Big data analysis
– a case study

Du, Haoran U5044502
Supervisor: Dr. Wang, Qing
COMP 8740: Software engineering project, Semester 1, 2014
Australian National University
Overview

• Introduction
• Methodology
• Experiments
• Conclusion
Piction company

- The Piction Digital Media eXchange (DMX) platform
- Australian War Memorial, Australian National Botanical Garden

**ABOUT PITION**
- Single integrated platform that addresses the requirements for digital resource control, management, commercialisation and distribution. From simple through to complex, Piction's platform can scale to your business requirements.
- Manage digital resources such as still images, PDFs, Word files, documents, audio, video, spatial and blobs.
- Rapid implementation and cost-effective solution.

**OUR SOLUTIONS**

- **Museums**
  Digital collection management & streamlined workflows.

- **Corporate**
  Brand Asset Management platform for managing brand resources.

- **Government**
  Digital Asset Management platform for managing digital resources such as images, video and audio.

- **Image Library**
  Image Library grade search and retrieval for complex unstructured data.
Customer requirements

A. Piction Server

B. Extraction Process

C. Extracted XML Files

D. Ingestion Process

E. Big Data Processing Tool
Objective:

Develop a recommender system tool for recommending products to customers effectively and efficiently
Data set

- 100360 distinct customers (big volume)
  72 different products (high dimension)
  2578273 distinct purchase

- The data is a mixture of real application data and testing data.
Motivation

• Avoid directly similarity computation
• Avoid exhaustive comparison
• Reduce dimensionality
• Maintain quality
Methodology

Original matrix:
- 237903
- 193841
- 733192
- ...
- 837493

Locality Sensitive Hashing

Fast Hamming Search

Candidate customers: those we should test similarity

Signature matrix:
- 101001
- 011010
- 010100
LSH for cosine distance

- Get a vector randomly
- Calculate the dot product of this random vector and the given vector
- If negative, 0 otherwise 1 — a signature
- If get k random vectors, a k-dimensional signature vector
Fast Hamming search

1. Permutated signature matrix
2. Sorting permutated signature vector lexicographically
3. Selecting neighbors:
Compared method

• The original naïve method
  – Calculate the cosine similarity between query customer and every other customer
  – Get weighted purchase time
  – Get the total purchase time for products
  – Sort all products by total purchase time
  – Output the top 3 products which the query customer not buy
Measurements

• Efficiency:
  – Time cost: the time taken to get the recommend results

• Result quality:
  – Mean: the mean of the real ranking number of the recommended product in naïve method
  – Standard variation: the standard variation of the real ranking number of the recommended products in naïve method
Quality
Simple Tool

![Commercial Recommendation Tool](image)
Conclusion

• LSH is an effective and efficient method. However,

• Integrate into Piction’s platform to test real-world data

• Test other algorithms such as SVD

• Improve the simple tool offering more functions
Q&A

u5044502@anu.edu.au