International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

Study on Applications of Information Filtering/ Retrieval Algorithms in Social Network

L.Rajeswari[1] and Dr.S.S.Dhenakaran[2]

[1]Programmer, Computer Centre, Alagappa University
Karaikudi 630 003 , Tamilnadu, India

[2] Professor, Computer Science and Engineering, Alagappa University
Karaikudi 630 003 , Tamilnadu, India

Abstract

Social Network [1] is defined as a network of interactions or relationships, where the nodes consist of actors and the edges consist of the relationships or interactions between these actors. Information retrieval is used to get the relevant information needed viewing and ignoring an irrelevant data. Information Filtering is a tool to help people to find the most valuable information. So that the limited time spent for reading/listening/viewing can be spent for the most interesting and valuable documents. This paper discusses some of the applications of Information Filtering/Retrieval Algorithms in the field of Social Network.

Keywords: Information Filtering, Information Retrieval, WebPages, Social Network Sites.

1. INTRODUCTION

The World Wide Web [2] is a collection of information resources on the Internet that are using the Hypertext Transfer protocol. It is repository of many interlinked hypertext documents accessed via the Internet. Web may contain text, images, video and other multimedia data. Social networks [1] have become very popular in recent years because of the increasing proliferation and affordability of internet enabled devices such as personal computers, mobile devices and other more recent hardware innovations like internet tablets. This is evidenced by the burgeoning popularity of many online social networks such as Twitter, Face book and Linked In. Social networks can be defined either in the context of systems such as Face book which are explicitly designed for social interactions or in terms of other sites such as Flicker which are designed for a different service such as content sharing but which allow an extensive level of social interaction also. Social networking sites [3] are the portals of entry into the Internet for many millions of users and they are being used both for advertisement as well as for ensuing commerce. A group of individuals with connections to other social worlds is likely to have access to a wider range of information. It is better for individual success to have connections to a variety of networks rather than many connections within a single network. Other Social networks such as You Tube and Google Video are used to share multimedia content and others such as Live Journal and BlogSpot are used to share blogs. The Social network itself is composed of links between users some sites allow users to link to any other user (without consent from the link recipient), while other sites follow a two-phase procedure that only allows a link to be established when both parties agree.

1.1 Social Network Usage

Online social network sites have become increasingly popular. More than 100 Social Network Websites are used by the people in the world. Among them, the most popular Social network sites are 15. The Worldwide Social Network users are given in Table - 1 and details in [4].

Table -1: Worldwide Social Network users

Social Network Websites	Users (Year 2014)
Face book	900,000,000
Twitter	310,000,000
Linked in	255,000,000
Printerest	250,000,000
Google Plus+	120,000,000
Tumbir	110,000,000
Instagram	100,000,000
VK	80,000,000
Flickr	65,000,000
Vine	42,000,000
Meetup	40,000,000
Tagged	38,000,000
Ask.frm	37,000,000
MeetMe	15,500,000
ClassMates	15,000,000

Leading Social Networks worldwide as of December 2014, ranked by number of active users (in millions) are depicted in the Figure 1 below and details in [5].

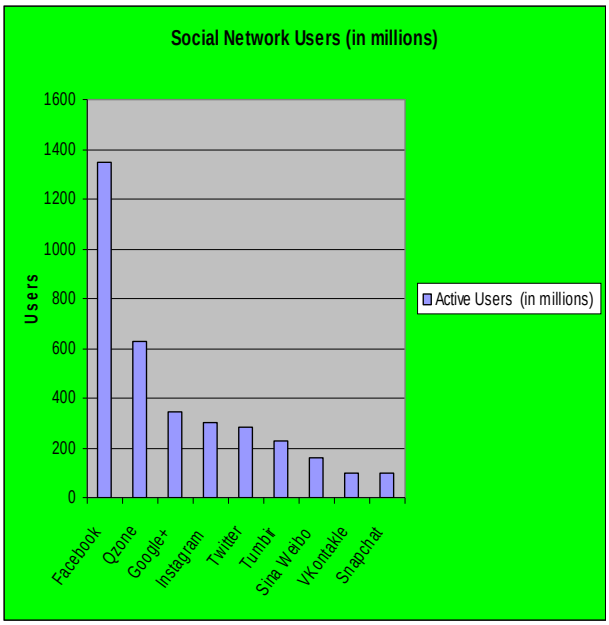

Figure 1 Social Network Users

1.2 Social Network Age Distribution

The age distribution data has been collected and calculated as what age distribution looks like across all the 19 sites counted together. The resulting chart is given in Figure -2 and details in [6].

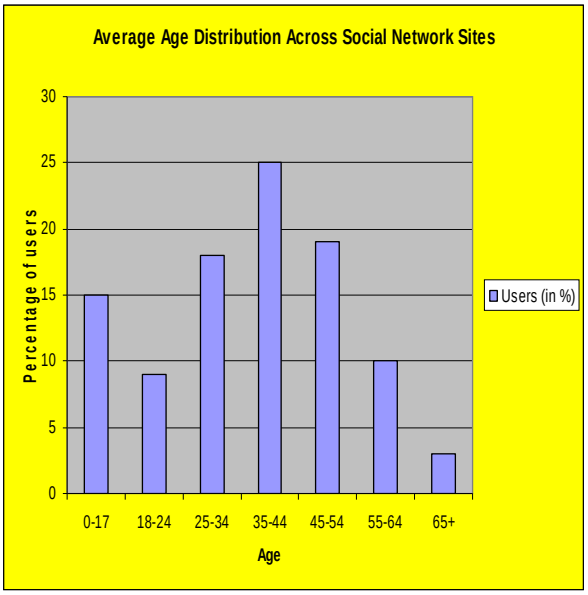

Figure 2 Age Distribution of Social Network Users

25% of the users of these sites are aged between 35 and 44, in other words the age group of 35 to 44 that dominates the social media sphere. Only 3% are aged at 65 or above.

A few observations:

- **The average age of social network user** is 37 years old.
- **Linked In**, with its business focus has a predictably high average user age as 44.
- **The average age of Twitter user** is 39 years.
- **The average age of Face book user** is 38 years.
- **The average age of My Space user** is 31 years.
- **Bebo** has been the youngest users as witnessed earlier with an average age of 28.

2BACKGROUND

2.1 Information Filtering

Information filtering is a name used to describe the variety of process involving in the delivery of information to people who needs it. Information retrieval has been characterized in a variety of ways ranging from a description of its goal, to relatively abstract models of its components and process. Information filtering deals with the delivery of information that the user is likely to find whether interesting or useful. An information filtering system assists users by filtering the data source and deliver relevant information to the users. For this, information has to be filtered.

2.2 Benefits of Information Filtering

On the Internet, almost anyone can easily publish anything they want at the low cost. This means that, a vast amount of information of varying quality is disseminated. There are lots of interesting things, but also lots of trash. Better ways at finding the most valuable information on the Internet and to avoid trash would very much enhance the value of the network. Since people downloads millions of messages and web documents every day and very often do not immediately get what they would mostly like to get, the gains through better filtering are enormous. Even a filtering with a 10% of efficiency gain, that gain would be a worth billions of dollar a year. Filters are also used to organize and structure information. People wants to read the most interesting messages and want to avoid having to read low-quality or uninteresting messages.

Filtering is often meant to imply the removal of data from an incoming stream rather than finding of data in that stream. In the first case, the users of the system see what is left after the data is removed; in the later case, they see the data that is extracted. Schools and universities select which information to teach the students based on scholarly criteria. The intention is again to help the customers, the students, to get the most out of a course. Political organizations select what information is discussed in their organizations and distributed to their members. The advantage with this filtering can be done in the background and that messages filtered away need never be downloaded to the client. For information filtering, many algorithms are applied nowadays. Some Information Filtering Algorithms are analyzed in this paper.

3ANALYSIS OF INFORMATION FILTERING/RETRIVAL ALGORITHMS

3.1 AdaRank: A Boosting Algorithm

In this method [7], first describe the general framework of learning to rank for document retrieval. In retrieval (testing), given a query, the system returns a ranking list documents in descending order of the relevance scores. The relevance scores are calculated with a ranking function (model). In learning (training), a number of

queries and their corresponding retrieved documents are given. Furthermore, the relevance levels of the documents with respect to the queries are also provided. The relevance levels are represented as ranks. The objective of learning is to construct a ranking function which achieves the best results in ranking of the training data in the sense of minimization of a loss function.

Suppose that $y = (r_1, r_2, \ldots\ldots\ldots, r_i)$ is a set of ranks, where i denotes the number of ranks. There exists a total order between the ranks

$r_i > r_{i-1} > \ldots\ldots\ldots > r_1$ where '>' denotes a preference relationship.

In training, a set of queries $Q = [q_1, q_2, \ldots\ldots\ldots q_m]$ is given. Each query q_i is associated with a list of retrieved documents $d_i = [d_{i1}, d_{i2}, \ldots\ldots\ldots d_{im(qi)}]$ and a list of labels $y_i = [y_{i1}, y_{i2}, \ldots\ldots\ldots y_{in(qi)}]$ where $n(q_i)$ denotes the sizes of lists d_i and y_i, d_{ij} denotes the j^{th} documents in d_i, and y_{ij} € Y denotes the rank of document d_{ij}.

The objective of learning is to create a ranking function f: X ←→ R such that, for each query, the elements in its corresponding document list can be assigned relevance scores using the function and then be ranked according to the scores. The queries are clustered into different groups based on the number of their associated documents pairs. Figure 3 below shows the distribution of the query groups.

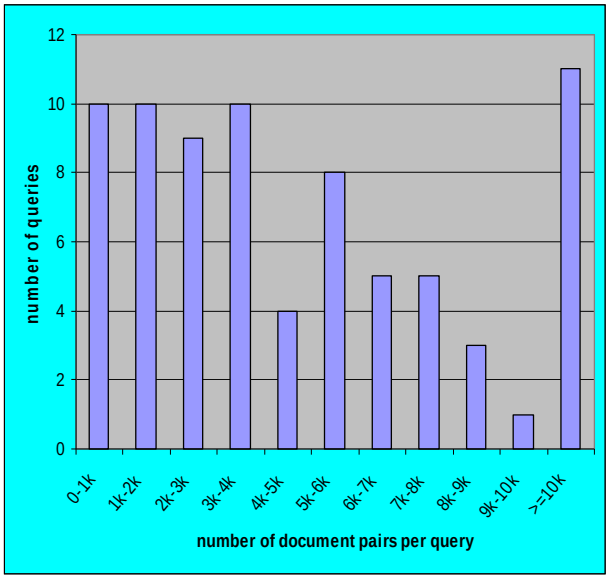

Figure 3 Distribution of queries with different number of document pairs

In Figure 3, for example '0-1k' is the group of queries whose number of document pairs is between 0 and 999. The number of document pairs really varies from query to query.

3.2 A New Information Filtering Method for WebPages

This method [8] is based on the use of Document Object Model (DOM) as the model for representing the WebPages. In this method, assume that a DOM tree is a data structure that represents each single element of a webpage with a node labeled with a text.

DOM tree: A Dome tree t = (V, E) is a tree whose vertices V are labeled nodes connected by a set of edges E.

Consider, l (n) → the label of a DOM node n

root(t) → the root of a DOM tree t

Webpage: A webpage is a pair (u, t) where u is an URL and t is a DOM tree.

Allow the user to specify complex queries that contain multiple words and metadata such as " " for exact search and Boolean operators (and, or, not) to produce combinations of texts that force a particular order of words.

Query: A query is a tuple (w, t, k, i) where w is a word that is associated to the information which is relevant for the user; t is an integer that represents the tolerance required in the search, k is a Boolean value that specifies whether the structure should be maintained and i is a Boolean value that specifies whether the filtering is inverse.

Algorithm: Information retrieval from single webpage
Input: A webpage p = (u, t) and a query q = (w, t, k, i)
Output: A webpage p' = (u, t')

Initialization: t = (v, e), t' = (Ø, Ø)

(1) key_nodes = {n ∈ v | l(n)=w}
(2) relevant_nodes = { n ∈ v | n - 'n' ∧ n' ∈ key_nodes}
(3) **if** (k=true)
(4) ancestors = {n ∈ v | $n_0 \rightarrow^* n \rightarrow^* n_1$ ∧ n_0=root(t) ∧ n_1 ∈ key_noes ⋃ relevant_nodes}
(5) **else**
(6) ancestors = Ø
(7) successors = {n ∈ v | $n_0 \rightarrow^* n$ ∧ n_0 ∈ relevant_nodes}
(8) **if** (i =T rue)
(9) final_nodes = {n ∈ v | n ∉ (successors ⋃ ancestors)}
(10) **else**
(11) final_nodes = successors ⋃ ancestors
(12) edges = {(n, n') ∈ e | n, n' ∈ final_nodes}

return p' = (u, (final_nodes, edges))

In this algorithm [8], it proceeds by

(i) Finding the *key nodes* for those labels is equal to the text specified by the user (w).

(ii) From these nodes, the *relevant nodes* are computed, which are those whose syntax distance to the relevant nodes is lower that the tolerance specified by the user (t).

The idea of using the tolerance as a measure of semantic relation is an important contribution of this technique.

3.3 Genetic Algorithm in Online Information Retrieval

3.3.1 Chromosome Representation

Online information retrieval using Genetic Algorithms (GA) [9] is based on vector space model. Within this model, both documents and queries are represented by the vector. A particular document is represented by vector of terms and a particular query is represented by vector of query terms.

A document vector (Doc) with a keywords and a query vector with m query terms can be represented as

Doc = {$term_1$, $term_2$, $term_3$, …….,$term_n$}

Query = {$qterm_1$, $qterm_2$, $qterm_3$, …….,$qterm_n$}

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.

Doc_1 = {Relational Database, Java, Web Technology, Cloud Computing, DBMS}

Doc_2 = {Artificial Intelligence, Internet, Visual Basic, Natural Language Processing}

Doc_3 ={C Programming, Expert System, Dot Net, Multimedia}

Doc_4 = {Fuzzy Logic, Neural Network, Cloud Computing}

Doc_5 = {Object-Oriented, DBMS, Java, Visual Basic}

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

Chromosome representation as in the Encode

Doc_1 = 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 1

Doc_2 = 1 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0

Doc_3 = 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0

Doc_4 = 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0

Doc_5 = 0 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0

These chromosomes are called initial population that feed into genetic operator process. The length of chromosome depends on number of keywords of documents retrieved from user query.

3.3.2 Crossover

Crossover is the genetic operator that mixes two chromosomes together to form new offspring. Crossover occurs only with some probability (crossover probability). GA constructs a better solution by mixture of good characteristics chromosomes together. Higher fitness chromosome have an opportunity to be selected more than that the lower ones; so good solutions always alive for the next generation. If the structures are represented as binary strings, crossover can be implemented by choosing a point at random called crossover point and exchanging the segments to the right of this point. Two chromosomes are crossover between position 5 and 11.

1 0 1 1 1 1 1 1 0 0 1 1 1 0 1

1 0 0 1 **1 0 0 1 1 1 1** 0 0 0 0

The resulting crossover yields two new chromosomes.

1 0 1 1 **1 0 0 1 1 1 1** 1 1 0 1

1 0 0 1 1 1 1 1 0 0 1 0 0 0 0

3.3.3 Mutation

Mutation involves the modification of the values of each gene of a solution with some probability (mutation probability). Chromosomes may be better or poorer than old chromosomes. If they are poorer than old chromosomes they are eliminated in selection step. For example, randomly mutate chromosome at position 10 results

1 0 1 1 1 1 1 1 0 ***0*** 1 1 1 0 1

1 0 1 1 1 1 1 1 0 ***1*** 1 1 1 0 1

3.3.4 Process of GA System

Step1. User enters query into the system

Step2. Match the keywords from user query with a list of keywords

Step3. Encode documents retrieved by the user query to chromosomes (initial population)

Step4. Population feed into genetic operator process such as selection, crossover and mutation.

Step5. Do step 4 until maximum generation is reached. An optimize query chromosome for document retrieval has got.

Step6. Decode optimize query chromosome to query and retrieve the document from database.

4.CONCLUSION

This paper discusses some of the applications of Information Filtering/Retrieval Algorithms in the field of Social Network. By analyzing the AdaRank, it optimizes a loss function that is directly defined on the performance measures. AdaRank offers several advantages: ease of implementation, theoretical soundness, efficiency in training and high accuracy in ranking. From the New Information Filtering Algorithm, it introduces a novel technique for information filtering that uses syntax distances to approximate semantic relations. The technique is able to work online and extract information from websites without any pre-compilation, labeling or indexing of the WebPages to be analyzed. From Genetic Algorithm, in the case of high precision documents prefer, the parameters will be high crossover probability and low mutation probability. While in the case of more relevant documents (high recall) prefer, the parameters will be high mutation probability and lower crossover probability. It is concluded that, an Information Filtering/Retrieving is very efficient and is very useful to filter the information from large data of WebPages.

References

[1] Charu C.Aggarwal, “An Introduction to Social Network Data Analytics”, IBM, T.J. Research Centre.

[2] Pooja Sharma et al., “Weighted Page Rank for Ordering Web Search Result”, IJEST, Vol.2(12), 2010, 7301-7310.

[3] Alan E.Mislove, “Online Social Networks: Measurement, Analysis and Applications to Distributed Information Systems”, in Ph.D. thesis, April 2009, Rice University.

[4] http://www.ebizmba.com/articles/social-networking- websites

[5] http://www.statista.com/statistics/272014/global-social-networks-ranked- by-number-of-users

[6] http://www.pingdom.com

[7] Jun Xu, Hang Li, “AdaRank: A Boosting Algorithm for Information Retrieval “, SIGIR 2007 Proceedings.

[8] Sergio Lopez, Josep Silva, "A New Information Filtering Method for WebPages", Universidad Politecnica de Valencia. Spain.
[9] Bangorn Klabbankoh, Ouen Pinngern, "Applied Genetic Algorithms in Information Retrieval ", King Mongkut's Institute of Technology Ladkrabng.