A Model For Structural Similarity Based On Random Walks On Hypergraphs

Special Thanks To: NSF, Massachusetts Institute of Technology, Texas State University

Kevin Tian, Department of Mathematics and Computer Science, Massachusetts Institute of Technology
Dr. Byron Gao, Department of Computer Science, Texas State University

Introduction

The study of determining relationships between objects, in particular assigning similarity scores between pairs of objects in one or several homogeneous domains. This problem has roots in the field of co-citation analysis, in which we do not assume any knowledge of properties of the objects which we are considering, but rather only know of relationships between said objects. We represent the objects in each of the homogeneous domains as vertices, but allow for the existence of hyperedges, in which arbitrarily-sized sets of vertices can interact.

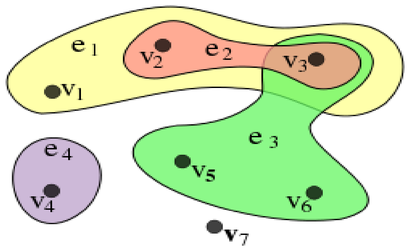
Background

The study of co-citation and the ideas behind it were introduced by Small and Marshakova in 1973. However, much influential work has been done in the field since then; in 1999, Jon Kleinberg published his influential HITS paper, which introduced the idea of hub and authority scores. PageRank built on this approach, by creating a ranking algorithm to measure importance in a set based on structural analysis. In 2001, the powerful SimRank algorithm was developed with ideas rooted in the aforementioned work. Its goal was to establish a function to measure structural similarity by creating an iterative algorithm which would converge to the desired values (pictured to the bottom). Our work greatly enhances the scope of SimRank by creating a new measure for structural similarity on hypergraphs, and extending the idea to span an arbitrary number of entity sets.

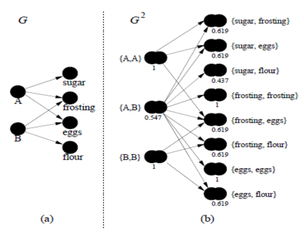
$$R_{k+1}(a,b) = \frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} R_k(I_i(a), I_j(b)) \qquad s(a,b) = \frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

Hypergraphs

Graph: a set of vertices V(*G*) connected by a set of edges E(*G*)
Hypergraph: Acts as an extension to the graph where edge set E(G) consist of edges e which act as a set of cardinality at least 2 vertices, with arbitrary size under that restriction.
Homogeneous Domain: Sets of objects which are comparable, or from the same category. Examples are objects from a user homogeneous domain, or foods, or restaurants, in a graph representative of a service which allows users to tag their restaurants and the types of food they ordered. It is only reasonable to compare objects from the same homogeneous domain.
k-partite Hypergraph on n Homogeneous Domains: A classical k-partite graph (where edges are purely binary) consists of k "parts" where no edge connects vertices from the same edge. We can extend this idea to a hypergraph, where each edge connects exactly (or at most) k vertices, no two of which are from the same homogeneous domain. It is reasonable to assume that k is no greater than n in this case. To the right we show an example of a 2-partite regular graph G, with G^2 also being a bipartite graph.

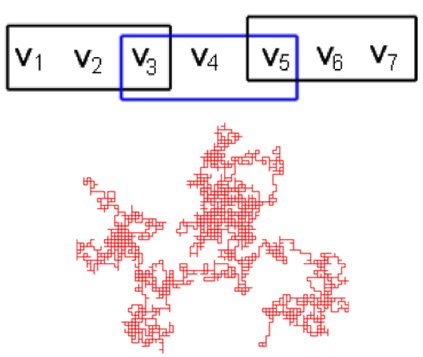


Hypergraph G consisting of the edge set V(G) = {v_1, v_2, v_3, v_4, v_5, v_6, v_7}, and the edge set (replacing vertices with their indices) {1, 2, 3}, {2, 3}, {3, 5, 6}, {4}. It is not a connected graph.



Random Walk on Hypergraph

A walk, more specifically a 1-walk, on a hypergraph consists of a sequence of vertices such that every two consecutive vertices is connected by a hyperedge. We will only be working with connected graphs, namely graphs in which there exists a walk from any vertex to any other vertex.
We will specifically be dealing with random walks on G and G^2, the graph consisting of ordered pairs (u, v) such that u and v are vertices in G, and there exists a hyperedge connecting two vertices in G^2 if and only if there exist hyperedges in G connecting their components.
A random walk is one in which at each vertex, the decision to extend to an arbitrary hyperedge is equally probable, and once the edge is selected, the decision to extend the walk to any of the other vertices along the hyperedge is equally probable. Pictured is a visual example of a 1-walk.



Our Work: Generalization of SimRank Method to Multiple-Domain Hypergraphs

One primary goal of our project was the extension of SimRank to provide a method to determine the similarity between two vertices in a homogeneous domain on a given hypergraph. We did so by proving the uniqueness of values satisfying a given set of equations (relationships among the similarity measures) as well as giving an algorithm that always converges to the desired values (which we also showed). Below, we include the system we developed.

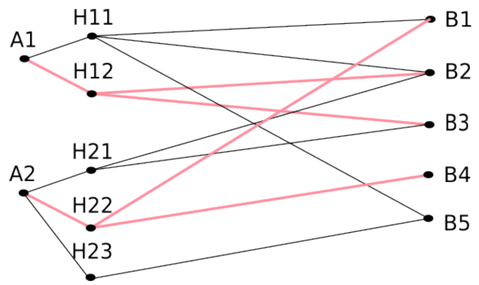
$$s(v_1,v_2) = \frac{c}{|E(v_1)||E(v_2)|}\sum_{E_1\in E(v_1)}\sum_{E_2\in E(v_2)}\frac{1}{(|E_1|-1)(|E_2|-1)}\sum_{u_1\in E_1}\sum_{u_2\in E_2} s(u_1,u_2)$$

Here, we include a portion of our paper which discusses the convergence of an algorithm based on the aforementioned system, as well as a proof of the uniqueness of the convergent values. The motivation for our method is because its recurrence relation is the same as that satisfied by the random walk model.

$$s_{k+1}(v_1,v_2) = \frac{c}{|E(v_1)||E(v_2)|}\sum_{E_1\in E(v_1)}\sum_{E_2\in E(v_2)}\frac{1}{(|E_1|-1)(|E_2|-1)}\sum_{u_1\in E_1}\sum_{u_2\in E_2} s_k(u_1,u_2)$$

First, we demonstrate the convergence of our above method. By the completeness of the real numbers, any nonempty bounded above set has a supremum (least upper bound), so it suffices to show that the sequence $s_0(v_1,v_2), s_1(v_1,v_2)\ldots$ is increasing and bounded, in which case the limit of the sequence will simply be its supremum. First, note that the sequence is bounded above by 1, which we can show inductively; clearly all of the $s_0(v_1,v_2)$ lie in the range $[0,1]$. Now, if all the $s_k(v_1,v_2)$ lie in $[0,1]$, it's clear that

$$s_{k+1}(v_1,v_2) \le \frac{c}{|E(v_1)||E(v_2)|}\sum_{E_1\in E(v_1)}\sum_{E_2\in E(v_2)}\frac{1}{(|E_1|-1)(|E_2|-1)}\sum_{u_1\in E_1}\sum_{u_2\in E_2} 1 = c < 1$$



We were also able to extend our idea to encompass a hypergraph on multiple entity sets, by using an extension of the random walk model we developed. In the extension, we would select a domain to "jump" to arbitrarily, whose intersection with the two randomly selected hyperedges were both nonempty. Among the other improvements we made to our results were the generalization of our function s: $V^2 \rightarrow [0, 1]$ to an arbitrary general similarity function from $V^k \rightarrow [0, 1]$ to measure the similarity between an arbitrary k-set of objects in the same homogeneous domain. The methods for convergence and uniqueness are analogous to the proof for the binary similarity calculation.

Future Work and Conclusions

In conclusion, our work holds significant value in acting as a more powerful and applicable tool to further the degree of accuracy and efficiency in computing structural similarity between objects in one or several domains, where we have no knowledge of the items' inherent properties. We manage to greatly improve the scope of the expected-meeting-distance random walk approach originally proposed by the influential SimRank paper, both by providing a model for hyperedges on multiple entity domains and by demonstrating the ability to generalize to an arbitrary number of items in the comparison function.
We hope to provide more experimental data to demonstrate the complexity and efficiency of our approach, and use it in practice to measure its effect on common applications of our method from web searches to uses along the lines of the original co-citation analysis.