Clustering and Traversals in Concept Association Networks

Arun R and **V Suresh** and **C E Veni Madhavan**
(arun_r, vsuresh, cevm)@csa.iisc.ernet.in
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, India

Abstract

We view association of concepts as a complex network and analyse its structure. We observe that concept association network is scale-free and has small-world properties. We also study two large scale properties of these networks — clusters and paths. First, we present an algorithm for clustering these networks which generate qualitatively better clusters than those generated by spectral clustering, a conventional mechanism used for graph partitioning. Next, we study paths generated by human traversals on these networks and contrast it with random walks and shortest distance paths. Our results are a first step towards viewing human cognitive abilities in the light of complex network analysis.

Concept Association Networks(CAN)

Concept associations can be intuitively understood as thoughts that occur in conjunction with each other. Typically, networks of such associations are built by presenting concepts to subjects and recording their output on the basis of *'what comes to your mind first'* in response to a cue. Such co-occuring cue-response concept pairs are considered to be *cognitively associated*.

Thus in a concept graph $G = (V,E)$, V the vertex set represents labelled nodes(concepts), and E the edge set represents co-occuring concepts. For our study we use word associations from USF Free Association Norm (http://w3.usf.edu/FreeAssociation) as it is more comprehensive than other databases and has also been studied earlier from a complex network prespective (Steyvers & Tenenbaum, 2005). Complex networks are graph abstractions to represent and analyse real world interacting systems like World-Wide-Web and social networks. A list of important structural properties of the concept network built based on this database is shown in table 1.

Table 1: Some salient network properties of CAN

nodes: 10618	avg degree: 12.01	max degree: 332
edges: 63788	edge density: 0.001131	
diameter: 7	$\gamma \sim 2.6$	CC: 0.1871

A power law degree distribution and high clustering coefficient(CC) are indicative of the similarity of concept association network to other widely studied complex networks such as World-Wide-Web, Social networks etc. (Albert & Barabási, 2002). Thus studying the properties of these concept interactions in the light of complex networks is justified. In this work we study two macro structres of concept networks, namely clusters –partitions of the network, and paths –traversals in the network, and relate its possible implications on cognition.

Clustering of Concepts - Algorithm

Clustering is an important aspect of generalization that helps in reducing intrinsically different things into broad groups for the sake of simplicity. Given that we learn concepts by relating to other similar concepts already known, it makes sense to cluster concept association networks into broader abstract entities. Such clusters would be useful if they can effectively represent human organisation of knowledge.

In this regard, we present a clustering algorithm and explain its usefulness in the context of cognition. We consider high degree hub nodes as the starting points. To begin with n hub nodes are labelled as belonging to its own cluster$\mathcal{C}_i$ ($i = 1$to n where $n = 10$ for this study). For each unlabelled node u in the graph its neighbourhood is explored to find the node with the highest degree v (say). If v is labelled, we assign the same label to u. If not, we perform the neighborhood exploration process on v. The recursion stops either when a hub node is hit or when no node with higher degree is present in the neighbourhood. In the former case, the node is assigned the hub node's cluster and in the latter, we assign it to a *default* cluster. Nodes assigned to the default cluster are finally assigned to the hub that is at the shortest distance in terms of path length. A stylistic version of the algorithm is given below.

$neigh(u)$: Set of nodes formed by the immediate neighbors of node u. $deg(u)$: degree of node u
$node_{degmax}(S)$: node with max degree in the set of nodes S. $label(u)$: label of node $u \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$
Init: Identify n hub nodes and label them $\mathcal{C}_1, \ldots, \mathcal{C}_n$
for each unlabelled node u
S1: let $v \leftarrow node_{degmax}(neigh(u))$
  if $deg(v) \geq deg(u)$
    if $label(v) = \mathcal{C}_i$, then $label(u) \leftarrow \mathcal{C}_i$
    continue
    if $label(v) \neq \mathcal{C}_i, \ \forall \ i \in \{1, \ldots, n\}$
      then $u \leftarrow v$, **GOTO S1**
  else $label(u) \leftarrow \mathcal{C}_0$

Comparison and Discussions

A comparison between spectral clustering (Ng, Jordan, & Weiss, 2001) and and our algorithm is shown in Figure 1 as log-log plots of cluster degree distribution. It is clear from

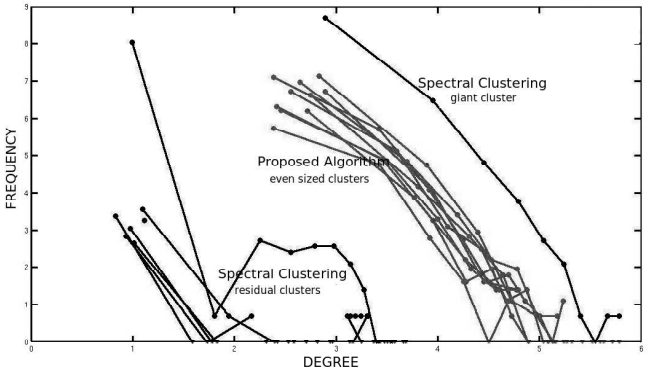

Figure 1: degree distribution of clusters: proposed Vs spectral

the figure that spectral clustering does not preserve scale-free characteristics within clusters. Moreover, the sizes of clusters are uneven. On the other hand, our algorithm splits the original graph into roughly equal sized clusters and each one has scale-free distribution with the same power-law exponent that applies to the whole graph. Thus the clusters from our approach are self-similar to the whole network. In effect our algorithm imparts a hierchical view to the whole network. This is in accordance with the the general hierarchical organisation of human knowledge.

Spectral clustering is a series of random walks to estimate cluster boundaries — walks are contained with in strongly connected components and rarely tend to take connecting bridges. One starts with various 'seeds' to begin the random walk and see the nodes that are reached eventually and thereby identify clusters. We believe that the difference in cluster properties between the two algorithms is because random walks are an unnatural means to navigate the cognitive space. This is further explored in the next section.

Concept Traversals - Observations

There are two extremes to (source,target) traversals: Shortest path from the source to the destination and Random walk starting from the source and proceeding till the target is reached. To quantify the properties of human generated paths[1], we compare them with both these extremes. It is intuitively clear that human generated paths must lie in the middle of these two strategic extremes. We identify a non-trivial property –the difference in degree of adjacent nodes in the paths– to offer a formal explanation for this intuition.

Figure 2 shows the difference of successive degrees of first two edges for shortest, random and human paths. Shortest paths show a steeper degree difference whereas for the random walk, the degree differences are smaller than those from human paths. The rationale for this is as follows.

[1]For our analysis, we asked 60 participants to perform concept traversals from source to targets for 183 concept pairs like (POWER,MONTH), (FAMILY,AREA) etc. Our observations are based on these subject generated paths.

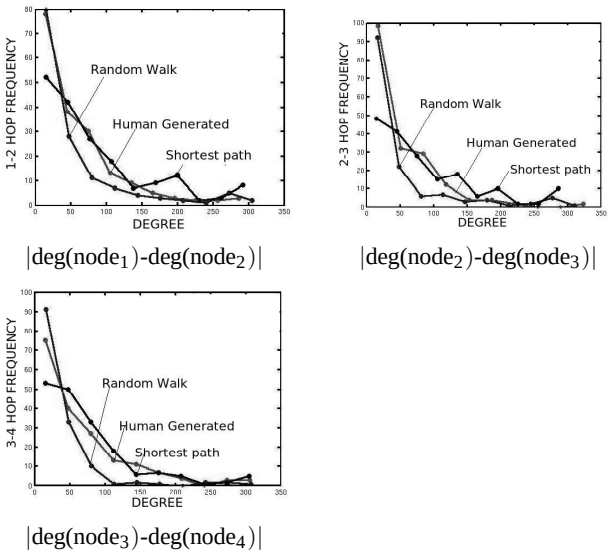

Figure 2: distribution of absolute degree difference

Given the structure of the concept graph –high CC implying dense neighbourhoods– degrees of successive nodes are expected to be similar. On such graphs random walks typically spends longer durations 'dwelling' in concept neighbourhoods rather than reaching the destination. Whereas shortest paths make incoherent conceptual jumps to reach the destination. In comparison, human traversals are a mix of smooth transitions and conceptual leaps and lie in the middle of these two extremes.

Conclusions

We proposed a clustering algorithm that exploits the structural properties of concept association network to produce self similar clusters that are arguably better than those produced by conventional clustering approaches. Then we compared concept traversals for human, random and shortest paths and quantified their differences in terms of the degree difference of adjacent nodes present in such paths. We observed that this network property can explain the intuitive idea that human paths are inbetween random and shortest paths –cogent yet amenable to conceptual leaps.

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems 14* (pp. 849–856). MIT Press.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic network: Statistical analyses and model of semantic growth. *Cognitive Science*, *29*(1), 41-78.