ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

Link Analysis in Wikipedia Using Subtree Mining with Relationships Extraction

Dr.K. Venkatasalam, Dr.P.Ramya

Department of Computer Science and Engineering, Mahendra Engineering College, Mallasamudram, Namakkal, India

Associate Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Mallasamudram, Tamil Nadu, India

ABSTRACT: Measuring relationships between pairs of objects in Wikipedia is focused . Wikipedia objects are often considered individual objects. Two sorts of relationships between objects exist: in Wikipedia, a certain relationship is pictured by one link between the two pages for the objects, associate degreed an implicit relationship is pictured by a link structure containing the two pages. A number of the antecedently planned ways for measurement relationships are cohesion-based ways, that underestimate objects having high degrees, though such objects can be necessary in constituting relationships in Wikipedia. The opposite ways are inadequate for measurement implicit relationships as a result of they use only one or two of the subsequent three necessary factors: distance, property, and cocitation. Here proposed a brand new methodology employing a generalized flow that reflects all the 3 factors and doesn't underestimate objects having high degree. This methodology will live the strength of a relationship a lot of fittingly than these antecedently planned ways do. Another outstanding facet of our methodology is mining elucidatory objects, that is, objects constituting a relationship.

I. INTRODUCTION

Several methods have been proposed for measuring the strength of a relationship between two objects on an information network. A Wikipedia is a type of content management system, it differs from a blog or most other such systems in that the content is created without any defined owner or leader, and wikis have little implicit structure, allowing structure to emerge according to the needs of the users. The encyclopedia project Wikipedia is the most famous wiki on the public web, but there are many sites running many 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. A wiki is not a carefully crafted site for casual visitors. Instead, it seeks to involve the visitor in an ongoing process of creation and collaboration that constantly changes the Web site landscape.

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. Typical keyword search engines can neither measure nor explain the strength of a relationship. The main issue for measuring relationships arises from the fact that two kinds of relationships exist: “explicit relationships” and “implicit relationships.”

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. Measuring a relationship on Wikipedia by reflecting all the three concepts: distance, connectivity, and cocitation. We measure relationship rather than similarities.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

II. EXISTING SYSTEM

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) ε E exists if and only if object u ε V has an explicit relationship to v ε V . Consider Wikipedia information network whose vertices are pages of Wikipedia and whose edges are links between pages. Concept "cohesion," exists for measuring the strength of an implicit relationship.

The cohesion based methods are used in calculating the relationship between objects. And it deals with measuring the strength of a relationship by counting all paths between two objects. It has a property that its value greatly increases if a popular object, an object linked from or to many objects, exists. PFIBF is unsuitable for measuring relationships in Wikipedia because of popular objects.

2.1 POPULAR OBJECTS IN WIKIPEDIA

An object linked from or to many objects, exists in wikipedia. This property is a defect for measuring the strength of a relationship. Both PFIBF and CFEC use the concept of popular objects for measuring the relations among Wikipedia pages.

The weight of a path becomes extremely small if a popular object exists in the path. The strength C(s, t) of the relationship between s and t is the sum of the weights of all paths from s to t. Figure 3.1 depicts two networks and all the paths between s and t.

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

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.

III. PROPOSED METHOD

The aim of this work is to measure relationships between pairs of objects in Wikipedia whose pages are often considered individual objects. We measure relationships rather than similarities. Assignment of the gain to each edge is important for measuring a relationship using a generalized maximum flow. The gain function is sufficient to measure relationships appropriately. Hence propose a generalized maximum flow-based method reflects all the three concepts and does not underestimates popular objects, in order to measure relationships on Wikipedia appropriately. 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 $\gamma(e)$. 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.

3.1 DISTANCE, CONNECTIVITY, COCITATION

The methods based on distance, a shorter path represents a stronger relationship. For evaluation, 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. Connectivity, a strong relationship is represented by many vertex disjoint paths from the source to the destination. The

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

number of vertex disjoint paths can be computed by solving a classical maximum flow problem. Therefore, it also can be used to estimate the connectivity.

Cocitation flow that emanates from the source into the destination, and therefore the flow seldom use an edge whose direction is opposite that from the source to the destination. On the other hand, we require use of both directions to estimate the cocitation of two objects.

3.2 CATEGORY GROUPING

A category representing a concept might have descendant categories each representing its sub concept. A part of descendant categories do not represent sub concepts of one. Categories are usually linked from more than three categories other than the categories. Grouping of data used in XSL Transformations that identifies keys in the results and then queries all nodes with that key. This improves the traditional alternative for grouping, whereby each node is checked against previous (or following) nodes to determine if the key is unique.

3.3 GAIN FUNCTION

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. We first specify a set Cs of categories to which s belongs. Similarly, we specify a set Ct for t. However, several categories contain too many unrelated pages. Such categories are unsuitable for grouping related objects.

IV. CONCLUSION

The proposed method of measuring the strength of a relationship between two objects on Wikipedia is efficient. By using a generalized maximum flow, the three representative concepts, distance, connectivity, and cocitation can be reflected. The existing method underestimates the objects having higher degree and ranking of objects are not good. Thus the relationship between objects are estimated wrongly. The proposed method does not underestimate objects having high degrees. It can obtain a fairly reasonable ranking according to the strength of relationships. Tree construction yields better ordering compared to previous methods. Due to enhanced category grouping of pages or objects and thus the most appropriate pages are retrieved and the computational cost and time taken are also minimized. It supports 3-hop implicit relations. Thus the elucidatory objects are mined.

REFERENCES

1. Xinpeng Zhang, Yasuhito Asano and Masatoshi Yoshikawa (2013) 'A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia' IEEE Transactions On Knowledge And Data Engineering, Vol. 25.
2. Christos Faloutsos, Mccurley.K.S, And Andrew Tomkins (2004) 'Fast Discovery Of Connection Subgraphs', Proc. 10th ACMSIGKDD Int'l Conf. Knowledge Discovery And Data Mining, Pp. 118-127.
3. Da´Niel Fogaras and Bala´Zs Ra´Cz (2007) 'Practical Algorithms And Lower Bounds For Similarity Search In Massive Graphs ', IEEE Trans. Knowledge Data Eng., Vol. 19, No. 5, Pp. 585-598.
4. Dat P.T. Nguyen, Yutaka Matsuo and Mitsuru Ishizuka (2007) 'Relation Extraction Fromwikipedia Using Subtree Mining',Proc. Twenty-Second Conference on Artificial Intelligence (AAAI-07)
5. Davide Buscaldi and Paolo Rosso (2006) 'Mining Knowledge From Wikipedia For The Question Answering Task',Proceedings of the Fifth International Conference on Language Resources and Evaluation.
6. David Milne and Ian H. Witten (2008) 'An Effective, Low-Cost Measure Of Semantic Relatedness Obtained From Wikipedia Links', Proc.AAAI Workshop Wikipedia And Artificial Intelligence: An Evolving Synergy.
7. Eneko Agirre, Enrique Alfonsecaz, Keith Hall, Jana Kravalova, Marius Pas and Aitor Soroa (2009) '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.
8. Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum (2007) 'Yago: A Core Of Semantic Knowledge Unifying Wordnet And Wikipedia', Proc. 16th Int'l Conf. World Wide Web Conf. (Www), Pp. 697-706.
9. Glen Jeh and Jennifer Widom (2002) 'Simrank: A Measure Of Structural-Context Similarity', Proc. Eighth ACM Sigkdd Int'l Conf. Knowledge Discovery And Data Mining, Pp. 538-543.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

10. Hai-Hsiang Yang, Chun-Yu Chen, Hahn-Ming Lee and Jan-Ming Ho (2008) 'EFS: Expert Finding System Based On Wikipedia Link Pattern Analysis', in Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics.
11. Hanghang Tong and Christos Faloutsos (2006) 'Center-Piece Subgraphs: Problem Definition And Fast Solutions', Proc. 12th ACM Sigkdd Int'l Conf. Knowledge Discovery And Data Mining, Pp. 404-413.
12. Jintao Tang, Ting Wang, Qin Lu, Ji Wang and Wenjie Li (2012) 'A Wikipedia Based Semantic Graph Model For Topic Tracking In Blogosphere',Proceeding of the Twenty-Second international joint conference on Artificial Pages 2337-2342.
13. Julian Szyma´Nski (2010) 'Mining Relations Between Wikipedia Categories '.
14. Majid Yazdani and Andrei Popescu-Belis (2010) 'A Random Walk Framework To Compute Textual Semantic Similarity: A Unified Model For Three Benchmark Tasks', Proc. IEEE Fourth Int'l Conf. Semantic Computing, Pp. 424-429.
15. Nakatani.M, Jatowt.Aand Tanaka.K, "Easiest-First Search: Towards Comprehension-Based Web Search," Proc. 18th ACM Conf. Information and Knowledge Management (Cikm), Pp. 2057- 2060.
16. Olena Medelyan, David Milne and Ian H. Witten (2008) 'Mining Domain-Specific Thesauri From Wikipedia: A Case Study', Proceeding WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence Pages 442-448.
17. Purnamrita Sarkar and Andrew W. Moore (2007) 'A Tractable Approach To Finding Closest Truncated-Commute-Time Neighbors In Large Graphs', Proc. 23rd Conf. Uncertainty In Artificial Intelligence (UAI).
18. Rudiger Gleim, Alexander Mehler and Matthias Dehmer (2009) 'Web Corpus Mining By Instance Of Wikipedia', ProceedingWAC '06Proceedings of the 2nd InternationalWorkshop onWeb asCorpus, Pages 67-74.
19. Silviu Cucerzan (2007) 'Large-Scale Named Entity Disambiguation Based On Wikipedia Data'.
20. Wangzhong Lu .J, Janssen .E, Milios.N and Japkowicz (2006)'Yongzheng Zhang (2006) 'Node Similarity In The Citation Graph', Knowledge and Information Systems, Vol. 11, No. 1, Pp. 105-129.
21. Yehuda Koren, Stephen C. North and Chris Volinsky (2006) 'Measuring And Extracting Proximity Graphs In Networks' , Proc. 12th ACM SIGKDD Int'l Conf.Knowledge Discovery And Data Mining, Pp. 245-255.