氏　　名	Le-Duc Tung
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1836 号
学位授与の日付	平成28年3月24日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	A Systematic Approach to Regular-Expression-Based Queries on Big Graphs
論 文 審 査 委 員	主　査　教授　胡　振江 教授　中島　震 教授　河原林　健一 助教　日高　宗一郎 助教　加藤　弘之 教授　岩崎　英哉　電気通信大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

Graphs have been increasingly important to represent data such as the World Wide Web, social networks, and biological networks. With the explosion of information, big data leads to big graphs. These big graphs are often stored in a distributed system, leading to many difficulties in proposing efficient algorithms for processing big graphs.

While many distributed programming models for graphs have been proposed, MapReduce and Pregel have been shown to be scalable to deal with big data as well as big graphs. Nonetheless, to obtain scalability, these models offer restricted forms in which users specify their programs. Hence, it is non-trivial for users to write their complicated programs as well as to obtain efficient ones. On the other hand, regular expression has been used as a powerful way to intuitively query data from graphs. Many useful applications of queries based on regular expressions have been discovered, such as, finding relationships in social networks, or finding chains of reactions in biological networks. However, evaluating regular expression-based queries on distributed graphs is non-trivial. First, regular expressions imply a highly sequential evaluation. Second, distributed evaluations often produce intermediate graphs whose size is larger than the input graph, and require a large amount of communications.

The objective of this dissertation is to bridge the gap between regular-expression-based queries and scalable distributed programming models. We study systematic approaches to build a general framework that automatically translates regular-expression-based queries into efficient distributed programs.

First, we focus on select-where regular path (SWRP) queries that return a graph constructed from subgraphs following paths whose labels spell a word in a regular expression. Queries can be nested and composed. We propose a structural-recursion based approach to translating SWRP queries into efficient programs in Pregel. SWRP queries are first translated into structural recursive functions on graphs. Then structural recursive functions are compiled into efficient programs in Pregel. The approach ensures that the sizes of intermediate graphs generated during the evaluation are minimized and close to the size of the final result. To the best of our knowledge, this is the first time a Pregel algorithm for SWRP queries is proposed.

Second, we propose a functional-based approach to further improve the performance of SWRP queries. We observe that there is a computation during the evaluation of SWRP queries takes more time than the other computations. This demands further refinement of our framework. We start with a more fundamental query that is a regular reachability (RR) query. An RR query is to decide whether two given vertices are connected or not by a directed path, where the concatenation of whose edge/node labels spells a word in a given regular expression. We propose a functional-based approach to a distributed evaluation of RR queries, which uses functions to encode mappings between sets of states in the automaton of the given regular expression. This approach exploits parallelism by processing a long path in a distributed manner, and it also reduces the computation and communication costs during the evaluation by encoding state transitions. Then we show how to apply this approach to improve the performance of the evaluation of SWRP queries.

Finally, we extend SWRP queries to support shortest-path conditions. We show that this extension requires us to solve an additional problem that is a shortest regular category-path (SRCP) query. An SRCP query is a variant of a constrained shortest path query whose constraints are expressed by a category-based regular expression. By using a dynamic programming formulation, we show that SRCP queries can be answered efficiently by a series of single source shortest path searches. This is useful because we can utilize fast single source shortest path algorithms that are optimized for different graphs (road networks, social networks, biological networks) and environments (shared or distributed memory).

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨
Summary of the results of the doctoral thesis screening

本論文は、ビッググラフを対象とする正規表現に基づく問い合わせの系統的処理法に関するものである。近年、ネットワーク構造を持つ大規模なグラフを処理対象とし、これらのグラフから新しい知識を獲得することを目的とするアプリケーションが増えている。グラフは典型的な半構造データであり、ソーシャルネットワーク、文献参照関係等の書誌情報ネットワーク、生命情報科学ネットワークなどがその代表例である。これらのグラフから有益な情報を抽出するために、正規表現に基づく問い合わせがよく用いられる。しかし、分散的なビッググラフを対象とする正規表現に基づく問い合わせを効率的に計算することは難しく、既存の並列化手法では入力グラフよりもはるかに大きい中間グラフの生成と大量の通信が必要となる。一方、ビッググラフ上の計算を分散並列環境で行うための効率的な基盤としては、Google によって提唱された Pregel モデルが広く認知されている。しかし Pregel は頂点中心の低レベルの計算に基づくモデルであり，必ずしも容易なプログラミングを提供するものではない。本論文では、正規表現に基づく問い合わせの利便性とPregelの効率性を融合し、正規表現に基づく問い合わせを効率的なPregelプログラムに自動的に変換するアルゴリズムを与え、ビッググラフから情報を効率的に抽出するための系統的な処理法を示している。

本論文は英語で記述されており、全 6 章から構成されている。

第 1 章は序論である。研究の背景、研究目的、主要な貢献など論文全体の構成を述べている。

第 2 章は基礎知識の紹介である。グラフモデル、正規表現とそれに基づくグラフ問い合わせ言語、Pregel 計算モデルについて議論している。

第 3 章では、SQL の文法に近い select-where regular path (SQRP)というグラフ上の問い合わせに着目し、SQRP を構造的再帰関数とその合成に変換することを経由して Pregel プログラムにコンパイルする方法を示している。抽象的計算を導入することにより不要な計算を削除することができ、大きい中間グラフが不要に生成されてしまうという問題を解決している。

第 4 章では、状態集合の遷移を関数でエンコードすることにより並列化を最大とし、SQRP を最適化する方法を示している。

第 5 章では、最短パス条件を記述可能とするための SQRP 拡張し、拡張された SQRP を Pregel プログラムへコンパイルする実現方法を議論している。

第 6 章は論文のまとめと今後の課題である。ビッググラフを対象に、提案した正規表現に基づく問い合わせの系統的処理法が新規的で有効であると結論づけるとともに、将来の課題を論じている。

(別紙様式 3)
(Separate Form 3)

なお、論文成果は海外論文誌論文 1 件（査読あり）、国際会議論文 3 件（査読あり）として公開されている。また、開発したシステムもウェブで自由にダウンロードできるようになっている。

審査会において、出願者は上述の内容に沿って説明を行い、そのあと審査委員との質疑応答を行った。質疑応答では、論文及び口頭発表の内容に関して、構造的再帰関数を用いる動機、実験データの規模、変換過程の正確さを中心に質問があり、的確な回答がなされた。

以上に基づき審査した結果、6 名の審査委員全員一致で,本学位請求論文は学位を授与するのに十分なレベルであるものと判定した。