ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Measure of Implicit Relationship between Words by Network Count

Keerthana.S.M[1], V.Rajakumareswaran[2]

PG Student / CSE, KSR Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India[1]

AP/ CSE, KSR Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India[2]

ABSTRACT: Searching Wikipedia is usually a better choice for a user to obtain knowledge of a single object than typical search engines. In Wikipedia, the knowledge of an object is gathered in a single page updated constantly by a number of volunteers. Two kinds of relationships exist between two objects in Wikipedia, an explicit relationship and an implicit relationship. Some of the previously proposed methods for measuring relationships are cohesion-based methods and the category grouping is not that much effective in calculating the relationship between pages. Thus this method does not obtain the accurate relationship between pages. So the proposed method makes use of the generalized maximum flow which calculates the gain function. The relationship and connectivy between objects is extracted using CFEC. The three important factors: distance, connectivity, and cocitation are calculated. The efficient category grouping makes use of tree construction. And is constructed by using the degree of relevance between the objects and the relationship is extracted. Thus the relevance calculation is effective in the proposed method than the typical search engine and other previous methods.

I. INTRODUCTION

Wikipedia, a collaborative Wiki-based encyclopedia, has become a huge phenomenon among Internet users. According to statistics of Nature, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica. It covers concepts of various fields such as Arts, Geography, History, Science, Sports, and Games. It contains more than 4,375,110 and it is becoming larger day by day while the largest paper-based encyclopedia Britannica contains only 4,094,309articles. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, sense disambiguation based on URL, brief link texts and well-structured sentences. Furthermore, Wikipedia contains quite a bit of structured information: it has a rich category structure, separate pages for ambiguous terms, and structured data for certain types of articles. Finally, it contains over 90million links between articles. Most of these links signify that some semantic relationship holds between the source and target concepts and can used to compute a measure of relatedness that typically outperforms traditional text similarity measures.

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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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.