ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

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

Gopika S J[1]
PG Scholar
Department of Computer Science, College of Engineering Perumon
Kerala - India

Sowmya K S[2]
Assistant Professor
Department Of Information Technology, College of Engineering Perumon,
Kerala - India

Abstract: ***The main focus of this method aims at measuring relationships between pairs of objects in Wikipedia whose pages can be regarded as individual objects. There are two types 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. The earlier cohesion based methods for measuring relationships 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. Then the new method using a generalized maximum flow is introduced which reflects all the three factors and does not underestimate objects having high degree. By experiments it is confirmed that this method can measure the strength of a relationship more appropriately than these previously proposed methods do. Another remarkable aspect of this method is mining elucidatory objects, that is, objects constituting a relationship. It proves that mining elucidatory objects would open a novel way to deeply understand a relationship.***

Keywords: link analysis, generalised flow, Wikipedia mining, relationships.

I. INTRODUCTION

Knowledge search has recently been extended to obtain knowledge of single object and relationships between multiple objects. Searching knowledge of objects using Wikipedia is one of the hottest topics in the field of knowledge search. In Wikipedia, the knowledge of an object is gathered in a single page updated constantly by a number of volunteers. Wikipedia also covers objects in a number of categories, such as people, science, geography, politic, and history.

A user also might desire to discover a relationship between two objects. For example, a user might desire to know which countries are strongly related to petroleum, or to know why one country has a stronger relationship to petroleum than another country. Typical keyword search engines can neither measure nor explain the strength of a relationship. The main issue for measuring relationships arises from the fact that two kinds of relationships exist: “explicit relationships” and “implicit relationships.” In Wikipedia, an explicit relationship is represented by a link. A user could understand the explicit relationship easily by reading the text surrounding the anchor text of the link. For example, an explicit relationship between petroleum and Gulf of Mexico might be represented by a link from Gulf of Mexico to petroleum. An implicit relationship is represented by multiple links and pages. For example, Gulf of Mexico is a major oil producer in the USA. This fact could be an implicit relationship represented by two links one from petroleum to Gulf of Mexico and other from Gulf of Mexico to USA. It is difficult for a user to understand an implicit relationship without searching a number of pages and links. Therefore it is an interesting task to mine and explain implicit relationships in Wikipedia.

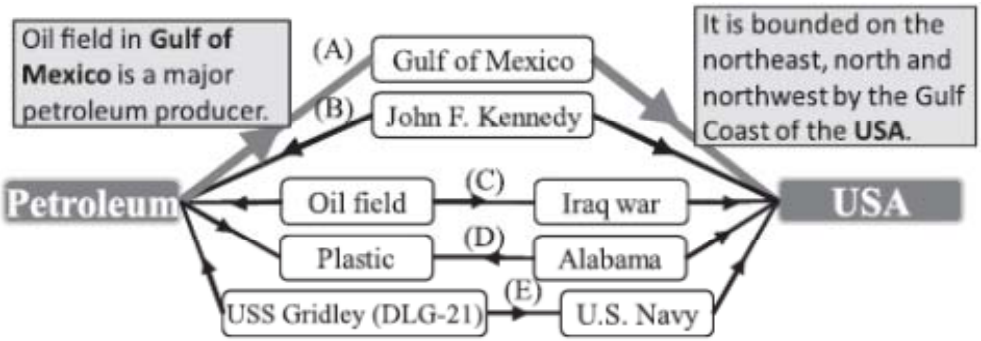

Fig 1: Explaining the Relationship between Petroleum and USA

II. Comparison between Cohesion based Methods and Generalised Maximum Flow

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) element of E exists if and only if object u element of V has an explicit relationship to v element of V . Here it defines 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. and PFIBF proposed by Nakayama et al. is based on cohesion. Here it does 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. 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. The method used here is "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. Here it proposes a heuristic gain function utilizing the category structure in Wikipedia. It is confirmed through experiments that the gain function is sufficient to measure relationships appropriately.

The method is evaluated using computational experiments on Wikipedia. First, select several pages from Wikipedia as source objects; and for each source object, then select several pages as the destination objects and 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 the method with those obtained by the "Google Similarity Distance" (GSD) proposed by Cilibrasi and Vita´nyi [7], PFIBF and CFEC, and ascertain that the rankings obtained by this method are the closest to the rankings obtained by human subjects. Especially, it is ascertain that only this 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. Here it says 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 $k > 1$. Fig. 1 depicts an example of a 3-hop implicit relationship between "Petroleum" and the "USA." This method can mine elucidatory objects constituting a relationship by outputting paths contributing to the generalized maximum flow, that is, paths along which a large relationship is a k-hop implicit relationship if the sub graph contains a path from s to t whose length is at least $k > 1$. Fig. 1 depicts an example of a 3-hop implicit relationship between "Petroleum" and the "USA." This 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. Then it will explain that mining elucidatory objects would open a novel way to deeply understand a relationship. Several semantic search engines [8] have been used for searching relationships between two objects, using a semantic

knowledge base [9] extracted from web or Wikipedia. However, the semantics in these knowledge bases, such as "isCalled," "type" and "subClassOf," are mainly used to construct 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. Here it do not utilize the semantic knowledge bases for measuring relationships.

The main contributions are as follows:

1. A detailed and methodical survey of related work for measuring relationships or similarities.
2. A new method using generalized maximum flow for measuring the strength of a relationship between two objects on Wikipedia, which reflects the three concepts: distance, connectivity, and co-citation.
3. Experiments on Wikipedia showing that our method is the most appropriate one.
4. Case studies of mining elucidatory objects for deeply understanding a relationship.

III. CONCLUSION

Here it 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, this method does not underestimate objects having high degrees.

It is ascertained that this method can obtain a fairly reasonable ranking according to the strength of relationships by this method compared with those by PFIBF , CFEC ,Particularly, this method is the only choice for measuring 3-hop implicit relationships and also confirmed that elucidatory objects are helpful to deeply understand a relationship.

The Demerits of earlier methods are PFIBF cannot distinguish a path containing a cycle from a path containing no cycle. PFIBF has a property that it estimates a single path, eg: (u,v) for multiple times. PFIBF is inappropriate for measuring a 3-hop implicit relationship. EC has the same drawback as PFIBF: it counts a path containing a cycle redundantly. (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. CFEC and PFIBF are unsuitable for measuring relationships in Wikipedia because of popular 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.

