VAST 2008 Challenge
Social Collaboration and Faction Discovery using Disagreement Graphs

Umang Sharan
INRIA Bordeaux
umangsh@gmail.com

Guy Melancon
INRIA Bordeaux
Guy.Melancon@labri.fr

ABSTRACT

This paper is a summary of the contest entry submitted to the VAST 2008 mini challenge 1. The primary task of the mini challenge was to use visual analytics to describe the groups and relationships of the editors of the *Paraiso (the movement)* wikipedia page based on the wikipedia edit logs. This paper summarizes the data analysis performed on the synthetic data set provided, describes the visualization algorithms and tools employed and the key observations from our analysis. We use Tulip [1] and GraphViz [2] for exploring the data set. Tulip is an extensive and flexible framework for visualizing large graphs providing the user an easy platform for exploring and manipulating large networks. Graphviz is an open source graph visualization software providing several interactive graphical interfaces and auxiliary tools for graph layouts.

Keywords: Visual analytics,VAST contest

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 INTRODUCTION

The goal of the Mini Challenge 1: Wiki Editors is to identify the factions represented in the edit logs for the wikipedia page *Paraiso (the movement)* associated with the Paraiso movement, and to comment whether the movement is involved in violent activities or not. The data set consists of 1009 wikipedia edits logged between 11^{th} August 2006 and 15^{th} January 2007. Given the quantity and nature of the data involved, we decided to employ Tulip and GraphViz to analyze the edit patterns and relationships. The rest of the paper is organized as follows: first, we discuss the data analysis task, then we describe the usefulness of Tulip and GraphViz for this challenge, followed by the results of our visualization algorithms.

2 DATA ANALYSIS

One of the most challenging aspects of the contest is the data analysis. Familiarization with the data required considerable effort. Apart from the wikipedia edit logs, a wikipedia discussion page was also provided as part of the data set. The wikipedia edit logs have a well defined grammar where each edit statement can be compiled using the following regular expression:

```
#[ ]+\(cur\) \(last\)[ ]+([0-9][0-9]:[0-9][0
-9]), ([\d]+) ([a-zA-Z]+) ([\d]+) ([\w.-]*)
\([\S]*[ \| \S]*\) [m ]?[\(.* bytes\)]?[ ]?(
\(.*\)?)
```

Scripts were written to extract the name, time stamp and the description message for each edit from the logs. Stemming, and term frequency (TF) [3] analysis were employed to filter relevant edits and extract useful information. On the other hand, the wikipedia discussion page had comments and opinions from users about different issues pertaining to the Paraiso movement. Therefore, there was no well defined grammar to automatically classify the comments apart from TF analysis and manual classification.

3 METHODOLOGY

We parsed the wikipedia edit logs and wikipedia discussion page to generate a *disagreement graph* between editors. We define an undirected graph $G = (V, E)$ as a *disagreement graph* where V denotes the nodes corresponding to the set of editors and E denotes the edges corresponding to the disagreements between two editors. Further, we define a weighing function $f : E \rightarrow \mathbb{N}$ on the edge set E where, $f(e) = n$ and $e(u, v) \in E$ implies that nodes u and v disagreed n times in the edit logs.

Figure 1 shows the disagreement graph extracted from the wikipedia edit logs in Tulip. The graph is loaded in the main window while properties/attributes associated with the nodes and edges in the graph are provided in the list view in the side for easy reference. Tulip also provides several graph layout algorithms for the best visualization (we use the GEM force-directed layout here).

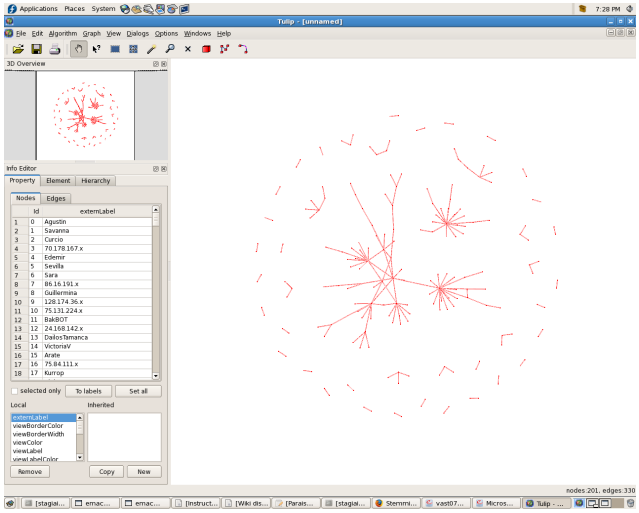

Figure 1: Wikipedia edit conflicts in Tulip

The notion of disagreement graph is a little counter-intuitive. However, the disagreement graph appeared the most logical representation because conflicts and arguments are easier to identify in edit logs based on keyword analysis.

4 OBSERVATIONS AND RESULTS

We expected the disagreement graph to be a bipartite graph between supporters and the opposers of the Paraiso movement. However, we found other factions as well. Figure 2 shows the disagreement graph in GraphViz. Apart from the strong supporters and strong opposers of the movement, we conjecture there is a faction of neutral wikipedia editors (moderators) which are mainly concerned with keeping the article conformant to the wikipedia standards. We also found several users that were automatic bots responsible for keeping in check drastic changes to the wiki page like page replacement, removal, etc. We also determined two types of spammers in the data set—neutral spammers who do random edits on the wiki page and

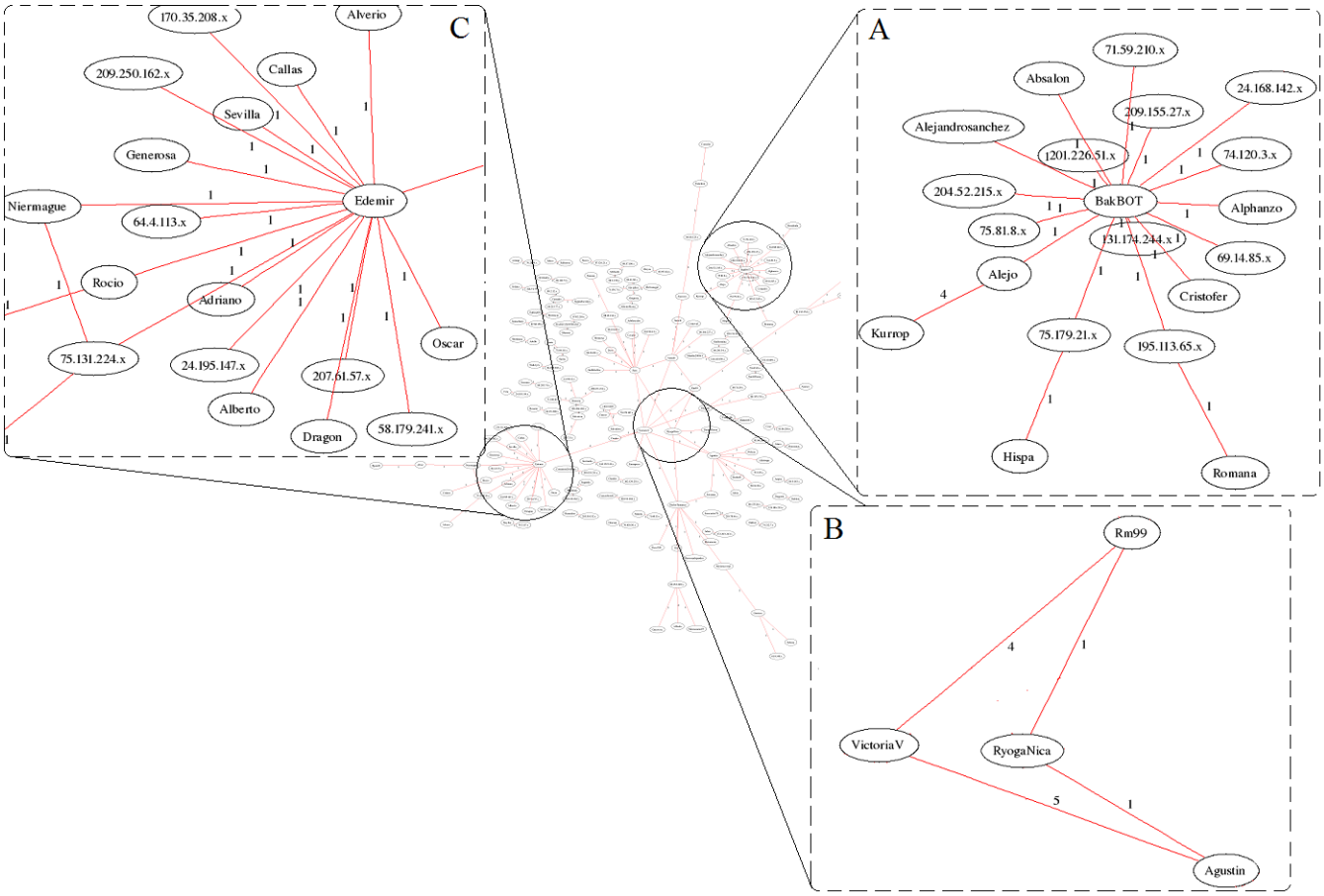

Figure 2: Disagreement graph between wikipedia editors

drastic spammers who make extreme changes like deleting the entire page. Such spammers appear to be strongly against the Paraiso movement but they edit the page only once—thus, they end up being classified as spammers rather than opposers of the movement. For example, in Figure 2 cluster *A* shows the disagreement cluster between a bot named BakBOT and a group of spammers like Absalon, Cristofer, etc. The following log snippet describes the disagreement between BakBOT and spammer Absalon.

```
# (cur) (last) 22:03, 3 September 2006 BakBOT
(Talk | contribs) (93,135 bytes) (Reverting
possible vandalism by Special:Contributions/
Absalon (see here). If this is a mistake,
report it. Thanks, BakBOT. (Bot))

# (cur) (last) 22:03, 3 September 2006 Absal
on (Talk | contribs) (129 bytes) (?Replaced
page with 'Well, this is simply a cult obses
sed with greed and the idea of more money. It
was created by some power crazy hispanic.')
```

Cluster *B* shows a bipartite graph between supporters and opposers of the Paraiso movement while cluster *C* shows the disagreements between a wikipedia moderator (neutral editor) and spammers. Contextual analysis was done to differentiate between bots and moderators—bots usually handle more extreme spammers which either destroy the wiki page or replace the entire wiki page while moderators are responsible for ensuring that the wiki article conforms to wikipedia's terms of usage and posting policy.

5 Conclusion

We combine some existing visualization tools like Tulip and GraphViz to explore relationships within the wikipedia data set. They provide the users with an overview of the entire data set facilitating exploratory data analysis. The extracted relationships modeled as the disagreement graph serve as a concrete visual analytic technique to assist investigators in analyzing wikipedia editor communities. The strength of our approach lies in the simplicity in exploring and navigating through the disagreement graph using Tulip and GraphViz.

One shortcoming with our approach is the data interpretation. Much work was done to extract information from all data sources without human intervention—however some sources like the wikipedia discussion page still required manual input. Future work in this direction would involve eliminating human intervention totally from the visualization and interpretation pipeline.

References

[1] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[2] J. Ellson, E. Gansner, E. Koutsofios, S. North, and G. Woodhull. Graphviz and dynagraph—static and dynamic graph drawing tools. In P. Mutzel and M. Jünger, editors, *Graph Drawing Software*, pages 127–148. Springer-Verlag, 2003.

[3] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.