Analysis of Implicit Relationships on Wikipedia through Elucidatory Objects

S. Reddy Mubaraq
M. tech student
Department of CSE
MITS, Madanapalle.

P. Rajarajeswari
Assoc. Professor
Department of CSE
MITS, Madanapalle

D. Vasumathi, PhD
Professor
Department of CSE
JNTU, Hyderabad

ABSTRACT

Wikipedia is a better option for a client to gain information of particular objects than other search engines. Previously, cohesion-based technique is used for analysis of implicit and explicit relationships between two objects but this technique is inadequate to analyze the relationships because it uses one or two concepts of connectivity, distance and co-citation. We proposed a method for analysis of relationships between objects by using generalized flow based technique which replicates all the concepts. We prove that our process is the most suitable for measuring the strength of relationships between source and destination objects by using elucidatory objects as well as grade the target objects by its strength. Here we quantitatively extract the elucidatory objects.

Keywords: Elucidatory objects, generalized flow, link analysis, Wikipedia mining, relationships

1. INTRODUCTION

Wikipedia is a better option to gain information of particular objects and relationships among several items such as places, peoples etc. than other search engines. Users want to know the relationships between objects. For example, a user want to know which countries are robustly related to petroleum and why one country has a stronger relationship to petroleum than other country. Therefore distinctive search engines neither compute nor elucidate the power of relationships, because there are two kinds of relationships they are implicit and explicit relationships. Explicit relationships is a single connection between pages a user can easily understand by reading of two pages ,but implicit relationships represents link structure contains multiple links and pages so user face difficulty while understand it. Therefore we introduced a method to analyze the implicit relationships in Wikipedia through *elucidatory objects* those are quantitatively extracted
Consider information system the same as a directed graph (V, E) where V and E are vertices and edges. Here we consider Wikipedia is an information network whose vertices and edges are pages, links between pages. To measure the relationship strength there are several methods have been proposed. Previous methods are unsuitable because which reflects only one or two concepts of connectivity, distance and co-citation. A user faces difficulties to understand implicit relationships.

We proposed a technique for analysis of implicit relationships using generalized flow based method which reflect all the three conditions connectivity, co-citation and distance, here we calculate the strength of relationships rather than similarity as discussed in [3] for example petroleum is not similar to USA ,although there exist some relationship among them. Generalized maximum flow introduced gain for each edge on the information network .The rate of flow is send along the edge and the flow is multiple by gain of the edging. The gain function utilizes the category grouping structure in Wikipedia to measure relationships and grade the objects by its strength .By experiments we compute that ranks obtain by our method are closest to the ranks given by humans by compared to ranks obtain by GSD proposed by [5] and PFIBF [2], CFEC [1].

2. RELATED WORK

We discuss some existing methods for measuring strength of relationships on Wikipedia. Erdos number [11] introduced numbering method to compute shortest path or distance from source to destination objects in order to compute strength of relationships but this is inadequate to measure strength of implicit relationships because it does not estimate the connectivity between objects. M.Yazadani[7],[8] proposed THT method to compute power of relationship by using average length of the paths connecting two items, smaller distance represent a larger similarity but this method also does not guess connectivity among two objects.

Connectivity exists in information network between vertex to vertex, objects has a stronger relationship if the connectivity from source to destination is large .The connectivity is equal to value of maximum flow on the edges but maximum flow does not estimate the distance since the quantity of flow along a path is free of the path length. W.Lu.J[9] projected a technique for calculating the power of relationships using maximum flow but this is difficult to measure distance between objects, the maximum flow value does not decreases the distance by setting only capacities on edges. So, we introduced generalized maximum flow by using gain function, it decreases if the distances become larger. H.D.White [10] proposed co-citation method; objects have strong relationships if the sum of items connected by both the two items is more.

Y.Koren [1] proposed CFEC technique uses cohesion based method to calculate strength of relationships by counting all paths among two things and its value is significantly increases if a popular object comes along the path that is an object is linked from or to various objects but this method does not estimate strength of implicit relationship effectively. Nakayama [2] proposed PFIBF method, it also uses cohesion based method, instead of counting every paths it counts path whose distance end to end is two. It cannot differentiate a path containing cycle from path containing no cycle; it counts some paths multiple times. Therefore, it is inappropriate to measure 3-hop implicit relationship.

J.Gracia [3] proposed a method to measure a relationship rather than similarity. For example, petroleum is not related to USA although there is a relationship exists among them. R.L.Cilibrasi [5] proposed co-occurrence method it estimates the power of relationships among two terms by including of WebPages containing both terms, but this is ineffective to measure the strength of implicit relationships. F.M.Suchanek [6] estimate the relationship based on semantic of objects rather than similarity. Several search engines has been used to measure relationship between objects from web or Wikipedia such as 'is called', 'type'. The above methods are unsuitable for measuring the power of relationship on Wikipedia because they reflect only one or two concepts of connectivity, distance and co-citation methods.

3. PROPOSED WORK

We propose a new technique to measure strength of relationship between things on Wikipedia is generalized flow based method. Let general network G= (V, E, s, t, μ, γ) be information network (V, E) where source s Є V and destination t Є V, μ is the capacity and γ is gain. The edge has a gain γ(e) > 0 ,the flow value send along the edge is multiple by γ(e) .Let f(e)≥0 is the flow of edge, and μ(e)≥0 is the capacity of edge .The capacity control f(e)≤μ(e) have to hold for each edge. For example, Fig.1 Shows the generalized maximum flow from source s to destination t, one unit of flow is send from source s to v1, i.e. f(s,v1)=1 , the sum flow is multiplied by γ(s,v1) when the flows occur at v1, 0.8 units occur at v1 like this 0.512 units occur at target t. The capacity control for edge e should hold before the gain is multiplied .f(s, v1) =1 ≤ μ(s, v1) have to grip. A better flow value represents stronger relationships.

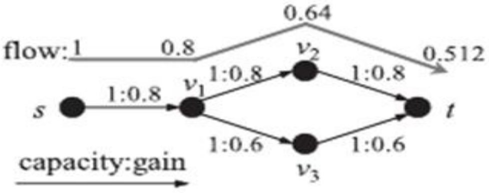

Figure1: A Generalized flow from s to t.

A generalized flow based technique reflects all three concept they are connectivity, distance and co citation. Distance, a shortest path denotes a stronger relation. We put γ(e)<1 for each edge e then a flow significantly decrease along a lengthy path. Connectivity, a stronger relationship is denoted by several disjoint paths form the source to target objects. Co-citation, actually flow sent form source to destination so the flow rarely uses an edge whose direction is reverse to that i.e. from the source to target. We require using both directions to guess the co citation of two items. Consider a relationship among object s to t, Figure2a.object u is co-cited by source and target, this co citation is denoted by two edges(s, u) and (t, u).Though, we are incapable to send a flow from source to target along the two edges, if we not reverse the direction of the edge (t, u) to (u, t).For that reason, we assemble the double network by putting a reversed edge to the existed edge in G. For example, Fig2 b represents a doubled network for Fig. 2a.

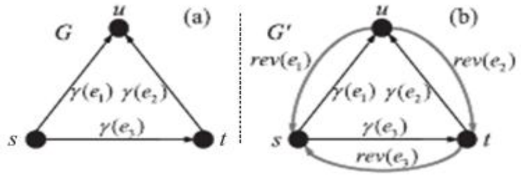

Figure 2. A Double Network Graph.

3.1 Gain Function

In order to establish the gain utility, we think what kind of explicit relationships are significant in constitute an implicit relationship. Assume an American politician A0 is trying to send a message to an India politician I0 in the real life. A0 does not have any explicit relationship to I0, but another American politician A1 and a Japan politician J0 has a relevant explicit relationship to I0. In this case, A0 ask A1 rather than Jo, to help transfer the message to I0 because A0 and A1 are belonging to same group i.e. American politician. Hence the explicit relationship among A1 and I0 as primarily significant in constitute the relationship among A0 and I0.A gain function utilizing the category grouping technique to estimate gain value from s to t. A cluster is a set of related or correlated objects such as Japan politicians, and Indian politicians' .For analyzing the implicit relationship we need to analyze explicit relationships between sources to destination objects.

Implicit relationship constitute of many significant explicit relationships are strong, in a network a path is collected of edges by huge gains can add to the value of flow in order to identify important explicit relationships to measure strength. So, we need to build group of objects in Wikipedia based on their category we divided and grouped the objects like class, sub class, for example category C object in Wikipedia has a set of sub categories of C. In Wikipedia, a page is attached to several categories. Suppose to find a relationship among s and t objects, it is simply to use every one of the categories billed to source or destination objects respectively .But a number of categories contain too a lot of unrelated object pages. For example, "Living People" category contains many people's completely independent to each others, such categories are inappropriate for grouping correlated objects. So we manually removed such categories.

3.2 Methodologies in Proposed System

We analyze the implicit relationships in Wikipedia as represented in the following method

1. Construct an information network G= (V, E, s, t, μ.γ) contain source and target to calculate the strength of relationship from s and t by the value of flow f (e)>0.

2. We introduce gain γ (e) >0 for each edge on the network

3. The rate of flow f send along the edge is multiplied by gain of the edge γ (e) >0.

Let μ (e)>0 is the edge capacity.

And the very edge must hold the capacity restriction f (e) ≤μ (e).

4. Gain utilizes the category grouping formation in Wikipedia.

Category grouping uses distance function d (e)

Gain for edge e depending on d (e) by two parameters α and β as

$$\gamma\ (e) = \alpha * \beta^{\ d(e)}\ ,0<\alpha<1.0<\beta\leq 1$$

5. Evaluate the rank by score given by users and can measure the power of relationship.

3.2.1 Evaluation of Ranking

A good evaluation method requires human subjects to measure relationships. For example, each of the participants read Wikipedia pages equivalent to or associated to the source and the target objects. Each user gives a numeral score among 0 to 10 separately to the others as the strength of relationship; a larger score represents a stronger relationship. We acquire ranks based on to the average of the scores specified by 10 users. We can view the rank by entering keyword. For example, Table 1 shows score given by users for countries for petroleum

Table1. Scores of Countries for Petroleum

Keyword	Country	Score
Petroleum	USA	7
Petroleum	China	6
Petroleum	Canada	5
Petroleum	Iran	9
Petroleum	Iraq	8

According to the score given by users we can estimate the ranks as shown in the Table 2.

Table 2. Ranking of Countries for Petroleum

Keyword	Country	Rank
Petroleum	Iran	1
Petroleum	Iraq	2
Petroleum	USA	3
Petroleum	China	4
Petroleum	Canada	5

4. CONCLUSION

We are planned a new technique for measuring the implicit relationships between objects by using maximum generalized flow which reflects all the three concepts co-citation, connectivity and distance. This technique uses Gain function by utilizing category grouping technique to measure the strength of relationships. In this method, we quantitatively evaluate the Elucidatory objects, which help to understand relationships. We plan to develop a tool for totally considerate the relationship by using elucidatory objects.

5. REFERENCES

[1] Y. Koren, S.C. North, and C. Volinsky, “Measuring and Extracting Proximity in Networks”, Proc. 12th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 245-255, 2006.

[2] M. Ito, K. Nakayama, T. Hara, and S. Nishio, “Association Thesaurus Construction Methods Based on Link Co-Occurrence Analysis for Wikipedia”, Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), pp. 817-826, 2008

[3] J. Gracia and E. Mena, “Web-Based Measure of Semantic Relatedness,” Proc. Ninth Int’l Conf. Web Information Systems Eng.(WISE), pp. 136-150, 2008.

[4] K.D. Wayne, “Generalized Maximum Flow Algorithm,” PhD dissertation, Cornell Univ., New York, Jan. 1999.

[5] R.L. Cilibrasi and P.M.B. Vitanyi, “The Google Similarity Distance,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.

[6] F.M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A Core of Semantic Knowledge,” Proc. 16th Int’l Conf. World Wide Web Conf. (WWW), pp. 697-706, 2007.

[7] M. Yazdani and A. Popescu-Belis, “A Random Walk Framework to Compute Textual Semantic Similarity: A Unified Model for Three Benchmark Tasks,” Proc. IEEE Fourth Int’l Conf. Semantic Computing (ICSC), pp. 424-429, 2010.

[8] P. Sarkar and A.W. Moore, “A Tractable Approach to Finding Closest Truncated-Commute-Time Neighbors in Large Graphs,” Proc. 23rd Conf. Uncertainty in Artificial Intelligence (UAI), 2007

[9] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang, “Node Similarity in the Citation Graph,” Knowledge and Information Systems, vol. 11, no. 1, pp. 105-129, 2006.

[10] H.D. White and B.C. Griffith, “Author Co-citation: A Literature Measure of Intellectual Structure,” J. Am. Soc. Information Science and Technology, vol. 32, no. 3, pp. 163-171, May 1981.

[11] “The Erdos Number Project”, http:// www.oakland.edu/emp/, 2012

[12] Xinpeng_Zhang “A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia”, Dept. of Social Inf., Kyoto Univ., Kyoto, Japan; Asano.Y; Yoshikawa.M