Mining Elucidate Objects and Analysis of Relationships on Wikipedia by using a GFBP Method

Venkata Vinay Kumar C
M.Tech, CSE, LBRCE, Mylavaram

M.A. Ranjit Kumar
Asst Professor, CSE, LBRCE Mylavaram

Sai Satyanarayana Reddy
Professor, CSE, LBRCE Mylavaram, India

Abstract
Evaluating the exact and correct relationships between sets of objects in the Wikipedia is the popular method in order to explain the strong and high relationships between objects. The relationships between two pairs of objects in Wikipedia are exists in two types. They are implicit relationship and another one is explicit relationship. The Implicit relationship in Wikipedia is denoted by a link structure comprising of two pages and an explicit relationship denoted by one link between pair of pages for the objects. Mining Elucidate objects is the popular way to know correct relationship between objects. The Elucidate objects are the main objects which constructs a strong relationship between pair of objects in Wikipedia. The previous methods including cohesion methods are insufficient in evaluate the two relationships because they make use only one or two of the features of the main three features: Path, link and reference. We propose a novel method using a generalized maximum flow pipe method which replicates all the three features. We confirm by experiments that this method can evaluate the strength of a relationship between objects more efficiently and Mine the Elucidate objects than the previous methods. Mining elucidate objects is the new way to understand a strong and high relationship between objects in Wikipedia.

Index Terms
Elucidate objects, Relationship analysis, Link structure, Generalized flow pipe, Wiki mining

1. Introduction

Analysis of relationships between objects has grown in the current period. Knowledge search has been researched to obtain exact knowledge of single object as well as different relationships between multiple objects such as people, species, countries, natural resources, places etc. While searching exact information in the form of web pages by using a keyword has been grown. Some times a user may need to use more than one keyword when searching the exact information. The most popular Encyclopedia namely Wikipedia is one of the popular topic in the field of knowledge searching in case of searching the relevant knowledge of different objects. In Wikipedia the relevant knowledge of a single object is collected in one page which was updated by many volunteers in order to add more data. While Wikipedia uses many objects in a number of different categories such as, people, history, biology, chemistry, Mathematics, science, countries, species etc. Many Typical search engines are not relevant in case of searching and obtaining knowledge of a single object when compared with the Wikipedia.

Discovering relationships between pair of objects is one of the hottest topics in field of knowledge search. A user might wish to find a relationship between pair of objects. For example, a user might wish to know which countries are strongly related to tourism or another example is to know why one country has a stronger relationship to particular natural resources than another country. The typical search engines can neither measure nor explain the strength of a various relationships between pair of objects. The main reason to measure the relationships arises from the fact that there exist two kinds of relationships. One is implicit relationships and another one is explicit relationships. An explicit relationship is denoted by one link between pair of pages for the objects in Wikipedia. For example, an explicit relationship between Tourism and Goa might be represented by one link from page “Tourism” to page “Goa”. User can understand its meaning by reading the text “Famous for it beaches and tourism is its primary Industry” surrounding the anchor text “Goa” and the implicit relationship in Wikipedia is denoted by a link structure comprising of two pages. For example, an implicit relationship between Goa and India can be represented by multiple links and pages are shown in Fig-1. In order to exist an implicit relationship between two objects, Elucidate Objects constitutes a strong relationship between pair of objects. Such types of objects enable us to explain the relationship between objects. For example, “Goa” is one of the elucidate objects. A user can easily understand an explicit relationship between two objects in Wikipedia. By observing the differences between two types of relationships, it is difficult for the user to perceive and find an implicit relationship and elucidate objects with out identifying a number of pages and links. Therefore, measuring and explaining the strength of an implicit relationship between pair of objects is an interesting problem in Wikipedia.

Manuscript received June 5, 2015
Manuscript revised June 20, 2015

Different Methods have been proposed for measuring the Strength of a relationship between two objects .For this an information network (V, E), a directed graph where V is a set of objects, an edge (u, v) $\in$ E exists if and only if object u $\in$ V has an explicit relationship to v $\in$ V . We can define a Wikipedia information network or a data knowledge network whose vertices are pages of Wikipedia and whose edges are links between pages. We propose a new method for measuring a relationship on Wikipedia by reflecting all the three concepts: path, link, and reference. We measure relationships rather than similarities. As discussed in [1], relationship is a more common concept than comparing with similarity. For example, it is hard to say Tourism is similar to India, but a relationship exists between Tourism and the India. The proposed method uses a generalized maximum flow pipe method [2], [3] on an data knowledge network to calculate 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. A gain is assigned for every edge on the network. The flow value is sent along an edge is multiplied by the gain of the edge. Mainly the allocation of the gain to each edge is important for measuring a strong relationship using a generalized maximum flow pipe method. We propose a heuristic gain function utilizing the category structure in Wiki. We confirm through experiments that the gain function is sufficient to measure strong relationships appropriately. Previously proposed methods then can be applied to Wikipedia by using a Wikipedia data knowledge network. Previously cohesion exists for measuring the strength of an implicit relationship. PFIBF–path Frequency inverse backward frequency proposed by Nakayama et al. [4], [5] and CFEC- Cycle Free Effective Conductance proposed by Koren et al. [6] are based on cohesion. We do not adopt the idea of cohesion based methods, because they always underestimates objects having high degrees although such objects could be important to constitute some relationships in Wikipedia, as we will explain in Section-2.Other previously proposed methods use only one or two of the three representative concepts for measuring a relationship: path, link and reference, although all the concepts are important factors for implicit relationships. Using all these concepts i.e., link, path and reference together would be more appropriate for measuring an implicit relationship and mining elucidate objects.

We calculate our method by using computational experiments on the encyclopedia Wiki. At first we select many pages from Wikipedia as our source objects and for each source object; we select many pages as the destination objects. Then we compute the strength of the relationship between a source object and each of its destination objects and finally rank the destination objects by the strength. Then by comparing the rankings acquiring by our method with those obtained by the path frequency inverse backward frequency and cycle-free effective conductance, Google Similarity Distance (GSD) Proposed by Cilibrasi and Vita´nyi [7], we determine that the rankings obtained by our method are the closest to the rankings obtained by human subjects. Especially, we determine that only our method can appropriately measure the strength of “3-hop implicit relationships” which abound in Wikipedia. In an data knowledge network, an implicit relationship between two objects s and t is represented by a sub graph containing s and t. We say that the implicit relationship is a k-hop implicit relationship if the sub graph contains a path from s to t whose length is at least k> 1. Fig. 1 depicts an example of a 3-hop implicit relationship between “Tourism” and the “India.”

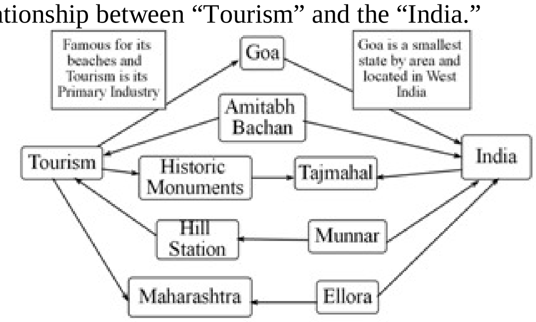

Fig 1: Explaining the relationship between Tourism and the India

Our method can mine elucidate objects which constitutes a relationship by out putting paths contributing to the generalized maximum flow pipe, i.e., paths along which a large amount of flow is sent. We will explain in Section 4 that mining elucidate objects would open a novel way to deeply understand a relationship. Several semantic search engines [8] have been used for Searching relationships between two objects, using a semantic knowledge base [9] extracted from web or Wikipedia. Mean while the semantics in different knowledge bases, such as “is called,” “type”, “subtype of” and “subclass of” are mainly used to construct a meaningful words for objects. Even though by using such semantic knowledge bases are still far from covering relationships existing in Wikipedia, such as “Goa” is a major primary industry in “Tourism”. The important contributions of this paper are listed as follows:

1. A detailed and methodical survey of related work for measuring relationships or similarities between objects (Section 2).
2. A new method using generalized maximum flow pipe procedure for measuring the strength of a relationship between two objects on Wikipedia, which reflects the three terms: path, link and reference (Section 3).
3. Experiments on Wikipedia showing that our method is the most appropriate one than previous methods (Section 5.2).

4. Mining elucidate objects for deeply understanding a relationship between two objects (Section 4).

2. RELATED WORK

We aim to measure the implicit relationships between pair of objects on the Wikipedia data knowledge network. Although a relationship between two objects is a more common concept than considering the term similarity, we discuss existing methods for measuring either relationships or similarities between objects in this section.

2.1 Path, Link and Reference

The concept hitting time [10], [11] from vertex s to vertex t is defined as the expected no of steps in a random flow 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 both s and t. P.Sarkar and A.W.Moore [11] proposed Truncated Hitting Time (THT) to calculate 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 link between two vertices. For example, suppose only $m \geq 1$ vertex disjoint paths of length k connect s to t. Truncated Hitting time computes the distance from s to t to be k for any m _ 1. We compare our method with THT through experiments .in Section 5.The Erdos number [12] used by mathematicians is based on path and co authorships. The legendary mathematician Paul Erdos has a number 0, and the people who co wrote a paper with Erdo¨ s have a number 1; the people who co wrote a paper with a person with a number 1 have a number 2, and so on. The Erdo¨ s number is the path, or the length of the shortest path, from a person to Erdo¨ s on an information network whose edge represents co authorship, a shorter path represents a stronger relationship.

The connectivity [2], mainly the vertex link 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 link from s to t is large. The link 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 path cannot be estimated by the maximum flow because the amount of a flow along a path is independent of the path length. Lu, Janssen, Milios [13] proposed a method for computing the strength of a relationship using a maximum flow pipe method. First they tried to estimate the path between two objects using a maximum pipe flow by setting the capacities of edges. However, the value of a maximum flow does not necessarily decrease by setting only capacities even if the path becomes larger. Their method cannot guess the path correctly by the value of the maximum pipe flow. As an alternative of setting capacities, we use a generalized maximum flow pipe method by setting every gain to a value which is less than 1. Therefore, the value of a maximum flow in our method decreases if the distance becomes longer.

Reference-based methods assume that two objects have a strong relationship if the number of objects linked by both the two objects is large [14]. On the other hand, Reference is a concept by which the strength is represented by the number of objects linking to both objects. The Google Similarity Distance (GSD) proposed by Cilibrasi and Vitanyi [7] can be regarded as a concurrence based method and the strength of a relationship is measured between two words by counting of web pages containing both words. That is, it implicitly regards the web pages as the objects linking to the two objects representing the two words. In an data knowledge network, an object linked by both objects becomes an object linking to the both if the direction of every edge is reversed. The concurrence can be treated as the reverse of the reference or co-citation. We then include concurrence based methods among Reference-based methods in this paper. Milne and Witten [15] also proposed methods measuring relation-ships between objects in Wikipedia using Wikipedia links based on Reference. Reference-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 (u.v) and (v,w).In contrast, reference-based methods can deal with two edges going into the same vertex, such as edges (u,v) and (w,v).Therefore, Reference-based methods are not suitable for measuring an implicit relationship. Furthermore, Reference-based methods cannot deal with 3-hop implicit relationships defined in Section 1 because these methods guess only relationships denoted by paths formed by two edges.

Sim-Rank, proposed by G. Jeh and J. Widom [16], is an extension of Reference-based methods. Sim-Rank employs recursive computation of co-cited objects , therefore it can deal with a path whose length is higher than 2, it can not deal with an implicit relationship “a friend of a close friend or a friend” similarly to Reference-based methods. If we define all edges as bidirectional, then Sim-Rank could measure the typical implicit relationship. However, we observed that Sim-Rank computes the strength of the relationship represented by a path constituted by an odd number of edges to be 0, even if all edges are bi-directional. Consider an example that Sim-Rank computes the strength of the relationship between u and w to be 0 if the relationship is represented by path (u,w) or (u,v0,v1,w). Such paths abound in the Wikipedia data knowledge network. Therefore, Sim-Rank is not suitable for measuring relationships on Wikipedia.

2.2 Interrelation

The Interrelation based methods are used to measure the strength of a relationship by calculating all paths between two different objects. The Interrelation based method was proposed by, L. Katz [17], Wasserman and K. Faust [18] and C.H.Hubbell [19]. Interrelation methods are also known as cohesion based methods. The cohesion method has a property that its value highly increases if a popular object i.e, an object linked from or too many objects, exists. As Listed and pointed in other researches [6], [4] and this property is not suitable for calculating the strength of a relationship. Many Interrelation based techniques mainly the PFIBF and CFEC explained in the following were proposed to incorporate this property.

Nakayama et al. [5], [4] proposed a cohesion based method named Path frequency inverse backward frequency (PFIBF). PFIBF approximately calculate all the paths whose length is at most $k > 0$ using the kth power of the adjacency matrix of an data knowledge network, instead of naming one by one all paths. However, in the kth power of the matrix and a path containing a cycle whose length is at most $k - 1$ would appear. Path frequency inverse backward frequency (PFIBF) can not distinguish a path containing a cycle from a path containing no cycle. For example, if $k \geq 3$ and two edges (u, v) & (v, u) exists, then PFIBF counts both path (u, v) and as well as path (u, v, u, v) consists a cycle (u, v, u). PFIBF has a property that it estimates a single path, e.g., (u, v) in the previous example, for repeated times. The length of a cycle must be at least two. No path containing a cycle appears if $k \leq 2$. In fact, PFIBF usually sets $k = 2$. Therefore, PFIBF is inappropriate for measuring three hop implicit relationships. However, a no of 3-hop implicit relationships exist in Wikipedia. The Effective Conductance proposed by P.G.Doyle & J.L.Snell [20] is a Interrelation-based method also. Effective Conductance has the same disadvantage as PFIBF and it counts a path containing a cycle redundantly. Y.Koren et al. [4] proposed the cycle free effective conductance based on EC by solving this drawback. For a positive integer k, CFEC name one by one only the k-shortest paths between s and t, instead of computing all the paths. The CFEC excludes a path contains a cycle, although it can not count all the paths. We explain below that CFEC and PFIBF are unsuitable for measuring relationships in Wikipedia because of popular objects.

2.2.1 High Popular Objects in Wikipedia

In contrast to the original cohesion method, PFIBF and CFEC do not estimate a popular object. CFEC defines the weight of path p = (s=v1, v2… vl = t) from s to t as

$$Wsum(V1) \,.\, \prod_{i=1}^{l-1} W(v\,i, v\,i+1)/Wsum(vi)$$

Where w (u, v) is the weight of edge (u, v) and wsum (v) is the sum of the weights of the edges going from vertex v. Therefore, the weight of a path becomes extremely small if a popular object exists in the path. The strength C(s, t) of the relationship between s and t is the sum of the weights of all paths from s to t. Fig. 3 depicts two networks and all the paths between source (s) and destination (t). Let the weight of every edge be 1. The wsum of each vertex is represented in the rectangle. The weight of each path is presented at the right side of the path. For the network G1 depicted in Fig. 3a, the wsum of s is 2, and the weight of path (s, v1, v2, t) is 1. C(s, t) for G1 is 2, which is equal to the connectivity between s and t. If we add two edges (v2, v3) and (v3, v2) to G1, then we obtain network G2 in Fig. 3b. Two vertices v2 and v3 become more popular in G2 than they are in G1, and C(s, t) decreases from 2 in G1 to 1.5 in G2. Consequently, Cycle Free Effective Conductance has the property that it could estimate the strength of a relationship smaller if the most popular objects are exists. Also, path frequency inverse backward frequency (PFIBF) has the identical property. The property is suitable for many types of different networks in which popular objects are considered as not important, such as stop words. However, this property would cause undesirable influences if popular objects might be important for a relationship. In Wikipedia, pages of famous people, species, history, places are written to be long and detail; these pages are linked from and linking to several different pages. Therefore, too many important popular objects existing on the Wikipedia data knowledge network represent famous people, places, species, history or events. Such important popular objects may be important to constitute some main relationships.

Let us consider the implicit relationship between the “Sushma” and “Sharif” depicted in Fig. 2. Modi was the Prime Minister of India and Sushma worked under the administration of Modi. Sharif and Haseena were the prime ministers of Pakistan and Bangladesh respectively. The numbers of objects which is linked to “Modi” in bidirectional way and “Sharif” are 1,299 and 389, respectively,in Wikipedia. CFEC and PFIBF allocate a less weight to path Pmodi containing “Modi” than that to path Phaseena containing “Haseena” because “Modi” is more popular, although path Pmodi would be not less important than path Phaseena in this example.The object popularity is essentially independent of the strength of a relationship in Wiki. We ascertain in Section 4 that CFEC and PFIBF are not suitable for measuring relationships on Wikipedia.

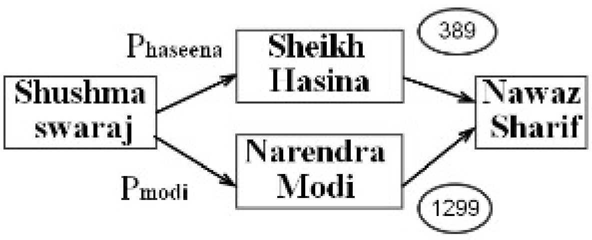

Fig.2. A Relationship between Sushma and Sharif

3. A Generalized Flow Based Pipe Method for measuring relationships in Wikipedia

The three basic concepts path, link and reference are important concepts for measuring relationships. Interrelation or cohesion based methods does not estimate popular objects. The popular objects may be important for constituting relationships in Wikipedia. We propose a generalized maximum flow based pipe method which reflects all the three concepts and does not underestimates popular objects, in order to measure different relationships on the encyclopedia Wikipedia appropriately.

3.1 Generalized Maximum Pipe Flow

The generalized maximum flow pipe problem is identical to the classical maximum flow problem except that every edge e has a gain or increase value γ (e)> 0. The flow value is sent along the edge e and it is multiplied by γ (e). Let f(e) > 0 be the flow f on edge e, and μ (e) $\geq$ 0 be the capacity of edge e. The capacity constraint f(e) $\leq$ μ (e) must hold for every edge e. The goal of the problem is to send a flow emanating from the source vertex which is subject into the destination vertex t to the highest level subject to the capacity constraints. Let a generalized flow based pipe network G = (V, E, s, t, μ, γ) be Information or data knowledge network (V,E) with the source s $\in$ V and the destination t $\in$ V , the capacity μ , and the gain γ . Fig 3b shows an example of a generalized maximum pipe flow on a generalized network. Flow is sent from the source s to v1 in the form of 1 unit, i.e. f(s, v1) =1the amount of the flow is multiplied by γ (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$\leq$ μ(s,v1) must hold. Now we propose a new method for calculating the strength of a relationship using the generalized maximum pipe flow. The value of flow f is defined as the total amount of f arriving at destination t. We use the value of a generalized maximum pipe flow emanating from s as the source into t as the destination in order to measure the strength of a relationship from object s to object t. A larger value signifies a stronger and important relationship. We treated the vertices in the paths composing the generalized maximum pipe flow as the objects constructing or constituting the relationship. We ascertain the claim that our method can reflect the three representative concepts explained in Section 2- path, link and reference also known as co citation.

At first, we consider the path , usually a shortest path denotes a higher relationship. In our method , we set γ (e) < 1 for every edge e, then a flow decreases along a long path. The shortest path contributes to the generalized maximum pipe flow by a larger amount than a long path does. A shorter path means a stronger and higher relationship in our method also.

Next, the Link Method in these methods a higher relationship is represented by many vertex disconnected paths from the source to the destination. The number of vertex disconnected paths can be computed by solving a classical maximum flow problem. The generalized maximum flow pipe problem is a general extension of the classical maximum flow problem.

Last one is the reference and also called as co citation at last. A flow emanates from the source into the destination and the flow uses an edge whose direction is opposite that from the source to the destination. We require using both of the directions to estimate the reference or co citation of two objects. We had considered the relationship between two objects s and t in the network presented in Fig. 4a. Object u is co-cited by s and t. This reference or co citation is represented by two edges (s, u) and (t, u). Unless we reverse the direction of the edge (t, u) to (u, t), we were unable to send a flow from s to t along the two edges. Therefore, we construct a doubled network by adding to every original edge in G a reversed edge whose direction is opposite to the original one. For example, Fig. 4b depicts the doubled network for the network presented in Fig. 4a. We present the definition of a doubled network.

Definition I. Let G = (V, E, s, t, μ, γ) be a generalized network, and rev: E $\rightarrow$ (0, 1] & be a reversed edge gain function for G. The doubled network G_{rev} = (V, E $'$, s, t, μ $'$, γ $'$) of G for rev is defined as follows: E$'$ consists of two types of edges: (1) every edge e(u, v) $\in$ E with μ $'$ (e(u, v))= μ (e(u, v))and γ $'$ (e(u, v)) = γ (e(u, v)); and (2) one reversed edge $e_{rev\ (v,\ u)\ for\ every\ edge\ e(u,\ v)}$ $\in$ E with

μ $'$($e_{rev\ (v,\ u))\ =}$ μ(e(u, v)) and γ $'$ $e_{rev(v,\ u))=rev(e(u,\ v\ )).}$

A flow on the original network satisfies the capacity constraint, that is, the flow is sent along each (u, v) by at most μ (e (u, v)). The constraint is satisfied on the doubled network if we introduce a new constraint f (e (u, v)) f (erev (v, u)) = 0 for flow f. The value of the generalized maximum flow pipe on a doubled network is unchanged even if the new constraint is introduced.

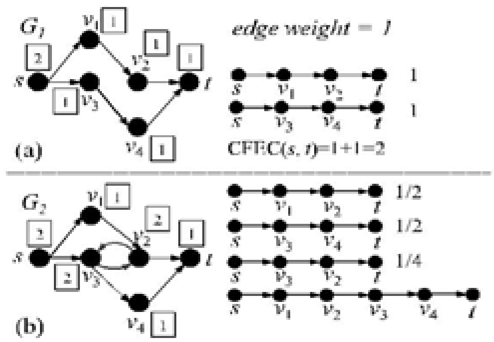

Fig.3a. Cycle Free Effective Conductance on two networks

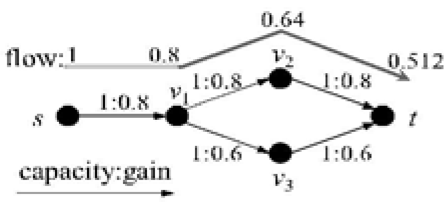

Fig.3b. Generalized Maximum Pipe Flow

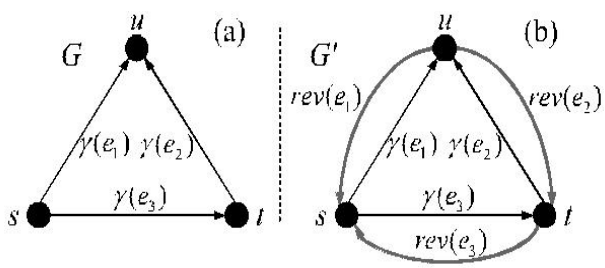

Fig 4. Double Network.

Theorem I. Let | f | be the value of a flow f, and Grev be a doubled network, and g be a generalized maximum flow in Grev. Let gc be a maximum flow in Grev satisfying the constraint that gc(e)gc(erev) =0 for each pair of the edges e and erev. Then, equation | g| = | gc| holds.

To prove this theorem, at first explain a proposition about a flow absorbing cycle [3]. If the product of the gains of the edges composing the cycle is less than 1then a cycle is called as flow absorbing.

Proposition I.. A generalized pipe flow can be converted into another generalized pipe flow containing no flow grasping or absorbing cycles by cancelling the flow absorbing cycles. Cancelling flow grasping cycles does not decrease the value of the flow.

Proof of Theorem I. Because introducing a constraint does not increase the value of the maximum flow, | g| ≥ |gc| . For each pair e and erev not satisfying the constraint g(e)g(erev) = 0 there is a flow grasping or absorbing cycle composed of e and erev. Then by deleting every such a flow grasping or absorbing cycle, we can obtain flow g′ satisfying g′ (e)g′ (erev) = 0 for every pair. Because gc is the maximum pipe flow which satisfies the constraint |gc| ≥ | g′|. On the other hand , | g′|≥ holds by Proposition I. Therefore, | g′| = |g| = |gc|. Now we can assess reference using a generalized maximum pipe flow on the doubled network.

3.2 Using a Gain or Growth Function for Wikipedia Network

In order to verify the growth function, we first consider what types of explicit relationships are important in constructing an implicit relationship. For example, suppose an Indian politician I0 is trying to send a message to a Pakistan politician P0 in the real life. I0 has no explicit relationship to P0, and another Indian politician I1 and an Bangladesh politician B0 have respective explicit relationships to P0. In this case, I0 would tend to ask I1, rather than B0, to help transferring the message to P0. I0 could contact I1 easily compared to P0 because I0 and I1 belong to the same group Indian politician. Then we regard the explicit-relationship between I1 and P0 as primarily important in constructing the relationship between I0 and P0. For the example depicted in Fig. 2, “Sushma” would send a message to “Sharif” through “Modi” rather than “Hasina,” an Bangladesh politician.

Let a “group” be a set of similar or related objects, such as Indian politicians, or Pakistan politicians. We embrace the following 3 assumptions, based on the conversation above, for investigating an implicit relationship between object sin group S(source) and object t in group T(destination) .

1. Explicit relationships between an object in S and an object in T are primarily important, such as that between “Modi” and “Sharif” in the example above.
2. Explicit relationships between objects in S or objects in T are secondarily important, such as that between “Sushma” and “Modi” in the example.
3. Explicit relationships connecting objects in other groups rather than S and T are unimportant, such as that connecting “Sushma” and “Hasina” in the example.

We have noticed a no of relationships in Wikipedia they including the Explicit and implicit relation ships and these suppositions have been correct in most of the cases. We will determine that these suppositions are very effective in measuring various relationships on Wikipedia in Section

5.3 through our experiments.Implicit relationships constructed of many important explicit relationships are very strong. In a generalized max flow pipe problem, a path comprise of edges with enormous gains can contribute to the value of a flow. Therefore, we assign a higher gain to edges denoting very important explicit relationships to measure relationships which are highly related to objects. In order to understand such a increase or gain assignment, we need to construct several groups of objects in Wikipedia. In Wikipedia, every page corresponding to an object belongs to at least one category. For example, the Pakistan politician "Sharif" belongs to the category Members of the Pakistan. Now, a group can be defined as the pages belonging to a same category. Mainly the categories can not be used as groups directly because the category structure of Wiki is too fractionalized. We combined the related various categories as groups at below.

3.2.1 The Relevant Category Grouping

A category ci representing a concept might have descendant categories each representing its sub concept. We should aggregate ci and its descendant categories as a group for ci. However, a part of descendant categories do not represent sub concepts of one denoted by ci. For a good example, The War category is a successor category of the Indian category. Such irrelevant inheritor categories should be excluded from the group for ci.
We have observed that most of the irrelevant descendant categories of ci are not direct children of ci, and such categories are usually linked from more than three categories other than kin-categories related to ci. Then we had decided to build a category group for a specified category ci in the following way. For category ci of Wiki, let A(ci) be the set of sibling categories of ci, parent categories of ci, grandparent categories of ci, and brother categories of the parents or the grandparents.
Categories in A(ci) are depicted by trapezoids in Fig. 5. Let D(ci) be the set of successor categories of ci, mainly which are illustrated by triangles in the Fig. 5. We regard A(ci) ∪ D(ci) ∪ {ci} is the set of kin categories of ci. Categories other than the kin categories are depicted by stars in Fig. 5. We then regard a category in D(ci) as an irrelevant descendant if the category is not a child of ci and is linked from more than three categories other than the kin categories of ci. Irrelevant descendants are depicted by filled triangles in Fig. 5. Let D′(ci) be a subset of D(ci), which is obtained by removing the irrelevant descendants from D(ci). Then, we define D′ (ci) ∪ {ci} as the category group for ci.

3.2.2 The Gain or Increase Function

At first we suggest or propose the increase function for the encyclopedia, Wiki. At first, consider a relationship between two different objects s and t, we construct two different sets S and T of objects that related to the same groups as s and t belongs to respectively in the following way. At first, we enumerate a set Cs of categories to which s relates. Similar way, we specify a set Ct for t.In Wiki, a page is allocated to several different categories. It is easy to use all the categories assigned to s or t as Cs or Ct respectively.

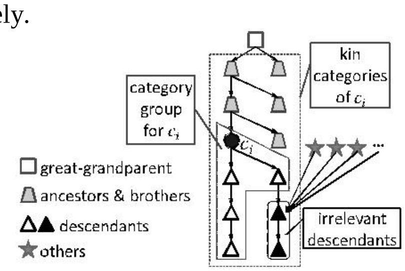

Fig.5. Grouping for category ci

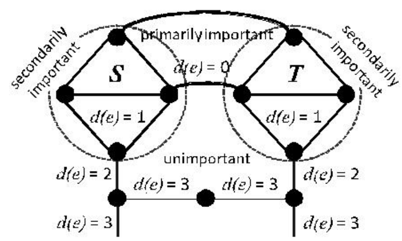

Fig. 6. Gain function

However, many categories contain too many unrelated pages. For example, the category "Alive people" for page "Narendra Modi" contains many people totally unrelated to each other. Such categories are not suitable for grouping different and high related objects. We can assume that such categories are manually deleted from Cs or Ct. In the previous experiments, we determine that using the assumption improves the correctness of our method slightly. Automatically it is possible to ascertain categories for pages which are alternative by using the query domain detection method proposed by the M.Nakatani et al. [21]. We then build a category group for every category in Cs. The set S for s consists of objects belonging to any category in the category groups for Cs. Similarly, we attain the set T for t.
The assumptions conferred in the beginning of this section can be formalized using S and T . The edges (u,v) such that $u \in S \wedge v \in T$ or $u \in T \wedge v \in S$ are the edges representing primarily important explicit relationships. The edges which represents the secondarily very important explicit relationships are inside S or T and the edges representing unimportant explicit relationships are outside S and T . Fig. 7 illustrates the three kinds of edges and reveals that edges distant from primarily important edges are not important. Then, we allocate the increase value or gain for an edge e=(u, v) depending on a distance function

d(e), defined as follows: if $u \in S \wedge v \in T$ or $u \in T \wedge v \in S$, then d(e)= 0; if $u \in S \wedge v \in S$ or $u \in T \wedge v \in T$, then d(e)= 1; otherwise, d(e) is set to 1 plus the number of edges, including e itself, in the shortest path from e to arbitrary vertex in S or T , computed by ignoring the directions of edges. Fig. 7 depicts the definition of d(e). We express the gain function for edge e depending on d(e) with two parameters α and β as

$$\gamma(e) = \alpha * \beta^{d(e)} \quad ,0 < \alpha < 1 \& 0 < \beta \le 1,$$

The opposite increase or gain function is denoted with parameter as rev (e) = $\lambda \times \gamma(e), 0 \le \lambda \le 1$
If the value of α is preset, a slighter β produces greater differences between the gains for edges representing mainly important explicit relationships and those for other edges. λ is used to adjust the importance of a reversed edge. We conduct experiments to ascertain α, β and λ in Section 5.3.

3.3 The Proposed Method Summary

We condense our method for calculating a relationship from source s to destination t as follows:

(i) Construct a generalized network G = (V, E, s, t, μ, γ) containing s and t from Wikipedia, by determining the parameters α and β. First we fix the capacity of every edge to 1.

(ii) Determine the parameter explained in section 3.2 for reversed edge gain rev for G, and construct the doubled network G_{rev} of G for rev.

(iii) Compute a generalized maximum pipe flow g in G_{rev}.

(iv) Let deg (o) indicates the number of objects which are connected from or to object o in Wiki. Outputting the value of the exact flow divided by $\sqrt{\text{deg}(s) * \text{deg}(t)}$ as the strength of the relationship.

(v) As those constructing the relationship, by outputting many paths contributing to the correct flow.

Computation on a large network is practically impossible. As discussed in [6], [16], only a part of the network is significant for measuring a relationship. For Wikipedia, we construct G at step 1 using pages and links within at most k hop links from source or destination in Wiki. By observing carefully the pages in Wikipedia exposed that several paths composed of three links are interesting for understanding a relationship. We are able to recognize some important paths comprised of four links between objects. Additionally, in initial experiments, we constructed G using three and four hop- links, individually and attaining the ranking according to the high strength of relationships calculated by our method. The ranking attained using four hop-links is almost identical to that obtained using three hop-links. Therefore, we set k = 3 at step 1.
Our method can be applied to both directed network and undirected network. For an undirected network, we set λ = 1 to use both directions of an edge equally. We construct the generalized network G for s and t using pages and links within at most 3 hop-links from s or t in Wikipedia. G becomes large if deg(s) or deg (t) is large, and vice versa. The size of G affects the value of the generalized maximum pipe flow and the value becomes high if the size is extended or large. The value of the flow becomes high or large if deg(s) or deg (t) is high. The high strength of the relationship between source s and destination t is expected to be non dependent of deg(s) and deg (t). Therefore, we decide to divide the value of the flow by function D(s, t) = $\sqrt{\text{deg}(s) * \text{deg}(t)}$ at step 4. We also tried several other functions such as D′ (s, t) = deg(s) $*$ deg (t) or D″(s, t) = log (deg(s) $*$deg (t)).In the initial experiments, we have observed that D(s, t) performs the best among all other functions, because D(s, t) represents the effect of the size of G on the value of the flow more closely than D′ or D″ does. Instead of D if we use D′, then the value of D′ excessively dominates the strength of a relationship, because the value raises much faster according to the increase of deg(s) and deg(t) than the effect of the size G does; on the other hand, the value of D′ ′ is too low to indicate the effect. In order to create a ranking according to the high intensity of relationships from a fixed source s to several destinations t's, we calculate the intensity of relationships by dividing the value of a flow by $\sqrt{\text{deg}(t)}$, because estimating deg(s) does not affect the ranking.

4 Mining Elucidate Objects

Mining Elucidate objects is the popular way to identify correct relationships between objects. The Elucidate objects are the main objects which constructs a strong relationship between couple of objects in the encyclopedia, namely the Wikipedia. Our proposed method outputs the topmost- k paths, say topmost-25 paths , for each and every relationship, primarily contributing to the generalized maximum pipe flow, that is, paths along which a large amount of the flow is sent. We discovered several examples in which elucidate objects are very interesting and meaningful for explaining higher and strongest relationships. Let we present one of these examples to show the possibility of elucidate objects for understanding various relationships.

Fig.7 shows five paths (A) to (E) contributing to the flow emanating from "Hinduism" into the "India." Hinduism originated from India and spread all over India as well as some places in the world. The Northern part of India in path (A) is a large geographic region of the India. Many Hindu saints from all over India and as well as from Asia are living in the region, and Hinduism is their primary religion. Rajinikanth in path (B) is a famous Indian actor as well as a Super star and practicing Yoga related to Hinduism. Iskcon is the famous Organization related to lord Krishna and in path (C) its head quarters is located at mayapur in the West Bengal State of India. Path (D) exists probably because many Hindu Naga saints from India as well as from some parts of the world are live in the region of Himalayas. Path (E) exists because the rate of Pilgrims and devotees in Tirumala is the highest among all the temples in India and too many temples exist there. By observing the above fig 7 we can recognize the five paths are helpful for us to understand the correct relationship between Hinduism and India.

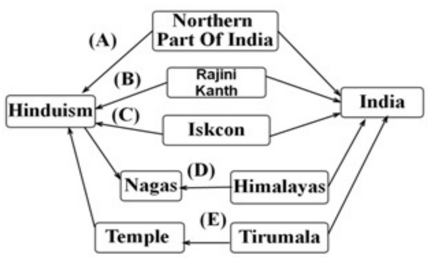

Fig. 7 Explaining the relationship between Hinduism and India

The methods proposed by Koren et al. [6] visualize a sub graph for explaining a relationship. However, their sub graphs tend to be complex. Hence a user still must investigate important paths in the sub graph to understand the different relationships and it is very easy for a user to understand a relationship which are explained by simple paths rather than a complex sub graph. As future work, we plan to utilize elucidatory objects to develop a system for explaining relationships.

5. Calculations and Experiments

In this section, we report experimental results. For this, we first match the rankings according to the high strength of relationships acquired by our method with those attained by PFIBF, CFEC, GSD and THT using human based ranking subjects in Section 5.2. Then the effects of changing the parameters of the increase or gain function are estimated in Section 5.4. We compare our method with other methods using the WordSim353 test collection [22]. In contrast to other methods, our method can output objects and paths establishing a relation. We also test such objects and paths are interesting to understand the high relationship.

5.1 Data Set and Environment

We perform experiments on a Indian Wikipedia data set (12080519 snapshot). 17,310,858 links appear in all of the related and unrelated pages. Delete pages that are not related to objects, such as each year, day, category, person list. Finally, we obtain 8,504,720 remaining links.

We use the rounded primal-dual algorithm [6] to compute an approximately maximum generalized pipe flow. For given approximation parameter $0 < \alpha < 1$, the algorithm outputs a generalized flow whose value is at least as much α times as the value of a generalized maximum flow, in O (n4 $\sqrt{\ }$ m (1- α)-1 log 2 B) time, where m is the number of edges, n is the number of vertices and log 2 B is the largest number of bits which is used to store gain value and high capacity. Our program is implemented in Java and performed calculations and experiments on a PC.

5.2 Assessment of Rankings

Always best calculation of methods measuring different relationships requires human based subjects, as performed in [5], [23], [1]. In this section, we first compare the rankings according to the strengths of relationships obtained by our method, Google Similarity Distance, PFIBF, CFEC and THT with those obtained by human subjects. For our method, we set the increase or gain function with α= 0:8, β= 0:8 and λ= 0:8, which are determined by the estimation of gain function described in Section 5.4.

5.2.1 Analysis of Relationships between Countries and Population

For our Experiment, we attain the rankings of the all 195 countries by using every method according to the strengths of their relationships with "Population" and it is very hard to find the truth for calculating these rankings. The Statistical methods for calculating the population of each country could be very helpful in estimating the rankings. We had create a statistics based ranking of the 195 countries according to the scores calculated by (1) using the statistics about population of the countries [24] the relationship between population and a country is not only dependent on its birth and death rates, census data . The statistics based ranking offers an objective way for calculating the rankings acquired by each and every method. In table 1, the top 10 countries in the rankings obtained by each method are presented. Our method yields the most similar ranking to the statistics based ranking; the top 10 countries of both rankings contain countries which would be strongly related to population. Especially, except our method, the two largest population countries in the world are “Japan” and “Russia” are not ranked in the top 10 by other methods

The population increase or growth rate can be defined as the rate at which the no of different individuals in a population raises in a given time period as a part of the initial population. The population growth rate value refers to the variation in population over a unit period of time, often stated as a percentage of the no of individuals in the population at the starting of that particular time period. This can be written as:

$$pop\ growth\ rate = \frac{P(t_2) - P(t_1)}{P(t_1)}$$

Usually, a positive growth rate denotes that the population is rising, while a negative growth ratio denotes the population is falling. A growth ratio belongs to 0 denotes that there were the same number of people at the two times a growth rate may be zero even when there are significant changes in the immigration rates, birth & death rates between the two times.We calculate the accuracy at the top n countries of a ranking, abbreviated to P@n, computed by | Sn | /n where Sn is the set of different countries appeared in both the ranking and the statistics based ranking.

TABLE 1. Rankings of Countries for Population

Ranking	Statistics-based	Ours3 hop	GSD	PFIBF2 hop	CFEC 3 hop k=1000				THT 3 hop Lmax=3
					ol	og	dl	dg	
1	China	Bangladesh	Vietnam	Brazil	Indonesia	Indonesia	Brazil	Brazil	Philippines
2	India	China	Brazil	Indonesia	Mexico	Mexico	Iran	Iran	Brazil
3	US	India	Indonesia	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	Malaysia
4	Indonesia	Indonesia	Mexico	Bangladesh	Brazil	Brazil	Indonesia	Indonesia	Vietnam
5	Brazil	China	Iran	Russia	Argentina	Argentina	Ethiopia	Philippines	Indonesia
6	Pakistan	Colombia	Colombia	Iran	Russia	Colombia	Philippines	Germany	Ethiopia
7	Nigeria	Mexico	Philippines	Argentina	Colombia	Philippines	Mexico	Mexico	Thailand
8	Bangladesh	Turkey	Kenya	Romania	Philippines	Egypt	Germany	Malaysia	Mexico
9	Russia	Brazil	Algeria	Colombia	Iran	Russia	Malaysia	Egypt	Iran
10	Japan	Spain	Hungary	Philippines	Ethiopia	Ethiopia	Argentina	Ethiopia	Sweden

Fig. 8 depicts P@10, P@20, and P@30 of all rankings. Our method first one is a 3 hop and our method second one is a 2 hop generate the highest accuracy. The accuracy of PFIBF (2 hop) is second highest, although that of path frequency inverse backward frequency (3 hop) is fairly worse. CFEC (2 hop) performs almost the same as cycle free effective conductance (3 hop). There are little differences in the accuracy of every variant of CFEC (3 hop). Therefore, both a doubled network and our gain function are ineffective for CFEC in this experiment. The accuracy of THT is not better than that of CFEC. The correctness of GSD are the worst here. The experimental results presented in Sections 5.2.1 imply that our method is the most suitable one for measuring the strength of a relationship in Wikipedia. Our method is the only choice for measuring 3-hop implicit relationships.

5.3 Relationships between Famous Persons

At first, we pick out 5 famous Indian and American as source objects from Indian Wikipedia, in order to enable the members to find relationships among the famous persons on Wiki and create suitable rankings. For each source (s), we select four famous persons related to the source as the destination (t) objects. We select only four destinations for each source (s) and for each of the 20 obtained pairs of a source and a destination (t), we compute the strength of the relationship from s to t using PFIBF, CFEC, GSD, THT and our method on the same data set explained in Section 5.1. We attain rankings according to the strengths. We search the web pages in the field of Indian Wikipedia using important keywords of the full names of these famous persons to compute GSD. For PFIBF, edge weight is allocated using the FB weighting method of its own [5]. For CFEC and THT, we implement them in four variants represented by the four symbols. They are ol, og ,dl ,dg.(o1) Compute them on the original network, and set the weight w(e) of every edge e to w(e)= 1, (og) Compute them on the original network, and set the weight w(e) of every edge e to w(e)= γ (e) using our increase or gain function.(d1) Compute them on the doubled network, and set the weight w(e) of every edge e to w(e)= 1, (dg) Compute them on the doubled network, set the weight w(e) of every edge e to w(e) = γ (e), and set the weight w(erev) of every reversed edge erev to w(erev)= rev(e), using our increase function. We compute THT for every value Lmax = 1, 2, . ,20 which is the maximum length of paths.

The rankings yielded by these interrelation methods are compared with those attained by human subjects. For examining each of the 20 relationships, each member read about five Wikipedia pages corresponding to or related to the s and t. Each member gives an integer score between 0 and 10, independently when compared to the others. A larger score represents a stronger relationship. By this we can find the strongest relationship and then we obtain rankings according to the average of the scores given by 5 members. Table 2 displays the rankings for the 5 sources. For each source (s), the ranking and the average score obtained by human subjects are written in the column Human an integer 1 to 4 is assigned as the ranking of the destination (t), a real no in parentheses is the result or score. The ranking and the strength obtained by our method, GSD, PFIBF and the four methods of Cycle free effective conductance and Truncated Hitting Time are written in the column Ours, PFIBF, GSD, CFEC and THT. The k hop written after the name of a method denotes that the method calculates a relationship between source s and destination t on the network constructed using at most k hop links from s and t.

5.4 Assessment of Gain or Increase Function

In this section, we evaluate the parameters α, β and λ for our gain function explained in Section 3.2. Let Þ (α, β, λ) be the correlation factor, which are averaged for the 20 relationships among famous persons depending on the values of parameters. Then the values of the parameters are set as α ∈ { 0.1, 0.2; . . . , 0.9}, β ∈ {0.1, 0.2, . . . ,1.0} and λ ∈ {0, 0.1, . . . ,1.0}. We compute Þ (α, β, λ) for all the possible 9 x 10 x 11 = 990 combinations of values. Let Þ (α = χ) be the average of Þ (α, β, λ) obtained by the combinations of fixing α = χ and varying β and λ. Þ (β= χ) and (λ = χ) are similarly defined. Table III presents the averages Þ (α = χ), Þ (β= χ), and Þ (λ = χ).

TABLE 2. Rankings of Famous Persons

Source	Destination	Human	Ours 3 hop	GSD	PFIBF 2 hop	CFEC 3 hop k=1000				THT dl 3 hop Lmax=3
						ol	og	Dl	dg	
Narendra Modi	L.K.Advani	1(9.2)	1(2.05)	2(0.46)	1(12.4)	1(1.02)	1(0.98)	1(1.90)	1(1.79)	1(2.98550)
	AB Vajpayee	2(8.6)	2(1.17)	3(0.37)	2(1.05)	4(0.22)	4(0.23)	2(1.82)	2(0.98)	4(2.99890)
	M. Singh	3(4.3)	3(1.10)	1(0.56)	3(1.02)	2(0.79)	2(0.75)	4(1.10)	4(0.78)	3(2.99678)
	Nawaz sharif	4(2.0)	4(0.79)	4(0.22)	4(1.00)	3(0.50)	3(0.42)	3(1.15)	3(0.86)	2(2.99452)
Sonia Gandhi	Rahul gandhi	1(8.5)	1(2.15)	1(0.47)	3(1.72)	1(1.20)	1(0.83)	1(1.50)	1(1.54)	1(2.98755)
	P.v n rao	2(6.2)	2(1.98)	3(0.38)	2(3.89)	3(0.30)	2(0.68)	3(0.73)	3(0.87)	2(2.98980)
	Sushma Swaraj	3(4.4)	3(0.99)	2(0.24)	1(7.29)	2(0.43)	4(0.13)	4(0.28)	4(0.80)	4(2.99878)
	M. Singh	4(3.1)	4(0.76)	2(0.24)	4(1.21)	4(0.12)	3(0.33)	2(0.99)	2(0.98)	3(2.99492)
Barack Obama	Mitt Romney	1(8.0)	1(1.99)	1(0.51)	2(8.42)	1(1.20)	1(0.78)	1(2.08)	1(0.97)	2(2.99187)
	George Bush	2(7.3)	2(1.19)	2(0.32)	1(15.1)	2(0.89)	2(0.71)	2(1.14)	2(0.83)	3(2.99374)
	Ben Bernanke	3(5.5)	4(0.87)	3(0.21)	3(4.27)	3(0.47)	3(0.47)	3(0.78)	3(0.77)	4(2.99650)
	D.D.Mckierman	4(3.1)	3(0.99)	4(0.12)	4(1.90)	4(0.32)	4(0.32)	4(0.15)	4(0.74)	1(2.99121)
Pranab Mukherjee	Sonia Gandhi	1(8.2)	1(2.80)	2(0.42)	1(8.41)	1(2.33)	1(0.80)	1(1.12)	1(1.49)	1(2.99112)
	M. Singh	2(6.4)	2(1.89)	3(0.33)	4(1.63)	3(0.27)	2(0.72)	3(0.89)	3(0.81)	4(2.99629)
	George Bush	3(5.2)	3(1.59)	4(0.19)	3(2.33)	2(1.89)	4(1.10)	4(0.77)	4(0.78)	3(2.99432)
	Hillary clinton	4(2.5)	4(0.89)	1(0.59)	2(4.98)	4(0.19)	3(0.21)	2(1.10)	2(0.87)	2(2.99367)
Hillary Rodham Clinton	Barack Obama	1(9.4)	1(3.46)	1(0.39)	1(6.07)	1(1.48)	1(1.0)	1(1.91)	1(1.19)	1(2.99121)
	Bill Clinton	2(8.1)	3(1.91)	3(0.12)	4(1.22)	4(0.12)	4(0.65)	4(0.87)	4(0.81)	4(2.99598)
	George Bush	3(7.8)	2(2.30)	2(0.24)	2(4.12)	2(0.59)	2(0.43)	2(1.29)	2(0.91)	2(2.99189)
	Richard Nixon	4(3.5)	4(1.15)	2(0.24)	3(2.08)	3(0.42)	3(0.28)	3(1.00)	3(0.87)	3(2.99250)

The differences between the averages are relatively small when χ is large. Therefore, our method is fairly robust against varying parameter values. The greatest average for a fixed α is Þ (α = 0.9) = 0.920, that for β is Þ (β= 0.8) = 0.913, and that for λ is Þ (λ = 1.0) = 0.893. The shadowed cells in the row "Þ (α = χ)" indicate that we could find no statistical significance among the distributions of Þ (α, β, λ) obtained by the combinations of fixing α = 0.7, 0.8 or 0.9, by fixing the Significance level to 0.05.The combinations producing very good and better results are α ∈ {0.7, 0.8, 0.9), β ∈ {0.7, 0.8, 0.9}, and λ ∈ {0.6, 0.7,1.0}.Similar combinations are obtained by evaluating the P@n of the ranking of countries for the 990 combinations of the parameters. We finally choose the combination α = 0.8, β= 0.8, and λ =0.8 which produces a medium result among the candidates. We attain the following observations:

(1)If β = 1, then the increase or gain function is insensitive to groups, build from the category structure of Wiki as explained in Section 3.2. Þ (β = 1) is inferior than the best average. Therefore, the category structure is necessary to our gain function.

(2)If λ = 0, then no reversed edges are used for measuring a relationship. Þ (λ=0) =0.810 is the foulest value in the bottom row. Therefore, reversed edges used for reflecting reference or co-citation is effective in measuring a higher and strongest relationship. Subsequently, the dual network is the best choice for measuring relationships than the original Wikipedia data knowledge network

TABLE 3. Average Correlation Coefficients with a Fixed Parameter

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\bar{\rho}(\alpha=\chi)$	-	0.705	0.811	0.855	0.878	0.891	0.901	0.908	0.914	0.920	-
$\bar{\rho}(\beta=\chi)$	-	0.778	0.805	0.829	0.850	0.870	0.889	0.905	0.913	0.910	0.899
$\bar{\rho}(\lambda=\chi)$	0.810	0.826	0.842	0.855	0.866	0.874	0.880	0.885	0.888	0.891	0.893

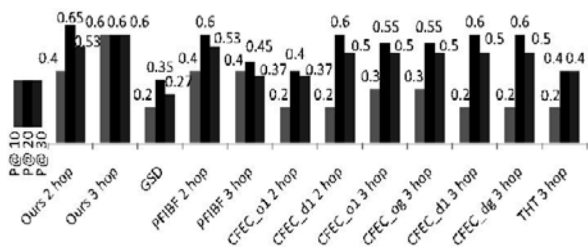

Fig. 8. Accuracy of rankings for Population.

6. Conclusion and Future Work

We have suggested a new method of measuring the strength of a relationship between two different objects on Wikipedia. By using a generalized maximum pipe flow, the three concepts, path, link and reference or co citation can be reflected in our method. Furthermore, our method estimate objects having high degrees. We have determined that we can obtain a fairly reasonable ranking according to the strength of relationships by our method compared with those by PFIBF [5], [4], CFEC [6], GSD [7], and THT [11]. Particularly, our method is the only choice for measuring 3-hop implicit relationships. Mining Elucidate objects is the popular way to identify correct relationships between objects. The Elucidate objects are the main objects which constructs a strong relationship between a pair of objects, we have also confirmed that elucidate objects are helpful to deeply understand a relationship. Some Future work remains. Elucidate objects constitutes a relationship between different pairs of objects. Evaluation of elucidate objects must be done in quantitatively manner. Though, relationships exist in various types between objects in Data and Knowledge field. Mining elucidate objects in case of various relationships between different objects must be done efficiently. Not only Mining the elucidate objects, we have to understand deeply the relationships exist in Wikipedia by using the elucidate Objects. For this we are developing efficient tools for the purpose of understanding the relationships existing in Wikipedia.

REFERENCES

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

[2] 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.
[3] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993.
[4] 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.
[5] 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.
[6] 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.
[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] P. Sarkar and A.W. Moore, "A Tractable Approach to Finding Closest Truncated-Commute-Time Neighbors in Large Graphs,"
[11] Proc. 23rd Conf. Uncertainty in Artificial Intelligence (UAI), 2007.
[12] "The Erdo¨s Number Project," http://www.oakland.edu/enp/, 2012.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] G. Jeh and J. Widom, "Simrank: A Measure of Structural-Context Similarity," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 538-543, 2002.
[18] L. Katz, "A New Status Index Derived from Sociometric Analysis," Psychometrika, vol. 18, no. 1, pp. 39-43, 1953.
[19] S. Wasserman and K. Faust, Social Network Analysis: Methods and Application (Structural Analysis in the Social Sciences). Cambridge Univ. Press, 1994.
[20] C.H. Hubbell, "An Input-Output Approach to Clique Identifica-tion," Sociometry, vol. 28, pp. 277-299, 1965.
[21] P.G. Doyle and J.L. Snell, Random Walks and Electric Networks, vol. 22. Math. Assoc. Am., 1984.
[22] M. Nakatani, A. Jatowt, and K. Tanaka, "Easiest-First Search: Towards Comprehension-Based Web Search," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 2057-2060, 2009.
[23] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, The WordSimilarity-353 Test Collection, 2002.
[24] W. Xi, E.A. Fox, W. Fan, B. Zhang, Z. Chen, J. Yan, and D. Zhuang, "Simfusion: Measuring Similarity Using Unified Relationship Matrix," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 130-137, 2005. "Country Ranks 2009," http://www.photius.com/rankings/ index.html, 2012.