Оптимизированный алгоритм поиска ученых на основе социального графа библиографической информации

Ибрагимов Рустам
группа 461
руководитель Нестеров В.М.

7 июня 2013 г.

Тема работы

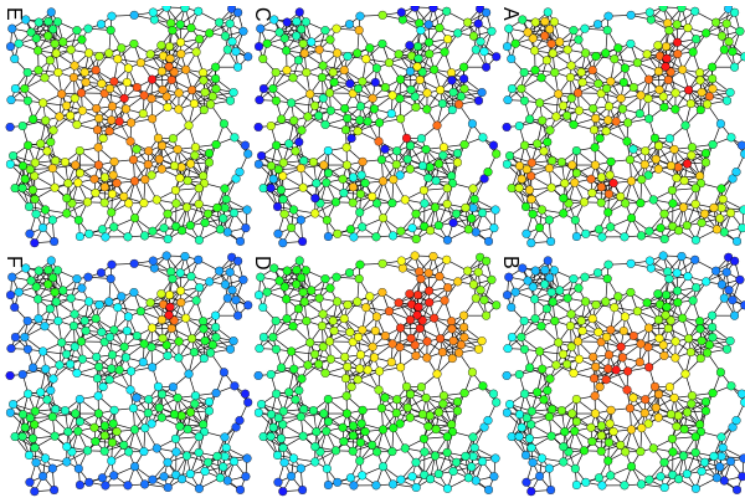
- Поиск и извлечение данных
- Научные сообщества
- Метрики для социального графа
- Машинное обучение

Постановка задачи

Целью данной работы является разработка алгоритма поиска ученых на социальном графе библиографической информации.

- Реализация вычисления метрик графа
- Разработка и реализация алгоритма поиска ученых на графе с помощью средств машинного обучения
- Оптимизация скорости работы на больших графах

Метрики



Метрики позволяют оценить значимость одной вершины социального графа и ее влияние на остальные вершины.
Данные значения можно использовать для выделения сообществ и составления рекомендаций пользователю.

Метрики

- Степенная метрика (degree centrality)

$$C_D(v) = deg(v)$$

- Метрика близости (closeness centraity)

$$C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)}$$

- Метрика промежуточности (betweenness centraity)

$$C_B(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Метрики

- Метрика Каца (Katz centraity)

$$\overrightarrow{C_K} = ((I - \alpha A^T)^{-1} - I)\overrightarrow{I}$$

$$C_K = \sum_{k=1}^{\infty} \alpha^k \sum_{j=1}^{N} \sum_{i=1}^{n} (A^k)_{ij}$$

- Альфа метрика (Alpha centraity)

$$x = (I - \alpha A^T)^{-1} e$$

Разложение матрицы

- LU-разложение
- LDL^T-разложение
- Разложение Холецкого

Численные библиотеки

- Matrix Toolkits Java
- Apache Commons
- OjAlgo
- Colt
- ParallelColt
- JAMA

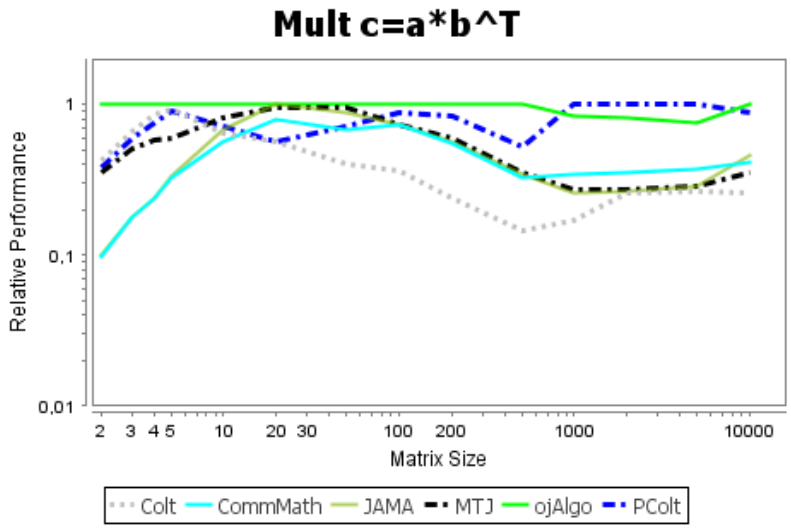
Сравнение библиотек

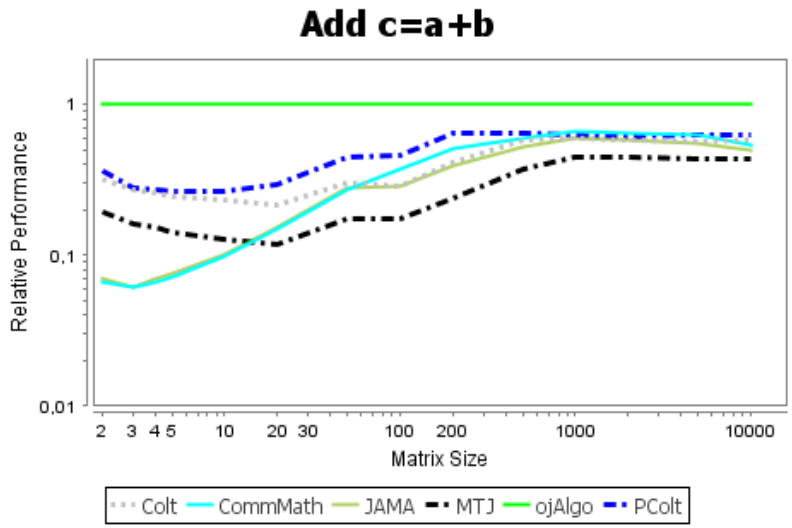
Критерии сравнения библиотек:

- Типы для хранения плотных и разреженных матриц
- Скорость работы при операциях с большими матрицами
- Реализация разложений

Сравнение библиотек

Java matrix benchmark - это инструмент для сравнения численных библиотек для языка программирования Java.





В результате были выбраны библиотеки OjAlgo и ParallelColt.

Этапы работы

1. Обработка поискового запроса
2. Построение исходного социального графа
3. Адаптация к входным данным
4. Вычисление метрик
5. Выделение групп
6. Построение результирующего социального графа

Адаптация к входным данным

Анализируя входную матрицу мы можем определить:

- Оптимальный тип хранения матрицы
- Оптимальный метод вычисления метрик

Входная матрица смежности может быть:

- Необратимой
- Обратимой
- Обратимой положительно-определенной

А также

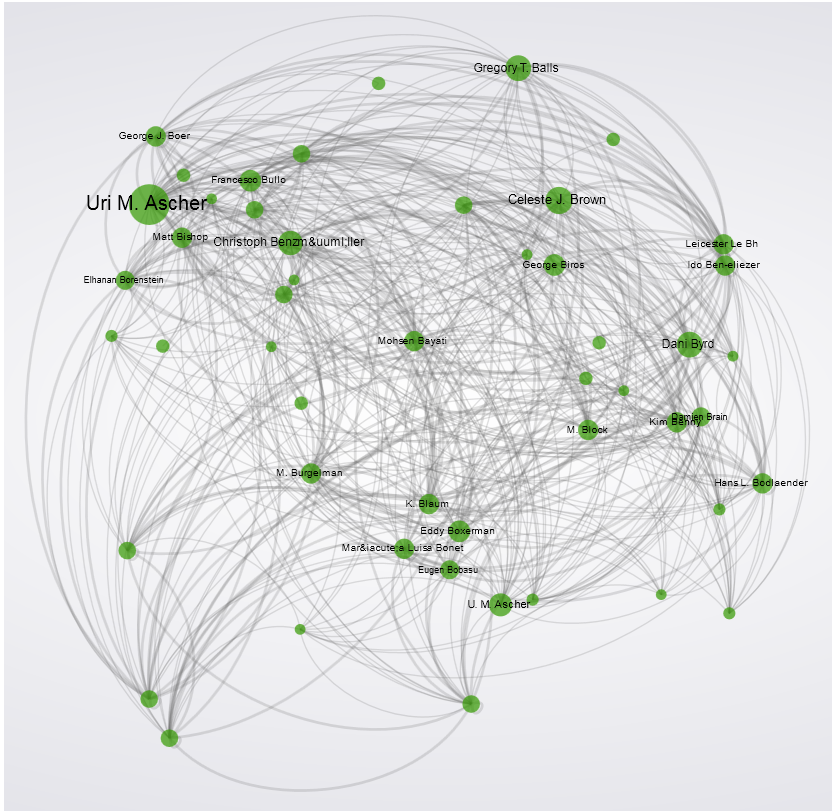
- Плотной
- Разреженной

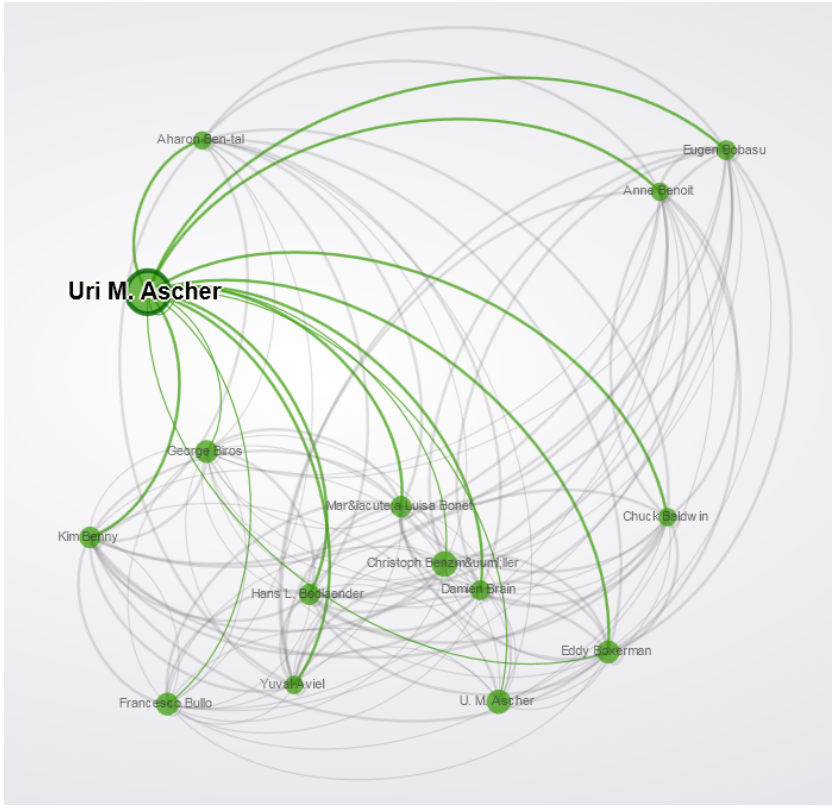
Выделение групп

WEKA - инструментарий для машинного обучения и data mining на языке Java.

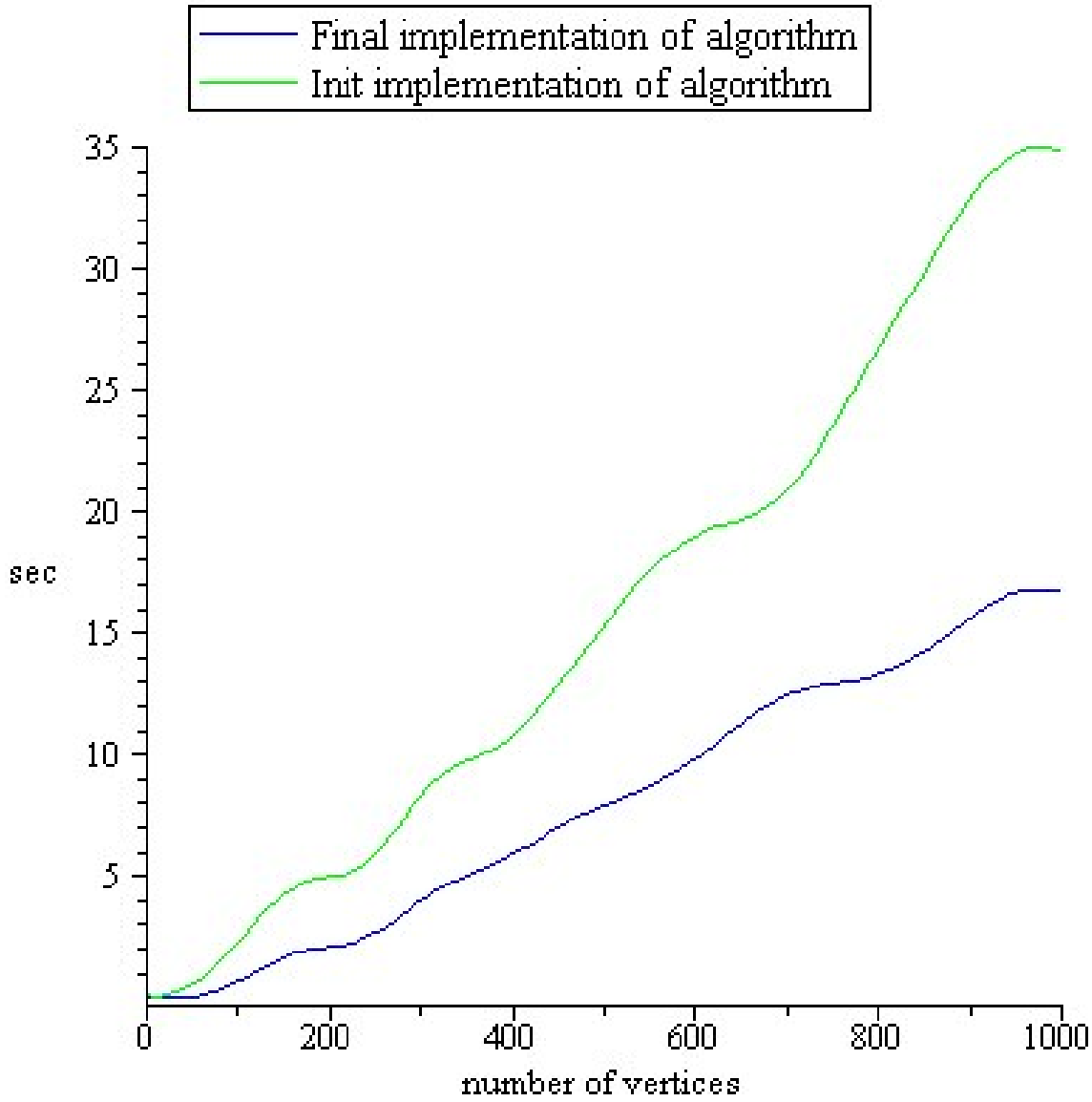
- Обучение с учителем
- Наивный байесовский классификатор
- Релевантность результата

Тестирование





Тестирование



Результаты

- Реализовано вычисление метрик, оптимизированное для больших графов
- С помощью средств машинного обучения разработан и реализован алгоритм поиска ученых