INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 3 ISSUE 1 –JANUARY 2015 - ISSN: 2349 - 9303

Enhancing Implicit Relations in Wikipedia Mining Using Object Relationship Technique

G.Shanmugapriya[1]
[1]B.S Abdur Rahman University, Computer Science,
sarushiya@gmail.com

S.Raja shaik[2]
[2]B.S Abdur Rahman University, Computer Science
srajashaik@gmail.com

Abstract— Relationships are there between objects in Wikipedia. Emphases on determining relationships are there between pairs of objects in Wikipedia whose pages can be regarded as separate objects. Two classes of relationships between two objects exist there in Wikipedia, an explicit relationship is illustrated by a single link between the two pages for the objects, and the other implicit relationship is illustrated by a link structure containing the two pages. Some of the before proposed techniques for determining relationships are cohesion-based techniques, and this technique which underestimate objects containing higher degree values and also such objects could be significant in constituting relationships in Wikipedia. The other techniques are inadequate for determining implicit relationships because they use only one or two of the following three important factor such as the distance, the connectivity and the cocitation.

Index Terms— Cocitation, Connectivity, Generalized flow, Measurement.

.

I.INTRODUCTION

In this decade for knowing about the things which we are not aware we depend on the search engine. There exists various search engines in that Wikipedia is one of most popular search engine and also the knowledge about a particular object in Wikipedia is brought in to a single wiki page (web page) and it is updated repeatedly by different volunteers. For searching a particular object we first enter the search string in the search engine and then we press search button and related server will display set of co citations in a wiki page.

A relationship is an association between two or more people that may range in duration from brief to enduring. Word Association is a common word game involving an exchange of words that are associated together. This association may be based on regular business interactions and might from some other type of social commitment. .The Interpersonal relationships will formed in the context of social network. The Interpersonal relationships were dynamic systems that will change continuously during their existence.

The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and also about Web based data warehousing. However, there is no established vocabulary, it will make leading to confusion when it comparing to research efforts. The word Web mining has been used in two distinct ways. The first type is called as Web content mining; the second type is called as Web usage mining. The Web mining defines the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of webdata.Web content mining, which is also known as text mining, is the second step in Web data mining. The web Content mining is the scanning and mining of text, words, images and graphs of a Web page to determine the relevance of the content to the search query.

The Web usage mining is the third category in web mining. This type will allow for the collection of Web access information for Web pages. And this usage data provides the paths leading to accessed Web pages. This information is often gathered routinely into access logs via the Web server.

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. Wiki promotes meaningful topic associations between different pages by making page link creation almost intuitively easy and showing whether an intended target page exists or not. Wiki is not a sensibly shaped site for visitors. Instead, it seeks to involve the visitor in an ongoing process of creation and collaboration that constantly changes the Web site landscape.

The nature of mining processes creates a potential negative impact on the environment both during the mining operations and for years after the mine is closed. This impact led to most of the world's nations adopting regulations to moderate the negative effects of mining operations. Safety has long been a concern a modern practices have improved safety in mines significantly. The data mining mission is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This involves using database techniques such as spatial indices. These such patterns can then be seen as a kind of summary of the input data, and might be used in further analysis or, for instance, in machine learning and of predictive analytics. For case, the data mining step might identify multiple groups in the data, which can then be used to obtain more exact prediction results by a decision support system.

Neither the data collection, data preparation, nor the result interpretation and reporting are part of the data mining step, but does belong to the overall knowledge discovery process as additional steps.

Data mining uses information from past data to analyse the outcome of a particular problem or situation that may happen. Data mining mechanism to analyse data stored in data warehouses that are used to store that data that is being researched. That specific data may come from all percentages of business, from the construction to the management. Directors also use data miniagree upon marketing strategies for their product. They can use data to associate and contrast among competitors. Data mining reads its data into real time analysis that can be used to rise sales, help new product, or obliterate product that is not value-added to the company.

Text analysis involves information recovery, verbal analysis, pattern acknowledgement, labelling/annotation, the information extraction, data mining techniques including link and association analysis, conception, and predictive analytics. The predominant goal is basically from short text into data for study, via suggestion of natural language processing (NLP) and analytical methods. A representative application is to scanning a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

In Wikipedia, the knowledge of an object is gathered in a single page updated constantly by a number of co-worker's. The Wikipedia also protects objects in a number of groups, such as society, knowledge, natural features, diplomatic, and past. Hence, keening Wikipedia is frequently a better choice for a user to obtain knowledge of a single object than typical search engines. A user also may desire to discover a relationship between two objects. The typical keyword search engines can neither measure nor explain the strength of a relationship. The foremost problem for measuring relationships arises from the fact that two kinds of relationships exist:“explicit relationships” and “implicit relationships.”

This paper proceeds as follows. The Division II which presents the related work and the global operations of system. The different metrics and components of the system are presented in Division III. The Division IV describes about the components. The Division V describes the conclusion and future works.

II.RELATED EFFORT

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

A. Dataset

Dataset in Wikipedia includes objects that have relationships between each other. The existing system implies that how to measure the strength of relationships among the pages.

1. Dataset Loading

A data set is a collection of data, it list out values for each of the variables. The query used to create a particular data set from the selected connection or flat file profile. Many data set definitions can be formed for the same profile in order to generate different data set instance. To improve categorization accuracy, irrelevant parameters and enduring data could be deleted from the data set.

2. Data Preprocessing

Data pre-processing is an important step in the data mining process .Analysing data that has not been screened for such problems can produce fake results. Thus, the quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy data and then knowledge discovery during the working out phase is added difficult. Data preparation and filtering steps would take considerable amount of processing time. The Data pre-processing method which also includes cleaning, normalization, alteration, selection, etc.

The product of data pre-processing is the final training place. The absent of attribute values, absent of certain attributes of interest, or might containing only aggregate data and reduce the volume but producing the same or similar analytical results.

B. Generalized Maximum Flow-based Method

We are having various methods for measuring the strength of the objects. Such as “cohesion” concept is used for measuring the strength of relationships. “CFEC” are based on the concept “cohesion”. One of them is generalized maximum flow. Cohesion based method is adequate because it does support higher degree objects and the other methods which proposed earlier will follow concepts like distance, connectivity, co citation. These three are important factors for implicit relationship but it is also in adequate for measuring the implicit relationships. Behind all these concepts generalized maximum flow was introduced for measuring the strength of relationships following the three factors such as the distance, the connectivity and the co citation. In the generalized maximum flow gain function is used for measuring the relationships.

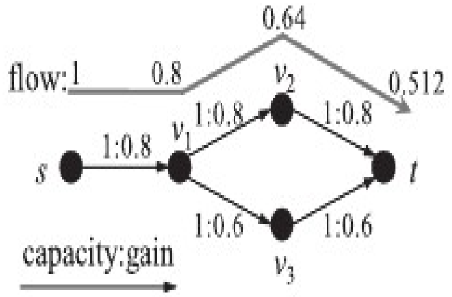


Fig 1.Generalized Maximum Flow

1. Distance

In this method distance calculates shorter path and it represents a stronger relationship. For this method, for every edge; then a flow considerably decreases including elongated path. A shorter path usually contributes to the generalized maximum flow by a greater amount than a long path does. So, a shorter path will exists means there is s stronger relationship also exists.

2. Connectivity

In social network analysis, cohesion based methods are used to measure the strength of relationship. The generalized maximum flow problem is a natural extension of the problem of classical maximum flow. So it also can be used to estimate the connectivity.

3. Cocitation

Cocitation related techniques assume that two nodes have a stronger relationship if the number of nodes linked by both the two nodes is large and at the other end co-occurrence is a concept by which the strength is represented by the number of nodes linking to the both objects. Further, the use of both directions is required to estimate the co citation of two objects.

III.METHODS TO MEASURE THE RELATIONSHIP

A. Relationship on Wikipedia Module

Propose a new method for measuring a relationship on Wikipedia by reflecting all the three concepts: distance, connectivity, and co-citation. I propose a new method for measuring the strength of a relationship using the method of generalized maximum flow. To measure the strength of a relationship from one object to another object, 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 consider the vertices in the paths composing the generalized maximum flow as the objects constituting the relationship. We ascertain that the claim that our method can reflect the three representative concepts.

B. Cycle-Free Effective Conductance (CFEC)

The cycle-free escape probability is the probability that a random walk originating will reached without visiting any node repeatedly. The transition from one state to another does not depend on the previous state and the transition could remain in same state. The proximity is the infinite number of attempts in networks that is made to reach from starting node to end node.

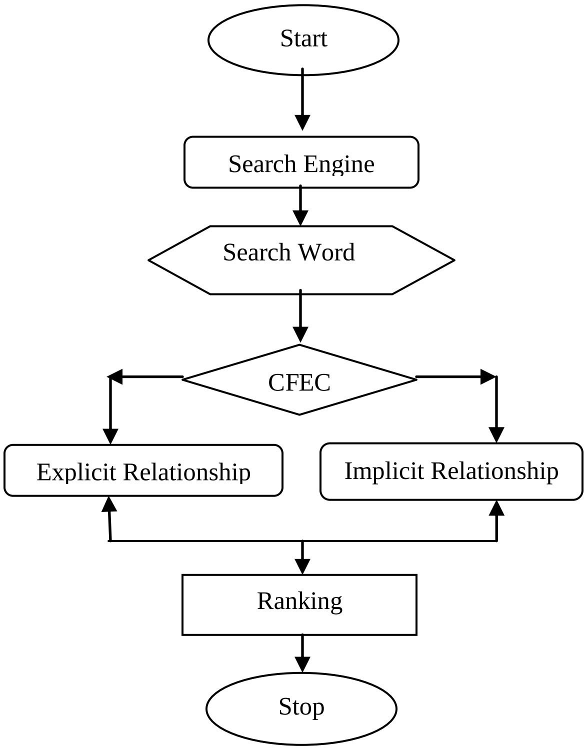


Fig 2. Work Flow Diagram For CFEC Networks

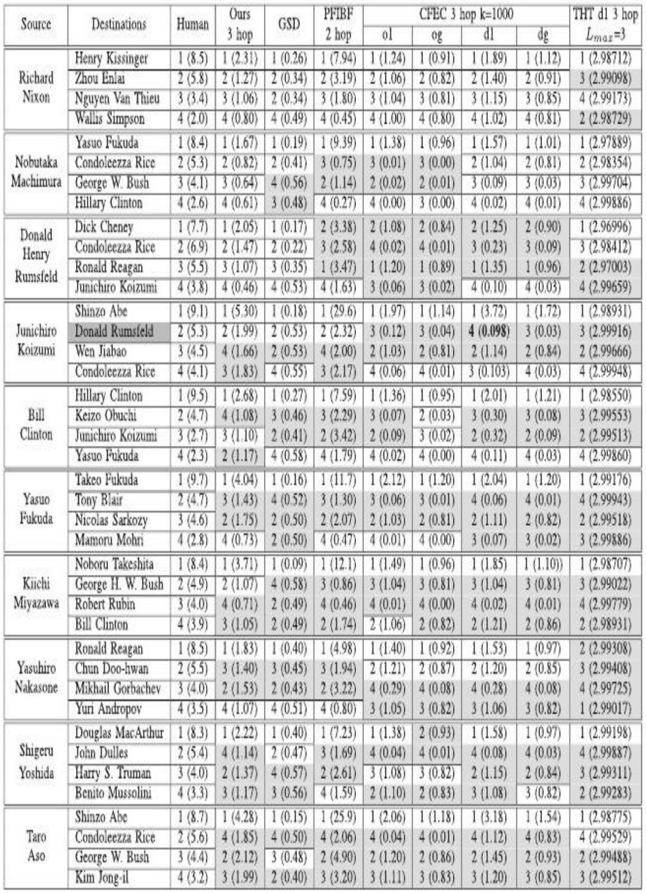
CFEC proximity allows to compute proximity graphs, that will defined as small portions of the network that are aimed at capturing a related proximity value. It is extended form of connection graph which is capable of presenting dense relationship between the objects of a network. It can presume relationship among more than two end points, the stretch to handle and would obtained by solving an tuneable optimization problem.

IV.RANKING

A ranking is a relationship between a set of items, for case take any two items, the first is either ranked as higher than as or lower than or equal to the second. By decreasing detailed measures to a order of ordinal numbers and rankings make it possible to estimate difficult information according to certain criteria.

In competition ranking, items that compare equal receive the same ranking number. The number of ranking numbers is missing out in this gap is one other less than the other number of items that it compared equal. The mission of distinct ordinal numbers to items that compare equal can be done at randomly as this gives stable results if the ranking is done at multiple times. Query-independent methods was made to measure the probable prominence of a page, independent of any consideration of how fit it matches the specific query.

TABLE 1. RANKING OF PERSONS



Source	Destinations	Human	Ours 3 hop	GSD	PFIBF 2 hop	CFEC 3 hop k=1000				THT d1 3 hop L_{max}=3
						ol	og	dl	dg	
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)

In the above example, for the source and the destination objects, select the famed person which is well-known by the participants and creating the rankings by their subjects. For each source, select four famed persons as the destination objects which related to the source. Here only four destinations for each source is selected and for each of the obtained pairs of a source and a destination, the strength of the relationship from the source to the destination using this method is computed.

V.RESULTS

The result arrives that enhancing the implicit relations on Wikipedia by using all the three methods of generalized flow based method which is not previously implemented and therefore the implicit relationship will not be underestimate the high degree values.

VI.CONCLUSION AND FUTURE WORKS

Thus this method can measure the strength of a relationship between two objects on Wikipedia and rank them. Some future challenges will remain and also interested in seeking possibilities of the elucidatory objects constituting a relationship mined by this method. This paper plans to evaluate the elucidatory objects.

VII.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, “AssociationThesaurus Construction Methods Based on Link Co-OccurrenceAnalysis for Wikipedia,” Proc. 17th ACM Conf. Information andKnowledge Management (CIKM), pp. 817-826, 2008.

[3] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia Mining for anAssociation 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 SemanticRelatedness,” 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,” PhDdissertation, Cornell Univ., New York, Jan. 1999.

[7] R.L. Cilibrasi and P.M.B. Vita´nyi, “The Google SimilarityDistance,” 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. IEEE24th Int’l Conf. Data Eng. (ICDE), pp. 953-962, 2008.

[9] F.M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A Core ofSemantic 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] L. Katz, “A New Status Index Derived from SociometricAnalysis,” Psychometrika, vol. 18, no. 1, pp. 39-43, 1953.

[12] S. Wasserman and K. Faust, Social Network Analysis: Methods andApplication (Structural Analysis in the Social Sciences). CambridgeUniv. Press, 1994.

[13] C. Faloutsos, K.S. Mccurley, and A. Tomkins, “Fast Discovery ofConnectionSubgraphs,” Proc. 10th ACM SIGKDD Int’l Conf.Knowledge Discovery and Data Mining, pp. 118-127, 2004.

[14] P.G. Doyle and J.L. Snell, Random Walks and Electric Networks,vol. 22. Math. Assoc. Am., 1984.

[15] M. Nakatani, A. Jatowt, and K. Tanaka, “Easiest-First Search:Towards Comprehension-Based Web Search,” Proc. 18th ACMConf. Information and Knowledge Management (CIKM), pp. 2057-2060, 2009.

[16] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G.Wolfman, and E. Ruppin, The WordSimilarity-353 Test Collection,2002.

[17] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pas¸ca, and A.Soroa, "A Study on Similarity and Relatedness Using DistributionalandWordnet-Based Approaches," Proc. 10th HumanLanguage Technologies: Ann. Conf. North Am. Chapter of the Assoc.Computational Linguistics (NAACL-HLT), pp. 19-27, 2009.

[18] D. Fogaras and B. Ra´cz, "Practical Algorithms and Lower Boundsfor Similarity Search in Massive Graphs," IEEE Trans. KnowledgeData Eng., vol. 19, no. 5, pp. 585-598, May 2007.

[19] W. Xi, E.A. Fox, W. Fan, B. Zhang, Z. Chen, J. Yan, and D. Zhuang,"Simfusion: Measuring Similarity Using Unified RelationshipMatrix," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research andDevelopment in Information Retrieval, pp. 130-137, 2005.

[20]"CountryRanks2009,"http://www.photius.com/rankings/index.html, 2012.