Prajna – Wiki Editors Challenge

Edward Swing*

Vision Systems & Technology, Inc.

ABSTRACT

The Prajna Project is a Java toolkit designed to provide various capabilities for visualization, knowledge representation, geographic displays, semantic reasoning, and data fusion. Rather than attempt to recreate the significant capabilities provided in other tools, Prajna instead provides software bridges to incorporate other toolkits where appropriate.

This challenge required the development of a custom application for visual analysis. By applying the utilities within the Prajna project, I developed a robust and diverse set of capabilities to solve the analytical challenge.

KEYWORDS: Information Visualization, Software Toolkit, Knowledge Representation

INDEX TERMS: D.2.11 [Software Engineering]: Software Architectures - Domain-specific architectures; [Computer Graphics]: Methodology and Techniques - Interaction Techniques.

1 INTRODUCTION

This challenge involved a set of records of Wikipedia edits for the Paraiso Movement page. This data was formatted in a standardized format used by Wikipedia, and spanned a period of several months. Supplemental information included the Wikipedia page for the Paraiso Manifesto, and part of the discussion page for the Paraiso Manifesto wikipedia page.

The challenge included questions about the factions who were active in the ongoing contentious editing of the Paraiso Movement page. We were asked to identify factions and key personnel within the group of editors, and determine whether the Paraiso movement was involved in violent activities.

VSTI develops software for diverse customers who face numerous analytical challenges. The solutions developed for this challenge should apply to other analytical challenges that our customers must face.

2 DEVELOPING THE SOLUTION

2.1 Analysis of the Problem

Unlike other challenges in this contest, this particular challenge did not suggest any immediate avenues for visual analysis. The wikipedia edit history contained many contentious comments, and numerous cases where editors from various factions reverted each other's contributions in an ongoing edit-war.

In addition, the various edits frequently included the topic that the editor had been working on. This provided our second possibility for analysis.

A third possibility involved applying text extraction techniques to the comments contained within the various edit records. Therefore, I began to research the current state of entity extraction utilities. Unfortunately, the pervasive slang, abbreviations, and jargon within the edit records prevented any application of entity extraction techniques.

2.2 Building with Prajna

The Prajna Project includes significant capabilities in a variety of technology areas. It also provides various software bridges to enable application developers to use specialized toolkits. Therefore, part of the design of the solution included selecting the appropriate elements of Prajna to include.

The first task for any analytical challenge is parsing the data. The wikipedia edit history contained enough structure that an automated process could parse. Initially, the parser was able to extract the comments, edit size, editor, and time of edit.

However, since I was interested in identifying the various edit wars and disputes, I needed to parse the comments to identify and extract reversions. This required developing a utility to parse the comment field of each edit. For edits containing the indications of a reversion, the parser attempted to identify whose edits had been reverted. This proved difficult, since various editors used different styles to convey this information. Fortunately, most editors used a small set of styles when reverting each other. I also identified several different types of reversions – standard reversions, good faith reversions, and cases where an editor undid a reversion.

Once I had extracted the various editors who disputed each other via reverting edits, I decided to plot this as a social network. Unlike most social networks where a connection indicates closeness, the connections in this network represented just the reverse. Editors connecting to each other in this network were arguing with each other.

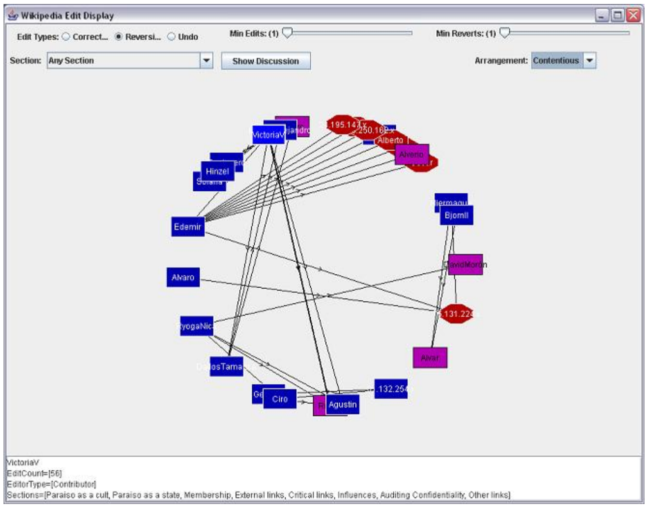
Furthermore, the set of reversions created a large number of disjoint sub-graphs. The sub-graphs also merged when different types of reversions were included or excluded from the network. Because of these factors, I rejected the use of a Force-Directed layout, and attempted to design alternate graph layout algorithms.

I developed one promising arrangement which separated the editors according to who they connected to. Two editors were grouped in the same faction if they had connections to the same editors. This was an attempt to represent agreement by common opponent. While this technique was not perfect, it did provide some interesting insights.

I added the ability to remove editors based upon the number of edits. Like most collaborative activities, the wikipedia edits had a small number of editors who contributed the majority of the content. Removing the editors with a low edit count also removed the vandals from the display.

To analyze the various topics, I developed filters to display only those comments for a particular section, or from a particular editor.

Email: deswing@vsticorp.com



Edit Types: ○ Correct... ● Reversi... ○ Undo
Min Edits: (1)
Min Reverts: (1)
Section: Any Section
Arrangement: Contentious

VictoriaV
EditCount=[56]
EditorType=[Contributor]
Sections=[Paraiso as a cult, Paraiso as a state, Membership, External links, Critical links, Influences, Auditing Confidentiality, Other links]

Display of one sub-graph of contentious editors

3 Performing the Analysis

Once we completed the development of the application, we applied the capabilities of the tools to the challenge itself.

In order to understand the social network, we examined several of the sub-graphs. One of the most active groups of editors also contained many of the editors with a high edit count. Therefore, this subgroup seemed to be a significant one. By examining this group, and its members, we revealed a significant amount of information about the social network of editors. Two of the principal editors, VictoriaV and Rm99, had a large count of reversions to each other. Clearly, these two were in opposing factions.

Another subgroup of interest contained a high number of vandals. After examining this subgroup, we identified that the vandals had largely been reverted by a single editor – BakBOT. This obviously represented an automated process which attempted to identify and correct vandalism attempts.

When we began to review and filter the comments by topic or by editor, we were disappointed. Because of the nature of the comments, many records had no topic identified. Others did not supply comments at all. We found it difficult to glean information without simply scanning the original series of edits.

We reviewed comments in the wikipedia discussion page to help analyze the roles that some of the principal editors played, and what factions they belonged to. This provided some indications of the various points of view held by various authors.

In order to answer the question about whether the Paraiso movement was involved in violent activities, we again turned to reviewing the individual edit records. Once again, the variability of the data prevented any automated analysis. Therefore, we scanned the edit records directly for clues.

4 Results

The factions of several personal key personalities were easy to deduce. We were able to identify the factions of other editors based upon the social graph. However, a significant amount of the identication still required reviewing the data manually.

To determine whether the Paraiso Movement was involved in violent activities, we tried some textual searches with violent terms. We expected, and found, nothing. However, we noticed several references to Paraiso activities in other countries. Notably, references to Paraiso prosecution in Belgium and Mexico gave us some evidence that Paraiso was involved in illegal activity. The final clue was an edit by Edemir, which referred to a confrontation with the Dept. of Health, followed by a vandalism asserting that Paraiso members had gunned down doctors and nurses. While other editors quickly removed this vandalism attempt, it strongly suggested that violent activities might be associated with Paraiso.

While this conclusion is far from iron-clad, and would not stand up to a court of law, we believe there is sufficient cause to investigate Paraiso activities further.

5 Conclusion

The tools developed for this challenge provided us most of the answers we sought. Furthermore, we were able to apply the principles and design of Prajna to this challenge, demonstrating its utility. The Prajna project attempts to provide a robust toolset, leaving the development of sophisticated visualization tools for other toolkits. In this fashion, Prajna may adopt the best visualization techniques by providing a software bridge to innovative toolkits.

VSTI is evaluating how to apply the techiques developed for this challenge to other projects. In order to verify the tool, I used it to parse other wikipedia edit histories. The edit history of the Scientology page contained a similar series of disputes and edit wars. This verified the robustness of the tool. VSTI has already integrated the wikipedia parser into other projects. Parsing semi-structured text is a common problem, and VSTI is continuing to research and develop tools to meet this challenge.

By providing an innovative architecture, which extends with software bridges to a variety of toolkits, Prajna avoids competing with the rapid pace of development across the spectrum of technology. Instead, Prajna offers developers the utilities to integrate new technology for knowledge representation in an intelligently designed architecture.

References

[1] E. Gamma, et al (the Gang-Of-Four), "Design Patterns: Elements of Reusable Object-Oriented Software", Addison-Wesley Professional, January 1995.

[2] D. Gotz, M. Zhou and V. Aggarwal, "Interactive Visual Synthesis of Analytic Knowledge", Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2006.

[3] J. Heer and M. Agrawala, "Software Design Patterns for Information Visualization", IEEE Transactions on Visualization and Computer Graphics (Proceedings), 2006.

[4] Wikipedia (web page): http://www.wikipedia.org