ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

A GENERALISED FLOW BASED METHOD FOR ANALYSIS OF IMPLICIT RELATONSHIPS ON WIKIPEDIA

M.Vinoth Kumar,
Research Scholar (ph.D),
Dr. MGR Educational and Research Institute,
Maduravoyal,
Chennai,Tamilnadu, India

Dr.K.Selvam,
Professor, Dept of Computer Applications,
Dr. MGR Educatinal and Research Institute,
Maduravoyal,
Chennai,Tamilnadu, India

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

Keywords: Explicit and Implicit Relationships, Distance, Connectivity, co citation, THT .

I INTRODUCTION

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

1.1 ANALYSIS OF IMPLICIT RELATIONSHIP ON WIKIPEDIA:

Searching Webpages containing a keyword has grown in this decade, while knowledge search has recently been

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

explain the strength of a relationship. The main issue for measuring relationships arises from the fact that two kinds of relationships exist they are "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.

It might be represented by links and pages depicted in Fig.1 for an implicit relationship between two objects, the objects have to 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) to E exists if and only if object u to V has an explicit relationship from v to V .

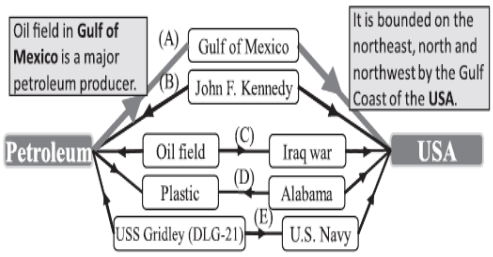

Fig. 1. Explaining the relationship between Petroleum and the USA.

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. Concept "cohesion" exists for measuring the strength of an implicit relationship.

CFEC proposed by Korean 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 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 like distance, connectivity, and co citation, 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 such as distance, connectivity, and co citation to check the implicit relationship. We measure relationships

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

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.

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 Calibres and Vita´nyi [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 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 > 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 "is called", "type" and "subclass of" 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. To understand the relationship deeply the elucidatory objects are examined in an appropriate manner.

1.2 MAJOR CONTRIBUTION METHODS

The major 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 (Section 3).

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

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).

These contributions are majorly used to find out the implicit relationship existing between each and every object in the Wikipedia. The similar relationships are identified by the methodical survey whereas the minimum and maximum strength between two objects depicts in the section 3.

The experiments and case studies of elucidatory objects are widely used to understand the relationships deeply without any confusion and congestion. Therefore all these four contributions are necessary to construct an implicit relationship in a generalized flow methodology. 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.

2. MEASURING IMPLICIT RELATIONSHIP BETWEEN TWO OBJECTS ON WIKIPEDIA

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, CO CITATION

The Erdo¨s number [10] used by mathematicians is based on distance and co authorships. The legendary mathematician Paul Erdo¨s 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 distance, 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. However, the Erdo¨s number is inadequate to represent the implicit relationship between a person and Erdo¨s because the number does not estimate the connectivity between them. The hitting time [12] from vertex s to vertex is defined as the expected number of steps in a random walk starting from s before t is visited for the first time.

Actually, the hitting time from s to t in a network represents the average length of all the paths connecting s and t. Sarkar and Moore [12] proposed “Truncated Hitting Time” (THT) to compute the average length of paths connecting two vertices whose length are at most Lmax only. A smaller distance represents a larger similarity.

THT does not estimate the connectivity between two vertices presents in the hitting time vertices. For example, suppose only m_1 vertex disjoint paths of length k connect s to t. THT computes the distance from s to t to be k for any m_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,

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

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 when compared with the similar one.

Co citation-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 Calibres and Vita´nyi [7] can be regarded as a co-occurrence based method; it measures the strength of a relationship between two words by counting of w 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 co citation. 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 co citation. Co citation-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 ðv;wÞ. In contrast, 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 ðu;wÞ or ðu; v0; v1; wÞ. Such paths abound in the Wikipedia information network. Therefore, SimRank is inappropriate for measuring relationships on Wikipedia.

Instead of setting capacities, we use a generalized maximum flow by setting every gain to a value less than one. Therefore, the value of a maximum flow in our method decreases if the distance becomes longer. Co citation-based methods assume that two objects have a strong relationship if the number of objects linked by both the two objects. Such paths bounds in the Wikipedia information network. Implicit relationship between "Petroleum" and the "USA." Our method can mine elucidatory objects constituting a relationship by outputting paths using the contribution methods.

By 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

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

elucidatory objects would open a novel way to deeply understand a relationship. Several semantic search engines have been used for searching relationships between two objects, using a semantic knowledge base extracted from web or 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 too 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_3 and two edges (u, v) and ðv; uÞ exist, then PFIBF counts both path (u, v) and path ðu; v; u; vÞ containing a cycle ðu; v; uÞ. 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_2.

In fact, PFIBF usually sets k ¼ 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. The property is suitable for several kinds of networks in which popular objects are considered as noise, such as stop words or portal sit.

$$\text{path } p = (s = v_1, v_2, \ldots, v_\ell = t) \text{ from } s \text{ to } t \text{ as}$$

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

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

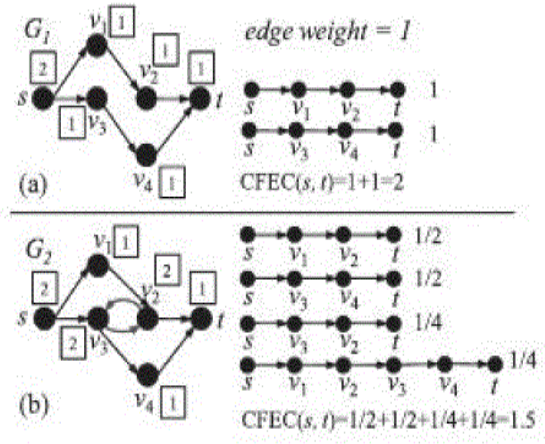

Fig.2 CFEC on two networks.

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 too 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.

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 Pbush containing "Bush" than that to path POlmert containing "Olmert" because "Bush" is more popular, although path PBush would be not less important than path POlmert in this example. There are many cases similar to this example in Wikipedia. , the three concepts, distance, connectivity, and co citation, are important concepts for measuring relationships; cohesion-based methods underestimate popular objects, although popular objects might be important for relationships in Wikipedia to represent the relationship between the rice and koizumi.

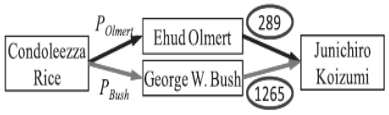

Fig 3. A Relationship between Rice and Koizumi

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. 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 ci.

We 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 of ci. The three categories are distance, connectivity and co citation. Where distance is the length of the shortest path and it represents a stronger relationship between the other remaining two categories equally all the time.

3. METHOD FOR ANALYSIS OF IMPLICIT RELATIONSHIPS USING GENERALIZED FLOW

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

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

smaller weight to path Pbush containing "Bush" than that to path POlmert containing where w(u, v) is the weight of edge (u,v) and wsumðvÞ 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 Cðs; tÞ 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 wsum of each vertex is written in the rectangle near the vertex.

The weight of each path is presented at the right side of the path. For the network G1 depicted in Fig.2a, the wsum of s is 2, and the weight of path ðs; v1; v2; tÞ is 1. Cðs; tÞ for G1 is 2, which is equal to the connectivity between s and t. If we add two edges ðv2; v3Þ and ðv3; v2Þ to G1, then we obtain network G2 in Fig.2b.

Two vertices v2 and v3 become more popular in G2 than they are in G1, and Cðs; tÞ decreases from 2 in G1 to 1.5 in G2. 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 Pbush containing "Bush" than that to path POlmert containing "Olmert" because "Bush" is more popular, although path PBush would be not less important than path POlmert 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. As discussed in Section 2, the three concepts, distance, connectivity, and co citation, 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.

In Wikipedia, to rank the relationships between two objects, the link will be measured among two pages and that represent the relationship. The Existing system could measures the relationship and it based on one or two methods of the following factors that is the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

distance, the connectivity and the co citation.

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 ðeÞ>0; the value of a flow sent along edge e is multiplied by ðeÞ. Let fðeÞ_0 be the flow f on edge e, and ðeÞ_0 be the capacity of edge e. The capacity constraint fðeÞ , ðeÞ 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, u, t, r) be information network ðV ; EÞ with the source s to V, the destination t to V, the capacity and the gain.

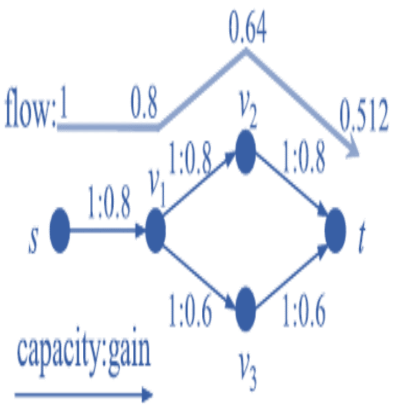

Fig.4 Generalized Maximum Flow

Fig.4 depicts an example of a generalized maximum flow on a generalized network. One unit of flow is sent from the source s to v1, i.e., fðs; v1Þ ¼ 1, the amount of the flow is multiplied by _ðs; v1Þ 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 ¼ (u, v) must hold before the gain is multiplied with fðs; v1Þ ¼ 1 _ðs; v1Þ 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 can reflect the three representative concepts explained in Section 2 distance, connectivity and co citation.

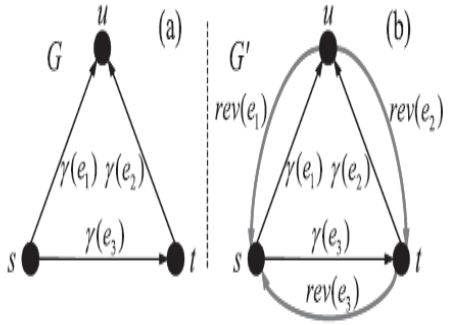

Fig. 5 A Doubled Network

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.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

Therefore, it also can be used to estimate the connectivity. We discuss the co citation 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 co citation of two objects. We consider the relationship between two objects s and t in the network presented in Fig.5a. Object u is co cited by s and t.

This co citation is represented by two edges ðs; uÞ and ðt; uÞ. However, we were unable to send a flow from s to t along the two edges, unless we reverse the direction of the edge ðt; uÞ to ðu; tÞ. 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 _ðe(u,v)Þ. The constraint is satisfied on the doubled network if we introduce a new constraint fðe(u,v)Þfðerevðv; uÞÞ¼ 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. That is the value of a flow f and G can be doubled by the network and g be a generalized maximum flow in G to satisfy the constrain equation.

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 tha $g_c(e)g_c(e_{rev}) = 0$ 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 cancelling the flow-absorbing cycles. Cancelling 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, $|g'| = |g| = |g_c|$. □

In order to determine the proposition, first consider what kinds of explicit relationships are important in constituting an implicit relationship. In a generalized max-flow problem, a path composed of edges with large scan contributes to the value of a flow. To

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

realize such a gain assignment, construct groups of objects in Wikipedia. Categories cannot be used as groups directly because the category structure of Wikipedia is too fractionalized. Mutation can result in several different types of change in sequences, Mutations in genes can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

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 A0 is trying to send a message to a Japanese politician J0 in the real life; A0 has no explicit relationship to J0, and another American politician A1 and an Israeli politician I0 have respective explicit relationships to J0.

In this case, A0 would tend to ask A1, rather than I0, to help transferring the message to J0. A0 could contact A1 easily compared to J0 because A0 and A1 belong to the same group "American politician." We therefore regard the explicit relationship between A1 and J0 as primarily important in constituting the relationship between A0 and J0.

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 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.

3.2.1 CATEGORY GROUPING

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

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 represented by ci. 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 ci. We 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. Therefore, we decide to construct a "category group" for a specified category ci in the following way.

For category ci of Wikipedia, let AðciÞ 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 AðciÞ are depicted by trapezoids in Fig6. A Wikipedia is a type of content management system, it differs most other such systems in that the content is created without any defined owner or leader, and Wikipedia have some implicit structure allowing to emerge according to the needs of the users. The Wikipedia is the most famous wiki on the public web, but there are many sites running different kinds of wiki software.

Let DðciÞ be the set of descendant categories of ci, which are depicted by triangles in Fig 6. We regard AðciÞ [DðciÞ [fcig is the set of kin categories of ci. Categories other than the kin categories are depicted by stars in Fig 6. We then regard a category in DðciÞ as an Irrelevant descendant if the category is not a child of ci and is linked from ci.

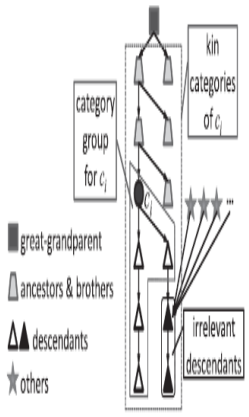

Fig. 6 Grouping for Category ci

Irrelevant descendants are depicted by filled triangles in Fig 6. Let D0ðciÞ be a subset of DðciÞ, which is obtained by removing the irrelevant descendants from DðciÞ. Then, we define D0ðciÞ [fcig as the category group for ci. 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.

3.2.2 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 Cs of categories to which s belongs. Similarly, we specify a set Ct 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 Cs or Ct, 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.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

Therefore, through the paper we assume that such categories are manually removed from Cs or Ct. 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].

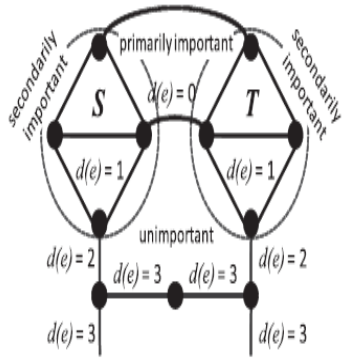

Fig. 7 Gain Function

We then construct 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 obtain the set T for t. The assumptions discussed in the beginning of this section can be formalized using S and T respectively every time.

The edges (u, v) such that u $u \in S \wedge v \in T$ 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.

In order to determine the gain function, first consider what kinds of explicit relationships are important in constituting an implicit relationship. In a generalized max-flow problem, a path composed of edges with large gain scan contributes to the value of a flow. To realize such a gain assignment, construct groups of objects in Wikipedia. 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 ¼ (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) ¼ 0;

if u € S ^ v € S or u 2 T ^ v 2 T, then $d(e) = 0$; if $u \in S \wedge v \in$ 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.

Categories cannot be used as groups directly because the category structure of Wikipedia is too fractionalized. Mutation can result in several different types of change in

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

sequences, Mutations in genes can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely.

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, μ, γ) containing s and t from Wikipedia, by determining the parameters and explained in Section 3.2. We set the capacity of every edge to one.2. Determine the parameter explained in Section 3.2 for reversed edge gain rev for G, and construct the doubled network Grev of G for rev.3. Compute a generalized maximum flow g in Grev.

2. Let deg ðoÞ 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.

3. 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 ¼ 3 at step 1. Our method can be applied to both directed network and undirected network.

For an undirected network, we set _ ¼1 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 degðsÞ or degðtÞ 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 degðsÞor degðtÞ is large. On the other hand, the strength of the relationship between s and t is expected to be independent of degðsÞ and degðtÞ.

Therefore, we decide to divide the value of the flow by function $D(s,t) = \sqrt{deg(s) * deg(t)}$ at step 2.

$D'(s,t) = deg(s) * deg(t)$ or $D''(s,t) = \log(deg(s) * deg(t))$.

We also tried several other functions such as 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 D0 or D00 does. If we use D0 instead of D, then the value of D0 excessively dominates the strength of a relationship, because the value increases much faster according to the increase of degðsÞ and degðtÞ than the effect of the size G does; on the other hand, the value of D00 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.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

We compute the strength of relationships by dividing the value of a flow by $\sqrt{deg(t)}$, because estimating $\sqrt{deg(t)}$, does not affect the ranking. [11] proposed a system in which FASTRA downloads and data transfers can be carried over a high speed internet network. On enhancement of the algorithm, the new algorithm holds the key for many new frontiers to be explored in case of congestion control. The congestion control algorithm is currently running on Linux platform. The Windows platform is the widely used one. By proper Simulation applications, in Windows we can implement the same congestion control algorithm for Windows platform also. The Torrents application which we are currently using can achieve speeds similar to or better than —Rapid share (premium user) application

VI 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 co citation, 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.

Thus this method can measure the strength of a relationship between two objects on Wikipedia and rank them. Furthermore, this method does not underestimate objects having high degrees. This paper plans to form a relation tree, a directed tree with a unique node corresponding to the most recent common ancestor of all the entities at the leaves of the tree.

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.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

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

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

[9] F.M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of
Semantic Knowledge," Proc. 16th Int'l Conf. World wide Web Conf.
(WWW), pp. 697-706, 2007.

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

[11] Christo Ananth, A. Ramalakshmi, S. Velammal,B. Rajalakshmi Chmizh, M. Esakki Deepana, "FASTRA –SAFE AND SECURE", International Journal For Technological Research In Engineering (IJTRE), Volume 1, Issue 12, August-2014,pp: 1433-14380.

[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 Co citation: A Literature
Measure of Intellectual Structure," J. Am. Soc. Information Science
and Technology, vol. 32, no. 3, pp. 163-171, May 1981.