ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia

G. Kumar
Associate Professor
Department of Computer Science Engineering,
Lords Institute of Engineering and Technology.

Ayesha Siddiqua
Department of Computer Science Engineering
Lords Institute of Engineering and Technology.

Abstract:

We focus on measuring relationships between pairs of objects in Wikipedia whose pages can be regarded as individual objects. Two kinds of relationships between two objects exist: in Wikipedia, an explicit relationship is represented by a single link between the two pages for the objects, and an implicit relationship is represented by a link structure containing the two pages. Some of the previously proposed methods for measuring relationships are cohesion-based methods, which underestimate objects having high degrees, although such objects could be important in constituting relationships in Wikipedia. The other methods are inadequate for measuring implicit relationships because they use only one or two of the following three important factors: distance, connectivity, and co-citation. We propose a new method using a generalized maximum flow which reflects all the three factors and does not underestimate objects having high degree. We confirm through experiments that our method can measure the strength of a relationship more appropriately than these previously proposed methods do. Another remarkable aspect of our method is mining elucidatory objects, that is, objects constituting a relationship. We explain that mining elucidatory objects would open a novel way to deeply understand a relationship.

Keywords:

Link analysis, generalized flow, Wikipedia mining, relationship.

1.INTRODUCTION:

1.1.Purpose:

Searching webpages containing a keyword has grown in this decade, while knowledge search has recently been researched to obtain knowledge of a single object and relationships between multiple objects, such as humans, places or events.

Searching knowledge of objects using Wikipedia is one of the hottest topics in the field of knowledge search. In Wikipedia, the knowledge of an object is gathered in a single page updated constantly by a number of volunteers. Wikipedia also covers objects in a number of categories, such as people, science, geography, politic, and history. Therefore, searching Wikipedia is usually a better choice for a user to obtain knowledge of a single object than typical search engines.A user also might desire to discover a relationship between two objects. For example, a user might desire to know which countries are strongly related to petroleum, or to know why one country has a stronger relationship to petroleum than another country. Typical keyword search engines can neither measure nor explain the strength of a relationship. The main issue for measuring relationships arises from the fact that two kinds of relationships exist: “explicit relationships” and “implicit relationships.” In Wikipedia, an explicit relationship is represented by a link. For example, an explicit relationship between petroleum and Gulf of Mexico might be represented by a link from page “Petroleum” to page “Gulf of Mexico.” A user could understand its meaning by reading the text “Oil filed in Gulf of Mexico is a major petroleum producer” surrounding the anchor text “Gulf of Mexico” on page “Petroleum.” An implicit relationship is represented by multiple links and pages. For example, an implicit relationship between petroleum and the USA might be represented by links and pages depicted in Fig. 1. For an implicit relationship between two objects, the objects, except the two objects, constituting the relationship is named elucidatory objects because such objects enable us to explain the relationship. For the example described above, “Gulf of Mexico” is one of the elucidatory objects. The user can understand an explicit relationship between two objects easily by reading the pages for the two objects in Wikipedia. By contrast, it is difficult for the user to discover an implicit relationship and elucidatory objects without investigating a number of pages and links. Therefore, it is an interesting problem to measure and explain the strength of an implicit relationship between two objects in Wikipedia.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

1.2.Scope:

Several methods have been proposed for measuring the strength of a relationship between two objects on an information network (V ;E) a directed graph where V is a set of objects; an edge (u; v) 2 E exists if and only if object u 2 V has an explicit relationship to v 2 V . We can define a Wikipedia information network whose vertices are pages ofWikipedia and whose edges are links between pages. Previously proposed methods then can be applied to Wikipedia by using a Wikipedia information network. A concept "cohesion," exists for measuring the strength of an implicit relationship.

CFEC proposed by Koren et al. [1] andPFIBF proposed by Nakayama et al. [2], [3] are based on cohesion. We do not adopt the idea of cohesion based methods, because they always punish objects having high degrees although such objects could be important to some relationships in Wikipedia, as we will explain in Section 2.2.Other previously proposed methods use only one or two of the three representative concepts for measuring a relationship: distance, connectivity, and cogitation, although all the concepts are important factors for implicit relationships. Using all the three concepts together would be appropriate for measuring an implicit relationship and mining elucidatory objects.

1.3Motivation;

We propose a new method for measuring a relationship on Wikipedia by reflecting all the three concepts: distance, connectivity, and cocitation. We measure relationships rather than similarities. As discussed in [4], relationship is a more general concept than similarity. For example, it is hard to say petroleum is similar to USA, but a relationship exists between petroleum and the USA. Our method uses a "generalized maximum flow" [5], [6] on an information network to compute the strength of a relationship from object s to object t using the value of the flow whose source is s and destination is t. It introduces a gain for every edge on the network. The value of a flow sent along an edge is multiplied by the gain of the edge. Assignment of the gain to each edge is important for measuring a relationship using a generalized maximum flow. We propose a heuristic gain function utilizing the category structure in Wikipedia. We confirm through experiments that the gain function is sufficient to measure relationships appropriately.

1.3.1 Definitions:

The main contributions of this paper are as follows:

1.A detailed and methodical survey of related work for measuring relationships or similarities
2.A new method using generalized maximum flow for measuring the strength of a relationship between two objects on Wikipedia, which reflects the three concepts: distance, connectivity, and cocitation
3.Experiments on Wikipedia showing that our method is the most appropriate one
4.Case studies of mining elucidatory objects for deeply understanding a relationship

System Model:

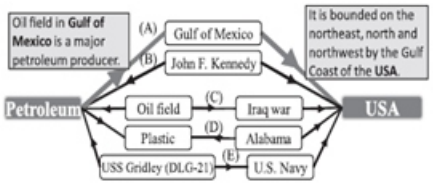

Fig. 1. Explaining the relationship between Petroleum and the USA.

2.RELATED WORK:

We aim to measure implicit relationships between two objects on the Wikipedia information network. Although relationship is a more general concept than similarity, we discuss existing methods for measuring either relationships or similarities, in this section.

2.1 Distance, Connectivity, Cocitation:

The Erdo¨s number [10] used by mathematicians is based on distance and coauthorships. The legendary mathematician Paul Erdo¨s has a number 0, and the people who cowrote a paper with Erdo¨s have a number 1; the people who cowrote a paper with a person with a number 1 have a number 2, and so on. The Erdo¨s number is the distance, or the length of the shortest path, from a person to Erdo¨ s on an information network whose edge represents coauthorship; a shorter path represents a stronger relationship. However, the Erdo¨s number is inadequate to represent the implicit relationship between a person and Erdo¨s because the number does not estimate the connectivity between them.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The hitting time [11], [12] from vertex s to vertex t is defined as the expected number of steps in a random walk starting from s before t is visited for the first time. Actually, the hitting time from s to t in a network represents the average length of all the paths connecting s and t. Sarkar and Moore [12] proposed "Truncated Hitting Time" (THT) to compute the average length of paths connecting two vertices whose length are at most Lmax only. A smaller distance represents a larger similarity. THT does not estimate the connectivity between two vertices. For example, suppose only m 1 vertex disjoint paths of length k connect s to t. THT computes the distance from s to t to be k for any m is greater than or equal to 1. The connectivity [5], more precisely the vertex connectivity, from vertex s to vertex t on a network is the minimum number of vertices such that no path exists from s to t if the vertices are removed. s has a strong relationship to t if the connectivity from s to t is large. The connectivity from s to t is equal to the value of a maximum flow from s to t, where every edge and vertex has capacity 1. However, the distance cannot be estimated by the maximum flow because the amount of a flow along a path is independent of the path length. Lu et al. [13] proposed a method for computing the strength of a relationship using a maximum flow. They tried to estimate the distance between two objects using a maximum flow by setting edge capacities. However, the value of a maximum flow does not necessarily decrease by setting only capacities even if the distance becomes larger.

Therefore, their method cannot estimate the distance successfully by the value of the maximum flow. Instead of setting capacities, we use a generalized maximum flow by setting every gain to a value less than one. Therefore, the value of a maximum flow in our method decreases if the distance becomes longer. co-citation-based methods in this paper. Milne and Witten [15] also proposed methods measuring relationships between objects in Wikipedia using Wikipedia links based on cocitation. Cocitation-based methods cannot deal with a typical implicit relationship, such as "person w is regarded as a friend by person v who is regarded as a friend by person u." This relationship is represented by the path formed by two edges ðu; vÞ and ðv; wÞ. In contrast, cocitationbased methods can deal with two edges going into the same vertex, such as edges ðu; vÞ and ðw; vÞ. Therefore, cocitationbased methods are inadequate for measuring an implicit relationship. Furthermore, cocitation-based methods cannot deal with 3-hop implicit relationships defined in Section 1 because these methods estimate only relationships represented by paths formed by two edges, as explained above.

2.2 Cohesion:

In the field of social network analysis, cohesion-based methods are known to measure the strength of a relationship by counting all paths between two objects. The original cohesion was proposed by Hubbell [17], Katz [18], Wasserman and K. Faust [19]. It has a property that its value greatly increases if a popular object, an object linked from or to many objects, exists. As pointed out in other researches [20], [1], [2], this property is a defect for measuring the strength of a relationship. Several cohesion-based methods, such as PFIBF and CFEC explained below, were proposed to dissolve this property.

Nakayama et al. [3], [2] proposed a cohesion-based method named PFIBF. Instead of enumerating all paths, PFIBF approximately counts paths whose length is at most k > 0 using the kth power of the adjacency matrix of an information network. However, in the kth power of the matrix, a path containing a cycle whose length is at most k 1 would appear. PFIBF cannot distinguish a path containing a cycle from a path containing no cycle. For example, if k > or = 3 and two edges (u, v) and (u, v) exist, then PFIBF counts both path (u; v) and path (u; v; u; v) containing a cycle (u; v; u). Consequently, PFIBF has a property that it estimates a single path, e.g., (u; v) in the above example, for multiple times. The length of a cycle is at least two. No path containing a cycle appears if k < or = 2. In fact, PFIBF usually sets k = 2.

Therefore, PFIBF is inappropriate for measuring a 3-hop implicit relationship. However, a number of 3-hop implicit relationships exist in Wikipedia. The "Effective Conductance" (EC) proposed by Doyle and Snell [21] is a cohesion-based method also. EC has the same drawback as PFIBF: it counts a path containing a cycle redundantly. Koren et al. [1] proposed cycle-free effective conductance (CFEC) based on EC by solving this drawback. For a positive integer k, CFEC enumerates only the k-shortest paths between s and t, instead of computing all paths. CFEC does not use a path containing a cycle, although it cannot count all paths.

2.2.1 Popular Objects in Wikipedia:

PFIBF and CFEC underestimate a popular object. CFEC defines the weight of path p=(s= v1, v2, . .vt =t) from s to t as

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

$$w_{sum}(v_1) \cdot \prod_{i=1}^{\ell-1} \frac{w(v_i, v_{i+1})}{w_{sum}(v_i)},$$

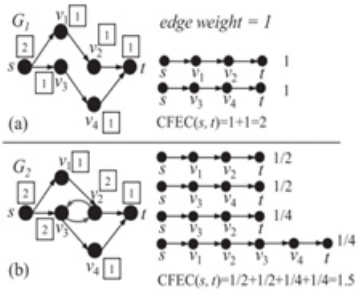

Fig. 2. CFEC on two networks.

Fig. 2 depicts two networks and all the paths between s and t. For simplicity, let the weight of every edge be one. The wsum of each vertex is written in the rectangle near the vertex. The weight of each path is presented at the right side of the path Let us consider the implicit relationship between the "Rice" and "Koizumi" depicted in Fig. 3. Bush was the President of the USA, and Rice worked under the administration of Bush. Koizumi and Olmert were the prime ministers of Japan and Israel, respectively. The numbers of objects linked from or linking to "Bush" and "Olmert" are 1,265 and 289, respectively, in Wikipedia.

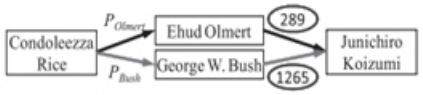

Fig. 3. A Relationship between Rice and Koizumi.

3 METHOD FOR MEASURING RELATIONSHIPS USING GENERALIZED FLOW:

Here we propose a generalized maximum flow-based method which reflects all the three concepts and does not underestimates popular objects, in order to measure relationships on Wikipedia appropriately.

3.1 Generalized Maximum Flow:

The generalized maximum flow problem is identical to the classical maximum flow problem except that every edge e has a gain ɣ(e) > 0; the value of a flow sent along edge e is multiplied by ɣ(e). Let f(e) ≥ 0 be the flow f on edge e, and μ(e)≥ 0 be the capacity of edge e. The capacity constraint f(e) ≤ μ(e) must hold for every edge e. The goal of the problem is to send a flow emanating from the source vertex s into the destination vertex t to the greatest extent possible, subject to the capacity constraints.

Let generalized network G =(V, E, s, t, μ,ɣ) be information network (V,E) with the source s belongs to V , the destination t belongs to V , the capacity , and the gain ɣ. Fig. 4 depicts an example of a generalized maximum flow on a generalized network. One unit of flow is sent from the source s to v1, i.e., f(s; v1) = 1, the amount of the flow is multiplied by ɣ (s; v1) when the flow arrives at v1. Consequently, only 0.8 units arrive at v1. In this way, only 0.512 units arrive at the destination t. The capacity constraint for edge e = (u, v) must hold before the gain is multiplied. F(s; v1) = 1 ≤ μ(s, v1) must hold, for example.

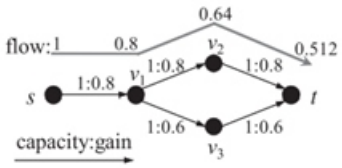

Fig. 4. Generalized maximum flow.

We discuss the cocitation at last. A flow emanates from the source into the destination, and therefore the flow seldom uses an edge whose direction is opposite that from the source to the destination. On the other hand, we require use of both directions to estimate the cocitation of two objects. We consider the relationship between two objects s and t in the network presented in Fig. 5a. Object u is cocited by s and t. This cocitation is represented by two edges (s; u) and (t; u).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

TABLE 1
Rankings of Persons

Source	Destinations	Human	Ours 3 hop	GSD	PFIBF 2 hop	CFEC 3 hop k=1000 o1	CFEC 3 hop k=1000 og	CFEC 3 hop k=1000 d1	CFEC 3 hop k=1000 dg	THT d1 3 hop L_{max}=3
Richard Nixon	Henry Kissinger	1 (8.5)	1 (2.31)	1 (0.26)	1 (7.94)	1 (1.24)	1 (0.91)	1 (1.89)	1 (1.12)	1 (2.98712)
	Zhou Enlai	2 (5.8)	2 (1.27)	2 (0.34)	2 (3.19)	2 (1.06)	2 (0.82)	2 (1.40)	2 (0.91)	3 (2.99098)
	Nguyen Van Thieu	3 (3.4)	3 (1.06)	2 (0.34)	3 (1.80)	3 (1.04)	3 (0.81)	3 (1.15)	3 (0.85)	4 (2.99173)
	Wallis Simpson	4 (2.0)	4 (0.80)	4 (0.49)	4 (0.45)	4 (1.00)	4 (0.80)	4 (1.02)	4 (0.81)	2 (2.98729)
Nobutaka Machimura	Yasuo Fukuda	1 (8.4)	1 (1.67)	1 (0.19)	1 (9.39)	1 (1.38)	1 (0.96)	1 (1.57)	1 (1.01)	1 (2.97889)
	Condoleezza Rice	2 (5.3)	2 (0.82)	2 (0.41)	3 (0.75)	3 (0.01)	3 (0.00)	2 (1.04)	2 (0.81)	2 (2.98354)
	George W. Bush	3 (4.1)	3 (0.64)	4 (0.56)	2 (1.14)	2 (0.02)	2 (0.01)	3 (0.09)	3 (0.03)	3 (2.99704)
	Hillary Clinton	4 (2.6)	4 (0.61)	3 (0.48)	4 (0.27)	4 (0.00)	3 (0.00)	4 (0.02)	4 (0.01)	4 (2.99886)
Donald Henry Rumsfeld	Dick Cheney	1 (7.7)	1 (2.05)	1 (0.17)	2 (3.38)	2 (1.08)	2 (0.84)	2 (1.25)	2 (0.90)	1 (2.96996)
	Condoleezza Rice	2 (6.9)	2 (1.47)	2 (0.22)	3 (2.58)	4 (0.02)	4 (0.01)	3 (0.23)	3 (0.09)	3 (2.98412)
	Ronald Reagan	3 (5.5)	3 (1.07)	3 (0.35)	1 (3.47)	1 (1.20)	1 (0.89)	1 (1.35)	1 (0.96)	2 (2.97003)
	Junichiro Koizumi	4 (3.8)	4 (0.46)	4 (0.53)	4 (1.63)	3 (0.06)	3 (0.02)	4 (0.10)	4 (0.03)	4 (2.99659)
Junichiro Koizumi	Shinzo Abe	1 (9.1)	1 (5.30)	1 (0.18)	1 (29.6)	1 (1.97)	1 (1.14)	1 (3.72)	1 (1.72)	1 (2.98931)
	Donald Rumsfeld	2 (5.3)	2 (1.99)	2 (0.53)	2 (2.32)	3 (0.12)	3 (0.04)	**4 (0.098)**	3 (0.03)	3 (2.99916)
	Wen Jiabao	3 (4.5)	4 (1.66)	2 (0.53)	4 (2.00)	2 (1.03)	2 (0.81)	2 (1.14)	2 (0.84)	2 (2.99666)
	Condoleezza Rice	4 (4.1)	3 (1.83)	4 (0.55)	3 (2.17)	4 (0.06)	4 (0.01)	3 (0.103)	4 (0.03)	4 (2.99948)
Bill Clinton	Hillary Clinton	1 (9.5)	1 (2.68)	1 (0.27)	1 (7.59)	1 (1.36)	1 (0.95)	1 (2.01)	1 (1.21)	1 (2.98550)
	Keizo Obuchi	2 (4.7)	4 (1.08)	3 (0.46)	3 (2.29)	3 (0.07)	2 (0.03)	3 (0.30)	3 (0.08)	3 (2.99553)
	Junichiro Koizumi	3 (2.7)	3 (1.10)	2 (0.41)	2 (3.42)	2 (0.09)	3 (0.02)	2 (0.32)	2 (0.09)	2 (2.99513)
	Yasuo Fukuda	4 (2.3)	2 (1.17)	4 (0.58)	4 (1.79)	4 (0.02)	4 (0.00)	4 (0.11)	4 (0.03)	4 (2.99860)
Yasuo Fukuda	Takeo Fukuda	1 (9.7)	1 (4.04)	1 (0.16)	1 (11.7)	1 (2.12)	1 (1.20)	1 (2.04)	1 (1.20)	1 (2.99176)
	Tony Blair	2 (4.7)	3 (1.43)	4 (0.52)	3 (1.30)	3 (0.06)	3 (0.01)	4 (0.06)	4 (0.01)	4 (2.99943)
	Nicolas Sarkozy	3 (4.6)	2 (1.75)	2 (0.50)	2 (2.07)	2 (1.03)	2 (0.81)	2 (1.11)	2 (0.82)	2 (2.99518)
	Mamoru Mohri	4 (2.8)	4 (0.73)	2 (0.50)	4 (0.47)	4 (0.01)	4 (0.00)	3 (0.07)	3 (0.02)	3 (2.99886)
Kiichi Miyazawa	Noboru Takeshita	1 (8.4)	1 (3.71)	1 (0.09)	1 (12.1)	1 (1.49)	1 (0.96)	1 (1.85)	1 (1.10))	1 (2.98707)
	George H. W. Bush	2 (4.9)	2 (1.07)	4 (0.58)	3 (0.86)	3 (1.04)	3 (0.81)	3 (1.04)	3 (0.81)	3 (2.99022)
	Robert Rubin	3 (4.0)	4 (0.71)	2 (0.49)	4 (0.46)	4 (0.01)	4 (0.00)	4 (0.02)	4 (0.01)	4 (2.99779)
	Bill Clinton	4 (3.9)	3 (1.05)	2 (0.49)	2 (1.74)	2 (1.06)	2 (0.82)	2 (1.21)	2 (0.86)	2 (2.98931)
Yasuhiro Nakasone	Ronald Reagan	1 (8.5)	1 (1.83)	1 (0.40)	1 (4.98)	1 (1.40)	1 (0.92)	1 (1.53)	1 (0.97)	2 (2.99308)
	Chun Doo-hwan	2 (5.5)	3 (1.40)	3 (0.45)	3 (1.94)	2 (1.21)	2 (0.87)	2 (1.20)	2 (0.85)	3 (2.99408)
	Mikhail Gorbachev	3 (4.0)	2 (1.53)	2 (0.43)	2 (3.22)	4 (0.29)	4 (0.08)	4 (0.28)	4 (0.08)	4 (2.99725)
	Yuri Andropov	4 (3.5)	4 (1.07)	4 (0.51)	4 (0.80)	3 (1.05)	3 (0.82)	3 (1.06)	3 (0.82)	1 (2.99017)
Shigeru Yoshida	Douglas MacArthur	1 (8.3)	1 (2.22)	1 (0.40)	1 (7.23)	1 (1.38)	2 (0.93)	1 (1.58)	1 (0.97)	1 (2.99198)
	John Dulles	2 (5.4)	4 (1.14)	2 (0.47)	3 (1.69)	4 (0.04)	4 (0.01)	4 (0.08)	4 (0.03)	4 (2.99887)
	Harry S. Truman	3 (4.0)	2 (1.37)	4 (0.57)	2 (2.61)	3 (1.08)	3 (0.82)	2 (1.15)	2 (0.84)	3 (2.99311)
	Benito Mussolini	4 (3.3)	3 (1.17)	3 (0.56)	4 (1.59)	2 (1.10)	2 (0.83)	3 (1.08)	3 (0.82)	2 (2.99283)
Taro Aso	Shinzo Abe	1 (8.7)	1 (4.28)	1 (0.15)	1 (25.9)	1 (2.06)	1 (1.18)	1 (3.18)	1 (1.54)	1 (2.98775)
	Condoleezza Rice	2 (5.6)	4 (1.85)	4 (0.50)	4 (2.06)	4 (0.04)	4 (0.01)	4 (1.12)	4 (0.83)	4 (2.99529)
	George W. Bush	3 (4.4)	2 (2.12)	3 (0.48)	2 (4.90)	2 (1.20)	2 (0.86)	2 (1.45)	2 (0.93)	2 (2.99488)
	Kim Jong-il	4 (3.2)	3 (1.99)	2 (0.40)	3 (3.20)	3 (1.11)	3 (0.83)	3 (1.20)	3 (0.85)	3 (2.99512)

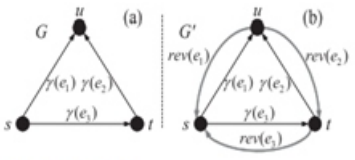

Fig. 5. A doubled network.

4 EXPERIMENTS AND EVALUATION:

4.1 Data Set and Environment:

Here, in our system, we perform experiments on a Japanese Wikipedia data set (20090513 snapshot). 27,380,912 links appear in all pages. We remove pages that are not corresponding to objects, such as each day, month, category, person list, and portal. Finally, we obtain 11,504,720 remaining links.

4.2 Evaluation of Rankings:

Table 1 presents the rankings for the 10 sources. For each source, the ranking and the average score obtained by human subjects are written in the column "Human;" an integer 1-4 is assigned as the ranking of the destination; a real number in parentheses is the score. Similarly, the ranking and the strength obtained by our method, GSD, PFIBF, the four methods of CFEC and THT, are written in the column "Ours," "GSD," "PFIBF," "CFEC," and "THT," respectively. "k hop" written behind the name of a method indicates that the method measures a relationship between source s and destination t on the network constructed using at most k hop links from s and t. Note that, GSD and THT use a smaller real number to represent a stronger relationship.

The shadowed cells for each method emphasize the difference between the ranking obtained by human subjects and that obtained by the method. Fig. 6 depicts the ratio r@deg of vertices having degree deg within each range in the 1,000 shortest paths used by CFEC to measure each of the four relationships whose source is "Koizumi." The 1,000 shortest paths for destination "Rumsfeld" contain much more popular objects than those for the other destinations do. Especially, 21.4 percent of the vertices for "Rumsfeld" have degree over 1,000.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

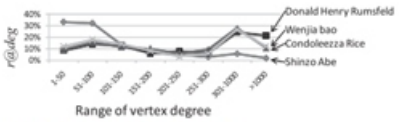

Fig. 6. Ratio of vertices by their degree.

Fig. 7 depicts the average Pdeg of the k1th to k1 + 99th shortest paths for each relationship, for k1 is 1, 101, ... , 901. The average Pdeg for "Rumsfeld" increases most rapidly along with the rising of k1 because many popular objects exist in these paths. Therefore, the weights of the paths for "Rumsfeld" become much smaller than those for the other destinations. Consequently, the relationship of "Rumsfeld" is underestimated by CFEC. We also observed similar results for other relationships underestimated by CFEC and PFIBF. Therefore, popular objects in Wikipedia cause undesirable influence on CFEC and PFIBF.

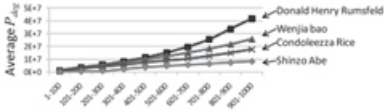

Fig. 7 Average P_{deg} of the k_1th to $k_1 + 99$th paths.

We also compute the Pearson's correlation coefficient between the obtained strength and the score given by the participants. For each method, Fig. 8 depicts the average correlation coefficient for the 10 sources. Note that, the bar "GSD" and "THT" indicates the absolute value of the coefficient for GSD and THT, respectively. The original coefficient for GSD and THT are negative because they gives smaller value to represent a stronger relationship.

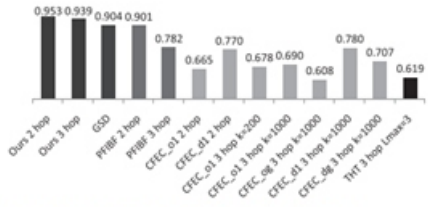

Fig. 8 Average correlation coefficient of each method.

Our methods (2 hop) and (3 hop) have the best two correlation coefficients: 0.953 and 0.939, respectively.

The coefficients of GSD and PFIBF (2 hop) are fairly good: 0.904 and 0.901, respectively. However, GSD cannot use three hop links by nature as explained in Section 2. The coefficient of PFIBF (3 hop) is fairly worse than that of PFIBF (2 hop).Therefore, GSD and PFIBF are unsuitable for measuring the strength of 3-hop implicit relationships. The coefficient of THT is even worse than that of PFIBF (3 hop). Moreover, GSD, PFIBF, and THT were unable to mine elucidatory objects constituting an implicit relationship, although our method can do so. The coefficients of the CFEC variants are much lower than those of other methods, except THT. For the same variant, the difference between the coefficients of CFEC (2 hop) and CFEC (3 hop) is very small; using k = 1,000 shortest paths performs slightly better than using k = 200. The variants (d1) and (dg) using doubled networks produce higher coefficients than the other variants. As discussed above, a doubled network is effective for CFEC. On the other hand, the variants (og) and (dg) using the gain function do not produce higher coefficients than (o1) and (d1), respectively. Therefore, our gain function is not effective for CFEC.

4.2.2 Relationships between Petroleum and Countries:

we obtain the rankings of the 192 countries according to the strengths of their relationships with "Petroleum" using each method as another experiment. It is difficult to find the ground truth for evaluating these rankings. However, the production and consumption of petroleum of each country could be helpful in estimating the rankings. We create a statistics-based ranking of the 192 countries according to the scores computed by using the statistics about the oil production and consumption of the countries.

$$score = \frac{oil\ production\ of\ a\ country}{oil\ production\ of\ the\ world} + \frac{oil\ consumption\ of\ a\ country}{oil\ consumption\ of\ the\ world}. \quad (1)$$

TABLE 2
Rankings of Countries for Petroleum

Ranking	statistics-based	Ours 3 hop	GSD	PFIBF 2 hop	CFEC 3 hop k=1000				THT 3 hop $L_{max} = 3$
					o1	og	d1	dg	
1	USA	Japan	Iraq	Iran	KSA	KSA	Iran	Iran	UAE
2	Russia	USA	Iran	KSA	Kuwait	Kuwait	Indonesia	Indonesia	Iran
3	China	Russia	KSA	Iraq	Iraq	Iraq	Iraq	Iraq	Romania
4	KSA	KSA	Kuwait	Japan	Iran	Iran	KSA	KSA	Iraq
5	Iran	China	Indonesia	Brazil	Egypt	Egypt	Norway	UAE	KSA
6	Canada	Libya	Libya	Indonesia	Brazil	Libya	UAE	Nigeria	Norway
7	Mexico	Kuwait	UAE	Egypt	Libya	UAE	Kuwait	Kuwait	Kyrgyzstan
8	Japan	UK	Pakistan	Turkey	UAE	Algeria	Nigeria	Romania	Kuwait
9	Brazil	Iran	Afghanistan	Libya	Indonesia	Brazil	Romania	Algeria	Indonesia
10	India	Bahrain	Singapore	UAE	Norway	Norway	Egypt	Norway	Tajikistan

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Although the relationship between petroleum and a country is not only dependent on its production and consumption of petroleum, the statistics-based ranking offers an objective way for evaluating the rankings obtained by each method. The top 10 countries in the rankings obtained by each method are presented in Table 2. Our method yields the most similar ranking to the statisticsbased ranking; the top 10 countries of both rankings contain countries which would be strongly related to petroleum, including petroleum producing countries such as "Saudi Arabia" and "Kuwait," and petroleum consuming countries such as "Japan" and "USA," in equilibrium. On the other hand, other methods rank few petroleum consuming countries strongly related to petroleum as the top 10 countries. Especially, except our method, the two largest consumer "USA" and "China" are not ranked in the top 10 by other methods.

We then evaluate the precision at the top n countries of a ranking, abbreviated to P@n, computed by |Sn|/n , where Sn is the set of countries appeared in both the ranking and the statistics-based ranking. Fig. 11 depicts P@10, P@20, and P@30 of all rankings. Similarly to the results of the first experiment depicted in Fig. 10, our method (3 hop) and our method (2 hop) generate the highest precision. The precision of PFIBF (2 hop) is second highest, although that of PFIBF (3 hop) is fairly worse. CFEC (2 hop) performs almost the same as CFEC (3 hop), similarly to the first experiment. There are little differences in the precision of every variant of CFEC (3 hop).

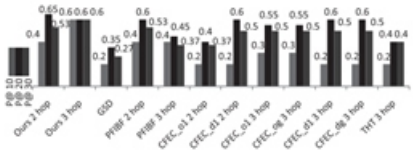

Fig. 8. Precision of rankings for petroleum.

4.3 Case Studies of Elucidatory Objects:

Our method outputs the top-k paths for each relationship, say top-30 paths, primarily contributing to the generalized maximum flow, that is, paths along which a large amount of the flow is sent. We call objects in such paths elucidatory objects constituting a relationship. We discovered several examples in which elucidatory objects are interesting and meaningful for explaining relationships. In this section, we present one of these examples to show the possibility of elucidatory objects for understanding relationships.Fig. 9 portrays five paths (A)-(E) contributing to the flow emanating from "Buddhism" into the "USA." Buddhism originated from India, extended around Asia, and spread further into Europe and to the USA. The Northern United States in path (A) is a large geographic region of the USA. Many immigrants from Southeast Asia are living in the region, and Buddhism is their primary religion. Richard Gere in path (B) is both a famous American actor and a practicing Buddhist. An Institute of Buddhist Studies in path (C) is located in the California State of the USA. Path(D) exists probably because many immigrants from Vietnam live in Los Angeles. About 85 percent of Vietnamese are Buddhist. Path(E) exists probably because the rate of Buddhist in Hawaii is the highest among all the states in the USA, and many temples exist there. These five paths are helpful for us to understand the relationship between Buddhism and the USA.

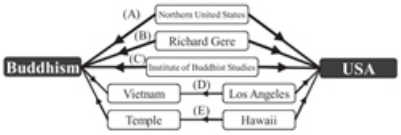

Fig. 9 . Explaining the relationship between Buddhism and the USA.

5 .CONCLUSION:

Here we did propose a new method of measuring the strength of a relationship between two objects on Wikipedia. With the help of a generalized maximum flow, the three representative concepts, distance, connectivity, and cocitation, can be reflected in our method. Furthermore, our method does not underestimate objects having high degrees. We have ascertained that we can obtain a fairly reasonable ranking according to the strength of relationships by our method. Particularly, our method is the only choice for measuring 3-hop implicit relationships. We have also confirmed that elucidatory objects are helpful to deeply understand a relationship. Some future challenges remain. We are also interested in seeking possibilities of the elucidatory objects constituting a relationship mined by our method. We plan to quantitatively evaluate the elucidatory objects. We are developing a tool for deeply understanding relationships by utilizing elucidatory objects.

REFERENCES :

[1]Y. Koren, S.C. North, and C. Volinsky, "Measuring and Extracting Proximity in Networks," Proc. 12th ACM SIGKDD Int'l Conf.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[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]K. Nakayama, T. Hara, and S. Nishio, "Wikipedia Mining for an Association Web Thesaurus Construction," Proc. Eighth Int'l Conf. Web Information Systems Eng. (WISE), pp. 322-334, 2007.

[4]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.

[5]R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993.

[6]K.D. Wayne, "Generalized Maximum Flow Algorithm," PhD dissertation, Cornell Univ., New York, Jan. 1999.

[7]R.L. Cilibrasi and P.M.B. Vita´nyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.

[8]G. Kasneci, F.M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum, "Naga: Searching and Ranking Knowledge," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 953-962, 2008.

[9]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.

[10]"The Erdo¨ s Number Project," http://www.oakland.edu/enp/, 2012.

[11]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.

[12]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.

[13]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.

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

[15]D. Milne and I.H. Witten, "An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links," Proc. AAAI Workshop Wikipedia and Artificial Intelligence: An Evolving Synergy, 2008.