A GLOBAL FLOW-BASED TECHNIQUE FOR ANALYSIS OF INHERENT INTERACTION ON WIKIPEDIA

[1]T.KIRAN KUMAR REDDY, [2]S.ABDUL JEELAN, [3]M.GIRI

[1]Student, Dept of Computer Science & Engineering, Sreenivasa Institute of Technology and Management Studies, Chittoor, India

[2]Assistant Professor, Dept of Computer Science & Engineering, Sreenivasa Institute of Technology and Management Studies, Chittoor, India

HOD, Assistant Professor ,Dept of Computer Science & Engineering, Sreenivasa Institute of Technology and Management Studies, Chittoor, India

Email: [1]kiranreddy541@gmail.com, [2]jeelan01@gmail.com, [3]prof.m.giri@gmail.com

Abstract: These documents address how to improve a website without introducing significant changes. Specifically, we propose a mathematical programming model to improve the user steering on a website while minimize alteration to its current structure. Results from extensive tests conducted on a publicly available real data set indicate that our model not only significantly improves the user steering with very few changes, but also can be successfully solved. We have also tested the model on large synthetic data sets to display that it scales up very well. In addition, we define two evaluation metrics and use them to assess the presentation of the improved website using the real data set. Evaluation results confirm that the user steering on the improved structure is indeed greatly enhanced. More interestingly, we find that heavily disoriented users are more likely to benefit from the improved arrangement than the less disoriented users. We propose a new method using a worldwide 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 constitute a relationship. We explain that mining elucidatory objects would open a novel way to deeply understand a relationship.

Keyword: Mathematical programming model, Large synthetic data, Two evaluation metrics

I. INTRODUCTION

In this paper, we are worried primarily with transformation approaches. The literature bearing in mind transformations approaches mainly focuses on developing methods to totally reorganize the link structure of a website. Although there are advocates for website reorganization approaches, their drawbacks are obvious. First, since a complete reorganization could radically change the location of familiar items, the new website may disorient users. Second, the reorganized website structure is highly unpredictable, and the cost of disorienting users after the changes remains unanalyzed. This is because a website's structure is classically designed by experts and bears business or organizational logic, but this logic may no longer exist in the new structure when the website is completely reorganized. Besides, no prior studies have assessed the usability of a completely reorganized website, leading to doubts on the applicability of the reorganization approaches. Finally, since website reorganization approaches could dramatically change the current structure, they cannot be frequently performed to improve the navigability. Recognizing the drawbacks of website reorganization approaches, we address the question of how to improve the structure of a website rather than reorganize it substantially. Specifically, we develop a mathematics programming (MP) model that facilitates user navigation on a website with negligible changes to its current structure. Our model is mainly appropriate for informational websites whose contents are static and comparatively stable over time. Examples of organizations that have informational websites are university, tourist attractions, hospitals, federal agencies, and sports organizations. Our model, however, may not be suitable for websites that purely use dynamic pages or have unbalanced contents. This is because a steady state might never be reached in user access patterns in such websites, so it may not be budding to use the weblog data to improve the site structure. The number of outward links in a page, i.e., the outdegree, is an important factor in modeling web structure. Prior studies typically model it as hard restriction so that pages in the new arrangement cannot have more links than a specified out-degree threshold, because having too many links in a page can cause information overload to users and considered undesirable. For instance, Lin uses 6, 8, and 10 as the out-degree threshold in experiments. This modeling approach, however, enforces severe restrictions on the new structure, as it prohibits pages from having more links than a specified threshold, even if adding these links may really facilitate user steering Our model formulates the out-degree as a cost term in the objective function to discipline pages that have more links than the threshold, so a page's out-degree may exceed the entrance the cost of adding such links can be justified. We perform extensive trial on a data set collected from a real website. The results indicate that our model can significantly improve the site

organization with only few changes. Besides, the optimal solutions of the MP model are successfully obtained, suggesting that our model is sensible to real-world websites. We also test our model with synthetic data sets that are considerably better than the real data set and other data sets tested in previous studies address website reorganization problem. The solution times are remarkably low for all cases tested, ranging from a fraction of second to up to 34 seconds. Moreover, the solution times are shown to increase logically with the size of the website, indicating that the proposed MP model can be easily scaled to a large extent. 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, 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" on an in order 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 target 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 widespread maximum flow. We propose a heuristic gain function utilizing the category configuration 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 more than a few pages from Wikipedia as our source objects; and for each source object, we choose several pages as the purpose objects. We then total the power of the relationship between a source object and each of its destination objects, and rank the destination objects by the strength. By compare the ranking obtained by our method with those obtained by the "Google Similarity Distance" (GSD) proposed by Cilibrasi and Vitanyi, PFIBF and CFEC, we ascertain that the ranking 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 in order 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 depicts an example of a 3-hop implicit relationship between "Petroleum" and the "USA." Our method can mine elucidatory matter constitute 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 that mining 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. 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 obtainable in Wikipedia, such as "Gulf of Mexico" is a major "petroleum" producer. We do not utilize the semantic knowledge bases for measure relationships in this paper.

II. OBSERVATIONS ON REDUCTION OF PROBLEM SIZE

The formulation has $|E|$ binary variables corresponding to the number of candidate links and $|T^R|$ constraints equivalent to the number of relevant mini sessions. While in practice the size of a website and the number of mini sessions obtain from server logs can be very large, it turns out that, in the situation of our problem, the formulation can be reduced to a significantly smaller one that can be quickly solved. We make several observations related to the problem size.

These observations together provide insights into why the problem size in our formulation can be considerably reduced and help explain the fast solution times in our experiments in the later sections. In fact, as will be shown later, our formulation has already taken steps to reduce the problem size.

2.1 Relevant Mini Sessions

Recall that a mini session is relevant only if its length is larger than the equivalent path threshold. Consequently, only relevant mini sessions need to be careful for improvement and this leads to a large number of irrelevant mini sessions (denoted as T^I) being eliminated from consideration in our MP model. In other words, define $T^I = T \setminus T^R$, any mini session $S \in T^I$ will not be considered in our formulation as the user navigation in S already meets the goal (set as path threshold).

As will be shown later, the choice of path entry can have significant impacts on relevant mini sessions. Generally, increasing the path threshold leads to a smaller number of relevant mini sessions while decreasing it has the opposite effect. For example, for the real data set used in the experiments (details are provided in Section), when the path threshold increases from, the number of related mini sessions reduces from several thousand to only a few hundred. Even for the case when $b = 1$, a large number of irrelevant mini sessions can be eliminate from consideration.

2.2 Relevant Candidate Links

Define $E' = \{(i,j) : i,j \in N\}$ as the possible links between all pages in a website with node set N. Theoretically, any link from E' can be careful in our decision problem without a preprocessing step, leading to a total number of $|N| \times |N|$ links (variables). This number can be very large even for a small website. Intuitively, not every link in E be used to improve user navigation. Recall that we term the links that can be selected to help user navigate as candidate links (denoted by E), which can be easily obtained from mini sessions. Thus, it follows that $x_{ij} = 0$ in the optimal $\forall (i,j) \in E' \backslash E$. In other words, non candidate links do not help improve user steering and hence need not center the formulation. It turns out that many candidate links can also be eliminated from reflection because they are not relevant to the decision for two reasons. First, given path thresholds, denote the set of candidate links for relevant mini sessions by E^{IM} and the set of candidate links for irrelevant mini sessions by E^{IM}. It follows that the candidate links in $E^{IM} \backslash E^{RM}$ are only for irrelevant mini sessions that need no development and hence can be eliminated from consideration. Second, not all candidate links in E^{RM} might be relevant to the decision. Particularly, for a mini session $S \in T^R$ with path threshold b, a link is said to be relevant to S if adding/improving it can help the user in S reach the target in no more than b paths, i.e., achieve the user navigation goal.

	n_1	n_2	n_3	n_4	n_5	n_6
n_1	0	0	0	1	1	0
n_2	1	0	0	1	1	1
n_3	1	1	0	0	0	0
n_4	1	1	1	0	1	0
n_5	1	1	1	1	0	1
n_6	0	1	1	1	0	0

The Connectivity matrix for illustrative examples.

in S. Thus, the candidate links applicable to the decision are those originating from the pages visited on the bth path or before. The other candidate links for S can be eliminated from consideration because selecting them cannot improve S to achieve the specified goal for user navigation. We term the set of contender links relevant to the decision the relevant candidate links, and we denote them by $E^R = \{(i,j) : (i,j) \in E \text{ and } \exists S \in T^R$ such that $i \in S,\ j = tgt(S), a^S_{ijkr} = 1 \text{ for } 1 \le k \le b_j, \text{ and } 1 \le r \le L_p(k,S)\}$. This leads to $x_{ij} = 0$ in the optimal solution $\forall (i,j) \in E^{RM} \backslash E^R$. Thus, the cardinality of the set of related candidate links $(|E^R|)$ could be relatively small even for a large website.

2.3 Dominated Mini Sessions

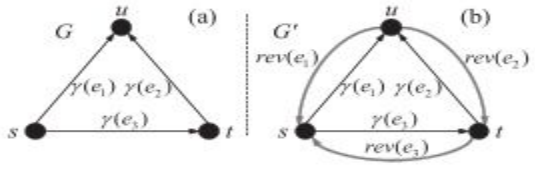
Another reason for the problem size decrease is that many applicable mini sessions "dominate" others with respect to relevant candidate links. Mini session S_q dominates mini session S_q if the set of relevant candidate links for S_q contains (at least) all relevant candidate links for S_p. This dominance is strict if there exists at least one candidate link that is relevant to S_q but is irrelevant to S_p. Therefore, when a mini session is improved in the new structure, the mini sessions that are dominated by this one are also improved. Consequently, the constraints corresponding to dominated mini sessions are redundant and can be eliminated from consideration in the MP model.

III. METHOD FOR MEASURING RELATIONSHIPS USING GENERALIZED FLOW

3.1Generalized 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 $f(e) \ge 0$ be the flow f on edge e, and $\mu(e) \ge 0$ be the capacity of edge e. The capacity constraint $f(e) \le \mu(e)$ must hold for every edge e. The goal of the problem is to send a flow emanate 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 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 $\gamma(s, v_1)$Þ when the flow arrives at v Consequently, only 0.8 units arrive at v 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. $\gamma(s,v_1)$ 1 $f(s,v_1) = 1 \le \mu(s,v_1)$ 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 might 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



A doubled network.

can reflect the three representative concepts explained in 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 quantity 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.

3.2 Gain Function for Wikipedia

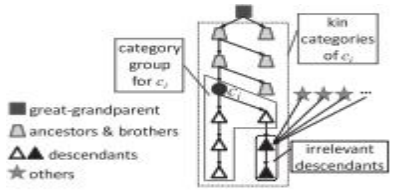
In order to determine the gain function, we think what kinds of explicit relationships are significant in constituting an implicit relationship. Suppose an American politician A_0 is trying to send a message to a Japanese politician J_0 in the real life; A_0 has no explicit relationship to J_0 another American politician A_0 and an Israeli politician I have respective explicit relationships to A_0 would tend to ask A_0, rather than I_0. In this case, A, to help transfer the message to J_0. A_0 could contact A_0 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 A_1 and J_0 as primarily important in constituting the relationship between A_0 and J . For the example depicted in, "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 conversation above, for analyzing an implicit relationship between objects 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

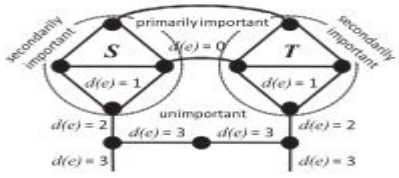


Grouping for category c

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

.3.2.1 Category Grouping

A category c_i representing a concept might have descendant categories each on behalf of 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



Gain function.

3.2.2 The Gain Function

We now propose the gain meaning 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_t or C_s, 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_t.In preliminary experiments, we determine 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 C_s. The set S for s consists of objects belonging to any category in the category groups for C. Similarly, we obtain the set T for t.

3.3 Summary of the Proposed Method

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

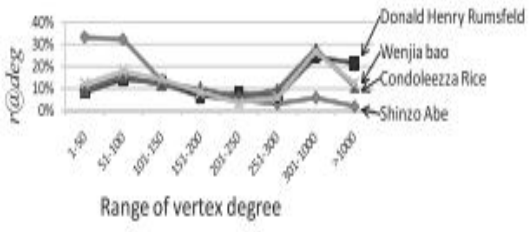
1. Construct a generalized network $G = (V, E, s, t, \mu, \gamma)$ containing s and t from Wikipedia, by determining the parameters α and β explained in Section. We set the capacity of every edge to one.
2. Determine the parameter λ explained in Section for inverted edge gain rev for G, and construct the doubled network G of G for rev.
3. Compute a generalized maximum flow g in $\bar{G}_{rev}$
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, 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 surveillance 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 compute 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 λ =1 to use both directions of an edge equally.

IV. EXPERIMENTS AND EVALUATION

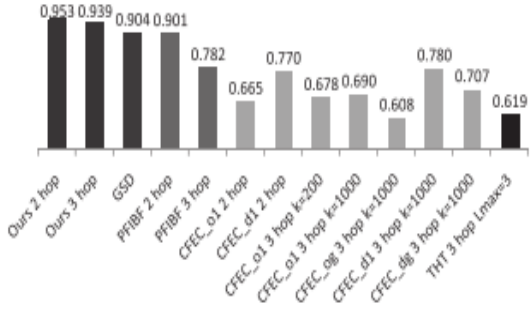
For the source and the destination objects, we select famous person known by the participants creating the rankings by their subjects. We first select 10 famous Japanese and American politicians as source objects from Japanese Wikipedia, in order to enables the participants to investigate relationships 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 destination for each source, 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 explain in Section .We then obtain rankings according to the strengths. 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. For CFEC and THT, we implement them in four variants represented by the following four symbols. (o1) Compute



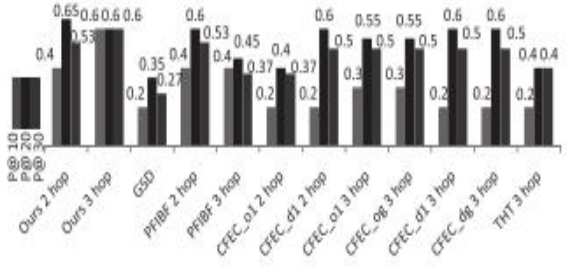
Ratio of vertices by their degree.

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.



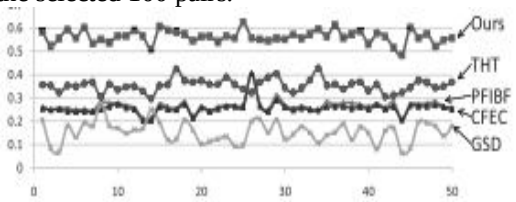
Average correlation coefficient of each method.

We then evaluate the precision at the top n countries of a ranking, abbreviated to P@n, computed by $\frac{|S-1|}{n}$, where S is the set of countries appear in both the ranking and the statistics-based ranking. Fig depicts P@10, P@20, and P@30 of all rankings. Similarly to the results of the first experiment depict in Fig., 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 accuracy of every variant of CFEC (3 hop). Therefore, both a doubled



network and our gain function are ineffective for CFEC in this experiment. The precision of THT is not better than that of CFEC. The precision of GSD are the worst here. Table presents both the Pearson's correlation coefficient and the Spearman's rank-order coefficient for each method. Our method shaped 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'scoe efficient 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 , GSD counts pages containing two words to measure the relationship between the two words. A word, specially 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 contain 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 in order 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.

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



Pearson's coefficients for each sample.

CONCLUSION

We have proposed a new method of measure 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 substance 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, PFIBF, CFEC, and THT. Particularly, our method is the only choice for measuring 3-hop implicit relationships. We have also confirmed that elucidatory substances are helpful to deeply understand a relationship. Some future challenges remain. We are also involved in seeking possibilities of the elucidatory objects constituting a relationship mined by our method. We plan to quantitatively evaluate the elucidatory objects. We are developing a tool for deeply understanding relationships by utilizing elucidatory objects.

REFERENCES

[1] Y. Koren, S.C. North, and C. Volinsky, "Measuring and Extracting Proximity in Networks," Proc. 12th ACM SIGKDD Int'l Conf. 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. 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 Erdos 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. 20572060, 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. Pas¸ca, 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. Ra cz, "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.

★★★