Розподілене обчислення генетичних алгоритмів за допомогою системи Hadoop. ("Distributed computing of genetic algorithms with HADOOP")

Автор: студент 6-го курсу, групи ДА-52м
ННК «ІПСА» НТУУ «КПІ»
Качко Микита Андрійович
Науковий керівник:
Харченко Костянтин Васильович

Ціль

- ✓ Дослідити наявні способи прогнозу поведінки деякої величини маючи попередні значення цієї величини.
- ✓ Запропонувати метод використання генетичних алгоритмів для отримання прогнозу поведінки вищезгаданої величини
- ✓ Дослідити та порівняти різні способи арпоксимації вихідних даних
- ✓ Дослідити пришвидшення обчислення об'ємних задач за допомогою HADOOP.

Актуальність

- ✓ Наявність більш швидкого та точного методу прогнозу поведінки деякої величини буде актуальним для абсолютно всіх галузей де таке знання буде приносити вигоду.
- ✓ Для більшості країн світу, наприклад, знання подальшої ціни на нафту може мати ключове значення у формуванні характеру зовнішньої політики
- ✓ Розподіл обчислень на декілька комп'ютерів зменшить час та зусилля, що необхідні для цього завдання.

Чому було обрано розподілення обчислень?

У протиставлення традиційній прямолінійній схемі викладення та обчислення задач мною були виявлені наступні особливості:

- ✓ В порівнянні з паралелізацією обчислень є достатньо гнучкими та в певній мірі вузло-незалежними.
- ✓ Нема необхідності розташовувати всі вузли в одному місті (без прив’язки до простору).
- ✓ Дозволяють працювати із величезними об’ємами даних.
- ✓ Додавання нових вузлів до системи не є складною задачею.

Які способи апроксимації досліджуються

Прогнозування ціни на нафту генетичним алгоритмом де хромосоми виступають у вигляді

- ✓ Коефіцієнтів багатостепеневого поліному
- ✓ Коефіцієнтів ряду Фур’є для косинусів

Як буде проводитися передбачення?

Мною було вирішено використовувати апроксимацію методом екстраполяції

Через фрактальний характер поведінки ціни, найбільш очевидним рішенням буде використовувати наближення ряду Фур'є для косинусів:

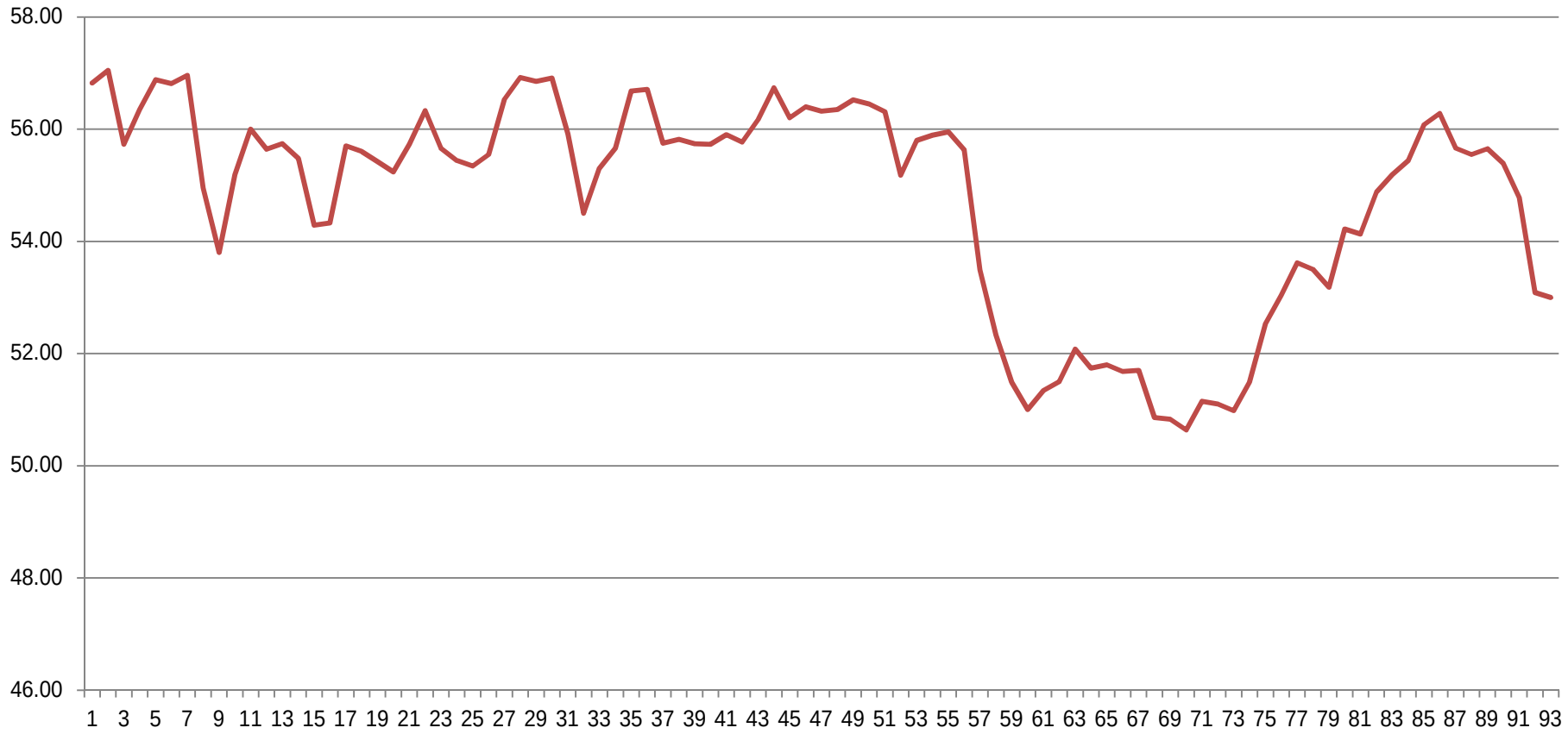
$$F(x_0, x_1, \ldots, x_n, i) = x_0 + x_1\cos(2\pi i) + \ldots + x_n\cos(2\pi n i).$$

Прогнозування даних буде задачею із лімітованим часом, тому було вирішено дослідити застосованність іншої апроксимації – ряду Тейлора

$$F(x_0, x_1, \ldots, x_n, i) = x_0 + x_1 i + \ldots + x_n\, i^n$$

Результати

Мною було вирішено використовувати в якості прикладу дані ціни за барель нафти за період 17 січня – 17 квітня 2017 року. Графік ціни виглядає наступним чином:



Результати

Нижче наведені таблиці порівнянь точності прогнозу прямої та розподіленої концепції для поліноміальної та тригонометричної апроксимації

Кількість поколінь	Прямолінійний спосіб	Розподілення обчислень
10	1.6831517159635392E37	4.21E+36
100	1.0971641732396855E35	2.74E+34
1 000	9.36782384466527E34	2.34E+34
10 000	2.109876515713443E33	5.27E+32
100 000	4.04468986317093E27	1.01E+27
1 000 000	1.1053724600213966E23	2.76E+22

Кількість поколінь	Прямолінійний спосіб	Розподілення обчислень
10	261723.8631681792	65430.96579
100	151035.88349620058	37758.97087
1 000	22420.05158651477	5605.012897
10 000	18645.83534393031	4661.458836
100 000	12221.89033896767	3055.472585
1 000 000	5913.1856239457	1478.296406

Результати

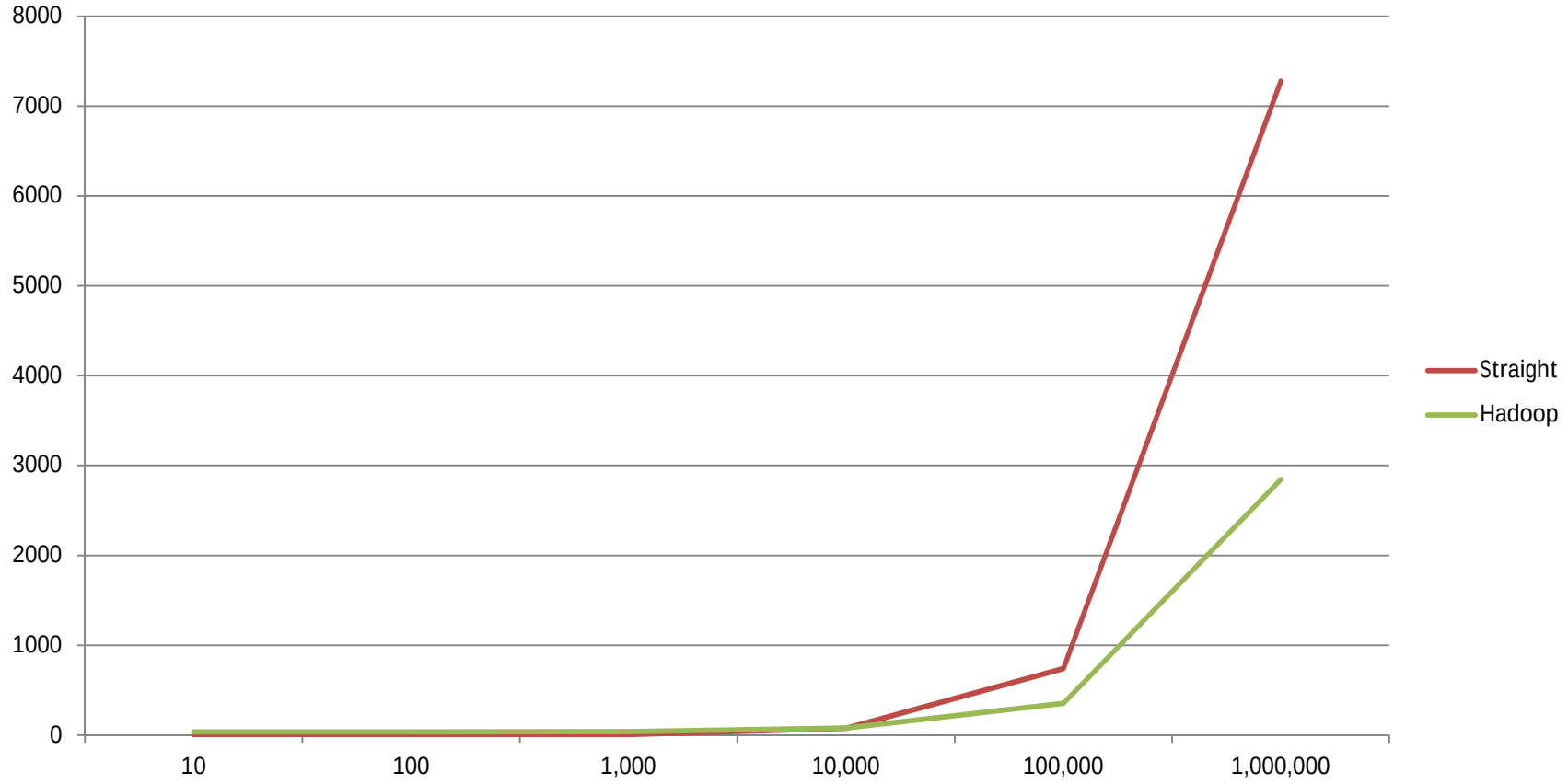
Нижче наведені таблиці порівнянь швидкості прогнозу прямої та розподіленої концепції для поліноміальної та тригонометричної апроксимації

Кількість поколінь	Прямолінійний спосіб, с	Розподілення обчислень, с
10	0.193	35.259
100	0.88	35.831
1 000	7.496	37.742
10 000	74.744	77.92
100 000	740.457	354.128
1 000 000	7276.355	2845.199

Кількість поколінь	Прямолінійний спосіб	Розподілення обчислень
10	0.144	34.529
100	0.597	34.8
1 000	5.031	36.742
10 000	48.785	50.92
100 000	481.503	164.821
1 000 000	4683.145	1754.317

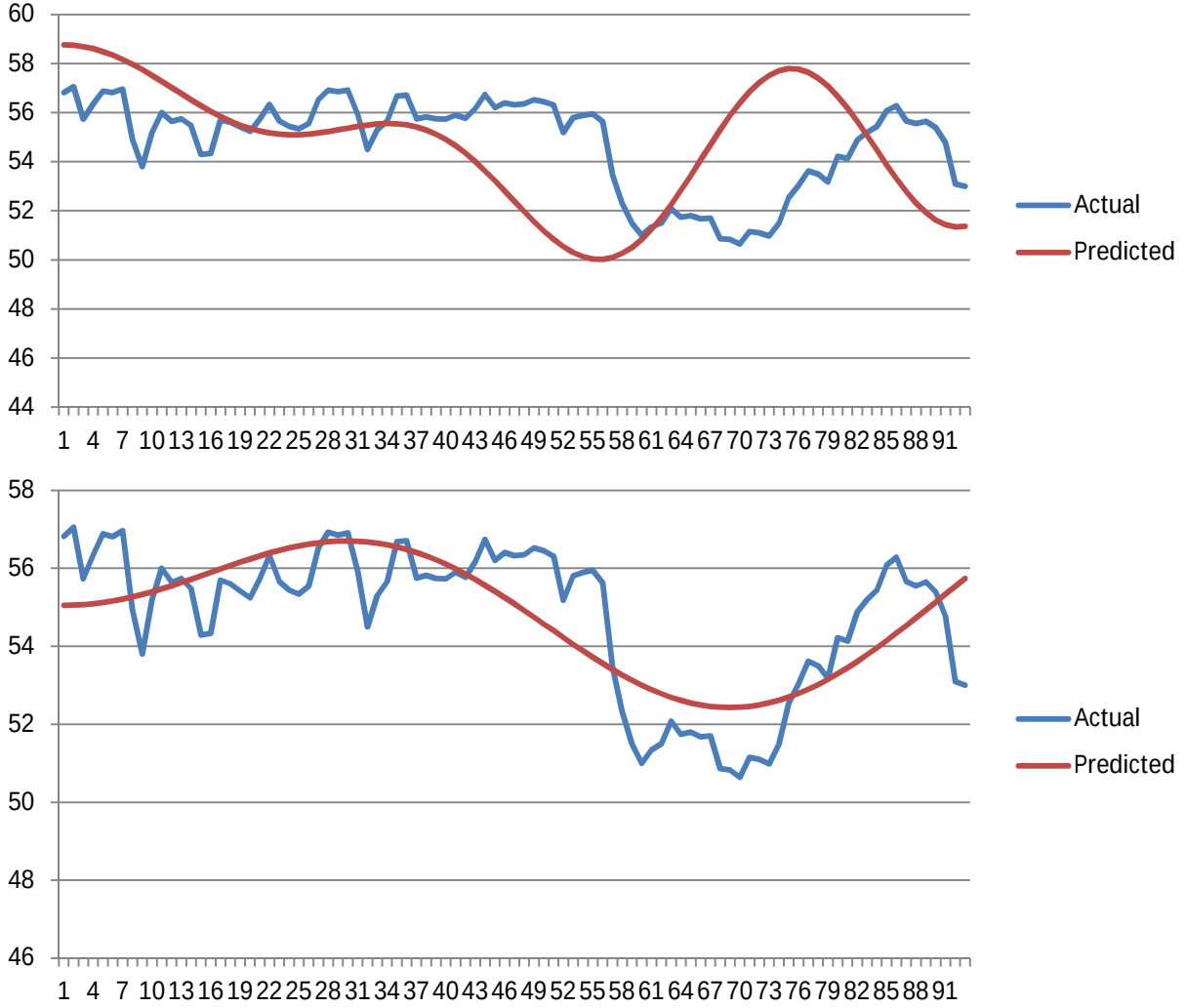
Результати

Нижче наведений графік порівнянь швидкості прогнозу прямої та розподіленої концепції для тригонометричної апроксимації



Результати

Нижче наведені графіки порівнянь точності прогнозу прямої та розподіленої концепції для тригонометричної апроксимації



Результати

- ✓ Було розглянуто два способи вирішення обчислювально складної задачі на прикладі яких можна виявити переваги та недоліки систем розподілених обчислень
- ✓ Було отриманого програмний продукт, за допомогою якого можна дослідити роботу системи розподілених обчислень HADOOP
- ✓ Отримані результати для подальших досліджень

В чому полягає відмінність підходів?

Мною були досліджені дві концепції в проведенні алгоритмічних обчислень – парадигма одного потоку та розподілення обчислень.

- ✓ Кожен вузол обічислювального кластеру виступає у якості самостійної одиниці, де незалежно від інших відбувається свій процес типового розвитку поколінь у генетичному алгоритмі
- ✓ Використання Hadoop для розподілення обчислень вводить певний рівень абстракції, що дозволяє відмежитися від типових проблем одно- та багатопоточної парадигми програмування, таких як розмеження адресного простору, синхронізація потоків, розділення навантаження, тощо.

Інноваційність запропонованого підходу

Для досягнення поставленої мети було розроблено новий інноваційний алгоритм що має наступні особливості:

- ✓ Мутація та генерація хромосом нових нащадків відбувається у вікні що звужується до значення хромосоми кращого нащадка
- ✓ У випадку використання розподілених систем кожен вузол самостійно виконує операції генетичних алгоритмів (пре-селекція), а потім кореневий вузол обирає кращий з нащадків
- ✓ Навіть кращий нащадок має змогу адаптуватися до нових даних шляхом корегування свого прогнозу на такий коефіцієнт що перетворює прогнозоване значення на поточне.

Висновки

- ✓ На прикладі прогнозу ціни за барель нафти, як і очікувалося, система добре себе проявила на великому об’ємі даних
- ✓ Апроксимація за допомогою екстраполяції степеневого поліному виявилася абсолютно не пристосованою для типової практичної ситуації на ринку нафти.
- ✓ Була перевірена і доведена фрактальна поведінка графіку ціноутворення на нафтовий продукт
- ✓ В подальшому запропонований підхід може бути використаний для передбачування таких непрогнозованих явищ як обвал курсу валюти або навіть економічна, похождення яких вважається випадковим.