Clustering using Small World Structure

Yutaka Matsuo National Institute of Advanced Industrial Science and Technology,
Aomi 2-41-6, Koto-ku, Tokyo, 113-0064 JAPAN

Abstract. Small world topology has been receiving much attention recent years. Many real-world graphs actually have a small world topology. In a small world graph, a shortcut plays an important role to make the world small; if shortcuts are eliminated, the graph is separated into clusters. In this paper, I propose a new clustering algorithm by eliminating "shortcuts," which helps a user understand the graph well.

1 Introduction

In the 1960s, Stanley Milgram showed that any two randomly chosen individuals in the United States are linked by a chain of six or fewer first-name acquaintances, known as "six degrees of separation." Watts and Strogatz defined what is small world, and showed some networks have small world characteristics [11]. Since their introduction, small-world networks and thir properties have received considerable attention. Numbers of networks are shown to have a small-world topology. Examples include social networks such as acquaintance networks and collaboration networks, technological networks such as the Internet, the World-Wide Web, and power grids, and biological networks such as neural networks, foodwebs, and metabolic networks. (For reference, see [4].) Matsuo et. al. showed that word co-occurrence in a technical paper also consists a small world graph [8].

In a small world graph, nodes are highly clustered yet the path length between them is small. Although some recent works have proposed different definitions of "small world" (for example, [6]), one by Watts and Strogats is appropriate to grab idea of node distance and clusters. They define the following two invariants [11]:

- 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 neighborhoods. For a node with k neighbors, 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.

In this paper, I propose a new method for detecting clusters based on the small world structure. In a small world network, shortcuts (or weak ties) play an important role to connect clusters (or communities). In other words, clusters already exist in the graph. Therefore, I eliminate some edges from the graph to make C and L large so that nodes are clustered and clusters are separated.

Finding clusters (or rather, "elicitation" of clusters when they exist in nature) is an important task when we try to understand the graph. A cluster often shows the particular context; for example, a cluster corresponds to a community in social networks, a Web page community in WWW, and a concept in a technical paper. Proper clustering may suggest us the context shared by the member of each cluster.

2 Related Works

Clustering is an important data exploration task in chance discovery[9] as well as in data mining, because it shows the overview of the data, and stimulates our interest when understanding the clusters.

The first hierarchical clustering dates back to 1951 by K. Florek, and since then there have been numerous modifications. One of the most widely used clustering methods is the single linkage clustering. It is a kind of hierarchical clustering, and its cluster relationships can be represented by a rooted tree, called dendrogram. A cluster is produced by cutting edges of the dendrogram with a threshold. However, application of only one threshold for all clusters would produce many too small clusters and a few large clusters.

To tackle this problem, a clustering method based on a linkage graph is proposed in [5] to cluster protein sequences into families. They formulate clustering as a kind of graph partitioning problem [2] of a weighted linkage graph and find *minimal cut* with consideration of balancing the size of clusters. The graph partitioning problem is of interest in areas such as VLSI placement and routing, and efficient parallel implementations of finite element methods, e.g., to balance the computational load and reduce communication time. [3] develops an algorithm to find communities on the Web by maximum flow / minimum cut framework. This algorithm performs well in practice i.e., under the condition that one couldn't have rapid access to the entire Web.

Another approach focuses on the *betweenness* of an edge in a linkage graph as the number of shortest paths between pairs of nodes that run along it [4]. An edge in a graph is iteratively removed if the betweenness of the edge is highest in order to find communities in social and biological networks.

3 Small World Clustering

I formalize my clustering algorithm, called *Small World Clustering*, as an optimization problem. It differs to the conventional graph partition problem [2] in that it uses the C and L for the measurement.

Definition 3.1 (Small World Clustering)
Given a graph $G = (V, E)$ and k where V is a set of nodes, E is a set of edges, and k is a positive integer, *Small World Clustering (SWC)* is defined as finding a graph G' such that k edges are removed from G so that

$$f = aL_{G'} + bC_{G'}$$

is to be maximized. ($L_{G'}$ and $C_{G'}$ are L and C for graph G' respectively, and a and b are constants.)

To deal with a disconnected graph, I extend the definition of L: An *extended* path length $d'(i,j)$ of node i and j is defined as follows.

$$d'(i,j) = \begin{cases} d(i,j), & \text{if } (i,j) \text{ are connected,} \\ n, & \text{otherwise.} \end{cases} \quad (1)$$

where n is a number of nodes in G.

The problem of finding an optimal connection among all possible pairs of nodes of a graph has been proven to be NP-complete. Therefore I consider an approximate algorithm for SWC as follows.

1. Prune an edge which maximize f iteratively until k edges are pruned.
2. Add an edge which maximize f. If an edge to be added is the same as the most previously pruned one, terminate.
3. Prune an edge which maximize f. Go to 2.

The second and third proceedures are optional. If clustering needs to be finished rapidly, these proceedures can be skipped. However in some cases, they provide a little better solution.

4 Examples

I show an example of SWC applied to a word co-occurrence graph. A word co-occurrence graph is constructed as follows [8];

1. pick up n frequent words as nodes,
2. calculate a Jaccard coefficient[1] for each pair of words, and add an edge if the coefficient is larger than a given threshold.

A word co-occurrence graph is shown to have small world characteristics [8]. Thus, clustering by SWC can be applied.

Fig. 1 is a word co-occurrence graph derived from a technical paper [10] with the single linkage clustering. We can see a big cluster, one little cluster, and six isolated nodes. Resulting in a big cluster and many isolated nodes is very common when single linkage clustering is applied. It is very difficult to grab the 'meaning' behind each cluster.

Fig. 2 shows a graph derived from the same paper but by SWC[2] instead of the single linkage clustering (with the same number of nodes and links). Four big clusters, three pairwise nodes and one single node are extracted. Because the nodes in a cluster is well connected (as C should be high) and clusters are properly separated (as L should be high), it is easier to grab the meaning of cluster; for example, the left upper cluster is words related to "extract author's basic concept," the center upper cluster is about "existing retrieval method", the left lower cluster is "the key concept of this paper", and the right big cluster is about "the proceedure of the new algorithm."

In the result graph, a lot of complete subgraphs (or cliques) and stars emerge. These are two frequent types of subgraphs when small worlds are generated artificially [7].

[1]The Jaccard coefficient is the number of sentences that contain both words divided by the number of sentences that contain either words.

[2]Constant a is set 1, and b is set 100.

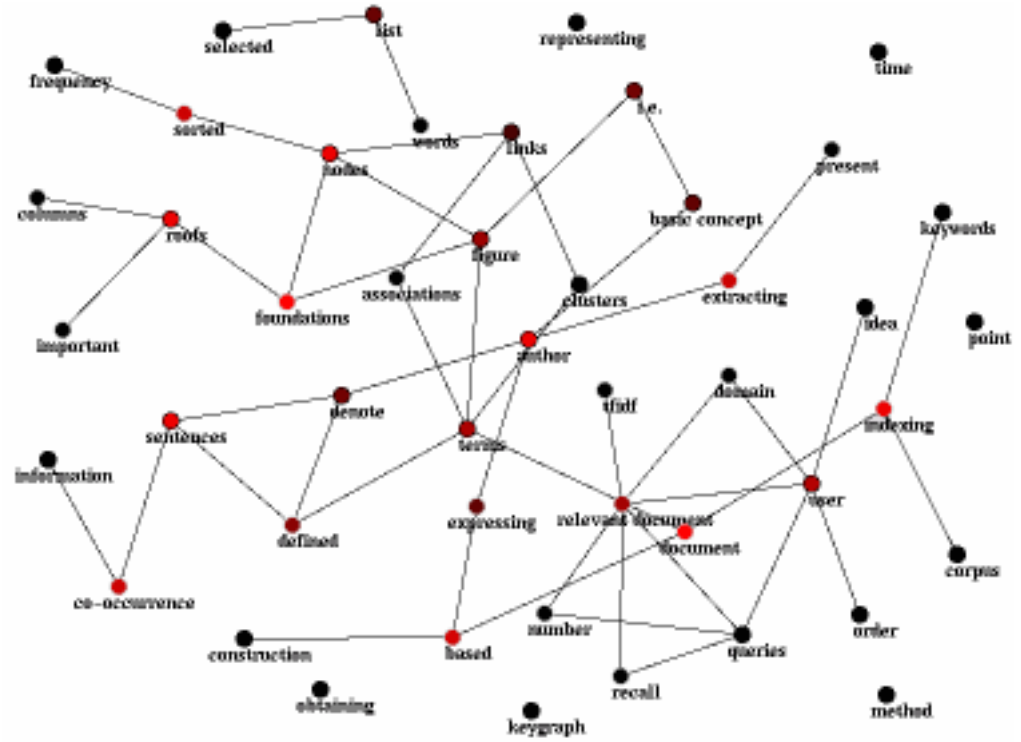

Figure 1: A word co-occurrence graph with a single linkage clustering. $C = 0.201$, $L = 12.1$.

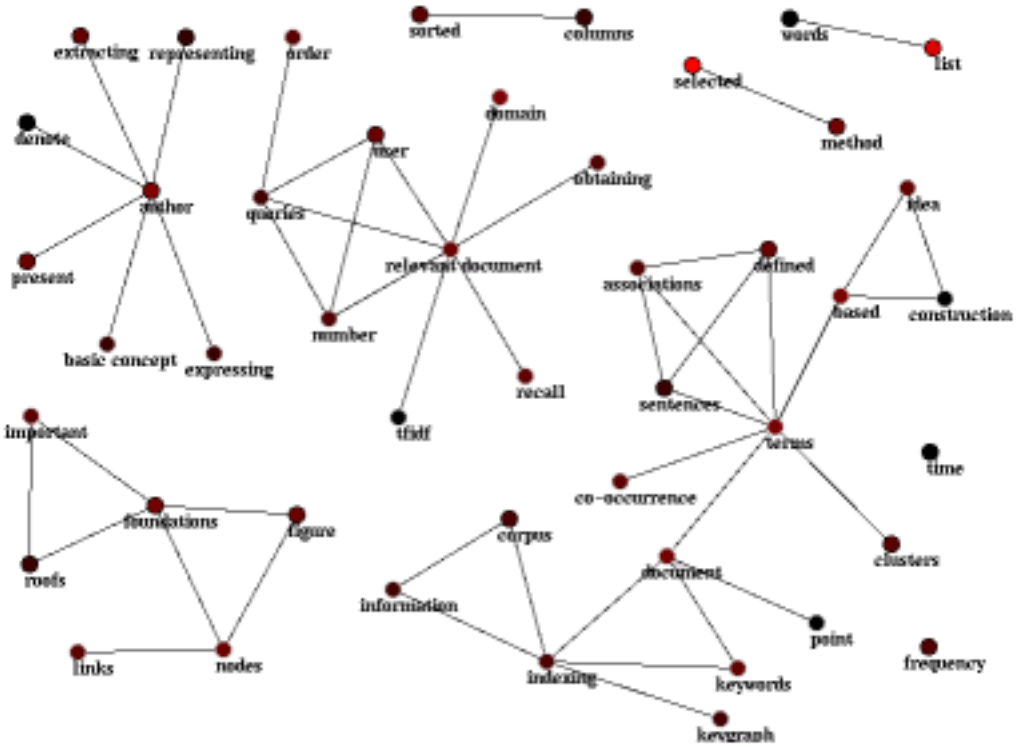

Figure 2: Clusters obtained by SWC. $C = 0.689$, $L = 18.3$

5 Discussion and Future Work

In this paper, I show a clustering algorithm which utilizes the small world structure of the graph. Quantative evaluation is ongoing research.

There is no consensus among the reserchers as to what constitutes a cluster. There is only some intuitive understanting: the intuitive idea behind clustering consists in condensing a subgraph into a single node, where the choice of the cluster is application-dependent [1]. From chance discovery point of view, if a user can find the novel meaning of clusters, the clustering algorithm is preferable. I will further investigate on what type of clustering can properly stimulate the imagination of a user.

References

[1] M. Ancona, W. Cazzola, E. Martinuzzi, P. Raffo, and I.B. Vasian. Clustering algorithms for the optimization of communication graphs. In *Proc. 4th Conf. Italo-Latino American of Industrial and Applied Mathematics*, 2001.

[2] P. Fjällström. Algorithms for graph partitioning: A survey. *Computer and Information Science*, 3, 1998.

[3] G. William Flake, S. Lawrence, and C. Lee Giles. Efficient identification of Web communities. In *Proc. ACM SIGKDD-2000*, pages 150–160, 2000.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *submitted to the Proceedings of National Academy of Sciences*, 2001.

[5] H. Kawaji, Y. Yamaguchi, H. Matsuda, and A. Hashimoto. A graph-based clustering method for a large set of sequences using a graph partitioning algorithm. *Genome Informatics*, 12:93–102, 2001.

[6] M. Marchiori and V. Latora. Harmony in the small-world. *Physica A*, 285:539–546, 2000.

[7] N. Mathias and V. Gopal. Small worlds: How and why. *Physical Review E*, 63(2), 2001.

[8] Y. Matsuo, Y. Ohsawa, and M. Ishizuka. KeyWorld: Extracting keywords from a document as a small world. In *Proceedings the Fourth International Conference on Discovery Science (DS-2001)*, 2001.

[9] Y. Ohsawa. Chance discoveries for making decisions in complex real world. *New Generation Computing*, to appear.

[10] Y. Ohsawa, N. E. Benson, and M. Yachida. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proc. Advanced Digital Library Conference (IEEE ADL'98)*, 1998.

[11] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.