Implementation of Graph in Page Rank Algorithm, Google

Muhammad S Maromi - 13510061[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13510061@std.stei.itb.ac.id

***Abstract*—The abstract is to be in fully-justified italicized text, at the top of the left-hand column as it is here, below the author information. The abstract is to be in 9-point, single-spaced type, and may be up to 8 cm long. Define all symbols used in the abstract. Do not cite references in the abstract. Do not delete the blank line immediately above the abstract; it sets the footnote at the bottom of this column. Leave two blank lines after the indexrder, separated by commas.**

I. Introduction

Google, almost everyone in the world know this search engine and most of them are using this engine. We often found some questions or problems in academic, government, sport etc. Beside that, there are many informations that can got from Google. By using this engine, any information can be accessed quickly. Just type the keyword, then only some seconds Google display thousands til millions of sites that contained keyword have been typed.

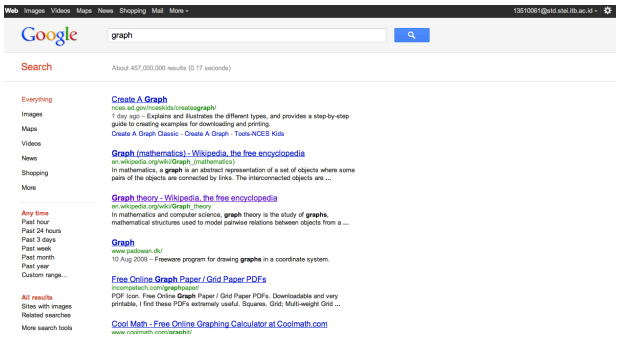

Figure 1

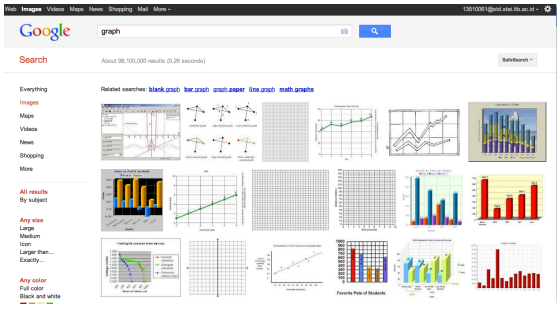

Figure 2

Furthermore, Google can display some images, videos, books and others which correspond with the keyword.

Until now, Google has had many features which very helpful. They're like Google Doc,Google Translate, Google Map etc. So, it's the reason why most of internet user in the world using this search engine rather than the other.

Now, we concern about the algorithm of Google search engine. Google success in large because of the clever algorithm, called "PageRank". This algorithm give rank to the webpages based on the priority that has many parameter. The rank value indicate the importance of a particular page. Webpage which has highest rank appear first and followed by webpages with lower rank.

The PageRank technique was result from mathematical algorithm based on graph created by all World Wide Web pages as vertexs and hyperlink as edges. Afterward, what was kind of this graph? Was it trending graph? And what was the simplicity of using graph in PageRank?

II. Graph Theory

A. History

Concept about graph arise first cause of the "Königsberg Bridge" Problem, that is how to pass every bridge just one time and back to original place.

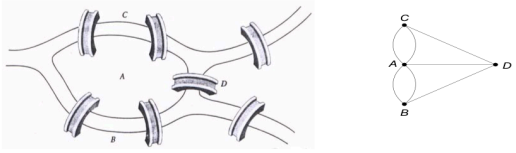

Figure 3

Graph which represents Königsberg Bridge was mainland as vertex and bridge as edge. Beside that, the problem of "Euler Trip", that is trip from one node, pass all segment only once and back to first node.

B. Definition

Graph G (V,E) is pair of two set

1. Set V which has elements called vertex or node.
2. Set E has elements called edge or segment.

Number of vertexs called order of graph, and

number of edges called size of graph.

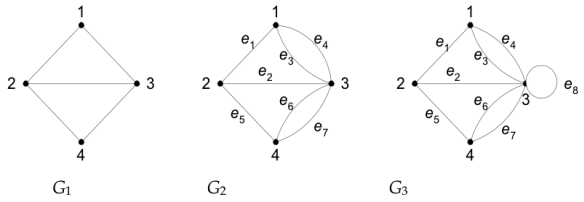

Figure 4

Bellow the description of figure 4

Graph G1

V = {1,2,3,4}

E = {(1,2),(1,3),(2,3),(2,4),(3,4)}

Graph G2

V = {1,2,3,4}

E = {(1,2),(1,3),(1,3),(2,3),(2,4),(3,4),(3,4)}

= {e_1, e_2, e_3, e_4, e_5, e_6, e_7}

Graph G3

V = {1,2,3,4}

E = {(1,2),(1,3),(1,3),(2,3),(2,4),(3,4),(3,4),(3,3)}

= {e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8}

At G2, side e_3 and e_4 named by **paralel edges**, because both sides connect two same nodes, viz node 1 and node 4.

At G3, side e_8 named by **self-loop** because begin and end in same node.

B. Definition

There are three ways to classification kind of graph

1. Presence of paralel edge or self-loop
   a. Simple graph
      Graph which doesn't have paralel edge.
   b. Multigraph
      Graph which have paralel edge.

2. Number of node
   a. Limited graph
      Graph which have n nodes and limited
   b. Unlimited graph
      Graph which have so many n node, unlimited

3. Number of node
   a. Directed graph
      Graph which sides have orientation.
   b. Undirected graph
      Graph which sides doesn't have orientation.

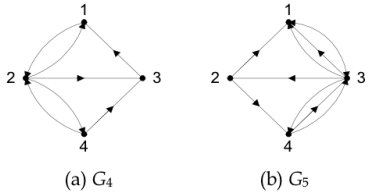

Figure 5

C. Terminology of Graph

1. Subgraph and Complement Subgraph
   For example G = (V,E) is a graph. $G_1 = (V_1,E_1)$ is subgraph from G if $V_1 \subseteq V$ and $E_1 \subseteq E$. Complement from subgraph G_1 toward graph G is graph G_2 such that $E_2 = E - E_1$ and V_2 is set of nodes that connect all sides of G_2.

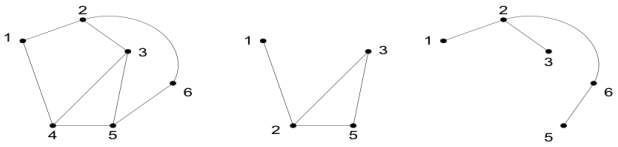

Figure 6. Graph G(left), G1(middle), G3(right)

2. Spanning Subgraph
   If E' contained all segments in E which both tip in V', then G' is subgraph that formed by V'.

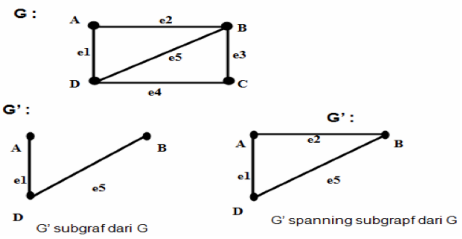

Figure 7.

3. Degree
   Degree is number of segments that connect node.

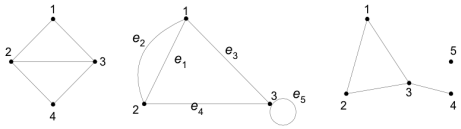

Figure 8. Graph G(left), G1(middle), G3(right)

Graph G1 : d(1) = d(2) = 2
d(2) = d(3) = 3

Graph G2 : d(1)=3
(side with multiple segment)
d(3) = 4
(side with self loop)

Graph G3 : d(5) = 0 (isolated node)
d(4) = 1 (last node)

4. Adjacency
   Two nodes adjacent if both of them direct connected. See figure 8.
   graph G1 : node 1 adjacents with node 2 and 3, node 1 not adjacents with node 4.

5. Incidence
   For any segment e = (v_j,v_k) called :
   e side with node v_j or
   e side with node v_k
   See figure 8.

   Graph G1 :
   segment (2,3) side with node 2 and node 3
   segment (2,4) side with node 2 and node 4
   but side (1,2) not side with node 4

6. Isolated Node
   Isolated node is node there is no node that side with. For example node node 5 in graph G3 (see figure 8).
7. Null Graph
   Graph which set of side is Null Set.

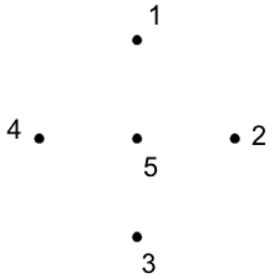

Figure 9.

C. Operation of Graph

$G_1 = (E_1,V_1)$, $G_2 = (E_2,V_2)$

1. Union $G_1 \cup G_2$ is graph with set of segments $E_1 \cup E_2$
2. Incision $G_1 \cap G_2$ is graph with set of segments $E_1 \cap E_2$
3. Difference G_1 - G_2 is graph with set of segments E_1 - E_2
4. Ring addition $G_1 \oplus G_2$ is graph with set of segments $(E_1 \cup E_2)$ - $(E_1 \cap E_2)$

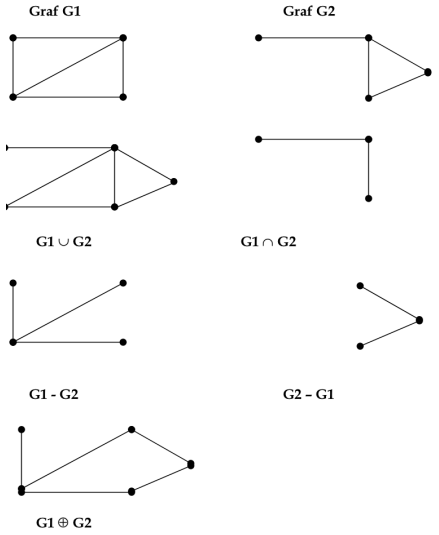

Figure 10

D. Decomposition

Graph is called decomposed into K and L if G = K ∪ L and K ∩ L = ∅. Example:

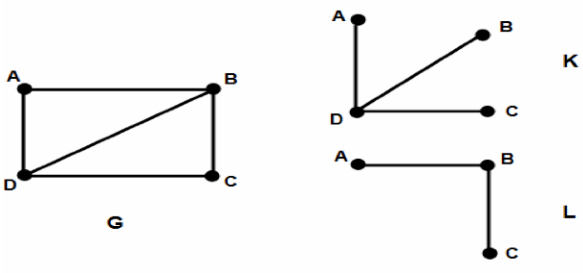

Figure 11

E. Deletion

1. Node Deletion
   Notation is G - {V}

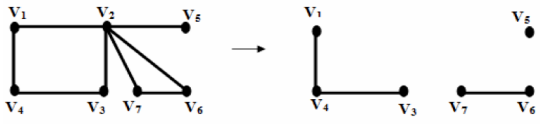

Figure 12. Deletion of node v_2

2. Segment Deletion
   Notation is G - {e}

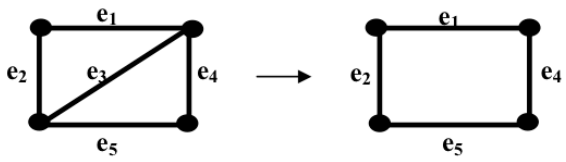

Figure 12. Deletion of segment e_3

F. Shorting

Shorting is delete node which connected by two segments, then connect points another tip from both segments

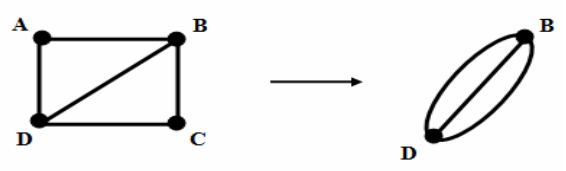

Figure 13. Shorting toward node A & C

G. Correlation

1. Walk
   Walk in graph is line of nodes and flit segments
   v_1, e_1, v_2, e_2,...., e_{n-1},v_n→e_i connect v_i and v_{i+1}
   wrote only line of segments or line of nodes.

In this, v_1 is called first node and v_n is called last node. *Close Walk* happen if $v_1 = v_n$ and *Open Walk* if connect v_1 and v_n. Length of walk is number of segment line.

2. Trail
   Trail is walk which all segments in that line are different.
3. Path
   Path is Walk which all node in that line are different.
4. (Circuit) Cycle
   Cycle is Trail which begin and end at the same node. Length of cycle is number of segment. Graph that doesn't include circuit called *acyclic.*

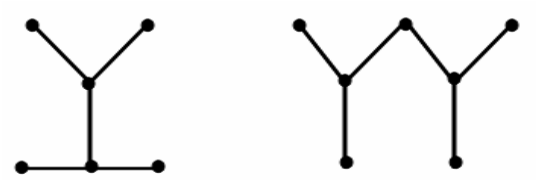

Figure 14.

III. Page Rank

The theory of graph already explained above, now turn to explain about the PageRank algorithm itself. The original PageRank algorithm was described by Lawrence Page and Sergey Brin. It is given by

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

where

- PR(A) is the PageRank of page A
- PR(Ti) is the PageRank of pages Ti which link to page A
- C(Ti) is the number of outbond links on page Ti
- d is damping factor which can be set between 0 and 1

So, first of all, we see that PageRank doesn't rank website a whole, but it rank every page individually. Further, the PageRank of page A is recursively defined by PageRanks of those pages which link to page A. The PageRank of pages Ti doesn't influence Pagerank of page A uniformly. Because the PageRank og page Ti has parameter C(T). More outbond link a page Ti, the less page A benefit from page Ti.

Finally, the sum of PageRank Ti that link to page A is multiplied with d which the value between 0 and 1. So, the benefit from another page linking to page A is reduced.

The characteristic of PageRank shal illustrated by a small example. Assume there are small pages A, B and C where page A links to page b and C, page B links to page C and page C links to page A. Then assume the damping factor d to 0.5. So, we get below PageRank equation:

$$PR(A) = 0.5 + 0.5\ PR(C)$$
$$PR(B) = 0.5 + 0.5\ (PR(A) / 2)$$
$$PR(C) = 0.5 + 0.5\ (PR(A) / 2 + PR(B))$$

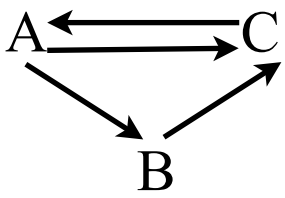

Figure 15.

Solve the equation above and get the PageRank values

PR(A) = 14/13
PR(B) = 10/13
PR(C) = 15/13

The example above is too easy compare to the real hyperlinks and billions of webpages registered. But, it 's helpful enough to make sense the whole algorithm of search engine, Google. Although it's not easy in practice.

IV. Implementation of Graph

Internet networks in the world were very large. The millions of websites and billions of webpage be a trouble find information that we don't know exactly the website address. But, with PageRank algorithm, the whole internet costumers could easily access information about news, sport, sains and others.

Further, the PageRank algorithm itself was still difficult to describe cause of the lot of webpages and hyperlinks. So, concept graph is going to helpful to describe whole webpages and hyperlinks to be a giant set of vertex and edge.

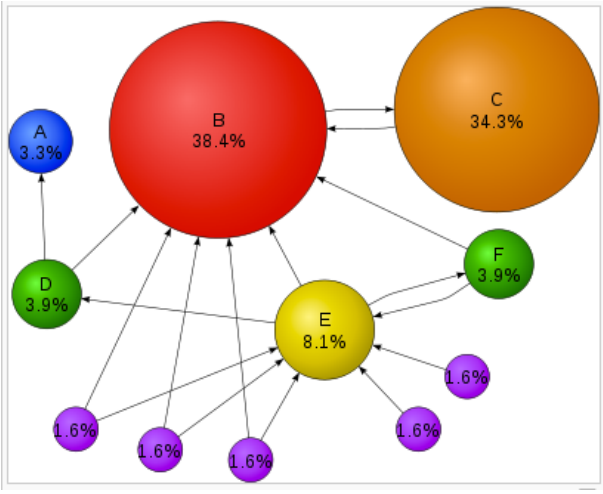

Figure 16

In search algorithm necessary to link as quick as possible. In graph theory, there is algorithm to find the simple way to a particular node, also if the the graph was weighted. It's same to the form of webpages in world wide web. The weight of webpage is correspond to PageRank. And the edge of direct graph correspond to webpage that link to othe page.

V. CONCLUSION

1. Google use PageRank algorithm to simply the search of particular webpage based on the keyword typed
2. The PageRank algorithm itself was simply described to using graph concept

REFERENCES

[1] Army N, *The Science of Search Engine Rangkings* (Book style). Proceton University Press. June 22, 2009. Ch 4

[2] http://en.wikipedia.org/wiki/Google_Search, on December 11, 2011

[3] http://en.wikipedia.org/wiki/PageRank on December 11, 2011