Yuan Tian
PhD Candidate
School of Information Systems, Singapore Management University

Research Statement

Integrated Data Analysis on Heterogeneous Data Source

1 Research Background

A crucial aspect common to many areas of our life is the huge amount of data generated daily. This data covering business, software development, sociology, and other important domains provides a rich source of useful information that can support decision making. This brings the need for efficient yet effective data analytics approaches.

A crucial challenge is the effective analysis of heterogenous data sources. In many scenarios, the situation is similar to the story of "The blind men and an elephant"[1]; often we have various pieces of information on a particular topic spread out in various data sources. Individually, the information might be of limited use and might even be contradictory to one another. However, by integrating them together we could get the complete picture which could guide a decision maker to make a wise decision. Data resides in different forms, ranging from unstructured data on the web to highly structured data in relational database systems. Many research studies often focus on only one kind of data source, either unstructured or structured. However, as the complexity of data increases, we could no longer only analyze one source of data independent from others as we would then miss much information just like the blind men in the story. Motivated by this, my research goal is to *holistically integrate and analyze heterogeneous data sources.*

2 Integrated Data Analytics

Traditionally, data source can be classified into three categories: structured data (e.g., graphs), unstructured data (e.g., text), and semi-structured data (e.g., XML). In my proposed research, I would focus on developing techniques that could holistically analyze structured and unstructured data. I plan to apply them on various domains including software engineering, and social network mining. In the following paragraphs, I highlight related studies that I plan to extend in my PhD work.

2.1 Structured Data Analysis

Structured data is one that is organized in a uniform and identifiable structure. The most common form of structured data is a database. I'm particularly interested on mining patterns from data that could be put into a set of transactions, a set of sequences, or a set of graphs.

The earliest study on pattern mining in a transaction database is by Agrawal and Srikant [1]. Many other studies extend their work to make it faster, or to mine a compact set of patterns [2]. The earliest study on sequential pattern mining is again by Agrawal and Srikant [3]. Many other studies have improved the performance of the previous algorithm by various heuristics and by mining a compact set of patterns [4, 5]. The earliest studies on mining information (i.e, pattern) from graph data were conducted

[1] This story is originated in India, where a group of blind men touch an elephant to learn what it looks like. Each one feels a different part, but only one part, such as the side or the tusk. They then compare notes with other and find that they are in complete disagreement.

by Cook and Holder [6] and Yoshida and Motoda [7] in the middle of the 1990's. Since then, a large number of papers in this topic have been published. e.g., [8].

In my PhD work, I plan to develop new techniques that could analyze heterogenous data that could combine two or more of the above data formats: transactions (or sets), sequences, and graphs. There are many applications where this integrated analysis is needed. For example, in the analysis of software systems, a program code is a graph, and a program execution traces is a sequence. In a social network, the friendship network is a graph, the sequence of activities inside the network is a sequence, and the properties of each user is a set.

2.2 Unstructured Data Analysis

Contrary to the definition of structured data, unstructured data refers to information that either does not have a pre-defined data model or does not fit well into relational tables. Typically, unstructured data is in the form of a text document. I'm particularly interested on applied text mining.

Text mining is a large and interesting research area that analyze and find knowledge from a collection of textual documents. The general framework for text mining contains two phases: *text refinement* that transforms the unstructured data into an intermediate form and *knowledge extraction* that deduces patterns or nuggets of knowledge from the intermediate form. I'm particularly interested on topic modeling, e.g., [9] and document similarity, e.g., [10].

In my PhD work, I plan to integrate pattern mining with some text mining approaches to analyze heterogenous data. Again there are many examples where this holistic approach would be needed. In a software development process, there are bug reports, commit logs, and comments which are textual documents; of course, there are also program code and execution traces that are graphs and sequences respectively. In a social network, there are textual contents exchanged between users and there are also links that connect users in the network forming a graph structure.

3 Past Research Work

In my previous work, I've analyzed structured data (i.e., program code), unstructured data (i.e., text), and a mixture of both. We have applied them to problems in software engineering and social network mining.

I have studied the problem of identifying bug fixing pathes which is to appear in the 34th ACM/IEEE International Conference on Software Engineering[2] [11]. Our work identifies patches that need to be propagated to stable versions of the Linux kernel. In this work, we use a framework that combines static program analysis and semi-supervised classification. It contains two main steps: the extraction of features from the data set (which is a set commit log (text), and program changes (graph)), and effective classification by using a combination of Learning from Positive-and-Unlabeled classifier (LPU) and Support Vector Machine (SVM).

In another work that is to appear in the 16th European Conference on Software Maintenance and Reengineering (ERA track)[3], I have built a technique that labels if a bug report is a duplicate or not [12]. We extract features based on document similarity

[2] Full paper, 11 pages.

[3] Emerging Research Achievement Track – 6 pages

and the concept of relative similarity. Next, we build an effective classifier based on these features.

I have also analyzed various software related microblogs in Twitter. Two papers are currently under submission on this topic.

4 Future Research Plans

In my previous studies, I benefited from collaborating with my supervisor Dr. David Lo. To extend the above studies, I would like to continue working with him. I first plan to study techniques that analyze structured data, especially in pattern (transaction, sequence, and graph) mining field. Next, I plan to develop better pattern mining techniques that can solve a domain specific need. In the unstructured data analysis area, I plan to read more on topic modeling and various measures of document similarity, and then develop domain specific techniques that can address peculiar textual documents, e.g., literals in a program code. Eventually, I plan to develop a holistic solution that combine structured and unstructured data analytics for domain-specific needs.

According to my good research performance so far and the rich research experience that SMU offers, I am confident that I will finally achieve my research goal.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] Mohammed Javeed Zaki. Mining non-redundant association rules. *Data Min. Knowl. Discov.*, 9(3):223–248, 2004.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *ICDE*, pages 3–14, 1995.

[4] Mohammed Javeed Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001.

[5] David Lo, Siau-Cheng Khoo, and Jinyan Li. Mining and ranking generators of sequential patterns. In *SDM*, pages 553–564, 2008.

[6] J.Cook and L.Holder. Substructured discovery using minimum description length and background knowledge. In *J.Artificial Intel. Research*, 1994.

[7] K.Yoshida, H.Motoda, and Indurkhya N. Graph based induction as a unified learning framework. In *J.of Applied Intel.*, 1994.

[8] Feida Zhu, Qiang Qu, David Lo, Xifeng Yan, Jiawei Han, and Philip S. Yu. Mining top-k large structural patterns in a massive network. *PVLDB*, 4(11):807–818, 2011.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *NIPS*, pages 601–608, 2001.

[10] C.Sun, D.Lo, S.Khoo, and J.Jiang. Towards more accurate retrieval of duplicate bug reports. In *ASE*, pages 253–262, 2011.

[11] Y.Tian, J.L.Lawall, and D. Lo. Indentifying linux bug fixing patches. In *International Conference on Software Engineering(ICSE)*, 2012.

[12] Y.Tian, C.Sun, and D. Lo. Improved duplicate bug report identification. In *European Conference on Software Maintenance and Reengineering (ERA track)*, 2012.