

## **Project Development Analysis of the OSS Community Using ST Mining**

Yongqin Gao  
University of Notre Dame  
ygao1@nd.edu

Greg Madey,  
University of Notre Dame  
gmadey@nd.edu

### **Abstract**

The OSS (Open Source Software) phenomenon is a novel, widely growing approach to develop both applications and infrastructure software recently. The fast growth of the community increases the interests in OSS related research. Accurate prediction of the project success is one of the interesting studies in OSS research. We propose to use the ST (Spatial Temporal) data mining techniques to predict the project success in the OSS community. ST mining has been studied in Euclidean distance based spatial systems like GIS, but to date has only received little attention in non-Euclidean network structured evolving system like the OSS community. In this paper, we introduce novel methods to project the evolving OSS community in a spatio-temporal data set and related ST mining algorithms to process the data set. Using ST mining techniques we propose, we are able to get the prediction of project success in the OSS community. We also present a detailed analysis and experimentally demonstrate the effectiveness and efficiency of these techniques in a real OSS community – SourceForge.net. The results show that our techniques can predict the project success and they are also useful in other non-Euclidean spatial systems.

#### Contact:

Yongqin Gao  
Dept. of Computer Science and Engineering  
University of Notre Dame  
Notre Dame, IN 46556

Tel: 1-574-631-7596  
Fax: 1-574-631-9260  
Email: [ygao1@nd.edu](mailto:ygao1@nd.edu)

Key Words: Open Source Software, ST mining

Support: This work was supported in part by the National Science Foundation (NSF), CISE/IIS-Digital Society & Technology, under Grant No. 0222829.

# Project development analysis of the OSS community using ST mining

**Yongqin Gao, Greg Madey**

Department of Computer Science and Engineering  
University of Notre Dame  
ygao1, gmadey@nd.edu

## 1 Introduction

Software is central to the functioning of modern society. The OSS (Open Source Software<sup>1</sup>) phenomenon is a novel, widely growing approach to develop both applications and infrastructure software. It exhibits many counter-intuitive properties and is not well understood. Open Source Software development, despite usually consisting of volunteers dispersed worldwide, now competes with commercial software firms. This competition is due in part to closed-source proprietary software being associated with risks to users. These risks, often referred to as a “software crisis,” can be summarized as systems that take too long to develop, are too costly and do not work very well when delivered. Over the following years after the software crisis was first addressed, many unsuccessful solutions [1] were proposed in an effort to solve the software crisis. But the approach promoted by Open Source Software (OSS) may finally present a good solution.

The purpose of Open Source Software was not to ensure distributing software to the end user without cost, but to ensure that the end user could use the software freely (no limitation on use or distribution, source code available and modifiable). Essentially, OSS exists in countless varieties today, each with its own unique history [4].

Linux, perhaps the best-known Open Source Software, began modestly in 1991, seven years after the founding of the Free Software Foundation. A recent survey carried out by the IDC<sup>2</sup> shows 49.8% growth in factory revenues and 51.4% growth in unit shipment for Linux servers year over year. “Linux server growth continued to accelerate,” said Jean S. Bozman, research vice president in IDC’s worldwide server group, “demonstrating that Linux servers are taking on important roles in IT customers’ computing infrastructure. What began with edge and web-centric workloads is branching out to include HPC and commercial workloads”.

Another notable OSS product is the Apache web-server, begun in February 1995. As of April 2005, Apache had achieved 69.32% of the web-server market with monthly increase about 0.25%, while market share of Microsoft’s web-servers has shrunk to 20.45%<sup>3</sup>.

Meanwhile, many OSS hosting sites emerged to support OSS development, such as Savannah<sup>4</sup>, Fresh-

---

<sup>1</sup>“Open source” is a certification mark owned by the Open Source Initiative (<http://www.opensource.org>).

<sup>2</sup>IDC is a premier global market intelligence and advisory firm in the information technology and telecommunications industries. Its homepage is <http://www.idc.com>

<sup>3</sup><http://www.netcraft.com/survey/>

<sup>4</sup><http://savannah.gnu.org>

meat<sup>5</sup>, SourceForge<sup>6</sup> and so on. SourceForge, as one of the largest and most famous OSS hosting web sites, offers features like bug tracking, project management, forum service, mailing list distribution, CVS and much more.

With the great and growing market share of several successful Open Source Software and fast expansion of the OSS communities, understanding the OSS movement came into the focus of many researchers. For computer science researchers, understanding the Open Source Software movement can help us improve current software engineering practices, both in quality and security.

The growing popularity of OSS recently has aroused interest in research about the OSS movement. Understanding the OSS movement can benefit many different domains like computer science, social science and economics. One of the interesting topics is the prediction of the successful projects. However, making prediction is not simple in the complex systems like the OSS community. SourceForge.net is one of the biggest OSS hosting sites in the OSS community. We will use the SourceForge community as the case study in this paper. SourceForge community is complex (as of April 2005, there are over 1,000,000 registered users, over 99,000 registered projects in the community, and there are also huge daily logs for activities and many statistical records since January 1999). We use the Spatio-temporal data mining techniques to tackle this problem.

ST (Spatio-Temporal) data mining is a new area in data mining research. The mixture of the complexity of spatial data and the difficulty in reasoning temporal information is the major challenge for ST data mining.

To the best of our knowledge, the ST data mining techniques have not yet been addressed in OSS research. Furthermore, the spatial and temporal properties in the SourceForge community are different from common spatio-temporal data. The space in the SourceForge community is a non-Metric space and the time in the SourceForge data consists of disconnected points defining a partial ordering of events. Until recently, there is few research about this kind of spatio-temporal data set. So the study of using ST data mining techniques in the SourceForge community can also expand the ST data mining research.

## 2 Related Works

There are many existing OSS research study the OSS community as a complex networked system. Jens [7] presented results of modeling the Open Source Software production process as a contest network, where similar vertices compete with each other and the winner will get some rewards. They suggested several possible individual motivations in the network – the active developers receive reputation and high investments and investors searching for highly talented applicants profit from the selection mechanism of the Open Source Software production process and finance it to receive inside information. Moreno and Faldani [12] used complex adaptive system theory to understand and analyze the Open Source Software community. Their paper concluded that the Open Source Software community is distinctive because it is neither controlled by a central authority that defines strategy and organization nor totally chaotic and it should be placed at a middle position between a planned community and a chaotic one. Their paper presented a description of the main characteristics of the functioning of the Open Source Software community regarding its organizational structure and development process. The concept of complex adaptive system was then introduced to simulate the OSS community. Using complex adaptive system theory, they interpreted the characteristics of the Open Source community.

Recently, researchers started to study the OSS community not only a complex networked system, but also an evolving system. Scacchi [14] concluded that OSS projects rely on electronic communication media,

---

<sup>5</sup><http://freshmeat.net>

<sup>6</sup><http://sourceforge.net>

virtual project management and version management mechanisms to coordinate globally dispersed software development efforts. He observed that OSS projects co-evolve with their development communities that reinvent and transfer software technologies as part of their community and project team building process. This research also confirmed the importance of complex structures and evolving properties in the OSS community. Madey and Gao [5] analyzed the empirical data they collected from SourceForge to obtain statistics and topological information of the Open Source Software developer collaboration network. They extracted the parameters of the evolution by inspecting the network over time. They also generated a model that depicts the evolution of this collaboration network. Degree distribution, diameter and clustering coefficient are frequent attributes used to describe a network and have been used ever since the beginning of small world network research. They also used these attributes to characterize the empirical data they collected from SourceForge. While other research tended to look at the network a single snapshot in its evolution, which means they all based their observations on network without respect to time. They were able to inspect the network with consideration of time using the empirical data collected over more than two years. A limitation of their research is that their studies are all based on non-automatic methods (e.g., statistical analysis). Since the data in the OSS community is huge, incomplete and noisy, non-automatic methods will become intractable to apply unaided. So we propose to develop and apply automatic methods like data mining to help with analyzing and understanding the OSS phenomenon.

As a data analysis technique, data mining can be used in social network research. The following are two examples of this research. Jensen and Neville [9] proposed the cross-disciplinary efforts and joint research efforts in machine learning, data mining and social network analysis. They argued that older data mining algorithms were developed to analyze propositional data, which are individual records that are assumed to be statistically independent to each other. But recent data mining algorithms focused on relational data where the relations among entities are central. Network data is typical relational data. So they concluded that such joint research in data mining and social network analysis will be very useful, especially in relational data. Kempe, Kleinberg and Tardos [10] studied the models for the processes by which ideas and influence propagate through a social network using data mining techniques. They used clustering techniques in data mining to aid their algorithm to discover the most influential nodes in the social network.

Data mining is a new and promising method recently used to study the OSS related research. Chawla, Arunasalam and Davis [3] reported their results of mining data acquired from SourceForge.net, the largest Open Source Software hosting website. In the process they introduced Association Rules Network(ARN), a (hyper-)graphical model to represent a special class of association rules. Using ARNs they discovered important relationships between the attributes of successful Open Source Software projects. They verified and validated these relationships using factor analysis, a classical statistical technique related to Singular Value Decomposition(SVD). This paper focused on the application of association rules method in the research of OSS. We will apply more techniques from data mining (clustering and classification) in the research of OSS. We will also develop new methods in the proposed research to support our research of OSS. Jensen and Scacchi [8] combined techniques from text analysis, link analysis and of repository usage and update patterns to discover software processes from OSS development web repositories. They believed that this discovery can help with the understanding of the process techniques that have led to their success. They only used classic feature based data mining methods in the research, which is inadequate in OSS research due to the existence of ample temporal information in the OSS community. We will develop new temporal related data mining techniques to study the OSS community.

In these studies, only classical feature based data mining techniques are involved. And these techniques are not efficient in the complex evolving system like the OSS community, so we will use Spatio-temporal data mining techniques in this study.

### 3 Our Approach

Research about spatio-temporal data mining has been triggered by the availability of gigantic volumes of geospatial data, often continually updated. Exploration of spatial data mining [6, 11] and temporal data mining [13] has received much attention independently in KDD and DM research community. Nevertheless, the combination of “spatial” and “temporal” makes the data mining procedure even more complicated. Recently, more and more researchers began to investigate spatio-temporal data mining [17, 2, 16, 15].

The OSS communities also have the spatial and temporal properties. The collaboration network inside the community is a graph without inherent distance definition. So it can only be described as an Euclidean space, which is different from the metric space we can see in geospatial information. Meanwhile, we also have log information for many entities in the community, which is the partial ordered (disconnected) temporal property of the OSS community. So the OSS community is a good case study for spatio-temporal data mining research due to its unique spatial and temporal characteristics and we will use the OSS community as case study to develop new techniques for spatio-temporal data mining. The focus of this paper will be the spatio-temporal data mining techniques used to predict the project success.

There are many different studies about the Open Source Software community. We only focus on the prediction of the successful project in the Open Source Software community. SourceForge.net is the study case used in this paper.

The collaboration relationships between the developers in the OSS community are the focus of our study. In this paper, we describe these relationships using networks.

1. *Bipartite network*: The OSS community can be presented as a graph  $G(V, E)$ , where  $V = V_p + V_u$ ,  $V_p = \{v_p | v_p \text{ is a project in the community}\}$  and  $V_u = \{v_u | v_u \text{ is a user in the community}\}$ ;  $E = \{e | e \text{ is an edge between } v_p \text{ and } v_u, \text{ when user } v_u \text{ participates in project } v_p\}$ .
2. *User network*: The OSS community can also be presented as a graph  $G(V, E, W)$ , where  $V = \{v | v \text{ is a user in the community}\}$ ,  $E = \{e | e \text{ is an edge of } (v_1, v_2), \text{ when } v_1 \text{ and } v_2 \text{ participate in one or more common project}\}$  and  $W = \{w_i | \text{for every } e_i \in E, w_i \text{ is the weight of the edge } e_i\}$ .
3. *Project network*: The OSS community can also be presented as the dual network of part (2).

Scientists and mathematicians have been studying such networks for some time. Different from the traditional networks computer scientists study, randomness is dominant in this kind of network, since there is no centralized control of the construction and evolution of the network. Due to this property, Making prediction about the development of the network or entities in the network is difficult. On the other hand, prediction about the development of the project is an interesting study in OSS research

In this paper, we will use the spatio-temporal data mining techniques to tackle this problem. Problem definition, algorithms and experiment results will be presented and discussed.

### References

- [1] F. Brooks. No silver bullet: essence and accidents of software engineering. *IEEE Computer Magazine*, pages 10–19, 1987.

- [2] E. Castro and M. Houle. Fast randomized algorithms for robust estimation of location. *Proc. International workshop on Temporal, Spatial and Spatio-temporal Data Mining, Lyon, France, 2000.*
- [3] S. Chawla, B. Arunasalam, and J. Davis. Mining open source software (oss) data using association rules network. *Proceeding of PAKDD conf.*, 2003.
- [4] G. Drummond. Open source software and documents: A literature and online resource review. [http : //www.omar.org/opensource/litreviewa](http://www.omar.org/opensource/litreviewa), 1999.
- [5] Y. Gao, V. Freeh, and G. Madey. Analysis and modeling of the open source software community. *Proceeding of NAACSOS conference, Pittsburgh, PA, 2003.*
- [6] J. Han, N. Stefanovic, and K. Koperski. Selective materialization: an efficient method for spatial data cube construction. *Proceedings of PAKDD 1998*, pages 144–158, 1998.
- [7] P. Jens. Network formation via contests: the production process of open source software. *Working paper, Johann-Wolfgang-Goethe University, Frankfurt, 2004.*
- [8] C. Jensen and W. Scacchi. Data mining for software process discovery in open source software development communities. *Workshop on Mining Software Repositories, Edinburgh, Scotland, 2004.*
- [9] D. Jensen and J. Neville. Data mining in social networks. *Symposium on Dynamic Social Network Modeling and Analysis, Washington DC, USA, 2002.*
- [10] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. *SIGKDD '03, Washington DC, USA, 2003.*
- [11] H. Miller and J. Han. Spatial clustering methods in data mining: a survey. *Geographic data mining and knowledge discovery, Taylor and Francis, 2001.*
- [12] M. Moreno and M. Faldani. Open source as a complex adaptive system. *Emergence*, 5(3), 2003.
- [13] J. Roddick and M. Spiliopoulou. a survey of temporal knowledge discovery paradigms and methods. *IEEE Transaction of Knowledge Data Engineering* 14, 4:750–767, 2002.
- [14] W. Scacchi. Free/open source software development practices in the computer game community. *IEEE Software, Special Issue on Open Source Software*, 21:59–67, 2004.
- [15] J. Sun, D. Papadias, Y. Tao, and B. Liu. Querying about the past, the present, and the future in spatial-temporal. *Proc. of ICDE 2004*, pages 202–213, 2004.
- [16] Y. Tao, J. Sun, and D. Papadias. Selectivity estimation for predictive spatial-temporal queries. *Proc. of ICDE 2003*, pages 417–428, 2003.
- [17] X. Ya. Research issues in spatio-temporal data mining. *Workshop in Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, 2003.*