lUser based Query suggestions using graphs

Mr. Chandan Wagh, Prof. Vivekkumar Shrivastava ITM,Bhilwara Rajasthan

Abstract— As numbers of users of Web are rising daily and data which are navigating worldwide are in unstructured form, there should be some method which can take out applicable structured data from formless web data. Today, numbers of web based method are available to suggest data to web users like suggesting songs, movies, images, software, and queries etc., but this suggestion cannot be correct because web data are unstructured. No matter which types of data we are using for suggestion, essentially these data can be cast into various types of Graphs. We devise new approach called User based Query Suggestion which uses Graph Diffusion algorithm with user profile hits to suggest accurately similar quires to user.

Index Terms— *Suggestion, query, diffusion, ranking, graph, hit, web usage data.*

1 INTRODUCTION

THE explosion of different types of information generated on the net., it is became very complex to deal with all information. Information available on the web is unstructed, so getting specific information as per user need is becoming complex task. In order to fulfill the need of net applications, suggestion techniques have become increasing essential.

Traditional suggestion techniques are based on the Collaborative filtering [2]. Collaborative automatically forecast the user interest by getting extra information from non like users and items. Collaborative filtering is being implemented in various online systems like movie lens, Netflix and Amazon.com. In classical collaborative filtering user item based matrix is essential which gives rating for items, but nearly all of the time rating data is unavailable.

There is a technique by which we can model unstructured data in the form of graph. By Graph suggestion, we can solve many suggestion problems but there is a challenge like user based query suggestion [1]. Amazon.com uses user based technique to suggest items to user. The practical analysis on several AOL (American Online) Data set shows that proposed technique is important to generate worth query suggestions.
Remaining paper is organized as follows: Related work is analyzed in section 2, section 3 shows diffusion model on undirected graph, section 4 shows proposed system. Section 5 shows practical analysis of model and user based query suggestion algorithm and in section 6, conclusion.

2 RELATED WORK

Suggestion represents filtering technique that present data in the form of queries, movies, books etc that is possibly of likeliness of users. We analyze several work related query suggestions, as well as collaborative filtering, click through data analysis.

2.1 Collaborative Filtering

Different types of collaborative filtering are studied: User based and Item based which comes under neighborhood based approach. In some system like Netnews and neighborhood based approach is widely used [3] [4]. In user based approach [5], forecast rating of user based on the rating of current users, item based approach [6] predict rating of users based on likeness on previous history of item purchased by users. These algorithm used Pearson Correlation Coefficient[6] and Vector space similarity algorithm [7]. Other approach is model based approach which use predefined model to guess data.

2.2 Query Suggestions

The technique being used in net based search engine like ask.com, yahoo.com etc. Query suggestion is somewhat similar to query expansion [8],query substitution[9], and query refinement [10]. These approach are used to know user searching behavior and to improve the result of search engine. In [11], query suggestions based on click data is suggested. In that, given query is feed to search engine and engine suggests the list of related queries based clustering algorithm but cons is that it neglect data set in query click bipartite graph and only take in account query from query log.

In [12], authors suggested, query suggestion using hitting time on the query click bipartite graph. Query suggestion also use page ranking[18] and HITS [17] algorithm. Page ranking is a value which represents how significant page is.

2.3 Click through Data

This is used to reduce query result. In [13], approach is introduced to learn retrieval functions by analyzing which link user has clicked. for that, it support Vector Machine approach. Web search log files are used to minimize the time to get result using cluster.

3 HEAT DIFFUSION MODEL

This section introduce different Heat Diffusion models. Basically heat diffusion is a physical phenomenon in which heat flows from high temperature to the low temperature. Heat diffusion technique are applied to domain such as classification and dimensionality reduction problems[19]. In [15], discrete diffusion kernel for categorized data and result showed that it is well for hypercube data. In [16], diffusion ranking algorithm proposed, it is used by heat diffusion process. In this paper, heat diffusion model is used to find similarity information widen on web graphs.

There are two approach of Diffusion, one is heat diffusion on undirected graph and other diffusion on directed graph.

These approach can be used for different application like to decrease web spamming of website or to get similar web pages on net. This paper presents method to find user based query suggestions based on heat diffusion.

The process of persuades others by people is similar to heat diffusion model. The heat flood through a geometric manifold with initial conditions that can be described by following equation:

$$\frac{\partial f(x,t)}{\partial t} - \Delta f(x,t) = 0,$$

$$f(x,0) = f_o(x) \qquad \text{.....(1)}$$

where f(x, t) is the temperature at the location x at time t, start with an initial distribution $f_o(x)$ at the time zero and Δf is the Laplace- Beltrami operator on a function f. sometimes, it is very difficult to represent web a geometry with the known dimension. This motivate to examine the heat flow on a graph. The graph is considered as an approximation to the actual manifold and so the heat flow on the graph is considered as a rough calculation to the heat flow on the manifold.

3.1 Diffusion on undirected graph

Consider an undirected graph G=(V,E), where V is a vertex set, and V = { v1, v2. . . v_n }. E = {(v_i, v_j) there is an edge from v1 to v2} is the set of all edges. The edge (v_1, v_2) is considered as a pipe that connects nodes v_i and v_j. The value fi (t) describes the heat at node vi at time t, beginning from an initial distribution of heat given by f_i (0) at time zero. f(t) denotes the vector consisting of f_i (t) . We construct our model as follows. Suppose, at time t, each node i receives an amount M (i, j, t, Δt) heat from its neighbor j during a period Δt. The heat M (i, j, t, Δt) should be proportional to the time period Δt and the heat difference f_j (t) – f_i (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, Δt) =α(fj(t) – fi(t))Δt, where α is the thermal conductivity-the heat diffusion coefficient. As a result, the heat difference at node i between time t + Δt and time t will be equal to the sum of the heat that it receives from all its neighbors. This is formulated as:

$$f_j(t+\Delta t) - f_i(t) / \Delta t = \alpha \sum_{j:(v_j,v_i)\in E} (f_j(t) - f_i(t)) \text{.....(2)}$$

where E is the set of edges. To find a closed form solution to Eq. (2), we express it in a matrix form:

$$f(t + \Delta t) - f(t) / \Delta t = \alpha H f(t) \text{..... (3)}$$

Solving this differential equation, we have:

$$f(t) = e^{\alpha t} f(0) \text{ .....(4)}$$

where d(v) denotes the degree of the node v, and $e^{\alpha tH}$ could be extended as:

$$e^{\alpha tH} = I + \alpha tH + (\alpha^2 t^2 / 2!)H^2 + (\alpha^3 t^3 / 3!)H^3 \text{ ..(5)}$$

The matrix $e^{\alpha tH}$ is called the diffusion kernel in the sense that the heat diffusion process continues infinitely many times from the initial heat diffusion. Example:

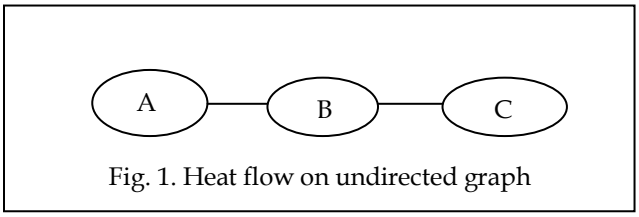

Fig. 1. Heat flow on undirected graph

Suppose, 5 unit of heat is given to node A, then as per diffusion process it will diffuse that heat to nearest node that is to node B, node C will get minimum heat as compared to node B because it far away from node A [1].

3.1 Thermal Conductivity

α plays an significant role in the enhancement of diffusion process. α is the thermal conductivity, i.e., the heat diffusion coefficient. If it has a high value, heat will diffuse very speedily. Otherwise, heat will diffuse gradually. In the severe case, if it is infinitely large, then heat will diffuse from one node to other nodes immediately.

4 PROPOSED SYSTEM

In [1], Authors presented module which purely based on Diffusion process which can be used to recommend the information. To consider personalized behavior of user in the form of profile we can extend concept of [1] to get better suggestions. Here in this proposed system, new user has to register on system, after login, system will store user history in log file with respect to session. By diffusion process we will get highest rank queries as suggestion but if any user hit some links then record of hitting links will be stored in log file. After getting highest rank values by diffusion process, our system compute user profile hitting record with diffusion process result, then we found out similarity between user query and recorded profiles of previous users hitting history finally most similar query will be recommended to user.

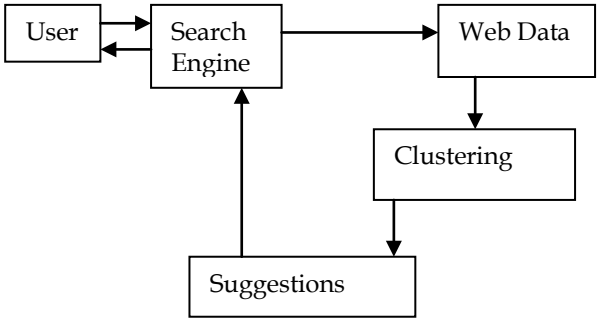

Fig. 2. Query suggestion system

5 PROPOSED ALGORITHM

In this paper, we are suggesting query using heat diffusion processes. Initially, we choose queries as the seeds for heat diffusion, denoted by a set S_k, and give a certain amount of heat h0 to each individual. At time zero of the heat diffusion process, we set $f_i(0) = h0$, where i ª Sk. As time elapses, the heat will diffuse through the sub graph. If the amount of heat of individual i at time t is greater than or equal to a threshold è, this individual i will be considered as having been successfully influenced by others heat, and will recommend.

6.1 User base Query Suggestion Algorithm:

Step-1: A bipartite graph G (V, E) consist of vertices as a set of URL or Query and edges as hyperlink. Directed edges are weighted using method introduced in [1].
Step-2: A query q in V, a sub graph is constructed using DFS (Depth First Search) in graph G. (We can decide sub graph nodes limit).
Step-3: Set thermal conductivity =1, and initial heat value of query f (0) =1. Start diffusion process using $f(1) = e^{\alpha t} f(0)$.
Step-4: Get highest value of queries k in vector f (1).
Step-5: Keep the hit record of user in p (Query, URL). Find similarity between queries using Cosine algorithm.
Step-6: Add vector f (1) with hit values of specified user.
Step-7: Output the highest value queries to users.

6.2 Results:

We have used AOL database for testing this Personalized Recommender system since it is freely available. We have compared our result with algorithm presented in [1]. Thermal conductivity value is set to 1(we can choose any value between 0 to 1 but for 1 it is giving better result). To compare our algorithm with other models we create set of 300 queries as the testing quires. From result our algorithm recommends queries which are exactly similar to test queries and semantically related queries. For a illustration, if test query is "Clinton" then system is recommending "Clintonct.com", "Hillary Clinton", "Bill Clinton" and "White house". Here the result semantically related to given query, addition to this if specific user hit any related query from result then that is recorded, this step continues for all user and if any other user hit similar query then base on cosine similarity from history highest rank query recommended to user. As data set is different from the data set search engines use so some time it is difficult to evaluate result quantitatively.

6.3 Impact of α:

This is very important parameter in our system. It control how fast heat will spread on the graph. We can observe that the best situation of á is 1. If we increase á then heat will spread faster but disadvantage is that unnecessary node will also get heat so unwanted link will get higher position in result so á value should be minimum and at 1 it is giving better result.

6.4 Impact of graph size:

Size of graph on web is very large because of this reason extract sub graph from original big graph. We have use size ofsub graph 500, 2000, 5000 and 10000. Due to diffusion process and personalized approach we get exactly similar queries. For 5000 and 10000 there is no change in result as node which are far away from initial node is not relevant. If size of sub graph is bigger than processing time increase.

6.5 Time complexity:

As Size of Web information is very large , the graph built upon it can become extremely large then complexity will become O(PM) which is large. Here P is any integer and M is no nodes in graph. To overcome this, we extract sub graph and it is constructed using DFS by this time complexity reduces.

6.6 Conclusion:

In this paper, a new approach to suggest queries on large web graph data is presented. Suggestions produced from user based suggestions are very similar to user queries. Practical analysis on very large scale data is showing good result of this framework.

REFERENCES

[1] JHao Ma, Irwin King, Senior Member, IEEE, and Michael Rung Tsong Lyu, Fellow, IEEE "Mining Web Graphs for Recommendations" Knowledge and Data Engineering, IEEE Transactions on)Volume: 24, Issue:) ISSN 1041-4347, June 2012.

[2] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google News Personalization: Scalable online collobrative filtering" 16th Inte'l Conf. world wide web, 271-280,2007.

[3] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Colloborative filtering recommender system" ACM Trans. Information system, vol. 22, no. 1, 5-53, 2004.

[4] G. Linden, B. Smith, and J. York, " Amazon.com recommendation: item to item collaborative filtering", IEEE Internet computing Vol 7 ,76-80,2003.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. riedl, "Grouplens: An open Architecture for collaborative filtering of Netnews", ACM conf. 1994,

[6] M. Deshpande and G. Karypis, "Item-Based Top-n Recommendation," ACM Trans. Information Systems, vol. 22, no. 1, pp. 143-177, 2004.

[7] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), 1998.

[8] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 7- 14, 2007.

[9] R. Kraft and J. Zien, "Mining Anchor Text for Query Refinement," WWW '04: Proc 13th Int'l Conf. World Wide Web, pp. 666-674, 2004.

[10] R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Current Q. Mei, D. Zhou, and K. Church, "Query Suggestion Using Hitting Time," CIKM '08: Proc. 17th ACM Conf. Information and Knowledge Management, pp. 469-477, 2008.

[11] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," KDD '02: Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 133-142, 2002.

[12] X. Wang and C. Zhai, "Learn from Web Search Logs to Organize Search Results," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf.

[13] R.I. Kondor and J.D. Lafferty, "Diffusion Kernels on Graphs and Other Discrete Input Spaces," ICML '02: Proc. 19th Int'l Conf. Machine Learning, pp. 315-322, 2002.
[14] H. Yang, I. King, and M.R. Lyu, "DiffusionRank: A Possible Penicillin for Web Spamming," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 431-438, 2007.
[15] Jon M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment." Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
[16] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107–117, 1998.
[17] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15 (6):1373–1396, 2003.
[18] J. Brown and P. Reinegen. Social ties and word-ofmouthreferral behavior. Journal of Consumer Research, 14 (3):350–362, 1987.
[19] N. Craswell and M. Szummer, "Random Walks on the Click Graph," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 239-246, 2007.