1. A Document as a Small World

Yutaka Matsuo[12], Yukio Ohsawa[23], and Mitsuru Ishizuka[1]

[1] University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
email: matsuo@miv.t.u-tokyo.ac.jp
[2] TOREST, Japan Science and Technology Corporation, Tsutsujigaoka 2-2-11, Miyagino-ku, Sendai, Miyagi, 983-0852 Japan
[3] University of Tsukuba, Otsuka 3-29-1, Bunkyo-ku, Tokyo 113-0012, Japan

A document is represented by a network; the nodes represent terms, and the edges represent the co-occurrence of terms. This paper shows that the network has the characteristics of being small world, i.e., highly clustered and short path length. Based on the topology, we can extract important terms, even if they are rare, by measuring their contribution to the graph being small world.

1.1 Introduction

Graphs that occur in many biological, social and man-made systems are often neither completely regular nor completely random, but have instead a "small world" topology in which nodes are highly clustered yet the path length between them is small [1.7, 1.5]. Watts and Strogatz have shown that a social graph (the collaboration graph of actors in feature films), a biological graph (the neural network of the nematode worm *C. elegans*), and a man-made graph (the electrical power grid of the western United States) all have a small world topology [1.7, 1.6]. World Wide Web also forms a small world network [1.1].

In this paper, we first show the graph derived from a document has the small world characteristics. Then we develop a new algorithm to find important terms by measuring a term's contribution to make the world small.

1.2 Small world

We treat an *undirected*, *unweighted*, *simple*, *sparse* and *connected* graph. (We expand to an *unconnected* graph in Section 1.4.) To formalize the notion of a small world, Watts and Strogatz define the clustering coefficient and the characteristic path length [1.7, 1.6]:

- The *characteristic path length*, L, is the path length averaged over all pairs of nodes. The path length $d(i,j)$ is the number of edges in the shortest path between nodes i and j.
- The *clustering coefficient* is a measure of the cliqueness of the local neighbourhoods. For a node with k neighbours, then at most ${}_kC_2 = k(k-1)/2$

edges can exist between them. The clustering of a node is the fraction of these allowable edges that occur. The clustering coefficient, C is the average clustering over all the nodes in the graph.

Watts and Strogatz define a small world graph as one in which $L \geq L_{rand}$ (or $L \approx L_{rand}$) and $C \gg C_{rand}$ where L_{rand} and C_{rand} are the characteristic path length and clustering coefficient of a random graph with the same number of nodes and edges. They propose several models of graphs, one of which is called β-Graphs. Starting from a regular graph, they introduce disorder into the graph by randomly rewiring each edge with probability p as shown in Fig.1.1. If $p = 0$ then the graph is completely regular and ordered. If $p = 1$ then the graph is completely random and disordered. Intermediate values of p give graphs that are neither completely regular nor completely disordered. They are small worlds.

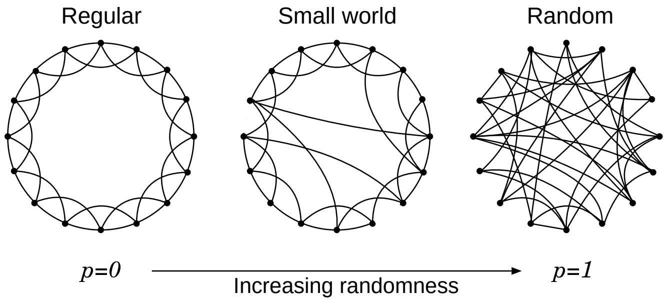

Fig. 1.1. Random rewiring of a regular ring lattice.

Walsh defines the proximity ratio $\mu = (C/L) \,/\, (C_{rand}/L_{rand})$ as the small-worldliness of the graph [1.5]. μ is larger than 1 in the graphs with a small world topology.

1.3 Term Co-occurrence Graph

A graph is constructed from a document as follows. We first preprocess the document by stemming and removing Salton's *stop words*. We apply n-gram to count phrase frequency. Then we regard the title of the document, each section title and each caption of figures and tables as a sentence, and exclude all the figures, tables, and references. We get a list of sentences, each of which consists of words (or phrases).

Then we pick up *frequent terms* which appear over a user-given threshold, f_{thre} times, and fix them as nodes. For every pair of terms, we count the *co-*

occurrence for every sentences, and add an edge if the Jaccard coefficient exceeds a threshold, J_{thre}[1].

Table 1.1 is statistics of the small-worldliness of 57 graphs, each constructed from a technical paper that appeared at the 9th international World Wide Web conference (WWW9) [1.8]. From this result, we can conjecture these papers certainly have small world structures. However, depending on the paper, the small-worldliness varies.

Table 1.1. Statistical data on proximity ratios μ for 57 graphs of papers in WWW9.

	L	L_{rand}	C	C_{rand}	μ
Max.	4.99	3.58	0.38	0.012	22.81
Ave.	5.36	—	0.33	—	15.31
Min.	8.13	2.94	0.31	0.027	4.20

We set $f_{thre} = 3$. We restrict attention to the giant connected component of the graph, which include 89% of the nodes on average. We exclude three papers, where the giant connected component covers less than 50% of the nodes. We don't show the L_{rand} and C_{rand} for the average case, because n and k differs dependent on the target paper. On average, $n = 275$ and $k = 5.04$.

One reason why the paper has a small world structure can be considered that the author may mention some concepts step by step (making the clustering of related terms), and then try to merge the concepts and build up new ideas (making a 'shortcut' of clusters). The author will keep in mind that the new idea is steadily connected to the fundamental concepts, but not redundantly.

1.4 Finding Important Terms

Admitting that a document is a small world, how does it benefit us? We try here to estimate the importance of a term, and pick up important terms, though they are rare in the document, based on the small world structure. We consider 'important terms' as the terms which reflect the main topic, the author's idea, and the fundamental concepts of the document.

Below we show a series of definitions to measure the importance of a term.

Definition 1.4.1. *An* extended *path length $d'(i,j)$ of node i and j is defined as follows.*

$$d'(i,j) = \begin{cases} d(i,j), & \text{if } (i,j) \text{ are connected,} \\ w_{sum}, & \text{otherwise.} \end{cases} \tag{1.1}$$

[1] In this paper, we set J_{thre} so that the number of neighbors, k, is around 4.5 on average. The Jaccard coefficient is simply the number of sentences that contain both terms divided by the number of sentences that contain either terms. This idea is also used in constructing a referral network from WWW pages [1.2].

where w_{sum} is a constant, the sum of the widths of all the disconnected subgraphs. $d(i, j)$ is a path length of the shortest path between i and j in a connected graph.

Definition 1.4.2. Extended characteristic path length L' *is an extended path length averaged over all pairs of nodes.*

Definition 1.4.3. L'_v *is an extended path length averaged over all pairs of nodes except node* v. L'_{G_v} *is the extended characteristic path length of the graph without node* v.

Definition 1.4.4. *The* contribution, CB_v, *of the node* v *to make the world small is defined as* $CB_v = L'_{G_v} - L'_v$.

If node v with large CB_v is absent in the graph, the graph gets very large. In the context of documents, the topics are divided. We assume such a term help merge the structure of the document, thus important.

1.5 Example

We show the example experimented on [1.4], i.e. the longer version of this paper. Table 1.2 shows the frequent terms and Table 1.3 shows the important terms measured by CB_v. Comparing two tables, the list of important terms includes the author's idea, e.g., *important term* and *contribution*, as well as the important basic concept, e.g., *cluster* and *coefficient*, although they are rare terms. However the list of frequent terms simply show the components of the papers, and are not of interest.

Table 1.2. Frequent terms.

Term	Frequency
graph	39
small	37
world	37
term	34
small world	30
node	29
paper	21
length	21
document	19
edge	19

Table 1.3. Terms with 10 largest CB_v.

Term	CB_v	Frequency
small	3.05	37
term	2.80	34
important term	1.93	7
contribution	1.64	6
node	1.00	29
make	0.82	6
cluster	0.57	15
graph	0.54	39
coefficient	0.52	8
average	0.50	8

Lastly, Fig. 1.2 shows the graphical visualization of the world of this paper. (Only the giant connected component of the graph is shown, though other parts of the graph is also used for calculation.) We can easily point out the terms without which the world will be separated, say *small* and *important term*.

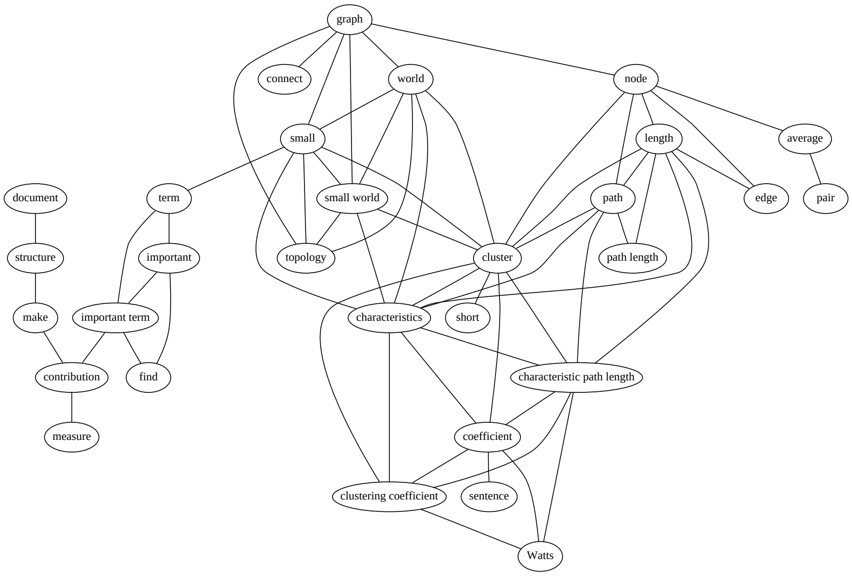

Fig. 1.2. Small world of the paper.

1.6 Conclusion

We expect our approach is effective not only to document indexing, but also to other graphical representations. To find out structurally important parts may bring us deeper understandings of the graph, new perspectives, and chances to utilize it. A change, which makes the world very small, may sometimes be very important.

References

1.1 R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the World Wide Web. *Nature*, 401, 1999.
1.2 H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magagine*, 18(2), 1997.
1.3 J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Technical Report TR 99-1776, Cornell University, 1999.
1.4 Y. Matsuo, Y. Ohsawa, M. Ishizuka. A Document as a Small World In *Proc. SCI-01*, Vol.8, pages 410–414, 2001
1.5 T. Walsh. Search in a small world. In *Proc. IJCAI-99*, pages 1172–1177, 1999.
1.6 D. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton, 1999.
1.7 D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 1998.
1.8 http://www9.org/.