does not consider about the efficiency of the results.	In future, it is evaluated using any one of retrieval algorithm.
Long-term Cross-Session Relevance Feedback Using Virtual Features	Peng-Yeng Yin, Bir Bhanu, Fellow, Kuang-Cheng Chang, Anlei Dong	MARCH 2008	It is the integration of long and short term learning experience and exploits the knowledge created by the multiple users.	Virtual feature updating algorithm is used. It Updates the long term learning experience.	It improves the performance while retrieving the images	-	-
Automated Ranking of Database Query Results	Sanjay Agrawal Surajit Chaudhuri Gautam Das	2003	Ranking and retrieving the information automatically without any feedback sessions	Index based threshold algorithm	The performances are strongly determined. Time and space require ments are linear with its data size	It is constant, not updating the future ranking. Sometimes it increases the workload in size	Application A burden to be reduced.
An Efficient k-Means Clustering Algorithm: Analysis and Implementation	Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu, Senior	JULY 2002	It computes the distance between the data points between the all centres. Comparing the old with new distance if it is less than or equal, the point in the same cluster.	Filtering algorithm	Implementation is easy compared to other algorithms. Efficiency is achieved by using constant data points.	The algorithm is quite complex. It has not significantly improved the faster running time.	-

split orthogonally in the longest side of the cell. For the given n points, it forms a Kd-binary tree. After constructing a binary tree using a filtering algorithm to implement and its use a less effective pruning method and compute a minimum or maximum distance between the each cell.

A Fuzzy Self Constructing Algorithm: In the proposed system, we use the Fuzzy self constructing algorithm. A uzzy is used for clustering the data sets in data mining. To find the initial data cluster center points is a challenged process. [14] This algorithm improves the efficiency of the process and speed up the calculation and automatically iterates each other every time. The existing K means algorithm is based on some constant factors; it does not automatically each other [12]. An algorithm consists of five steps:

Step 1: Choose a number of clusters.

Step 2: Assign randomly to each point coefficients for being in the clusters.

Step 3: Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than, the given sensitivity threshold):

Step 4: Compute the centroid for each cluster, using the formula above.

Step 5: For each point, compute its coefficients of being in the clusters, using the formula above.

The fuzzy c-means algorithm is very similar to the k-means algorithm. The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum and the results depend on the initial choice of weights. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise [13-19].

CONCLUSION

As a concluding remark all of the above mentioned techniques deals with information retrieval process in web. The proposed approach enhances the efficiency of few of the techniques discussed above. A new approach called “A fuzzy self constructing algorithm” which is used for clustering the user feedbacks, based on ranking was introduced. The user feedbacks are converted into pseudo documents. After clustering, each cluster can be considered as a user search goal. This algorithm improves the speed of the calculations and reduces the computation time to enhance the efficiency. This will automatically iterate every time and reduces the running time [20-24].

REFERENCES

1. [Online]. Available: Introduction to Information Retrieval, Jian-Yun Nie University of Montreal Canada.
2. An Efficient k-Means Clustering Algorithm: Analysis and Implementation Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman and Angela Y. Wu, Senior Member, IEEE, Ieee Transactions on Pattern Analysis and Machine Intelligence, 24(7), JULY 2002.
3. [online]. Available: Prof. Navneet Goyal BITS, Pilani. PPt.
4. An Efficient Fuzzy c-means Clustering Algorithm. Ming-chuan and Don-lin Yang. Dept of Information Technology, FengChia.
5. An effective approach for increasing the efficiency of inferring user search goals with feedback sessions. B. Saranya, G. Sangeetha, Valliammai Engineering College, Chennai.
6. WebCap: Inferring the user’s Interests based on a Real-Time Implicit Feedback. Nesrine zemrili, Informationsystemdepartment,978-1-4673-2430-4/12-2012.
7. A Collaborative Decentralized Approach to Web Search Athanasios Papagelis and Christos Zaroliagis, *Member, IEEE, IEEE* Transactions on Systems, Man and Cybernetics—part A: Systems and Humans, 42(5), September 2012.
8. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, Member, IEEE, IEEE Transactions on Knowledge and Data Engineering, 23(7), JULY 2011.

9. Correspondence Falcons Concept Search: A Practical Search Engine for Web Ontology Yuzhong Qu and Gong Cheng IEEE Transactions on Systems, Man and CYB ERN 2011.
10. One Size Does Not Fit All: Towards User and Query Dependent Ranking For Web Databases Aditya Telang, Chengkai Li, Sharma Chakravarthy Department of Computer Science and Engineering, University of Texas at Arlington July 16, 2009.
11. Long-Term Cross-Session Relevance Feedback Using Virtual Features Peng-Yeng Yin, Bir Bhanu, Fellow, IEEE, Kuang-Cheng Chang and Anlei Dong, IEEE Transactions on Knowledge and Data Engineering, 20(3), March 2008.
12. A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE and Zhaohui Zheng.
13. Automated Ranking of Database Query Results, Sanjay Agrawal, Surajit Chaudhuri, Gautam Das,Microsoft Research. Aristides Gionis Computer Science Dept, Stanford University, Proceedings of the 2003 CIDR Conference.
14. Kumaravel, A., 2013. An Application of Non-uniform Cellular Automata for Efficient Cryptography, Xplore, pp: 1200 -1205.
15. Kumaravel, A., 2013. Routing Algorithm over Semi-regular Tessellations, Xplore, pp: 1180-1184.
16. Kumaravel, A., 2013. Algorithm for Automaton Specification for Exploring Dynamic Labyrinths, Indian Journal of Science and Technology, 6(5).
17. Kumaravel, A., 2013. Application of Non-uniform Cellular Efficient Cryptography Automata, Indian Journal of Science and Technology, 6(5S): 4561-4566.
18. Kumaravel, A., 2013. Introducing an Efficient Programming Paradigm for Object-oriented Distributed Systems, Indian Journal of Science and Technology, 6(5S): 4597-4603.
19. How to Use Search Engine Optimization Techniques to Increase Website Visibility John B. Killoran, Ieee Transactions on Professional Communication, Vol. 56, No. 1, March 2013.
20. Shafaq Sherazi and Habib Ahmad, 2014. Volatility of Stock Market and Capital Flow Middle-East Journal of Scientific Research, 19(5): 688-692.
21. Kishwar Sultana, Najm ul Hassan Khan and Khadija Shahid, 2013. Efficient Solvent Free Synthesis and X Ray Crystal Structure of Some Cyclic Moieties Containing N-Aryl Imide and Amide,Middle-East Journal of Scientific Research, 18(4): 438-443.
22. Pattanayak, Monalisa. and P.L. Nayak, 2013. Green Synthesis of Gold Nanoparticles Using Elettaria cardamomum (ELAICHI) Aqueous Extract World Journal of Nano Science and Technology, 2(1): 01-05.
23. Chahataray, Rajashree. and P.L. Nayak, 2013. Synthesis and Characterization of Conducting Polymers Multi Walled Carbon Nanotube-Chitosan Composites Coupled with Poly (P-Aminophenol) World Journal of Nano Science and Technology, 2(1): 18-25.
24. Parida, Umesh Kumar, S.K. Biswal, P.L. Nayak and B.K. Bindhani, 2013. Gold Nano Particles for Biomedical Applications World Journal of Nano Science and Technology, 2(1): 47-57.