DOI: 10.21917/ijsc.2013.0080

AN EFFECTIVE RECOMMENDATIONS BY DIFFUSION ALGORITHM FOR WEB GRAPH MINING

S. Vasukipriya[1] and T. Vijaya Kumar[2]
Department of Information Technology, Bannari Amman Institute of Technology, India
E-mail: [1]vasukipriyacse@gmail.com and [2]vishal_16278@yahoo.co.in

Abstract
The information on the World Wide Web grows in an explosive rate. Societies are relying more on the Web for their miscellaneous needs of information. Recommendation systems are active information filtering systems that attempt to present the information items like movies, music, images, books recommendations, tags recommendations, query suggestions, etc., to the users. Various kinds of data bases are used for the recommendations; fundamentally these data bases can be molded in the form of many types of graphs. Aiming at provided that a general framework on effective DR (Recommendations by Diffusion) algorithm for web graphs mining. First introduce a novel graph diffusion model based on heat diffusion. This method can be applied to both undirected graphs and directed graphs. Then it shows how to convert different Web data sources into correct graphs in our models.

Keywords:
Recommendation System, Web Mining and Heat Diffusion

1. INTRODUCTION

The volatile growth of web information has not only created a crucial challenge for search engine companies to handle huge measure data, but also increased the difficulty for a user to achieve his information need. Since user-generated information is more freestyle and less structured, this increases the difficulties in mining useful information from these data sources. In order to realize the information needs of Web users and develop the user knowledge in many Web applications, Recommender Systems have been well designed in academia and widely deployed in numerous Areas.

A recommender system is capable to automatically give personalized recommendations based on the chronological record of customer's actions. These activities are usually signified by the connections in a user-item bipartite graph [4]. The collaborative Filtering (CF) is the most successful technique in the design of recommender systems [5], where a user will be recommended items that societies with similar tastes and preferences liked in the earlier. Even though its success, the performance of CF is strongly limited by the sparsity of data resulted from: (i) the enormous number of items far beyond user's ability to evaluate even a small fraction of them; (ii) users do not incentive wish to rate the purchased or viewed items.

The first one is the ambiguity which normally exists in the natural language. Queries having ambiguous languages may confuse the algorithms which do not fulfill the information needs of users. The second challenge is how to takings into account the personalization structures. Personalization is wanted for many consequences where different users have different information needs. The adoption of personalization will not only filter out unrelated information to a person, but also provide more particular information that is increasingly appropriate to a person's benefits. The last challenge is that it is time consuming and inefficient to design different recommendation procedures for different recommendation tasks. Essentially, most of these recommendation problems have common structures, where a common framework is wanted to unify the recommendation tasks on the Web [2].

Aim at solving the problems analyzed above, to suggest a common framework for the recommendations on the Web. This framework is made upon the heat diffusion on both undirected graphs and directed graphs, and has numerous benefits.

- It is a common method, which can be used to several recommendation responsibilities on the Web.
- It can offer latent semantically related results to the original information essential.
- This model offers a natural treatment for personalized recommendations.
- The planned recommendation algorithm is accessible to very large data sets.

2. RELATED WORK

In this section, review several work related to recommendation, containing query suggestion techniques, collaborative filtering, and click through data analysis. Since there are four major stages in recommendation system for web mining. Divide the system in to four models like Populating Data and Preprocessing, Query Processing, Graph Construction, and Performance Analysis. Each model has its own design issue and task to perform. The diagram represents the flow among these models in Fig.1.

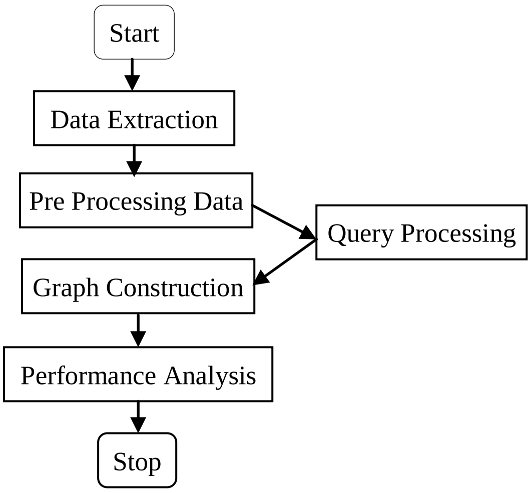


Fig.1. General Flow Diagram

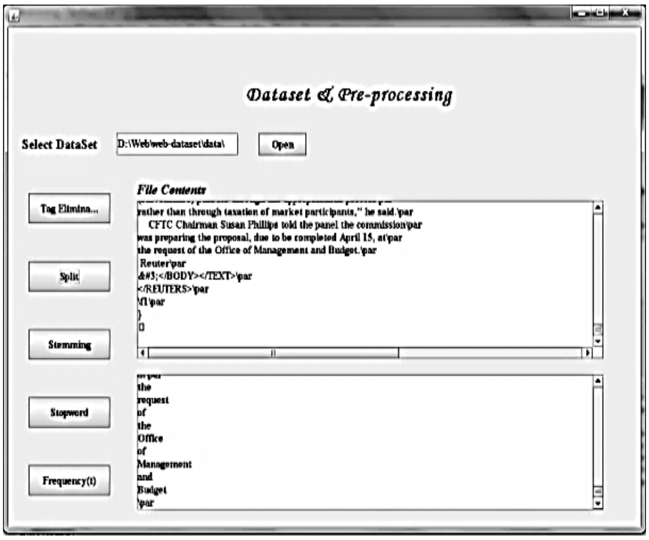
This framework is gives the heat diffusion on both undirected and directed graphs. Using this modules utilize the information effectively and efficiently has become more critical. In general method shows many recommendation tasks for web. Provide a latent semantically relevant result to the original information need and also having personalized recommendations. The designed recommendation algorithm is scalable to very large data sets [3].

3. SYSTEM MODEL

3.1 DATA PREPROCESSING

The first and foremost step in our system is data set extraction. Construct the query suggestion graph based on the click through data of the AOL search engine. Click through data record the activities of Web users, it shows their interests and the latent semantic relationships between users and queries as well as queries and clicked Web documents. Dataset is based on an existing a data source. A dataset specifies query parameters, query, filters, and a field collection. And also specify data options, such as collation, case, width and accent, for the data retrieved from the data source.

Every line of click through data contains: a user ID, a query issued by the user, a URL on which the user clicked, the rank of that URL, and the time at which the query was submitted for search. That data set is the raw data recorded by the search engine, and contains a lot of noise which will potentially affect the effectiveness of our query suggestion algorithm. So, in this module filter the data by only keeping those frequent, well formatted, English queries.



D:\Web\web-dataset\data\

```
rather than through taxation of market participants," he said.\par
   CFTC Chairman Susan Phillips told the panel the commission\par
was preparing the proposal, due to be completed April 15, at\par
the request of the Office of Management and Budget.\par
 Reuter\par
&#3;</BODY></TEXT>\par
</REUTERS>\par
\f1\par
}
□
```

```
the
request
of
the
Office
of
Management
and
Budget
\par
```

Fig.2. Data Pre-Processing Method

In pre-processing method want to removes the nose data from our dataset like

- Tag Elimination
- Splitting Word
- Stemming Word
- Stop Word

Tag Elimination is designed to removing unnecessary tagging and untagging operation from automatically generated programs. Split word is used to split the paragraph into word and this word is used for next Pre-Processing methods. Stemming is a fundamental step in pre-processing data preceding the tasks of information retrieval, and text mining processing. Stop word is used to filter out Articles, Prepositions, Conjunctions and Pronouns Words that occur in the document. Such words have no values for retrieval purpose. This pre-processing method is shown in Fig.2.

3.2 QUERY PROCESSING

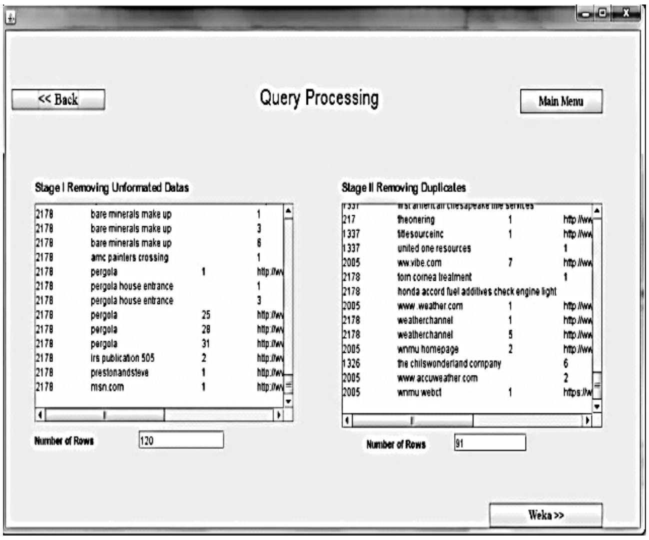
In this module want to implement query suggestion algorithm for graph construction in the next module. This query suggestion algorithm contains diffusion methodology to suggest recommendation. To get the exact query for search means to calculate and get the pre-processed results for

- Query Set
- Redundant
- New Query Set
- Query Diffusion

Query suggestion is associated to query substitution or query expansion, which extends the unique query with original search terms to narrow down the range of the search [1]. But dissimilar from query expansion, query suggestion ambitions to suggest full queries that have been expressed by previous users so that query integrity and coherence are preserved in the suggested queries in Fig.3.

Data cleaning is the very important process in the data mining process and also improves the quality of the data. It is used to detection the irrelevant data and removing it. Data quality problem is solved by using data cleaning method.

In Query processing removes the unformatted data and duplicates data. In this data set having query Id, URL, rank and time. Based on this data set removes the unformatted data in the data set. Calculating the rank values for every values for URL. Based on that rank values removes duplicates are removed from data set. And also shows the number of rows removed in this process. Select the Query search for calculating the optimization values for that query search.


<< Back
Query Processing
Main Menu
Stage I Removing Unformated Datas
Stage II Removing Duplicates
Weka >>

```
2178    bare minerals make up               1
2178    bare minerals make up               3
2178    bare minerals make up               6
2178    amc painters crossing               1
2178    pergola                    1        http://ww
2178    pergola house entrance              1
2178    pergola house entrance              3
2178    pergola                    25       http://ww
2178    pergola                    28       http://ww
2178    pergola                    31       http://ww
2178    irs publication 505        2        http://ww
2178    prestonandsteve            1        http://ww
2178    msn.com                    1        http://ww
```

Number of Rows: 120

```
1337    first american chesapeake title services
217     theonering                 1        http://ww
1337    titlesourceinc             1        http://ww
1337    united one resources                1
2005    ww.vibe.com                7        http://ww
2178    torn cornea treatment               1
2178    honda accord fuel additives check engine light
2005    www .weather.com           1        http://ww
2178    weatherchannel             1        http://ww
2178    weatherchannel             5        http://ww
2005    wnmu homepage              2        http://ww
1326    the chilswonderland company         6
2005    www.accuweather.com                 2
2005    wnmu webct                 1        https://w
```

Number of Rows: 91

Fig.3. Removing Unformatted and Duplicates Data

3.3 GRAPH CONSTRUCTION

Heat diffusion method is the used for graph contraction. Design novel graph diffusion model is based on that heat diffusion method. This model is used for both directed graphs and undirected graphs and also shows the infer parameter on the graph structure, analyze the complexity of model [6].

3.3.1 Diffusion on Undirected Graphs:

In undirected graph $G = (V, E)$ where V is the vertex set, and E is the edges. There is an edge between v_i to v_j is the set of all edges. The edge is considered as a pipe that connects nodes v_i and v_j. The value $f_i(t)$ defines the heat at node v_i at time t, start from an initial distribution of heat given by $f_i(0)$ at time zero. $f(t)$ means the vector consisting of $f_i(t)$. This is formulated as,

$$\frac{f_i(t+\Delta t)-f_i(t)}{\Delta t} = \alpha\sum{}_{j:}(v_j,v_i)\in E\big(f_j(t)-f_i(t)\big), \tag{1}$$

where, E is set of an edges. Express it in a matrix form

$$\frac{f_i(t+\Delta t)-f_i(t)}{\Delta t} = \alpha(H-D)\,f(t) \tag{2}$$

where,

$$H_{ij} = \begin{cases} 1, & (v_i,v_j)\in E \text{ or } (v_j,v_i)\in E \\ 0, & i=j \\ 0, & otherwise \end{cases} \tag{3}$$

$$D_{ij} = \begin{cases} d(v_i), & i=j \\ 0, & otherwise \end{cases} \tag{4}$$

where, $d(v_i)$ is the degree of node v_i. From the definition, the matrix D is a diagonal matrix.

$$H_{ij} = \begin{cases} \dfrac{1}{d(v_i)}, & (v_i,v_j)\in E \\ 0, & i=j \\ 0, & otherwise \end{cases} \tag{5}$$

$$D_{ij} = \begin{cases} 1, & i=j \\ 0, & otherwise \end{cases} \tag{6}$$

In the limit $\Delta t \rightarrow 0$, this becomes,

$$\frac{d}{dt}f(t) = \alpha t(H-D)\,f(t). \tag{7}$$

Solving this differential equation, it has

$$f(1) = e^{\alpha(H-D)}f(0), \tag{8}$$

$$e^{\alpha(H-D)} = I + \alpha(H-D) + \frac{\alpha^2}{2!}(H-D)^2 + \frac{\alpha^3}{3!}(H-D)^3 + .. \tag{9}$$

The matrix $e^{\alpha(H-D)}$ is called the diffusion kernel in the sense that the heat diffusion process continues infinitely many times from the initial heat diffusion.

Using Eq.(3) and Eq.(4) calculate the matrix values for the given graph. Unit of heat is calculated using Eq.(8) for each node present in the graph. The entries in matrix H-D are shown in Fig.4.

$$H-D = \begin{bmatrix} -1 & \frac{1}{3} & 0 & 0 \\ 1 & -1 & 1 & 1 \\ 0 & \frac{1}{3} & -1 & 0 \\ 0 & \frac{1}{3} & 0 & -1 \end{bmatrix}$$

Fig.4. Matrix Values

Example is shown in Fig.5. Initially, at time zero, suppose node 2 is given 3 units of heat, and node 1 is given 1 units of heat; the vector $f(0)$ equals [1,3,0,0,0]T.

This is also indicates that if a node has additional paths connected to the heat source, it will potentially get more heat. This is a perfect property for recommending relevant nodes on a graph.

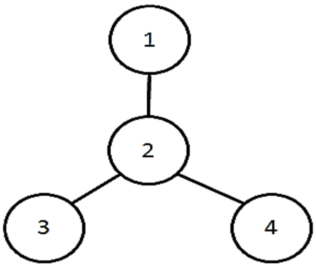
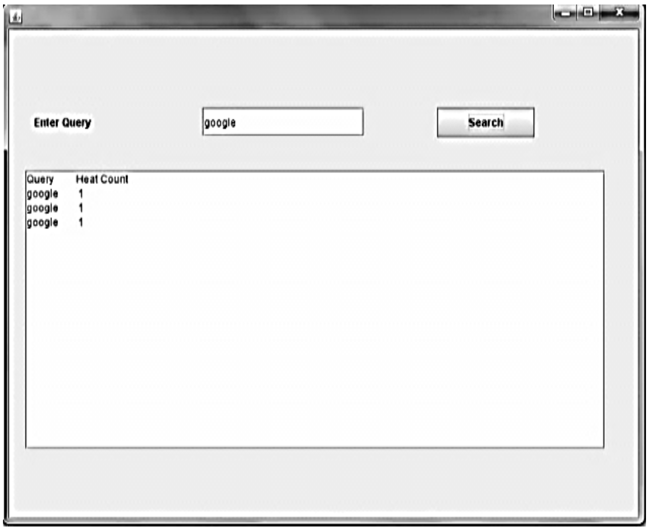


Fig.5. Simple Heat Diffusion Graph

3.3.2 Diffusion on Directed Graphs:

The above heat diffusion model is designed for undirected graphs, but in many times, the Web graphs are directed, especially in online recommender systems or knowledge sharing sites. Every user in knowledge sharing sites typically has a trust list. The users in the faith list can influence this user deeply.

These relationships are directed since user is in the faith list of user b, but user b might not be in the trust list of user a. At the same time, the extent of trust relations is different since user u_i may trust user u_j through trust achieve 1 while trust user u_k only with trust score 0.2. Hence, there are different weights associated with the relations [7]. Heat count value is shown in Fig.6.



google

Query	Heat Count
google	1
google	1
google	1

Fig.6. Heat Count Values

Consider a directed graph, $G = \{V, E, W\}$ where; V is the vertex set, $V = \{v_1, v_2, \ldots. v_n\}$; $W = w_{ij}$, where w_{ij} is probability that edge (v_i, v_j) exists or the weight that is associated with this edge and $E = (v_i, v_j)$, an edge from v_i to v_j and $w_{ij} > 0$ is the set of all edges.

At the same time, node v_i diffuses DH $(I, t, \Delta t)$ sum of heat to its subsequent nodes. We assume that

- The heat DH $(I, t, \Delta t)$ must be proportional to the time period Δt.
- The heat DH $(I, t, \Delta t)$ must be proportional to the heat at node v_i.
- Each node has the same ability to diffuse heat.
- The heat DH $(I, t, \Delta t)$ must be proportional to the weight. It is assigned in between node v_i and its following nodes.

This is formulated as,

$$\frac{f_i(t+\Delta t)-f_i(t)}{\Delta t} = \alpha(-\tau_i f_i(t) + \sum_{j:}(v_j, v_i) \in E \frac{w_{ij}}{\sum_{k(j,k)\in E} W_{ij}} f_i(t) \tag{10}$$

where, τ_i is a flag to identify whether node v_i has any out links. To obtain,

$$f(1) = e^{\alpha(H-D)} f(0), \tag{11}$$

where,

$$H_{ij} = \begin{cases} \dfrac{W_{ij}}{\sum_{k:(j,k)\in E} W_{jk}}, & (v_j, v_i) \in E \\ 0, & i = j \\ 0, & otherwise \end{cases} \tag{12}$$

and

$$D_{ij} = \begin{cases} \tau_i, & i = j \\ 0, & otherwise \end{cases} \tag{13}$$

4. RESULT ANALYSIS

Web graphs are normally very huge; we will perform our algorithm on a subgraph extracted from the original graph. Hence, it is necessary to evaluate how the size of this subgraph affects the recommendation accuracy. The Fig.7 shows the performance changes with different subgraph sizes. We observe that when the size of the graph is very small, the performance of our algorithm is not very good since this subgraph must ignore some very relevant nodes. When the size of subgraph is increasing, the performance also increases.

From the outcome, observe that our recommendation algorithm not only suggest queries which are factually similar to the test queries, but also provides latent semantically related recommendations. For example, if the trial query is a technique, like "java," means it recommends "sun Microsystems" and "virtual machine". The last suggestion is the company who own the Java Platform, and the former suggestion is a key element of the Java programming language. They both have high latent semantic relations to the query "java."

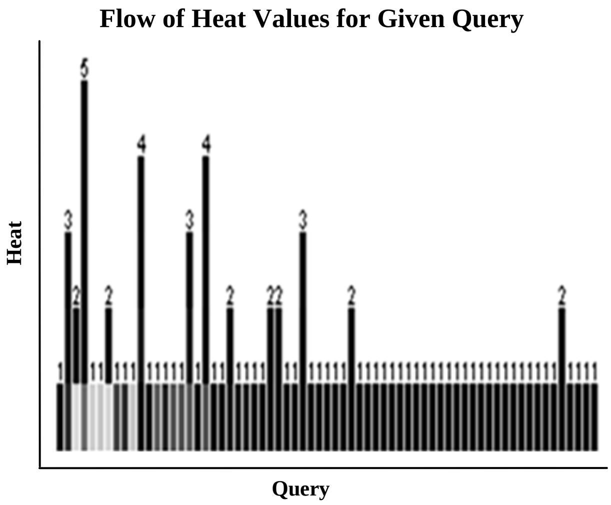


Fig.7. Heat Changes with Query

Algorithm is very efficient and can be applied to large data sets. This algorithm has similar difficulty with FRW and BRW methods. The computation time for the query suggestion task of these three methods. SimRank is not well-ordered since it has a high computational difficulty. It takes more time to calculate a query suggestion task in data set.

5. CONCLUSION

Recommendation System is used in web graphs using heat diffusion. This general framework which can basically be adapted to most graphs for recommendation tasks are query suggestion and personalized recommendations. The generated suggestions are semantically related to the input. It is important for Web 2.0 related applications since user-generated information is more freestyle and less structured, which increases the difficulties in mining useful information from these data sources. In order to satisfy the information requirements of Web users and improve the user experience in various Web applications.

REFERENCES

[1] Hang Cui, Ji-Rong Wen, Jian-Yun Nie and Wei-Ying Ma, "Query Expansion by Mining User Logs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 1–11, 2003.

[2] Yong Zhen Guo, K. Ramamohanarao and L. A. F. Park, "Personalized PageRank for Web Page Prediction Based on Access Time-Length and Frequency", *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 687–690, 2007.

[3] Daniel Fogaras and Balazs Racz, "Practical Algorithms and Lower Bounds for Similarity Search in Massive Graphs" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 5, pp. 585 – 598, 2007.

[4] Ming-Sheng Shang, Yan Fu and Duan-Bin Chen, "Personal Recommendation using Weighted Bipartite Graph Projection", *International Conference on Apperceiving Computing and Intelligence Analysis*, pp. 198 – 202, 2008.

[5] Jinbo Zhang, Zhiqing Lin, Bo Xiao and Chuang Zhang, "An Optimized item-based collaborative filtering Recommendation Algorithm", *Proceedings of the IEEE International conference on Network Infrastructure and Digital Content*, pp. 414 – 418, 2009.

[6] K. Kazama, M. Imada and K. Kashiwagi, "Characteristics Estimation of Information Sources by Information Diffusion Analysis", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 484 – 491, 2010.

[7] Hao Ma, Irwin King and Michael Rung-Tsong Lyu, "Mining Web Graphs for Recommendations", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 6, pp. 1051–1064, 2012.