General Framework on Mining Web Graphs for Recommendations

K.Stephey Chrysolite1† and B.Krishna Sagar2††,

University of Jawaharlal Nehru Technology Anantapur, College Madanapalle Institute of Technology and science,
ANDHRA PRADESH, INDIA

Summary

As the exponential explosion of various contents generated on the Web, Recommendation techniques have become increasingly indispensable. Innumerable different kinds of recommendations are made on the web every day, including movies, music, books, images, books recommendations, query suggestions, tags recommendations, etc. No matter what types of data sources are used for the recommendations, essentially these data sources can be modelled in the form of various types of graphs. In this paper, aiming at providing a general framework on mining Web graphs for recommendations, we first propose a novel diffusion method which propagates similarities between different nodes and generates recommendations; then we illustrate how to generalize different recommendation problems into our graph diffusion framework. The proposed framework can be utilized in many recommendation tasks on the World Wide Web, including query suggestions, tag recommendations, expert finding, image recommendations, image annotations, etc.

Key words:

Recommendation, Diffusion, Query Suggestion, Image Recommendation

1. Introduction

With the diverse and explosive growth of Web information, how to organize and utilize the information effectively and efficiently has become more and more critical. This is especially important for Web 2.0 related applications since user generated information is more free-style and less structured, which increases the difficulties in mining useful information from these data sources. In order to satisfy the information needs of Web users and improve the user experience in many Web applications, Recommender Systems, have been well studied in academia and widely deployed in industry.

Typically, recommender systems are based on Collaborative Filtering, this is a technique that automatically predicts the interest of an active user by collecting rating information from other similar users or items. The underlying assumption of collaborative filtering is that the active user will prefer those items which other similar users prefer.

Fortunately, on the Web, no matter what types of data sources are used for recommendations, in most cases, these data sources can be modeled in the form of various types of graphs. If we can design a general graph recommendation algorithm, we can solve many recommendation problems on the Web. However, when designing such a framework for recommendations on the Web, we still face several challenges that need to be addressed. The first challenge is that it is not easy to recommend latent semantically relevant results to users. The second challenge is how to take into account the personalization feature. The last challenge is that it is time-consuming and inefficient to design different recommendation algorithms for different recommendation tasks. Actually, most of these recommendation problems have some common features, where a general framework is needed to unify the recommendation tasks on the Web. Moreover, most of existing methods are complicated and require to tune a large number of parameters.

In this, aiming at solving the problems analyzed above, we propose a general framework for the recommendations on the Web. This framework is built upon the heat diffusion on both undirected graphs and directed graphs, and has several advantages: (1) It is a general method, which can be utilized to many recommendation tasks on the Web; (2) It can provide latent semantically relevant results to the original information need; (3) This model provides a natural treatment for personalized recommendations; (4) The designed recommendation algorithm is scalable to very large datasets.

2. Related Work

Recommendation on the Web is a general term representing a specific type of information filtering technique that attempts to present information items (queries, movies, images, books, Web pages, etc.) that are likely of interest to the users. In this section, we review several work related to recommendation, including collaborative filtering, query suggestion techniques, image recommendation methods, and click through data analysis.

Manuscript received September 5, 2015
Manuscript revised September 20, 2015

2.1 Collaborative Filtering

With the diverse and explosive growth of Web information, how to organize and utilize the information effectively and efficiently has become more and more critical. In order to satisfy the information needs of Web users and improve the user experience in many Web applications, Recommender Systems, have been well studied in academia an widely deployed in industry. Typically, recommender systems are based on Collaborative Filtering, which is a technique that automatically predicts the interest of an active user by collecting rating information from other similar users or items.

2.1.1Overview of the Collaborative Filtering Progress

The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the user's previous likings and the opinions of other like-minded users. In a typical CF scenario, there is a list of m users U={u1,u2,...um} and a list of n items I={i1,i2,...in}. Each user ui has a list of items Iui, which the user has expressed his/her opinions about. Opinions can be explicitly given by the user as a rating score, generally within a certain numerical scale, or can be implicitly derived from purchase records, by analyzing timing logs, by mining web hyperlinks and so on. Note that Iui ⊆ I and it is possible for Iui to be a null-set. There exists a distinguished user Ua ∈ U called the acive user for whom the task of a collaborative filtering algorithm is to find an item likeliness that can of two forms.

- **Prediction** is a numerical value $P_{a,j}$P$_{a,j,}$, expressing the predicted likeliness of items $i_j \notin I_{u_a}$ for the acive user u_a . This predicted value is within the same scale (e.g., from 1 to 5) as the opinion values provided by u_a.
- **Recommendation** is a list of N items, $I_r \subset I$, that the active user will like the most. Note that the recommended list must be on items not already purchased by the active user, i.e., $I_r \cap I_{u_a} = \emptyset$. This interface of CF algorithms is also known as Top N recommendation.

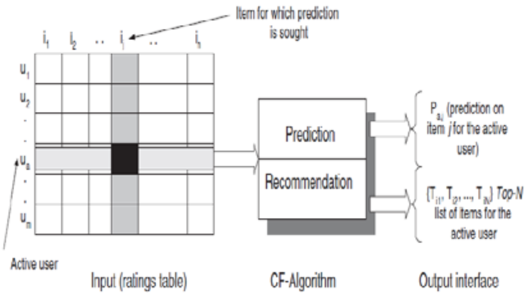

Item for which prediction is sought
i_1
i_2
..
i_j
..
i_n
u_1
u_2
u_a
u_m
Active user
Prediction
Recommendation
$P_{a,j}$ (prediction on item j for the active user)
$\{T_{i1}, T_{i2}, ..., T_{iN}\}$ Top-N list of items for the active user
Input (ratings table)
CF-Algorithm
Output interface

Fig. 1 The Collaborative Filtering Process

Figure 1shows the schematic diagram of the collaborative filtering process. CF algorithms represent the entire m × n user-item data as a rating matrix, $\mathcal{A}$ Each entry a$_{i,j}$ in $\mathcal{A}$ represents the preference score (ratings) of within a numerical scale and it can as well be 0 indicating that the user has not yet rated that item.

2.2 Query Suggestion

In order to recommend relevant queries to Web users, a valuable technique, query suggestion, has been employed by some prominent commercial search engines, such as Yahoo, Live Search, Ask and Google.

The goal of query suggestion is similar to that of query expansion, query substitution and query refinement, which all focus on understanding users' search intentions and improving the queries submitted by users. Query suggestion is closely related to query expansion or query substitution, which extends the original query with new search terms to narrow down the scope of the search. But different from query expansion, query suggestion aims to suggest full queries that have been formulated by previous users so that query integrity and coherence are preserved in the suggested queries . Query refinement is another closely related notion, since the objective of query refinement is interactively recommending new queries related to a particular query.

Actually, several different ranking methods using random walks can also be employed into the query suggestion tasks on a query-URL bipartite graph, including PageRank, HITS, etc. Typically, query suggestion is based on local (i.e., search result sets) and global (i.e., thesauri) document analysis, or anchor text analysis. However, these traditional methods have difficulty summarizing the latent meaning of a Web document due to the huge noise embedded in each Web page. Moreover, this noise is not easily removed by machine learning methods. In order to avoid these problems, some additional data sources are likely to be very helpful to improve the recommendation quality. In fact, clickthrough data is an ideal source for mining relevant queries.

2.3 Clickthrough Data Analysis

In the field of clickthrough data analysis, the most common usage is for optimizing Web search results or rankings. Besides ranking, clickthrough data is also well studied in the query clustering problem. Query clustering is a process used to discover frequently asked questions or most popular topics on a search engine. This process is crucial for search engines based on question-answering. Recently, clickthrough data has been analyzed and applied to several interesting research topics, such as Web query hierarchy building and extraction of class attributes.

2.4 Image Recommendation

Besides query suggestion, another interesting recommendation application on the Web is image recommendation. Image recommendation systems, like Photoree, focus on recommending interesting images to Web users based on users' preference. Normally, these systems first ask users to rate some images as they like or dislike, and then recommend images to the users based on the tastes of the users. In the academia, few tasks are proposed to solve the image recommendation problems since this is a relatively new field and analyzing the image contents is a challenge job. However, since it is a context-based method, the computational complexity is very high and it cannot scale to large datasets. While in our framework proposed in this paper, by diffusing on the image tag bipartite graph with one or more images, we can accurately and efficiently suggest semantically relevant non personalized or personalized images to the users. In general, comparing with previous work, our work is a general framework which can be effectively, efficiently and naturally applied to most of the recommendation tasks on the Web.

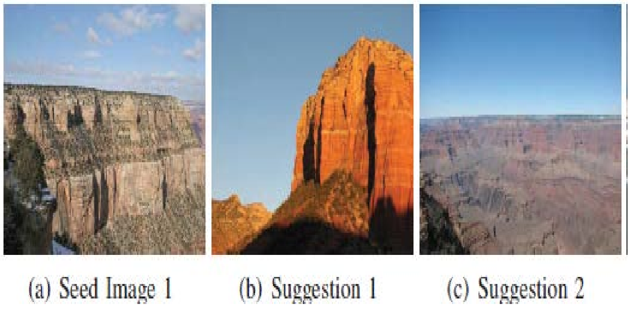

(a) Seed Image 1 (b) Suggestion 1 (c) Suggestion 2

Fig 2: Examples for Image Recommendations

3. Diffusion on Graphs

Heat diffusion is a physical phenomenon. In a medium, heat always flows from a position with high temperature to a position with low temperature. Recently, heat diffusion based approaches have been successfully applied in various domains such as classification and dimensionality reduction problems.

In this paper, we model diffusion of innovations as processes of heat diffusion. Actually, the process of people influencing others is very similar to the heat diffusion phenomenon. In a social network, the innovators and early adopters of a product or innovation act as heat sources, and have a very high amount of heat. These peoples start to influence others, and diffuse their influence to the early majority, then the late majority. Finally, at a certain time point, heat is diffused to the margin of this social network, and the laggards adopt this product or innovation.

3.1 Diffusion on Undirected Graphs

Consider an undirected graph *G* = (*V*,*E*), where *V* is the vertex set, and *V* = *{v*1, *v*2, . . . , *vn}*. *E* = *{*(*vi*, *vj*) | there is an edge between *vi* to *vj}* is the set of all edges. The edge (*vi*, *vj*) is considered as a pipe that connects nodes *vi* and *vj* . The value *fi*(*t*) describes the heat at node *vi* at time *t*, beginning from an initial distribution of heat given by *fi*(0) at time zero. f(*t*) denotes the vector consisting of *fi*(*t*).
We construct our model as follows. Suppose, at time *t*, each node *i* receives an amount *M*(*i*, *j*, *t*,Δ*t*) of heat from its neighbor *j* during a time period Δ*t*. The heat *M*(*i*, *j*, *t*,Δ*t*) should be proportional to the time period Δ*t* and the heat difference *fj*(*t*) − *fi*(*t*). Moreover, the heat flows from node *j* to node *i* through the pipe that connects nodes *i* and *j*. Based on this consideration, we assume that *M*(*i*, *j*, *t*,Δ*t*) = *α*(*fj*(*t*) − *fi*(*t*))Δ*t*, where *α* is the thermal conductivity-the heat diffusion coefficient. As a result, the heat difference at node *i* between time *t*+Δ*t* and time *t* will be equal to the sum of the heat that it receives from all its neighbors. This is formulated as:

$$\frac{f_i(t+\Delta t) - f_i(t)}{\Delta t} = \alpha \sum_{j:(v_j,v_i)\in E} (f_j(t) - f_i(t)), \qquad (1)$$

where *E* is the set of edges. To find a closed form solution to Eq. (1), we express it in a matrix form:

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

Where

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

And

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

where d(vi) is the degree of node vi. From the definition, the matrix D is a diagonal matrix. In order to generate a more generalized representation, we normalize all the entries in matrices H and D by the degree of each node. The matrices H and D can be modified to

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

And

$$D_{ij} = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise}. \end{cases} \quad (6)$$

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

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

Solving this differential equation, we have:

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

where d(v) denotes the degree of the node v, and eα(H−D) could be extended as:

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

The matrix eα(H−D) is called the diffusion kernel in the sense that the heat diffusion process continues infinitely many times from the initial heat diffusion.

3.2 Diffusion on Directed Graphs

The above heat diffusion model is designed for undirected graphs, but in many situations, 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 trust list can influence this user deeply. These relationships are directed since user a is in the trust 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 ui may trust user uj with trust score 1 while trust user uk only with trust score 0.2. Hence, there are different weights associated with the relations. Based on this consideration, we modify the heat diffusion model for the directed graphs as follows.
Consider a directed graph G = {V, E,W}, where V is the vertex set, and V = {v1, v2, . . . , vn}. W = {wij | where wij is the probability that edge (vi, vj) exists} or the weight that is associated to this edge. E = {(vi, vj) | there is an edge from vi to vj and wij > 0} is the set of all edges. On a directed graph G(V,E), in the pipe (vi, vj), heat flows only from vi to vj . Suppose at time t, each node vi receives RH = RH(i, j, t,Δt) amount of heat from vj during a period of Δt. We make three assumptions: (1) RH should be proportional to the time period Δt; (2) RH should be proportional to the heat at node vj ; and (3) RH is zero if there is no link from vj to vi. As a result, vi will receive $\sum_{j:(v_j,v_i)\in E} \sigma_j f_j(t)\Delta t$ amount of heat from all its neighbours that point to it.
At the same time, node vi diffuses DH(i, t,Δt) amount of heat to its subsequent nodes. We assume that: (1) The heat DH(i, t,Δt) should be proportional to the time period Δt; (2) The heat DH(i, t,Δt) should be proportional to the heat at node vi; (3) Each node has the same ability to diffuse heat; (4) The heat DH(i, t,Δt) should be proportional to the weight assigned between node vi and its subsequent nodes. As a result, node vi will diffuse $\alpha w_{ij} f_i(t)\Delta t / \sum_{k:(i,k)\in E} w_{ik}$ amount of heat to each of its subsequent nodes vj , and each vj should receive $\alpha w_{ij} f_i(t)\Delta t / \sum_{k:(i,k)\in E} w_{ik}$ amount of heat from node vi. Therefore $\sigma_j = \alpha w_{ji} / \sum_{k:(j,k)\in E} w_{jk}$. In the case that the outdegree of node vi equals zero, we assume that this node will not diffuse heat to others. To sum up, the heat difference at node vi between time t+Δt and t will be equal to the sum of the heat that it receives, deducted by what it diffuses. This is formulated as

$$\frac{f_i(t+\Delta t) - f_i(t)}{\Delta t} = \alpha\left(-\tau_i f_i(t) + \sum_{j:(v_j,v_i)\in E} \frac{w_{ji}}{\sum_{k:(j,k)\in E} w_{jk}} f_j(t)\right), \quad (10)$$

where *τi* is a flag to identify whether node *vi* has any outlinks. Solving it, we obtain

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

Where

$$H_{ij} = \begin{cases} w_{ji}/\sum_{k:(j,k)\in E} w_{jk}, & (v_j, v_i) \in E, \\ 0, & i = j, \\ 0, & \text{otherwise}, \end{cases} \quad (12)$$

And

$$D_{ij} = \begin{cases} \tau_i, & i = j, \\ 0, & \text{otherwise}. \end{cases} \quad (13)$$

4. Conclusion

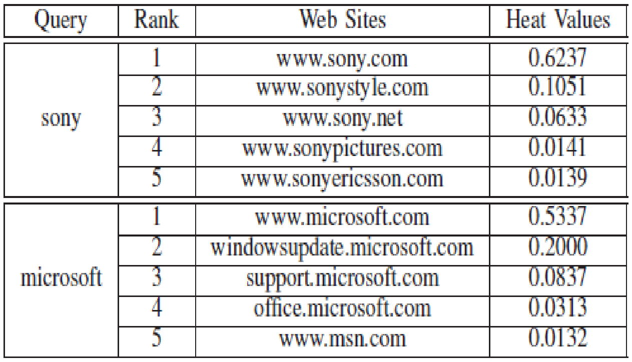
We present a novel framework for recommendations on large scale Web graphs using heat diffusion. This is a general framework which can basically be adapted to most of the Web graphs for the recommendation tasks, such as query suggestions, image recommendations, personalized recommendations, etc. The generated suggestions are semantically related to the inputs. The experimental analysis on several large scale Web data sources shows the promising future of this approach.

5. Future Work

5.1 Search Results Improvement

We are going to use heat values in query suggestion. These values not only can be used in query suggestions, but also are very informative in the advertisement when customers bid for query terms. Actually, since the diffusions are between all the nodes in the graph (including the nodes representing queries and the nodes representing URLs), all the URLs also have heat values. Hence, it is easy to infer that, for a given query, after the diffusion process, the heat values of URLs represent the relatedness to the original query, which can also be employed as the ranking of these URLs. Table 1shows search results improvement.

Table 1: Search Results Improvement



Query	Rank	Web Sites	Heat Values
sony	1	www.sony.com	0.6237
	2	www.sonystyle.com	0.1051
	3	www.sony.net	0.0633
	4	www.sonypictures.com	0.0141
	5	www.sonyericsson.com	0.0139
microsoft	1	www.microsoft.com	0.5337
	2	windowsupdate.microsoft.com	0.2000
	3	support.microsoft.com	0.0837
	4	office.microsoft.com	0.0313
	5	www.msn.com	0.0132

5.2 Social Recommendation

Since our model is quite general, we can apply it to more complicated graphs and applications, such as Social Recommendation problem. Recently, as the explosive growth of Web 2.0 applications, social-based applications gain lots of traffics on the Web. Social recommendation, which produces recommendations by incorporating users' social network information, is becoming to be an indispensable feature for the next generation of Web applications.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst., 30(1-7):107–117, 1998.

[2] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In Proc. of WWW, pages 271–280, Banff, Alberta, Canada, 2007.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46(5):604–632, 1999.

[4] A. Kohrs and B. Merialdo. Clustering for collaborative filtering applications. In Proc. of CIMCA, 1999.

[5] H. Ma, I. King, and M. R. Lyu. Effective missing data prediction for collaborative filtering. In Proc. of SIGIR, pages 39–46, Amsterdam, The Netherlands, 2007.

[6] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proc. of WWW, pages 285–295, Hong Kong, Hong Kong, 2001. ACM.

[7] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Webpage summarization using clickthrough data. In Proc. of SIGIR, pages 194–201, Salvador, Brazil, 2005.