Graph-Based Approach to Extracting Semantic Clusters from Behavioral Data

Chester Parrott, Doris Carver

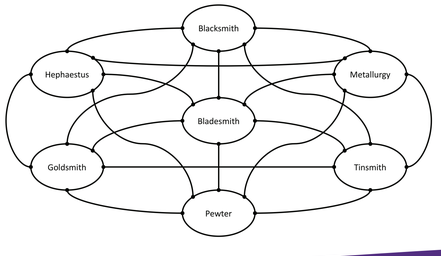
Introduction

The website *Wikipedia.com*, has become a rich source of information for users of the site and researchers alike. Users find value from the content in the millions of articles which *Wikipedia.com* hosts while researchers find value from not only the content but also the structure and organization of these articles. In Software Engineering, there is value in finding collections of articles sharing domain content. Previous efforts at semantic grouping of articles on *Wikipedia.com* have focused on using the structure, organization, and content of the articles.

We present a method of using anonymous user behavioral data (collected by *Wikipedia.com)* to group together articles which have shared domain content - semantic clusters.

Wikipedia as a Graph

We view the articles on *Wikipedia.com* as *nodes* in a *graph*. If an article, say A_1 (on *Blacksmithing)* has web links to other k other articles (say A_2, A_3, A_4, ... , A_k), we view those links to be *edges* in our graph between: (A_1 and A_2), (A_1 and A_3), (A_1 and A_4), etc. As our goal is to extract semantic clusters of articles, the volume of edges per node can result in the seeming untenability of the graph as a computational data structure. The *Wikipedia.com* clickstream data provides a solution for reducing the number of links between articles.



Behavioral Data

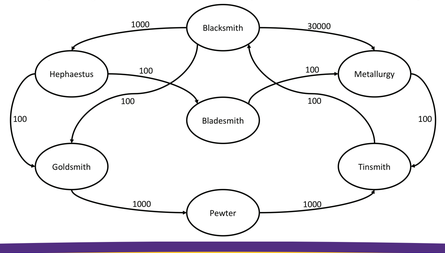
The first behavioral clickstream dataset for Wikipedia was released by the Wikimedia Foundation for the month of January 2015. Sporadic monthly releases have since followed which represent the aggregated traffic of all articles for the observed month. These data include the name of the '**Previous**' article from which the browsing behavior is initiated, the '**Current**' article which is being navigated to, and the volume of traffic associated with the browsing behavior ('**N**'). We can then compute **P(C|P)**, the conditional probability that given we have previously observed article **P** we now observe article **C**.

TABLE I
EXAMPLE WIKIPEDIA CLICKSTREAM DATA

Previous	Current	N	P(C|P)
Blacksmith	Hephaestus	1000	0.0322
Hephaestus	Goldsmith	100	0.500
Goldsmith	Pewter	1000	1.000
Pewter	Tinsmith	1000	1.000
Tinsmith	Blacksmith	100	1.000
Hephaestus	Bladesmith	100	0.500
Blacksmith	Goldsmith	100	0.0032
Blacksmith	Metallurgy	30000	0.9646
Metallurgy	Tinsmith	100	1.000
Bladesmith	Metallurgy	100	1.000

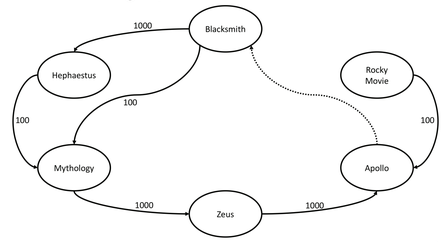
Clickstream Graph

By constructing a *graph* using only the behavioral data (instead of ALL article links), the resulting *graph* has a considerably smaller number of *edges*. Using Tarjan's **STRONGCONNECT** algorithm, we extract *connected components* from our behavioral *graph*. However, **STRONGCONNECT** has difficulty extracting semantic clusters from large connected components (as *Wikipedia.com* users may still browse all primary, secondary, and tertiary links on an article, these links may still be present and obstruct the purity of the resulting connected components.)



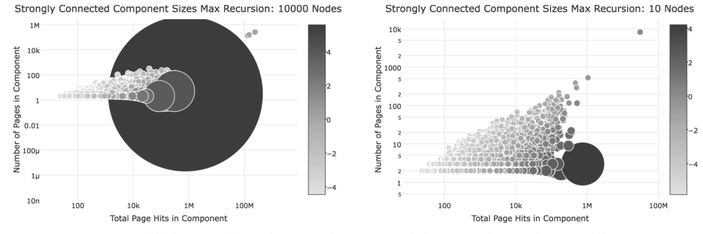
Depth Constraint

The **STRONGCONNECT** algorithm explores the clickstream graph through a graph exploration mechanism known as *depth-first search.* We found that due to the number and variety of links in Wikipedia.com articles, lengthy chains of links could result in the extraction of strongly connected components containing many unrelated articles. For instance, the graph below could result in a cluster of articles containing *Blacksmith*, *Hephaestus*, *Mythology*, *Zeus*, and *Apollo* (but not *Rocky Movie*.) We thus constrain the depth to which the **STRONGCONNECT** algorithm searches the graph.



Results

Depth Constraint allows us to reduce the size of the resulting connected components by two orders of magnitude. These connected components cluster together fewer articles which are semantically unrelated, resulting in a more homogeneous set of clusters.



This method can be used in conjunction with classic text-based approaches to semantic clustering of *Wikipedia.com* articles.

LSU College of Engineering
School of Electrical Engineering and Computer Science
Division of Computer Science and Engineering