Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Qualitative implicit Relationship Identification through Cut Detection on Wikipedia

P.Naga laxmi*
Department of CSE
IARE,JNTUH,HYD-43,A.P, India

Dr.Nagu.Chandra Sekhar Reddy
Professor ,Department of CSE
IARE,JNTUH,HYD-43,A.P, India

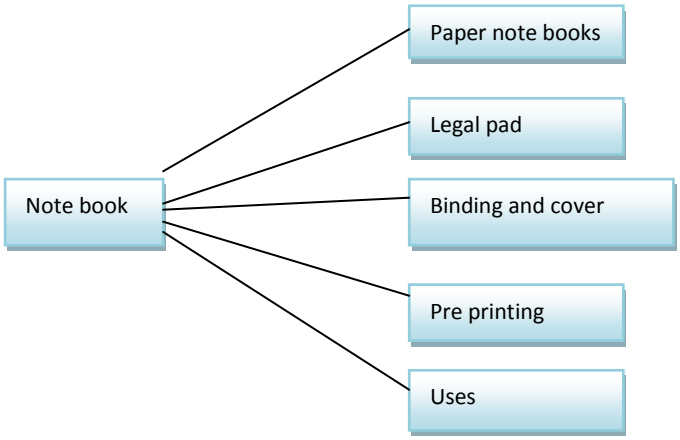
P. Ila Chandana Kumara
Asst.,Prof., Department of CSE
IARE,JNTUH,HYD-43,A.P, India

Abstract - We do use Wikipedia for searching knowledge of objects; there exists two types of relationships in between the wiki pages- explicit and implicit relationship. Explicit relationship is represented by a single link in between two wiki pages for the nodes and implicit relationship is represented by a link structure in between the objects. In the earlier cases cohesion based methods are used for measuring the implicit relationships but it inadequate for measuring implicit relationships. In the existing system generalized maximum flow method is used for measuring the implicit relationship. But by using this we may not get quality and quantitative results. In order to overcome this drawback in the proposed system we partition the co citation into high probability and low probability links using partition algorithms and it eliminates the low probability links and can access the high probability links to reach destination in a short span of time.

Index Terms – Link analysis, generalized maximum flow, Wikipedia mining, relationship

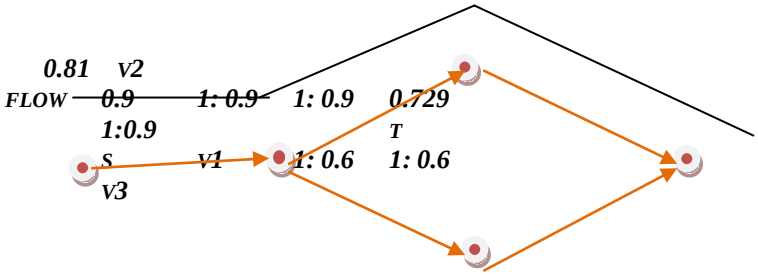
INTRODUCTION

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. in wikipedia the knowledge about a particular object in 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, after that we press search button and related server will display set of co citations in a wiki page. with the help of set of co citations we reach the destination node. consider for example we want to search about the object "notebook" – then we enter the sting "notebook" in the search engine. then the server displays the related co citations.



in the above diagram it displays set of co citations related to the object note book. similarly we get set of co citations for other kind objects also. consider a user want to measure relationship in between any of the objects i.e. in the typical cases a user might desire to calculate the relationship (relationship calculates whether the wiki pages are related or not) in between two wiki pages. here comes the fact that there exist two types of relationships - "explicit relationship" and "implicit relationship". a user might not easily differentiate these two relationships. explicit relationship has only a single link i.e. a single link from one web to the other web page. in the above link from note book web page to uses web page is an explicit relationship and implicit relationship is represented by a link structure and it has many intermediate nodes. these intermediate nodes are called elucidatory objects. a user might not differentiate elucidatory objects and implicit relationships. 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" (proposed by koren et al), "pfibf" (proposed by nakayama et al) are based on the concept "cohesion". one of them is generalized maximum flow. cohesion based method is adequate because it does support high degree objects. and other methods proposed earlier 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. after all these concepts generalized maximum flow was introduced for measuring the strength of relationships following the three factors distance, connectivity and co citation. in the generalized maximum flow gain function is used for measuring the relationships.



in the above the diagram depicts the generalized maximum flow. in diagram in order to reach destination from the source s to t we are having two paths with some gain values. s-v1-v2-t is one path, its path value is 0.729 and s-v1-v3-t is other path and its path value is 0.288. greatest path value is considered as the best search, so in the diagram s-v1-v2-t is considered as the best path. but drawback is that it may not give quantitative and qualitative results and it time taking. in the proposed system we over come draw back using partition algorithm. with this proposed system we get quantitative and qualitative results

Related work:

In this section let us see earlier techniques for calculating relationships in between two co citations on the Wikipedia – a search engine

DISTANCE, CONNECTIVITY, CO CITATION:

In the earlier erdos number (which was introduced by a famous mathematician PaulErdo) was used for calculating the distance. A source co citation has erdos number as 0, the next intermediate node of source co citation has erdos number as '1', next intermediate node has erdos number as '2' etc, this erdos number represents the shortest path to reach from source co citation to the destination co citation, and this shortest path is considered as the strongest relation ship. But the erdos number is inadequate to represent the implicit relationship as it does not estimate the connectivity in between two objects. The hitting time from the source co citation 'A' to source co citation 'B' is defined as the expected number of steps in reaching randomly from A to B. sarkae, Moore proposed THT (truncated hitting time) to calculate the average length of paths between source object to destination object. A smaller distance value represents larger similarity. This THT is also inadequate to represent connectivity between two co citations. For effectively calculating connectivity between source node A to source node B we have to remove minimum number of vertices such that no path exists from A to B. If the connectivity from A to B is large then A is having strong relationship with that of B. the connectivity value between A to B is considered as the value of maximum flow Where Vertex and Edge capacity is equal to 1. The distance estimated by maximum flow may not lead to the correct path. In order to over come this draw back Lu et al proposed a technique for calculating the strength of relationship. He calculated the distance between two nodes using a maximum flow value by setting edge capacities. How ever the maximum flow value does not change by setting edge capacities. Thus this method does not calculate distance effectively with the value of maximum flow. Instead of setting capacities we use generalized maximum flow by setting every gain value less than 1. Thus the value of maximum flow in our method decreases, if distance value becomes longer.

Co citation:

Co citation 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. Google similarity distance was proposed by cilibrasi and vitanyi was regarded as a co occurrence based technique. This technique measures the strength of a relationship between two words by counting of web pages containing both the words i.e. it implicitly regards the WebPages as nodes linking to the nodes representing the two words. In a network containing information, a node linked by both nodes becomes a node linking to the both if the direction of every edge is reversed. Thus the co occurrence can be treated as the reverse of the co citation. Milan and Witten also proposed techniques for measuring relation ships in between words in Wikipedia using Wikipedia links based on co citation.co citation related techniques cannot deal with a typical implicit relationship, such as "friend of A= friend of =B= friend of C". (A, B) and (B, C) and the relationship represents the path formed by 2 edges. In contrast the co citation related methods are in adequate for calculating implicit relationship. Moreover, co citation – related methods cannot deal with three hops (jumps) implicit relationships as already defined because these methods estimate only relationships represented by two edges as stated before. Jon and wisdom proposed simrank, it is an extension of co cited objects, and therefore it can deal with a path whose length is longer than two, although it cannot deal with implicit relationship. "A friend of 'A' = friend of 'C' " similarly to co citation based method if we define all the edges as bidirectional, then simrank could measure typical implicit relationship. But we have seen that simrank computes only the strength of the relationship represented by a path constituted by an odd number of edges to be 0, even if all the edges are bidirectional. Consider simrank computes the strength of the relationship is represented by path (A, C) or (A, B1, B2, C). Such paths abandon the Wikipedia information network. Therefore simrank is inadequate for measuring relationships on Wikipedia.

COHESION:

In social network analysis, cohesion based methods are used to measure the strength of relationship by counting all paths between two objects. Hubbel and Katz, Wassermann and K.faerst originally proposed co citation. But it has a property that it value increases for popular object, an object linked to one or to many objects exists. But it is a defect for measuring the strength of a relationship. PFIBF and CFEC- methods of cohesion are explained below .PFIBF- a cohesion based method was proposed by nakayama et al. PFIBF counts paths whose length is at most i>0 using ith power of the adjacency matrix of an information network. In the matrix if the ith power contains path cycle of almost (i-1). Drawback of PFIBF is that it can not differentiate a path containing cycle and path with no cycle. Consider for i>=3 we get two number of edges (a, b) and (b, a), such that PFIBF counts the path (a, b) and (a, b, b, a) is forming a cycle (a, b, a). if i<= then these exists no cycle, thus PFIBF is in adequate for measuring the implicit relationships. Next for measuring implicit relationships effective conductance was proposed by Doyle and Snell but it also faces same drawback. In order to over come the above drawback Korean et al. proposed CFEC (cycle free effective conductance) based on effective conductance. In measuring the implicit relationship CFEC does not traverse a path containing a cycle, though it won't count all the paths.

In the above all the cohesion based methods are in adequate for measuring implicit relationships in Wikipedia. In order to overcome the drawback generalized maximum flow based method was proposed, which supports all the 3 concepts and it does not criticize any major object in the process of measuring implicit relation ships. In the generalized maximum flow every edge e is contain gain γ (e)>0, flow value of edge e is multiplied by γ (e).consider the flow value of edge e, f (e)>=0 and capacity μ (e)>=0. F (e)<=μ(e) must follow for every edge e. in the generalized maximum flow at a greatest extent we reach source vertex to destination vertex. Value of f be the is defined as the total amount of f arriving at destination

PROBLEM STATEMENT*:*

Using generalized flow method we may not get quality and quantitave results.

Existing system implementation steps are:

Step1: enter the object name to be searched in Wikipedia.

Step2: Wikipedia displays the related wiki pages.

Step3: In the obtained wiki pages find out the source node.

Step4: after selecting source node traverse possible paths to reach destination nodes.

Step5: which contains less distance and more connectivity of object path is to be selected.

Step6: apply gain function on the object (selected) path.

Proposed statement:

Proposed system implementation steps are:

Step1: enter the object name to be searched in Wikipedia.

Step2: Wikipedia displays the related wiki pages.

Step3: in the step 3 it displays two kinds of links.

i) High probability of links.

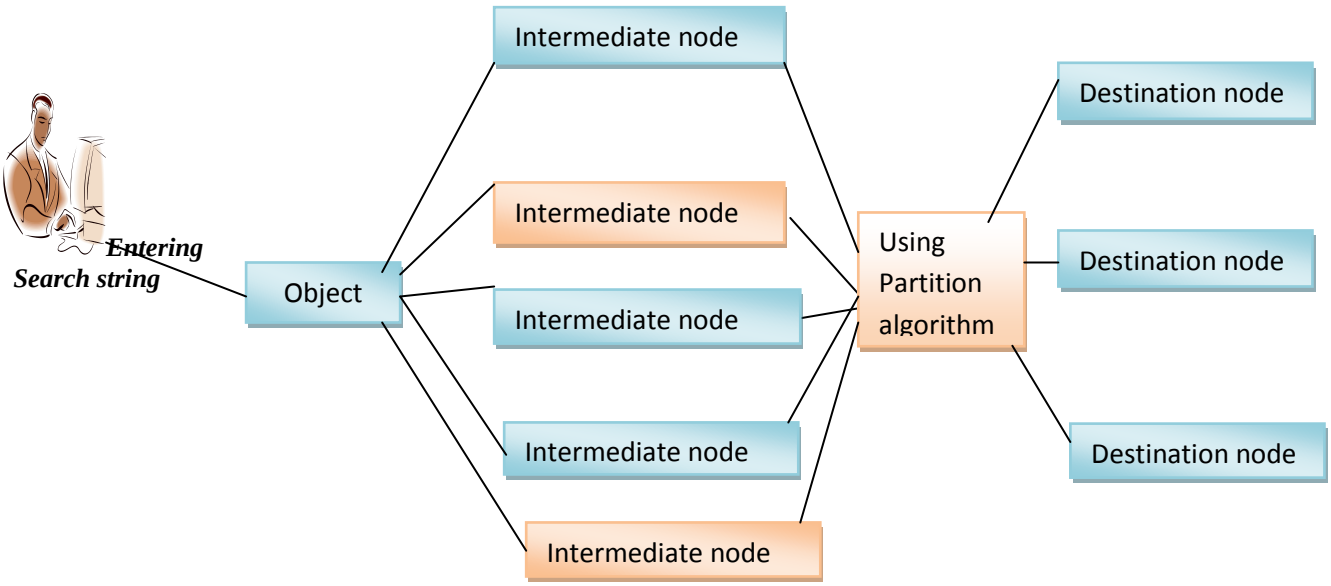
ii) Low probability of links.

Step4: eliminate the low probability of links.

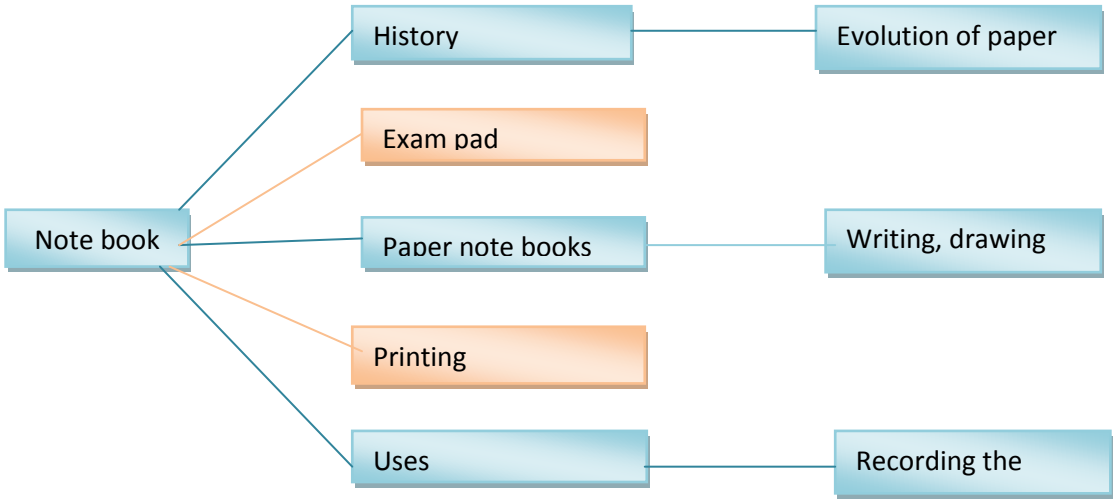
Step5: with the high probability of links we reach destination through less number of paths.

Step6: supports faster knowledge search results and it is time saving.

PROPOSED ARCHITECTURE:



EXAMPLE



Semantic Deep Web Relationships Extraction Algorithm or Qualitative relationship identification through Cut Detection approaches or partitioning approaches

Partitioning algorithm on objects relationship:

Input*: dataset (organized documents of content=Wikipedia pages), query, probability*

Output*: Qualitative objects detection*

step1*: Enter the topic name (a)*
step2*: search in dataset and define the related Wikipedia pages (n -> a)*
step3*: in total number of Wikipedia pages consider the first Wikipedia page as a source page and next define the destination page (a1 be source, a2 be the next intermediate node, a (n) be destination node)*
step4*: in first page many number of objects are available, calculate the link probability using the connectivity*

First page: object1 : mapping: different pages of objects

If there exists relationship then: 1, otherwise: 0
(p(a1)=1) =>repeat ; (p(a2)=0)=> exit)

Identify the probability: 1

Repeatedly identify the relationship with objects: probability is increases
step5*: display the different objects of link probabilities*
step6*: apply the threshold detect the high link probability nodes (partitioning operation)*
step7*: final we produce the quality objects links*

```
Topic name=a;
'n' be a data set , ( a,a1,a2,a3 ……a(n)) -> n;
//(a1, a2, ……..an) are intermediate nodes;
for( i = 0; i<=n; i++)
{
If(p(ai)==1)
Repeat;
If(p(ai)==0)
Break;
}
```


PARTITIONING ITERATIVE ALGORITHM:

After completion partitioning algorithm now we checks the relationship
=>consider the necessary relationships objects and eliminates the objects without relationship
=>identify the degree related to each and every object
=>categorize the objects as important objects

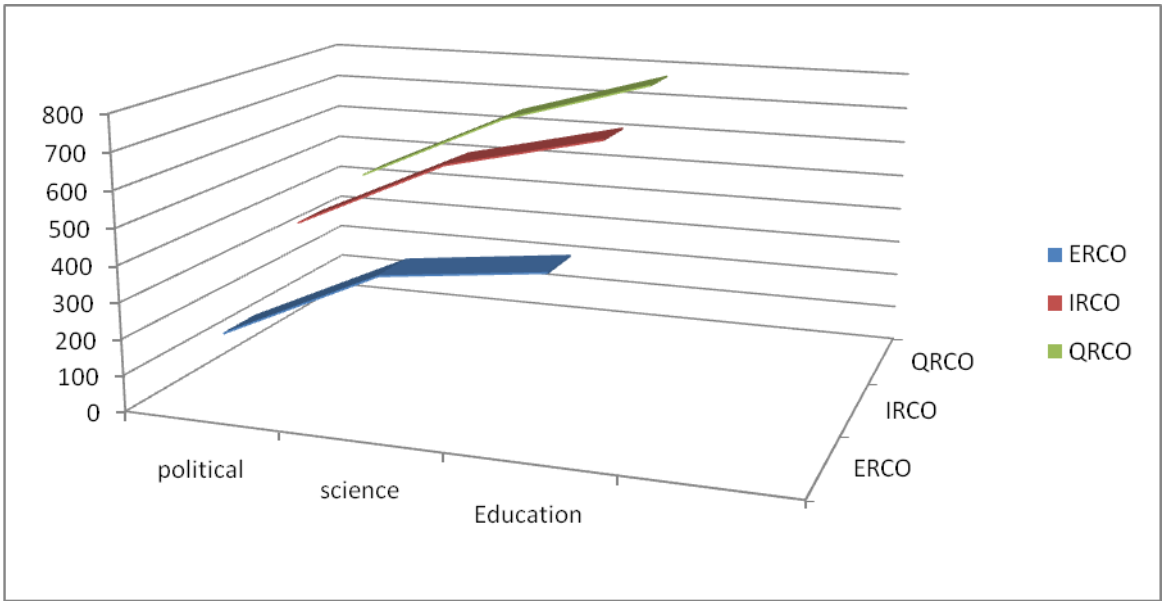
GRAPH*:*

Consider that we are searching objects related to politics, science, and education;

For politics we are considering 200 number of explicit relation co citation, 400 number of implicit co citation, out these two relations considering 450 qualitative relation co citations.

Similarly for science 400 number of explicit relation co citation, 600 number of implicit co citation, out these two relations considering 650 qualitative relation co citations.

Similarly for education 450 number of explicit relation co citation, 700 number of implicit co citation, out these two relations considering 777 qualitative relation co citations.



CONCLUSION:

In the proposed system we over come the drawback of the existing system i.e. instead of directly applying generalized maximum flow in a particular path, we partition the co citations into high probability link and low probability links and eliminate the low probability links, reach the destination object in a short span of time.

REFERENCES *:*

[1] Y. Korean, 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. Garcia 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 REFERENCES , 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 Erdo¨ s 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.

AUTHORS

	P.Naga laxmi is student of institute of aeronautical engineering, Hyderabad, AP, INDIA. She has received B.Tech Degree computer science and engineering, M.Tech Degree in computer science and engineering. Her main research interest includes data mining, Databases and DWH.
	Dr.Nagu.Chandra Sekhar Reddy working as a Professor ,CSE Dept., in Institute of aeronautical engineering, JNTUH, Hyderabad, Andhra Pradesh, India. He has received B.E, M.Tech, & PhD in CSE.
	Ms. Ila Chandana kumari is working as Associate Professor at institute of aeronautical engineering, Hyderabad, AP, INDIA. she has received M.Tech Degree in CSE.

