



ОПТИМИЗАЦИЯ ИТЕРАТИВНЫХ АЛГОРИТМОВ ОБРАБОТКИ ГРАФОВ НА МОДЕЛИ MAPREDUCE.

В. К. Варанкин

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Введение. Итеративные алгоритмы представляют собой важный класс задач, применяемых во многих областях человеческой деятельности: веб-поиск, анализ социальных сетей, медицина и многих других. Многие итеративные алгоритмы обработки графов обладают так называемыми слабыми вычислительными зависимостями [1]. То есть, изменение состояния одной вершины непосредственно затрагивает только вершины, соседние с ней. На данный момент использование данной особенности в некоторых промышленных системах обработки данных, является проблематичным [2]. В докладе представлена модель итеративного процесса, учитывающая описанную особенность алгоритмов обработки графов.

Модель вычислений. Для эффективного исполнения данного типа алгоритмов заменим функцию $f(s)$, обрабатывающую все множество S , на две другие функции:

$$\begin{aligned} W_{i+1} &= c(S_i, W_i, G) \\ S_{i+1} &= u(S_i, W_i) \end{aligned}$$

где функция c вычисляет множество W - множество данных, кандидатов, которые могут повлиять на промежуточный результат S_{i+1} . Множество G представляет собой все множество вершин и дуг, описывающих исходный граф (не меняется в процессе итераций). Функция u производит обновление S на основании множества W . Тогда алгоритм примет следующий вид:

$$\begin{aligned} &\text{while } t(W_i) \text{ do} \\ &S_{i+1} = u(S_i, W_i) \\ &W_{i+1} = c(S_i, W_i, G) \end{aligned}$$

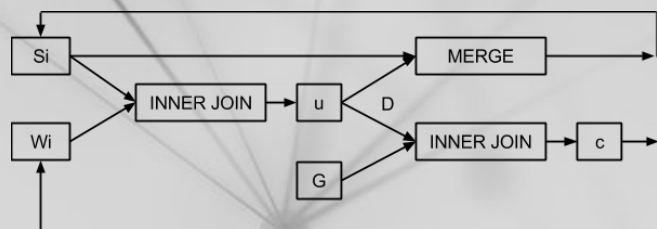


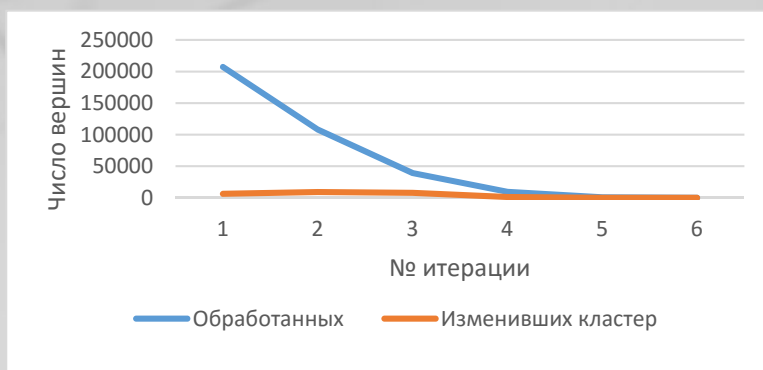
Рис. 1. Схема алгоритма

Выводы. Предложенная модель позволяет снизить объем данных, необходимых для вычисления алгоритма, а также количество передаваемых данных по сети, которая является наиболее узким местом в распределенных системах обработки данных [4].

Эксперимент. В качестве демонстрации потенциала выборочной обработки данных рассмотрим пример исполнения алгоритма соединенных компонентов на графе “Взаимного голосования пользователей Википедии” [3].

Граф представляет собой совокупность направленных дуг из вершины А в В, означающих, что пользователь А проголосовал за пользователя В. Количество вершин в графе - 7115, количество дуг - 103689.

Реализованный алгоритм, помимо основной своей функции, также подсчитывает количество потенциальных (тех, что могли произойти) и реальных (тех, что произошли) изменений кластеров вершин графа на каждой из итераций. В результате его работы получаем 2 графика, демонстрирующих зависимость числа обработанных вершин от номера итерации:



Список литературы:

1. Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. GraphLab: A new framework for parallel machine learning. In UAI, pp. 340-349, 2010
2. А. В. Митяков, Татаринов Ю. С. Подходы к эффективному исполнению итеративных алгоритмов на модели MapReduce. // Известия ЛЭТИ, Вып. 2, 2014 С. 30-41.
3. SNAP: Network datasets: Wikipedia vote network. <http://snap.stanford.edu/data/wiki-Vote.html>
4. А. В. Митяков, Татаринов Ю. С. Подход к ускорению итеративных алгоритмов на модели MapReduce // Известия ЛЭТИ, Вып. 8, 2014 С. 27-31.