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

Xinpeng Zhang, *Member*, *IEEE*, Yasuhito Asano, *Member*, *IEEE*, and Masatoshi Yoshikawa

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 cocitation. 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.

Index Terms—Link analysis, generalized flow, Wikipedia mining, relationship

1 Introduction

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.

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 $(u, v) \in E$ exists if and only if object $u \in V$ has an explicit relationship to $v \in V$. We can define a *Wikipedia information network* whose vertices are pages of Wikipedia 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] and PFIBF 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

• *The authors are with the Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan. E-mail: xinpeng.zhang@db.soc.i.kyoto-u.ac.jp, {asano, yoshikawa}@i.kyoto-u.ac.jp.*

Manuscript received 18 Feb. 2010; revised 14 Oct. 2011; accepted 16 Oct. 2011; published online 10 Nov. 2011.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-02-0099. Digital Object Identifier no. 10.1109/TKDE.2011.227.

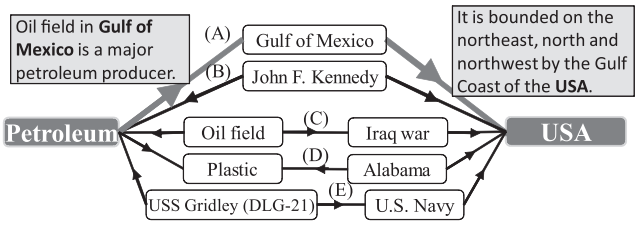

Fig. 1. Explaining the relationship between Petroleum and the USA.

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 cocitation, 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.

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.

We evaluate our method using computational experiments on Wikipedia. We first select several pages from Wikipedia as our source objects; and for each source object, we select several pages as the destination objects. We then compute the strength of the relationship between a source object and each of its destination objects, and rank the destination objects by the strength. By comparing the rankings obtained by our method with those obtained by the "Google Similarity Distance" (GSD) proposed by Cilibrasi and Vitányi [7], PFIBF and CFEC, we ascertain that the rankings obtained by our method are the closest to the rankings obtained by human subjects. Especially, we ascertain that only our method can appropriately measure the strength of "3-hop implicit relationships" which abound in Wikipedia. In an information network, an implicit relationship between two objects s and t is represented by a subgraph containing s and t. We say that the implicit relationship is a *k-hop implicit relationship* if the subgraph 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 "Petroleum" and the "USA."

Our method can mine elucidatory objects constituting a relationship by outputting paths contributing to the generalized maximum flow, that is, paths along which a large amount of flow is sent. We will explain in Section 4.5 that mining elucidatory 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. However, the semantics in these knowledge bases, such as "isCalled," "type" and "subClassOf," are mainly used to construct an ontology for objects. Such semantic knowledge bases are still far from covering relationships existing in Wikipedia, such as "Gulf of Mexico" is a major "petroleum" producer. We do not utilize the semantic knowledge bases for measuring relationships in this paper.

The main contributions of this paper are as follows:

1. A detailed and methodical survey of related work for measuring relationships or similarities (Section 2).
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 (Section 3).
3. Experiments on Wikipedia showing that our method is the most appropriate one (Section 4.2).
4. Case studies of mining elucidatory objects for deeply understanding a relationship (Section 4.5).

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 Erdös number [10] used by mathematicians is based on distance and coauthorships. The legendary mathematician Paul Erdös has a number 0, and the people who cowrote a paper with Erdös 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 Erdös number is the distance, or the length of the shortest path, from a person to Erdös on an information network whose edge represents coauthorship; a shorter path represents a stronger relationship. However, the Erdös number is inadequate to represent the implicit relationship between a person and Erdös because the number does not estimate the connectivity between them. 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 $L_{\max}$ only. A smaller distance represents a larger similarity. THT does not estimate the connectivity between two vertices. For example, suppose only $m \geq 1$ vertex disjoint paths of length k connect s to t. THT computes the distance from s to t to be k

for any $m \geq 1$. We compare our method with THT through experiments in Section 4.

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.

Cocitation-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, co-occurrence is a concept by which the strength is represented by the number of objects linking to both objects. The "Google Similarity Distance" proposed by Cilibrasi and Vitányi [7] can be regarded as a co-occurrence based method; it measures the strength of a relationship between two words by counting of webpages containing both words. That is, it implicitly regards the webpages as the objects linking to the two objects representing the two words. In an information network, an object linked by both objects becomes an object linking to the both if the direction of every edge is reversed. Therefore, co-occurrence can be regarded as the reverse of the cocitation. We then include co-occurrence-based methods among 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 (u, v) and (v, w). In contrast, cocitation-based methods can deal with two edges going into the same vertex, such as edges (u, v) and (w, v). Therefore, cocitation-based 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.

SimRank, proposed by Jeh and Widom [16], is an extension of cocitation-based methods. SimRank employs recursive computation of cocited objects, therefore it can deal with a path whose length is longer than two, although it cannot deal with a typical implicit relationship "a friend of a friend" similarly to cocitation-based methods. If we define all edges as bidirectional, then SimRank could measure the typical implicit relationship. However, we observed that SimRank 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 bidirectional. For example, SimRank computes the strength of the relationship between u and w to be 0 if the relationship is represented by path (u, w) or (u, v_0, v_1, w). Such paths abound in the Wikipedia information network. Therefore, SimRank is inappropriate for measuring relationships on Wikipedia.

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 $k \geq 3$ and two edges (u, v) and (v, u) 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 $k \leq 2$. 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.

We below explain that CFEC and PFIBF are unsuitable for measuring relationships in Wikipedia because of popular objects.

2.2.1 Popular Objects in Wikipedia

In contrast to the original cohesion, PFIBF and CFEC underestimate a popular object. CFEC defines the weight of path $p = (s = v_1, v_2, \ldots, v_\ell = t)$ from s to t as

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

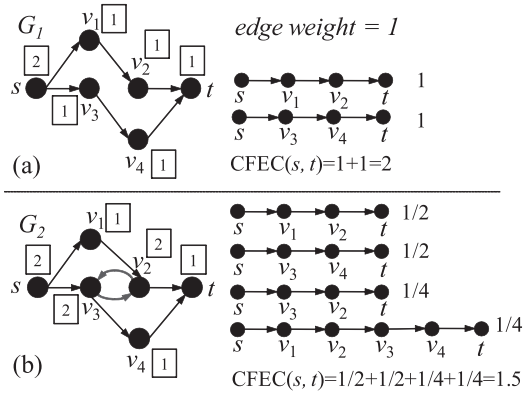

Fig. 2. CFEC on two networks.

where $w(u,v)$ is the weight of edge (u,v) and $w_{sum}(v)$ 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. 2 depicts two networks and all the paths between s and t. For simplicity, let the weight of every edge be one. The w_{sum} 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. For the network G_1 depicted in Fig. 2a, the w_{sum} of s is 2, and the weight of path (s, v_1, v_2, t) is 1. $C(s,t)$ for G_1 is 2, which is equal to the connectivity between s and t. If we add two edges (v_2, v_3) and (v_3, v_2) to G_1, then we obtain network G_2 in Fig. 2b. Two vertices v_2 and v_3 become more popular in G_2 than they are in G_1, and $C(s,t)$ decreases from 2 in G_1 to 1.5 in G_2. Consequently, CFEC has the property that it could estimate the strength of a relationship smaller if popular objects exist. Similarly, PFIBF has the same property.

The property is suitable for several kinds of networks in which popular objects are considered as noise, such as stop words or portal sites. However, this property would cause undesirable influences if popular objects might be important for a relationship. In Wikipedia, pages of famous people, places or events, are written to be long and detail; these pages are linked from and linking to many other pages. Therefore, many popular objects existing on the Wikipedia information network represent famous people, places or events. Such popular objects might be important to some relationships. 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. CFEC and PFIBF assign a smaller weight to path P_{Bush} containing "Bush" than that to path P_{Olmert} containing

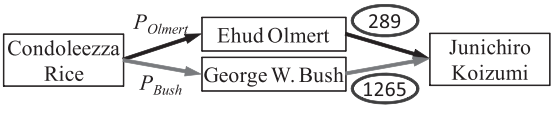

Fig. 3. A Relationship between Rice and Koizumi.

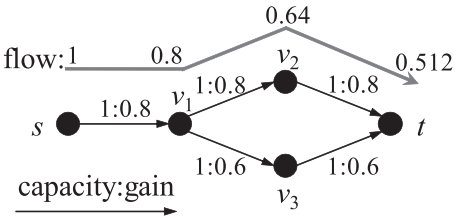

Fig. 4. Generalized maximum flow.

"Olmert" because "Bush" is more popular, although path P_{Bush} would be not less important than path P_{Olmert} in this example. There are many cases similar to this example in Wikipedia. Therefore, the popularity of an object is essentially independent of the strength of a relationship in Wikipedia. We ascertain in Section 4.2 that CFEC and PFIBF are unsuitable for measuring relationships on Wikipedia.

3 Method for Measuring Relationships Using Generalized Flow

As discussed in Section 2, the three concepts, distance, connectivity, and cocitation, are important concepts for measuring relationships; cohesion-based methods underestimate popular objects, although popular objects might be important for relationships in Wikipedia. Therefore, 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 $\gamma(e) > 0$; the value of a flow sent along edge e is multiplied by $\gamma(e)$. Let $f(e) \geq 0$ be the flow f on edge e, and $\mu(e) \geq 0$ be the capacity of edge e. The capacity constraint $f(e) \leq \mu(e)$ 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* $G = (V, E, s, t, \mu, \gamma)$ be information network (V, E) with the source $s \in V$, the destination $t \in V$, the capacity μ, and the gain γ. 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 v_1, i.e., $f(s, v_1) = 1$, the amount of the flow is multiplied by $\gamma(s, v_1)$ when the flow arrives at v_1. Consequently, only 0.8 units arrive at v_1. 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. $f(s, v_1) = 1 \leq \mu(s, v_1)$ must hold, for example.

We propose a new method for measuring the strength of a relationship using the generalized maximum flow. The value of flow f is defined as the total amount of f arriving at destination t. To measure the strength of a relationship from object s to object t, we use the value of a generalized maximum flow emanating from s as the source into t as the destination; a larger value signifies a stronger relationship. We regard the vertices in the paths composing the generalized maximum flow as the objects constituting the relationship. We qualitatively ascertain the claim that our method

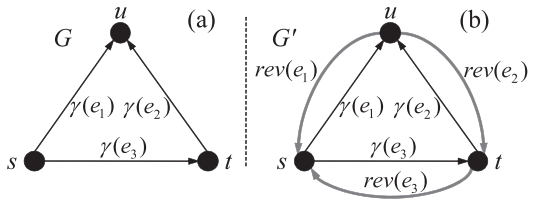

Fig. 5. A doubled network.

can reflect the three representative concepts explained in Section 2: distance, connectivity, and cocitation.

We first discuss the distance. In the methods based on distance, a shorter path represents a stronger relationship. For our method, we set $\gamma(e) < 1$ for every edge e; then a flow considerably decreases along a long path. A short path usually contributes to the generalized maximum flow by a greater amount than a long path does. Therefore, a shorter path means a stronger relationship in our method also.

We then discuss the connectivity. In methods based on connectivity, a strong relationship is represented by many vertex disjoint paths from the source to the destination. The number of vertex disjoint paths can be computed by solving a classical maximum flow problem. The generalized maximum flow problem is a natural extension of the classical maximum flow problem. Therefore, it also can be used to estimate the connectivity.

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). However, we were unable to send a flow from s to t along the two edges, unless we reverse the direction of the edge (t, u) to (u, t). 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. 5b depicts the doubled network for the network presented in Fig. 5a. We present the definition of a doubled network.

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

A flow on the original network satisfies the capacity constraint, that is, the flow is sent along each (u, v) by at most $\mu(e(u, v))$. The constraint is satisfied on the doubled network if we introduce a new constraint $f(e(u, v))f(e_{rev}(v, u)) = 0$ for flow f. Fortunately, the value of the generalized maximum flow on a doubled network is unchanged even if the new constraint is introduced.

Theorem 1. *Let $|f|$ be the value of a flow f, and G_{rev} be a doubled network, and g be a generalized maximum flow in G_{rev}. Let g_c be a maximum flow in G_{rev} satisfying the constraint that $g_c(e)g_c(e_{rev}) = 0$ for each pair of the edges e and e_{rev}. Then, equation $|g| = |g_c|$ holds.*

To prove this theorem, we explain a proposition about a *flow-absorbing cycle* [6]. A cycle is called flow absorbing if the product of the gains of the edges composing the cycle is less than 1.

Proposition 1. *A generalized flow can be converted into another generalized flow containing no flow-absorbing cycles by canceling the flow-absorbing cycles. Canceling flow-absorbing cycles does not decrease the value of the flow.*

Proof of Theorem 1. Because introducing a constraint does not increase the value of the maximum flow, $|g| \geq |g_c|$. For each pair e and e_{rev} not satisfying the constraint $g(e)g(e_{rev}) = 0$, there is a flow-absorbing cycle composed of e and e_{rev}. By canceling every such a flow-absorbing cycle, we can obtain flow g' satisfying $g'(e)g'(e_{rev}) = 0$ for every pair. Because g_c is the maximum flow satisfying the constraint, $|g_c| \geq |g'|$. On the other hand, $|g'| \geq |g|$ holds by Proposition 1. Therefore, $|g'| = |g| = |g_c|$. □

Therefore, we can estimate cocitation using a generalized maximum flow on the doubled network.

3.2 Gain Function for Wikipedia

In order to determine the gain function, we consider what kinds of explicit relationships are important in constituting an implicit relationship. Suppose an American politician A_0 is trying to send a message to a Japanese politician J_0 in the real life; A_0 has no explicit relationship to J_0, and another American politician A_1 and an Israeli politician I_0 have respective explicit relationships to J_0. In this case, A_0 would tend to ask A_1, rather than I_0, to help transferring the message to J_0. A_0 could contact A_1 easily compared to J_0 because A_0 and A_1 belong to the same group "American politician." We therefore regard the explicit relationship between A_1 and J_0 as primarily important in constituting the relationship between A_0 and J_0. For the example depicted in Fig. 3, "Rice" would send a message to "Koizumi" through "Bush" rather than "Olmert," an Israeli politician.

Let a "group" be a set of similar or related objects, such as American politicians, or Japanese politicians. We adopt the following three assumptions, based on the discussion above, for analyzing an implicit relationship between object s in group S and object t in group T.

1. Explicit relationships between an object in S and an object in T are primarily important, such as that between "Bush" and "Koizumi" in the example above.
2. Explicit relationships between objects in S or objects in T are secondarily important, such as that between "Rice" and "Bush" in the example.
3. Explicit relationships connecting objects in other groups rather than S and T are unimportant, such as that connecting "Rice" and "Olmert" in the example.

We have observed a number of relationships in Wikipedia, and these assumptions have been true in most

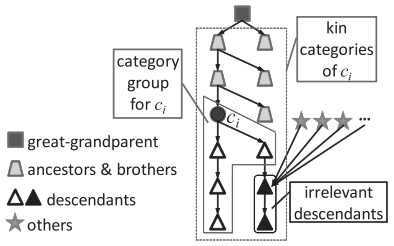

Fig. 6. Grouping for category c_i.

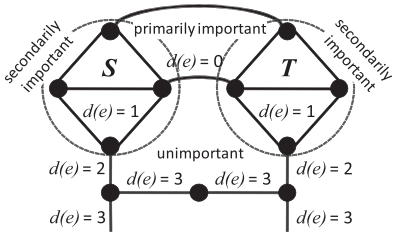

Fig. 7. Gain function.

cases. We will ascertain that these assumptions are effective in measuring relationships on Wikipedia in Section 4.3 through experiments.

Implicit relationships constituted of many important explicit relationships are strong. In a generalized max-flow problem, a path composed of edges with large gains can contribute to the value of a flow. Therefore, we assign a larger gain to edges representing important explicit relationships to measure relationships. To realize such a gain assignment, we need to construct groups of objects in Wikipedia. In Wikipedia, the page corresponding to an object belongs to at least one category. For example, the Japanese politician "Junichiro Koizumi" belongs to the category "Members of the Diet of Japan." We then could define the pages belonging to a same category as a group. However, categories cannot be used as groups directly because the category structure of Wikipedia is too fractionalized. Therefore, we aggregate related categories as groups at below.

3.2.1 Category Grouping

A category c_i representing a concept might have descendant categories each representing its sub concept. We should aggregate c_i and its descendant categories as a group for c_i. However, a part of descendant categories do not represent sub concepts of one represented by c_i. For example, "The Pacific War" category is a descendant category of the "Thailand" category. Such irrelevant descendant categories should be excluded from the group for c_i.

We observed that most of the irrelevant descendant categories of c_i are not direct children of c_i, and such categories are usually linked from more than three categories other than kin categories of c_i. Therefore, we decide to construct a "category group" for a specified category c_i in the following way. For category c_i of Wikipedia, let $A(c_i)$ be the set of sibling categories of c_i, parent categories of c_i, grandparent categories of c_i, and brother categories of the parents or the grandparents. Categories in $A(c_i)$ are depicted by trapezoids in Fig. 6. Let $D(c_i)$ be the set of descendant categories of c_i, which are depicted by triangles in Fig. 6. We regard $A(c_i) \cup D(c_i) \cup \{c_i\}$ is the set of kin categories of c_i. Categories other than the kin categories are depicted by stars in Fig. 6. We then regard a category in $D(c_i)$ as an irrelevant descendant if the category is not a child of c_i and is linked from more than three categories other than the kin categories of c_i. Irrelevant descendants are depicted by filled triangles in Fig. 6. Let $D'(c_i)$ be a subset of $D(c_i)$, which is obtained by removing the irrelevant descendants from $D(c_i)$. Then, we define $D'(c_i) \cup \{c_i\}$ as the *category group* for c_i.

3.2.2 The Gain Function

We now propose the gain function for Wikipedia. Given a relationship between two objects s and t, we construct two sets S and T of objects belonging to the same groups as s and t belongs to, respectively, in the following way. We first specify a set C_s of categories to which s belongs. Similarly, we specify a set C_t for t. In Wikipedia, a page is allocated to several categories. It is simple to use all the categories allocated to s or t as C_s or C_t, respectively. However, several categories contain too many unrelated pages. For example, category "Living people" for page "George W. Bush" contains many people totally unrelated to each other. Such categories are unsuitable for grouping related objects. Therefore, through the paper we assume that such categories are manually removed from C_s or C_t. In preliminary experiments, we ascertain that using the assumption improves the precision of our method slightly. Alternatively, it is possible to determine categories for pages automatically using the query domain detection method proposed by Nakatani et al. [22]. We then construct a category group for every category in C_s. The set S for s consists of objects belonging to any category in the category groups for C_s. Similarly, we obtain the set T for t.

The assumptions discussed 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 representing secondarily 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 unimportant. Therefore, we assign the 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

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

and the reverse gain function is represented with parameter λ as

$$rev(e) = \lambda \times \gamma(e), 0 \leq \lambda \leq 1.$$

If the value of α is fixed, a smaller β produces larger differences between the gains for edges representing

primarily important explicit relationships and those for other edges. λ is used to adjust the importance of a reversed edge. We conduct experiments to determine α, β, and λ in Section 4.3.

3.3 Summary of the Proposed Method

We summarize our method for measuring a relationship from s to t as follows:

1. Construct a generalized network $G = (V, E, s, t, \mu, \gamma)$ containing s and t from Wikipedia, by determining the parameters α and β explained in Section 3.2. We set the capacity of every edge to one.
2. Determine the parameter λ explained in Section 3.2 for reversed edge gain rev for G, and construct the doubled network G_{rev} of G for rev.
3. Compute a generalized maximum flow g in G_{rev}.
4. Let $deg\ (o)$ denote the number of objects linked from or to object o in Wikipedia. Output the value of the flow divided by $\sqrt{deg(s) * deg(t)}$ as the strength of the relationship.
5. As those constituting the relationship, output several paths contributing to the flow.

Computation on a large network is practically impossible. As discussed in [1], [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 s or t in Wikipedia. Careful observation of pages in Wikipedia revealed that several paths composed of three links are interesting for understanding a relationship, although we were able to find few interesting paths composed of four links. Furthermore, in preliminary experiments, we constructed G using three and four hop links, separately, and obtained the ranking according to the strength of relationships computed by our method. However, the ranking obtained using four hop links is almost identical to that obtained using three hop links. Therefore, we usually set $k = 3$ at step 1.

Our method can be applied to both directed network and undirected network. For an undirected network, we set $\lambda = 1$ 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 flow; the value becomes large if the size is large. Consequently, the value of the flow becomes large if $deg(s)$ or $deg(t)$ is large. On the other hand, the strength of the relationship between s and t is expected to be independent of $deg(s)$ and $deg(t)$. Therefore, we decide to divide the value of the flow by function $D(s,t) = \sqrt{deg(s) * deg(t)}$ 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 preliminary experiments, we observed that $D(s,t)$ performs the best among all 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. If we use D' instead of D, then the value of D' excessively dominates the strength of a relationship, because the value increases 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 small to represent the effect. For creating a ranking according to the strength of relationships from a fixed source s to several destinations, we compute the strength of relationships by dividing the value of a flow by $\sqrt{deg(t)}$, because estimating $deg(s)$ does not affect the ranking.

4 Experiments and Evaluation

In this section, we report experimental results. We first compare the rankings according to the strength of relationships, obtained by our method with those obtained by GSD, PFIBF, CFEC, and THT using human subjects, in Section 4.2. We then estimate the effects of varying the parameters of the gain function in Section 4.3. In Section 4.4, we compare our method with other methods using the WordSim353 test collection [23], [24]. In contrast to other methods, our method can output objects and paths constituting a relation. We also examine that such objects and paths are interesting to understand the relationship described in Section 4.5.

4.1 Data Set and Environment

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.

We use the rounded primal-dual algorithm [6] to compute an approximately maximum generalized 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 $O(n^4\sqrt{m}(1-\alpha)^{-1}\log_2 B)$ time, where m is the number of edges, n is the number of vertices, and $\log_2 B$ is the largest number of bits to store each capacity and gain. We implemented our program in Java and performed experiments on a PC with four 3.0 GHz CPUs (Xeon), 64 GB of RAM, and a 64-bit OS (Windows 2008 Server).

4.2 Evaluation of Rankings

A good evaluation of methods measuring relationships always requires human subjects, as performed in [3], [25], [4]. In this section, we first compare the rankings according to the strengths of relationships obtained by our method, GSD, PFIBF, CFEC and THT, with those obtained by human subjects. For our method, we set the gain function with $\alpha = 0.8$, $\beta = 0.8$, and $\lambda = 0.8$, which are determined by the estimation of gain function described in Section 4.3.

4.2.1 Relationships between People

For the source and the destination objects, we select famous person known by the participants creating the rankings by their subjects. We first select 10 famous Japanese and American politicians as source objects from Japanese Wikipedia, in order to enables the participants to investigate relationships among the persons on Wikipedia and create appropriate rankings. As the destination objects for each source, we select four famous persons related to the source. We select only four destinations for each source,

TABLE 1
Rankings of Persons

Source	Destinations	Human	Ours 3 hop	GSD	PFIBF 2 hop	CFEC 3 hop k=1000				THT d1 3 hop
						o1	og	d1	dg	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)

because we preliminarily observed that participants sometimes wavered in their judgments for five or more destinations. For each of the 40 obtained pairs of a source and a destination, we compute the strength of the relationship from the source to the destination using our method, GSD, PFIBF, CFEC, and THT, on the same data set explained in Section 4.1. We then obtain rankings according to the strengths. We search webpages in the domain of Japanese Wikipedia using keywords of the full names of these persons to compute GSD. For PFIBF, edge weight is assigned using the FB weighting method of its own [3]. For CFEC and THT, we implement them in four variants represented by the following four symbols. (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 $w(e) = \gamma(e)$ using our gain function described in Section 3.2; (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 $w(e) = \gamma(e)$, and set the weight $w(e_{rev})$ of every reversed edge e_{rev} to $w(e_{rev}) = rev(e)$, using our gain function. We compute THT for every value $L_{\max} = 1, 2, \ldots, 20$ which is the maximum length of paths explained in Section 2.1.

We compare the rankings yielded by these methods with those obtained by human subjects. For examining each of the 40 relationships, each participant read about five Wikipedia pages corresponding to or related to the source and the destination. Each participant gives an integer score between 0 and 10, independently to the others, as the strength of a relationship; a larger score represents a stronger relationship. We then obtain rankings according to the average of the scores given by 10 participants.

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

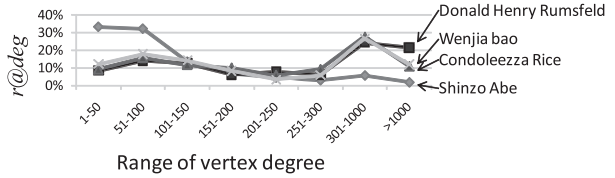

Fig. 8. Ratio of vertices by their degree.

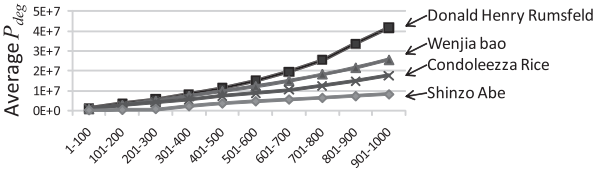

Fig. 9. Average P_{deg} of the k_1th to $k_1 + 99$th paths.

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.

The rankings obtained by PFIBF (3 hop) are much worse than those obtained by PFIBF (2 hop). Therefore, we describe the rankings of PFIBF (2 hop) only. We use $k = 1{,}000$ shortest paths for all the four variants of CFEC, through Section 4. Through the experiments for THT, the combination of variant (d1), using 3 hop links, and $L_{\max} = 3$, produced the best results among all the possible combinations. Therefore, we describe only the results of this combination later. One of the reasons why THT does not work well when $L_{\max} > 3$ is that THT estimates a path shorter than $L_{\max}$ for multiple times if $L_{\max} > 3$, similarly to PFIBF as discussed in Section 2.2. The rankings obtained by our method are the closest to those obtained by human subjects. However, some rankings created by other methods are inferior. For example, for "Donald Henry Rumsfeld," PFIBF, all the four variants of CFEC rank "Ronald Reagan" first, although the participants rank him third. For "Kiichi Miyazawa," GSD ranks "George H. W. Bush" fourth, although the participants rank him second. For "Yasuhiro Nakasone," THT ranks "Yuri Andropov" first, although he is ranked fourth by the participants. The variants of CEFC using double networks, (d1) and (dg), produce better rankings than the other variants, (o1) and (og), for some sources. For example, for source "Nobutaka Machimura," (d1) and (dg) produce the same ranking as the one obtained by human subjects, although (o1) and (og) do not. Especially, (o1) and (og) estimate the strength for the destinations "Condolezza Rice" and "George W. Bush" to be extremely small, because the original networks have few directed paths from the source to the destinations. In contrast, (d1) and (dg) are able to use a lot of paths, because the double network is constructed by adding the reverse edge to every edge of the original network.

Next, we discuss how popular objects affect CFEC. We confirmed that CFEC tends to assign a destination a extremely low score if the network between the destination and its source contains many popular objects. One of the examples is found in the relationships of source "Junichiro Koizumi." For "Junichiro Koizumi," CFEC (d1) ranked destination "Donald Rumsfeld" fourth, although human subject ranked it second. That is, CFEC underestimates the relationship between "Junichiro Koizumi" and "Donald Rumsfeld."

Fig. 8 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.

CFEC defines the strength of a relationship as the sum of the weights of the $k = 1{,}000$ shortest paths; the weight of path $p(s, \ldots, t)$ is defined as $1/P_{deg}$, where P_{deg} is the product of the degrees of the vertices except s in p. Fig. 9 depicts the average P_{deg} of the k_1th to $k_1 + 99$th shortest paths for each relationship, for k_1 is $1, 101, \ldots, 901$. The average P_{deg} for "Rumsfeld" increases most rapidly along with the rising of k_1 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 as claimed in Section 2.2.1.

We also compute the Pearson's correlation coefficient between the obtained strength and the score given by the participants. For each method, Fig. 10 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.

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

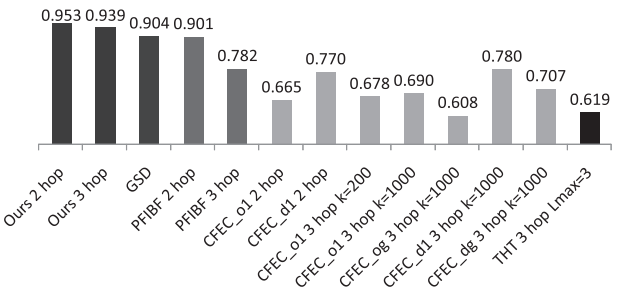

Fig. 10. Average correlation coefficient of each method.

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

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. We discuss the effectiveness of a doubled network and our gain function for our method in Section 4.3.

In addition to the methods appeared on Table 1, we compute the coefficients for SimRank [16] using parameters $C = 0.8$ and $K = 5$. The coefficients of SimRank (3 hop) using original networks and using doubled networks are 0.35 and -0.16, respectively. The results accord with the discussions presented in Section 2.1 that SimRank is inappropriate for measuring relationships in Wikipedia, even after using the doubled network.

It took 102 s to compute the generalized maximum flow using three hop links for the 40 relationships described above. The time for computing PFIBF (3 hop) is 400 s, which is about four times longer than our method. For computing CFEC (3 hop), using 200 shortest paths and 1,000 shortest paths took 91 and 5,631 s, respectively. The computing time of the generalized maximum flow was experimentally proportional to the number of edges, vertices, and types of edge gains. After SimRank was proposed by Jeh and Widom [16], several approaches were proposed to compute SimRank scalablely, such as the one proposed by Fogaras and Rácz [26]. While we focus on the accuracy problem of our method in this paper, one of our future work will be to construct a scalable method based on the generalized max flow by applying ideas used in the approaches.

4.2.2 Relationships between Petroleum and Countries

As another experiment, we obtain the rankings of the 192 countries according to the strengths of their relationships with "Petroleum" using each method. 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 (1) using the statistics about the oil production and consumption of the countries [27]

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

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 statistics-based 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 $\frac{|S_n|}{n}$, where S_n 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). Therefore, both a doubled

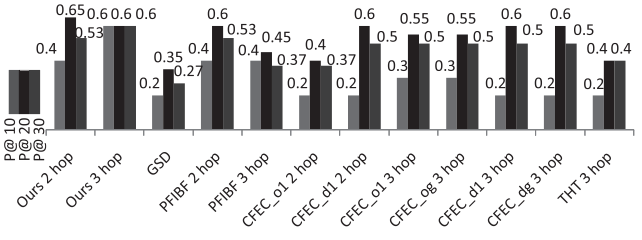

Fig. 11. Precision of rankings for petroleum.

TABLE 3
Average Correlation Coefficients with a Fixed Parameter

χ	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

network and our gain function are ineffective for CFEC in this experiment. The precision of THT is not better than that of CFEC. The precision of GSD are the worst here.

The experimental results presented in Sections 4.2.1 and 4.2.2 imply that our method is the most appropriate one for measuring the strength of a relationship in Wikipedia. Particularly, our method is the only choice for measuring 3-hop implicit relationships.

4.3 Estimation of Gain Function

In this section, we evaluate the parameters α, β, and λ for our gain function explained in Section 3.2. Let $\rho(\alpha, \beta, \lambda)$ be the correlation coefficient, averaged for the 40 relationships among politicians described in Section 4.2, depending on the values of parameters. We set the values of the parameters as $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, $\beta \in \{0.1, 0.2, \ldots, 1.0\}$ and $\lambda \in \{0, 0.1, \ldots, 1.0\}$. We compute $\rho(\alpha, \beta, \lambda)$ for all the possible $9 \times 10 \times 11 = 990$ combinations of values. Let $\bar{\rho}(\alpha = \chi)$ be the average of $\rho(\alpha, \beta, \lambda)$ obtained by the combinations of fixing $\alpha = \chi$ and varying β and λ. $\bar{\rho}(\beta = \chi)$ and $\bar{\rho}(\lambda = \chi)$ are similarly defined. Table 3 presents the averages $\bar{\rho}(\alpha = \chi)$, $\bar{\rho}(\beta = \chi)$, and $\bar{\rho}(\lambda = \chi)$. The differences between the averages are relatively small when χ is large. Therefore, our method is fairly robust against varying values of the parameters. The highest average for a fixed α is $\bar{\rho}(\alpha = 0.9) = 0.920$, that for β is $\bar{\rho}(\beta = 0.8) = 0.913$, and that for λ is $\bar{\rho}(\lambda = 1.0) = 0.893$. The shadowed cells in the row "$\bar{\rho}(\alpha = \chi)$" indicate that we could find no statistical significance among the distributions of $\rho(\alpha, \beta, \lambda)$ obtained by the combinations of fixing $\alpha = 0.7$, 0.8 or 0.9, by setting the significance level to 0.05. The shadowed cells in the two bottom rows have similar indication. Therefore, candidate combinations producing good results are $\alpha \in \{0.7, 0.8, 0.9\}$, $\beta \in \{0.7, 0.8, 0.9\}$, and $\lambda \in \{0.6, 0.7, ..., 1.0\}$. Similar candidate 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 $\alpha = 0.8$, $\beta = 0.8$, and $\lambda = 0.8$ which produces a medium result among the candidates.

In addition, we obtain the following observations:

- If $\beta = 1$, then the gain function is insensitive to groups, constructed from the category structure of Wikipedia as explained in Section 3.2. $\bar{\rho}(\beta = 1) = 0.899$ is worse than the best average. Therefore, the category structure is essential to our gain function.
- If $\lambda = 0$, then no reversed edges are used for measuring a relationship. $\bar{\rho}(\lambda = 0) = 0.810$ is the worst value in the bottom row. Therefore, reversed edges used for reflecting cocitation are effective in measuring a relationship. Consequently, the doubled network is a better choice for measuring relationships than the original Wikipedia information network.

TABLE 4
Correlation Coefficient for Each Method

Coefficient	Ours 3 hop	PFIBF 2 hop	CFEC_o1 3 hop k=1000	GSD	THT 3 hop L_{max}=3
Pearson	0.56	0.26	0.25	-0.15	-0.35
Spearman	0.60	0.47	0.48	0.25	0.41

4.4 Test Collection Measuring Word Relatedness

The WordSim353 test collection contains 353 word pairs for measuring word similarity or relatedness [23]. For every pair, a score representing similarity or relatedness is given as the ground truth. WordSim353 was used for evaluating relatedness between words in Wikipedia in [2]. Recently, Agirre et al. [24] classified the data set into two subsets, one for evaluating similarity, and the other for evaluating relatedness. In this section, we evaluate our method, GSD, PFIBF, CFEC, and THT using the latter subset which contains 252 word pairs. As discussed in [2], the test collection contains words which do not appear as titles of pages, except disambiguation pages, in Wikipedia. We select 130 word pairs which could be mapped to pages in an English Wikipedia data set (20100312 snapshot) for evaluation, similarly to the experiments conducted in [2]. For each method, we first compute the strength of the relationship between every pair of words; we then compute correlation coefficients between the obtained strengths and the ground truth.

Table 4 presents both the Pearson's correlation coefficient and the Spearman's rank-order coefficient for each method. Our method produced the highest Pearson's coefficient 0.56 and the highest Spearman's coefficient 0.60, both of which are much higher than those of other methods. PFIBF (2 hop) performs almost the same as CFEC (3 hop) using our doubled networks. THT produced a better Pearson's coefficient than PFIBF and CFEC did, while its Spearman's coefficient is worse than those of PFIBF and CFEC. Both coefficients of GSD are worst. As discussed in Section 2, GSD counts pages containing two words to measure the relationship between the two words. A word, especially a common noun, could be a part of a phrase representing a different object. For example, "life" is a part of "life hack." WordSim353 contains many common nouns, such as "life," "market," and "star." Therefore, GSD has a problem that it counts pages containing a word which are not necessarily pages containing the object represented by the word. The other methods do not suffer from such a problem because they use the Wikipedia information network to identify each object distinctly.

The test collection contains words in various categories, such as "OPEC," "Music," and "Ear." To verify whether our method is robust enough to measure relationships between objects of diverse kinds, we do the following processes 50 times, similarly to the experiments described in [2].

1. Randomly sample 100 pairs from the 130 word pairs explained above.
2. Compute the coefficients for each method using the selected 100 pairs.

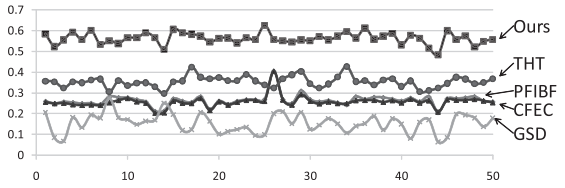

Fig. 12. Pearson's coefficients for each sample.

Figs. 12 and 13 depict the Pearson's coefficients and the Spearman's coefficients, respectively, of each method for each of the 50 samples. For GSD and THT, Fig. 12 depicts the absolute values of their coefficients, similarly to Fig. 10. Our method produces the highest coefficients for all the samples. The coefficients produced by PFIBF and CFEC are always almost the same. The Spearman's coefficients produced by THT are always worse than those of PFIBF and CFEC. However, the Pearson's coefficients of THT are better than those of PFIBF and CFEC for 49 samples. GSD performed inferiorly than other methods. Therefore, we conclude that our method is robust toward objects of various kinds, and the experimental results presented in Table 4 are valid.

4.4.1 A Case Study: Relationships between Planet and Other Words

We conduct a case study using a ranking of relationships between a fixed source and several kinds of objects as destinations. However, evaluation of such a ranking is a difficult task. We use WordSim353 again because it gives the ground-truth strength of some relationships which could be utilized for evaluating such a ranking.

The 130 word pairs explained in Section 4.4 contain 179 words. We first select "Planet" from the 179 words, we then rank the other 178 words according to the strengths of their respective relationships with "Planet." Five of the 130 word pairs consist of "Planet" and each of "Galaxy," "Constellation," "Astronomer," "Space," and "People" whose ground-truth strength are 8.11, 8.06, 7.94, 7.92, and 5.75, respectively. Table 5 presents the top 20 words ranked by each method. At a glance, the ranking obtained by GSD is inappropriate. For example, "Zoo" which would not have a strong relationship with planet, is ranked second by GSD. The other four rankings obtained by our method, PFIBF, CFEC, and THT are similar, and most of the top 20 words in all the four rankings seem have strong relationships with planet. "Galaxy," "Constellation," "Astronomer," and "Space" appear in all the four rankings, and "People" whose ground truth is much lower does not exist in the four rankings. However, the rankings of PFIBF, CFEC, and THT would be slightly inferior than our ranking. "Flight" ranked 18th by our method does not appear in the rankings obtained by other methods, although its relationship with "Planet" would be stronger than the relationships between "Planet" and each of "Music," "Television," "Child," and "Confidence" appearing in the other rankings.

TABLE 5
Rankings of Words for Planet

	Ours 3 hop	PFIBF 2 hop	CFEC 3 hop k=1000	GSD	THT 3 hop L_{max}=3
1	Atmosphere	Atmosphere	Constellation	Peace	Constellation
2	Mars	Constellation	Astronomer	Zoo	Astronomer
3	Day	Galaxy	Atmosphere	Health	Atmosphere
4	Galaxy	Astronomer	Galaxy	Sea	Galaxy
5	Astronomer	Mars	Day	Galaxy	Space
6	Constellation	Water	Water	Mars	Scientist
7	Life	Physics	Mars	Scientist	Water
8	Water	Day	Life	Atmosphere	Life
9	Energy	Energy	Space	Physics	Day
10	Space	Life	Nature	Weapon	Mars
11	Nature	Nature	Weather	Surface	Weather
12	Weather	Space	Scientist	Mind	Nature
13	Proton	Weather	Physics	Observation	Child
14	Physics	Proton	World	Laboratory	Planning
15	Month	Scientist	Energy	Astronomer	War
16	Scientist	Month	Computer	Constellation	World
17	Observation	Computer	Television	Dawn	Music
18	Flight	Chemistry	Music	Disaster	Physics
19	Reason	Observation	Chemistry	Flight	Computer
20	Chemistry	Music	Proton	Mouth	Confidence

4.5 Case Studies of Elucidatory Objects

For each relationship, our method outputs the top-k paths, 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. 14 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

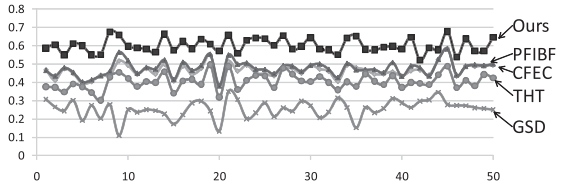

Fig. 13. Spearman's coefficients for each sample.

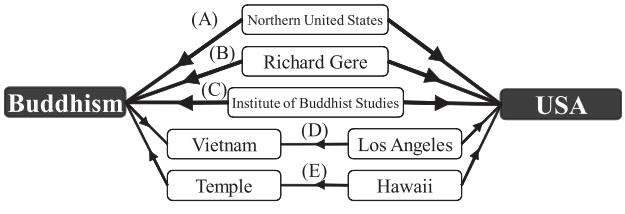

Fig. 14. Explaining the relationship between Buddhism and the USA.

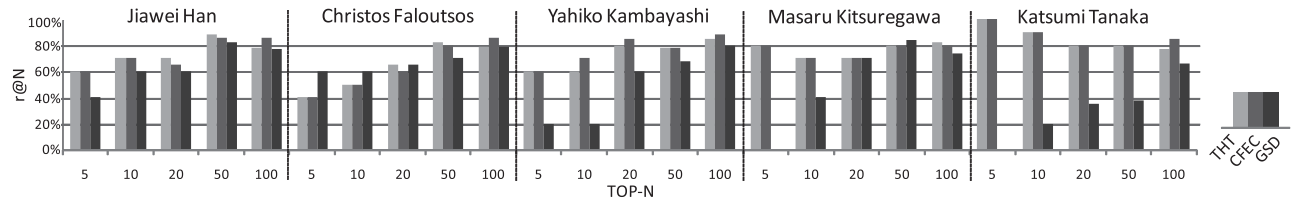

Fig. 15. Comparing rankings obtained on DBLP.

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.

The methods proposed by Faloutsos et al. [20], Tong and Faloutsos [28] and Koren et al. [1] visualize a subgraph for explaining a relationship. However, their subgraphs tend to be complicated, hence a user still must investigate important paths in the subgraph to understand the relationship. It is usually easier for a user to understand a relationship explained by simple paths rather than a complicated subgraph. As future work, we plan to utilize elucidatory objects to develop a system for explaining relationships.

4.6 Measuring Relationships on DBLP

In this section, we confirm that our method could be applied to other data sets, such as a DBLP network. A DBLP network consists of two types of edges: (e1) from a author to his/her papers; (e2) from a paper to the papers it cites. We conduct experiments to demonstrate that our method could measure relationships on a DBLP network as other methods could.

Let A_1 be a set of coauthors of given author a, A_2 and A_3 be a set of coauthors of every author in A_1 and A_2, respectively. We first randomly select 50 authors from each of A_1, A_2, and A_3. We then rank the 150 authors according to the strengths of their relationships with a, computed by each of our method, THT, CFEC, and GSD. We do not apply PFIBF [3] to DBLP because PFIBF is designed for Wikipedia only.

We set gain 0.9 for edges (e1) and their reversed edges; and we set gain 0.3 and 0.15 for edges (e2) and the reversed edges of (e2), respectively. The gain settings are based on the idea that a coauthorship is usually much stronger than the relationship between two authors a_s and a_t where the paper of a_s cites the paper of a_t. We set the capacity of every edge to one.

To measure the relationship between two authors, we construct a doubled network using authors, papers, and edges connecting the authors within 6 hops. We then compute our method, THT setting $L_{\max} = 10$, and CFEC on the doubled networks. For THT and CFEC, we set the weight to 1 for every edge, because setting the weight to the gain described above produces slight differences in their rankings.

It is difficult even for humans to evaluate a ranking according to the relationships between authors. Therefore, we examine how similar our ranking and the other rankings are. For author a, we compute the ratio $r@N$ of authors appearing in the top-N ranked by our method which also exist in that ranked by each of the other methods. Fig. 15 indicates $r@N$, where $N \in \{5, 10, 20, 50, 100\}$, for five authors, "Jiawei Han," "Christos Faloutsos," "Yahiko Kambayashi," "Masaru Kitsuregawa," and "Katsumi Tanaka." As a result, our method, THT and CFEC generated similar rankings for all the five authors. Especially, their top-5 authors are identical for "Tanaka." Where $N = 50$ or $N = 100$, around 80 percent of the authors appearing in the top-N of the ranking of our method also appear in that of THT or CFEC for each of the five authors. Similarly, at least 60 percent of the authors are common where $N \in \{5, 10, 20\}$ except for "Faloutsos." In contrast, the rankings of the three methods are less similar to those of GSD.

The results above give evidence that our method performs similarly to THT and CFEC on DBLP. Therefore, our method is possible to be applied to several kinds of data sets other than Wikipedia. Another candidate data set is a social network. Objects and edges in most social networks are much simpler than those in Wikipedia. Consequently, a gain function for a social network would be simpler than our function for Wikipedia, like our function for DBLP.

5 Conclusion

We have proposed a new method of measuring the strength of a relationship between two objects on Wikipedia. By using 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 compared with those by GSD [7], PFIBF [3], [2], CFEC [1], and THT [12]. 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.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (21013026) from The Ministry of Education, Culture, Sports, Science and Technology(MEXT), and Grant-in-Aid for Scientific Research(B) (20300036), and Grant-in-Aid for Young Scientists (B) (23700116) from Japan Society for the Promotion of Science (JSPS).

REFERENCES

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

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

[3] 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. Vitányi, "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 Erdös 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.

[16] 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.

[17] C.H. Hubbell, "An Input-Output Approach to Clique Identification," *Sociometry,* vol. 28, pp. 277-299, 1965.

[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. Faloutsos, K.S. Mccurley, and A. Tomkins, "Fast Discovery of Connection Subgraphs," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 118-127, 2004.

[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] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A Study on Similarity and Relatedness Using Distributional and Wordnet-Based Approaches," *Proc. 10th Human Language Technologies: Ann. Conf. North Am. Chapter of the Assoc. Computational Linguistics (NAACL-HLT),* pp. 19-27, 2009.

[25] 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.

[26] D. Fogaras and B. Rácz, "Practical Algorithms and Lower Bounds for Similarity Search in Massive Graphs," *IEEE Trans. Knowledge Data Eng.,* vol. 19, no. 5, pp. 585-598, May 2007.

[27] "Country Ranks 2009," http://www.photius.com/rankings/index.html, 2012.

[28] H. Tong and C. Faloutsos, "Center-Piece Subgraphs: Problem Definition and Fast Solutions," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 404-413, 2006.

Xinpeng Zhang received the BS degree from the School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2004, and the MS and PhD degrees in information science from the Graduate School of Informatics, Kyoto University, Japan, in 2009 and 2012, respectively. Since January 2012, he has been a researcher at NICT. His research interests include graph data mining, information retrieval, and natural language process. He is a member of the IEEE, IEEE Computer Society, and DBSJ.

Yasuhito Asano received the BS, MS, and DS degrees in information science, the University of Tokyo in 1998, 2000, and 2003, respectively. In 2003-2005, he was a research associate in the Graduate School of Information Sciences, Tohoku University. In 2006-2007, he was an assistant professor in the Department of Information Sciences, Tokyo Denki University. He joined Kyoto University in 2008, and he is currently an associate professor in the Graduate School of Informatics. His research interests include web mining, network algorithms. He is a member of the IEEE, IEICE, IPSJ, DBSJ, and OR Soc. Japan.

Masatoshi Yoshikawa received the BE, ME, and PhD degrees from the Department of Information Science, Kyoto University, in 1980, 1982, and 1985, respectively. From 1985 to 1993, he was with Kyoto Sangyo University. In 1993, he joined the Nara Institute of Science and Technology as an associate professor in the Graduate School of Information Science. From April 1996 to January 1997, he was in the Department of Computer Science, University of Waterloo as a visiting associate professor. From June 2002 to March 2006, he served as a professor at Nagoya University. From April 2006, he was a professor at Kyoto University. His current research interests include XML database, databases on the web, and multimedia databases. He is a member of the ACM and IPSJ.