Clustering in Concept Association Networks

Arun R., V. Suresh, and C.E. Veni Madhavan

Department of Computer Science and Automation, Indian Institute of Science,
Bangalore 560 012, India

{arun_r,vsuresh,cevm}@csa.iisc.ernet.in

Abstract. We view association of concepts as a complex network and present a heuristic for clustering concepts by taking into account the underlying network structure of their associations. Clusters generated from our approach are qualitatively better than clusters generated from the conventional spectral clustering mechanism used for graph partitioning.

Keywords: complex-networks; clustering; concept associations; clustering coefficient; power-law distribution; scale-free networks; small-world networks.

1 Introduction

Studies on complex networks –networks arising from real-world interactions– have shed light towards understanding universal principles behind the dynamics of these diverse systems [1]. We view human concepts as a system of interacting entities and study their network structure. We observe that concepts and their associations have network properties that are similar to other well established networks like WWW and social networks. We then show that understanding the network structures of concepts is useful in grouping similar concepts and that these groupings are better than those generated by a conventional clustering method like spectral clustering [7]. This work is organized as following sections: In the next section we give an overview of some developments in complex network models and related work. This is followed by our empirical observations on concept association networks. We then describe our clustering heuristic and compare it with an existing approach following which we present our concluding remarks.

2 Complex Networks: Models and Related Work

The first model to explain successfully the properties of social networks is the Small-World Network Model [4] by introducing the notion of clustering coefficient to describe the common contacts shared in human friendships. Following this Albert and Barabási [3] explained the structure and evolution of WWW and other networks by accounting for the presence of hubs (nodes that interact with a large number of other nodes in the network) through the preferential attachment

model. Networks obeying such growth models are known as power law or scale-free networks; their degree distributions obey the relation $p(k) \sim k^{-\gamma}$, where $p(k)$ is the probability of a node to have degree k and γ controls the number of hubs.

Word Web –a network of interacting words based on their co-occurrence in a sentence– is seen to have power-law degree distribution and small-world characteristics [5]. WordNet¹ and word associations from USF Free Association Norm² –the same database which we use in this work– are shown to be scale-free, small-worlds and possess high clustering coefficients [6]. Dorogovtsev and Mendes [2] study the evolution of languages by using a preferential growth model in conjunction with evolving links among existing nodes.

3 Concept Association Networks(CAN)

Concept associations can be intuitively understood as thoughts that occur in conjunction with each other. In this work individual words are considered to be ‘concepts’; words that occur in the mind as a response to a cue word are considered to be concepts that are *cognitively associated* with the it. Thus a concept graph $G = (V, E)$, is a set of labelled nodes V representing concepts, and a pair of them are connected by an edge if they have a cognitive association. The set of edges are collectively represented as E . For our study we use word associations available from USF Free Association Norm.

It is evident that these networks, from the way they are built, are directed and weighted –MILK might evoke CAT, but CAT may evoke RAT more than MILK; however, we choose to treat CAN as an undirected, unweighted graph as it facilitates an easy first-cut analysis. Insights gained from this study would be put to use in the analysis of the more generic graph representations. An illustration of concept graph studied in this work is given in Figure 1. We note here that CAN is fundamentally different from word co-occurrence networks and semantic networks like WordNet. The latter are as a result of the structure of the language which may be studied and understood through the network properties. However they may not shed light on cognitive association of concepts which are interactions that are fundamentally independent of the structure of the language. In the rest of the section we give some important structural properties of concept association networks by treating the above mentioned USF Free Association Norm as their representative. A list of these properties are presented below:

nodes: 10618; **edges:** 63788; **avg degree:** 12.01; **max degree:** 332; **diameter:** 7; **power-law exponent:** $\gamma \sim 2.6$; **CC:** 0.1871

Diameter. Distance of the longest of all the shortest paths in the graph. Typically this is log order of the size of the graph. This means that the number of concepts can grow exponentially without changing the diameter much. However this is not unique for concept associations, this is true of any random graph

¹ <http://wordnet.princeton.edu>

² <http://w3.usf.edu/FreeAssociation>

Fig. 1. An illustration of unweighted and undirected concept associations and its degree distribution

wherein the connections are unmotivated. However, unlike random graphs, short diameters arise from the hubs nodes and their effect on the overall topology of the graph as described below.

Degree Distribution. Concept association networks are scale-free and that they have a power-law distribution with the approximate value of γ being 2.6. This is shown in Figure 1 as a log-log plot. This could imply that a preferential attachment model as in other scale free networks like (eg: WWW) is responsible for the structure and evolution of concept associations. The high degree nodes –hubs– are some of the salient concepts associated with human life: FOOD, MAN, GOD, MONEY, MARRIAGE etc. Scale-free behaviour of concept associations imply that humans arrive that few thoughts more often than most other thoughts. This could imply that we tend to associate new concepts in terms of these salient concepts and use them as landmarks to navigate the mind-scape. Another important consequence of the presence of hubs in the graph is a hierarchical organization of the network. This is the case if the subgraph induced by percolating from a hub node in turn has a scale-free distribution. This property is explored in the next section in the context of clustering.

Clustering Coefficient(CC). This property of nodes ensures that local communities are densely connected while still being able to communicate with the rest of the graph. CC of a node ranges from 0 to 1. In natural language, this is a measure of how many of our friends know each other. High CC may not necessarily mean isolated communities and hence longer diameters. Presence of a few random links across such local communities are enough to ensure small diameters.

For CAN we observe high CC for relatively small degree nodes while hub nodes have low CC indicating that the neighborhood of hubs are sparse. This indicates

that the hub nodes mainly serve as switches to keep the network connected rather than ensure similarity of nodes. The average CC observed is 0.18. This implies that in a concept neighborhood almost 20% of the concepts are associated with each other; this is a fairly high value. We rationalize this in the following way: there is an incentive to associate concepts as dense neighbourhoods as it adds to subtle variations in terms of associations –subtlety is directly related to richness of concepts. Long range links across dense neighborhoods are the ones that make the concept graph navigable with relatively short path lengths and hence help in relating distant concepts. Thus there is a cohesion between these two different forces –*abstraction* and *detail* provided respectively by hubs and densely connected neighbourhoods– resulting in forging associations that are rich and diverse.

4 Clustering of Concepts

In the following, we present our heuristic for clustering by taking into account the scale-free property and the high CC of this network and compare properties of clusters generated with those that result from spectral clustering.

4.1 Our Clustering Heuristic

We consider hub nodes as the starting point of this heuristic. To begin with n hub nodes are labelled as belonging to its own cluster $\mathcal{C}_i$. For each unlabelled node u in the graph, we explore its neighborhood and find the highest degree node v . If it is labelled $\mathcal{C}_i$, u is assigned the same label. If v is unlabelled and if $\text{degree}(v) \geq \text{degree}(u)$ we proceed to explore the neighborhood of v in order to assign it a label. Another possibility is v is unlabelled and $\text{degree}(v) < \text{degree}(u)$. In this case we assign u to a default cluster called $\mathcal{C}_0$. After the nodes are exhausted, this cluster is handled separately. Ideally we expect very few nodes to be in $\mathcal{C}_0$. If not, a possible reason could be that some hub nodes that were not considered as separate clusters have ended up in $\mathcal{C}_i$. Another reason could be the presence of small components that are detached from the main giant cluster. These can be rectified easily by including these hubs as separate clusters. Another way to resolve the nodes in the default cluster would be by labelling them nodes based on the shortest path to one of the hubs of the initialization step. In our experiments we encountered only few nodes in the default cluster and used this rule to resolve them.

The logic behind our heuristic is the following: As mentioned in the context of clustering coefficient, nodes with similar degrees in the neighborhood tend to give variations to a common theme whereas higher degree nodes tend to represent abstraction of concepts. Our heuristic ensures that label in a hub node percolates down to lower degree nodes in accordance with the above hypothesis.

4.2 Comparison between Our Heuristic and Spectral Clustering

A comparison between spectral clustering [7] and our heuristic is shown in Figure 2 as log-log plots of cluster degree distribution. For our discussions we have taken the

$n = 10$. As mentioned above, hubs correspond to concepts that are indicative of high recall. We are interested in 1. identifying a small number of leading concepts which serves as abstractions for other concepts, and 2. illustrating the nature of such clusters. In line with the hierarchical nature of concept associations, rather than having too many clusters, we prefer to cluster concepts into small number of clusters which lend themselves to sub-clustering. This apart there is no significance to our choice of n .

Fig. 2. Degree distribution of clusters: heuristic Vs spectral. Closely occurring plots in the middle correspond to clusters generated by our heuristic.

It is clear from the figure that spectral clustering does not preserve scale-free characteristics within clusters. Moreover, the sizes of clusters are uneven. A giant cluster is formed out of approximately 6000 nodes and another large cluster with around 3000 nodes; remaining 1000 odd nodes form 8 small clusters. Scale-free property is preserved only for the giant cluster³ for the simple reason that it is almost the entire graph. Unlike this, our heuristic splits the original graph into roughly equal sized clusters and each one has scale-free distribution with the same power-law exponent that applies for the whole graph. Thus the clusters from our approach are self-similar to the whole network. One could argue that a ‘cognitive structure’ is preserved in the clusters generated by our approach. Further our heuristic lends itself to power-law preserving recursive sub-clustering thereby imparting a hierarchical view to the whole network.

4.3 Discussions

The difference in the cluster properties generated by our heuristic and spectral clustering may be as a result of the latter’s nature. In essence spectral clustering is a series of random walks to estimate cluster boundaries –walks are contained within strongly connected components and rarely tend to take connecting

³ Clustering this further with the same method results in unevenly fragmented clusters and do not conform to scale-free distribution.

bridges. One starts with various ‘seeds’ to begin the random walk and sees the nodes that are reached eventually and thereby identify clusters. We believe that the difference is due to the fact that random walks are an unnatural means to navigate cognitive associations. Further studies are required to quantify this and are part of our on-going efforts.

5 Conclusions

In this work we studied an important aspect of concept associations — grouping of similar concepts. We showed that the structural properties of CAN’s help group similar and related concepts better than a conventional mechanism like spectral clustering. Towards this end we proposed a clustering heuristic that accomplishes this task and explained the rationale behind its design. As concept associations are a manifestation of human cognition and thought process, we believe that further and deeper study of their network representations will enhance our understanding of cognition.

References

1. Dorogovtsev, S.N., Mendes, J.F.F.: *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Inc., Oxford (2003)
2. Dorogovtsev, S.N., Mendes, J.F.F.: Language as an Evolving Word Web. *Proceedings of The Royal Society of London* 268(1485), 2603–2606 (2001)
3. Albert, R., Barabási, A.-L.: Emergence of Scaling in Random Networks. *Science* 289, 509–512 (1999)
4. Watts, D., Strogatz, S.: Collective dynamics of ‘Small-World’ Networks. *Nature* 293, 440–442 (1998)
5. Cancho, R.F.I., Sole, R.V.: The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences* 268, 2261–2266 (2001)
6. Steyvers, M., Tenenbaum, J.B.: The Large-scale Structure of Semantic Network: Statistical analyses and model of semantic growth. *Cognitive Science* 29(1), 41–78 (2005)
7. Andrew, Y., Ng, M.I.: Jordan and Yair Weiss: On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems*, vol. 14, pp. 849–856. MIT Press, Cambridge (2001)