CSIT5210

Data Mining and Knowledge Discovery

Overview

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse

Teaching Mode

- In this semester,
  - We will teach this course in the physical classroom.
  - In the add/drop period (1-14 Sept), we will have the mixed-mode teaching (i.e., both online and physical).
  - If you want to attend the online class from 15 Sept to 30 Sept, you need to contact the MSc IT program office for approval.

Teaching Mode

- After this period, depending on the official class list of this course, we will switch to the face-to-face teaching mode only (without any online component).

Teaching Mode

- Since the face-to-face mode is our focus in this semester, I will focus on face-to-face students.
- As suggested by the university, there is only limited interaction (or even no interaction) with online students.

Teaching Mode

- If you are students “currently” outside Hong Kong, please send an email to me with the email subject “CSIT5210: About Online Student Arrangement” which includes the following.
  - Where are you now?
  - Do you plan to come back to Hong Kong?
    If yes, please give the tentative date that you will come to HK and the tentative date that you will go to the UST campus (just after the quarantine arrangement)

Course Details

- Reference books/materials:
  - Papers

Course Details

- Data Mining: Concepts and Techniques. Jiawei Han and Micheline Kamber. Morgan Kaufmann Publishers (3rd edition)
- Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, Vipin Kumar Boston : Pearson Addison Wesley (2006)

Course Details

- Grading Scheme:
  - Assignment 30%
  - Project 30%
  - Final Exam 40%

Assignment

- If the students can answer the selected questions in class correctly,
  - for each corrected answer,
    I will give him/her a coupon
  - This coupon can be used to waive one question in an assignment
  - which means that s/he can get full marks for this question without answering this question

Assignment

- Guideline
  - For each assignment, each student can waive at most one question only.
    - s/he can waive any question he wants and obtain full marks for this question (no matter whether s/he answer this question or not)
    - s/he may also answer this question. But, we will also mark it but will give full marks to this question.
  - When the student submits the assignment,
    - please staple the coupon to the submitted assignment
    - please write down the question no. s/he wants to waive on the coupon

Project

- Each project is completed by a group.
- The number of students in a group is 5-6.
- The duration of each presentation is at most 20 minutes (not including the Q&A session)

Project

- Project Type (One of the following)
  - Survey

    Your group only needs to read about 2~5 papers
  - Implementation-oriented Project

    Your group only needs to read about 1~2 papers
  - Research-oriented Project

    You can read some papers and conduct research

Project

- Project Type (One of the following)
  - Survey

    1. Proposal
    2. Presentation
    3. Final report

    Full Score = 80%

  - Implementation-oriented Project

    Full Score = 90%

    1. Proposal
    2. Presentation
    3. Final report
    4. Coding

  - Research-oriented Project

    1. Proposal
    2. Presentation
    3. Final report (containing your proposed methodology)
    4. Coding (if any)

    Full Score = 100%

Project

- Project Topic
  - Some pre-selected topics/papers
  - Your own choice
- For fairness, please do not choose the topic which is closely related to your own research

Project

- More details could be found in the Canvas webpage.

Exam

- This is an on-site exam.
- You are allowed to bring a calculator with you. The list of permitted calculators is consistent with the HK Exam Authority as follows. https://www.hkeaa.edu.hk/DocLibrary/IPE/cal/CAL2019.pdf
- Please remember to prepare a calculator for the exam

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

1. Association

Customer	Apple	Orange	Milk
Raymond	Apple	Orange	
Ada		Orange	Milk
Grace	Apple	Orange	
…	…	…	…

We are interested in the items/itemsets with frequency >= 2

Items/Itemsets	Frequency
Apple	2
Orange	3
Milk	1
{Apple, Orange}	2
{Orange, Milk}	1

Frequent Pattern
(or Frequent Item)

Frequent Pattern
(or Frequent Item)

Frequent Pattern
(or Frequent Itemset)

1. Association

Customer	Apple	Orange	Milk
Raymond	Apple	Orange	
Ada		Orange	Milk
Grace	Apple	Orange	
…	…	…	

We are interested in the items/itemsets with frequency >= 2

Items/Itemsets	Frequency
Apple	2
Orange	3
Milk	1
{Apple, Orange}	2

Association Rule:

1. Apple → Orange
(100% customers who buy apple will probably buy orange.)

2. Orange → Apple
(67% customer who buy orange will probably buy apple.)

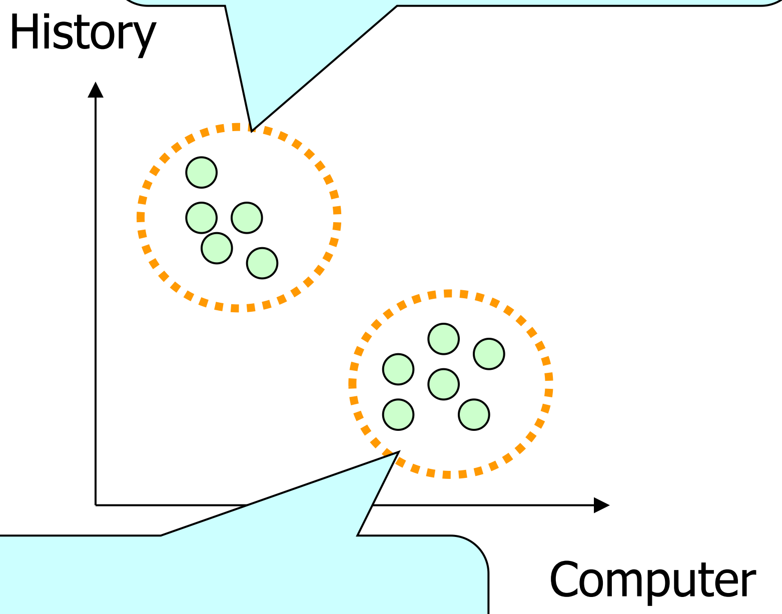
Problem: to find all frequent patterns and association rules

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

2. Clustering

	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
…	…	…



Cluster 2
(e.g. High Score in History and Low Score in Computer)

Cluster 1
(e.g. High Score in Computer and Low Score in History)

Problem: to find all clusters

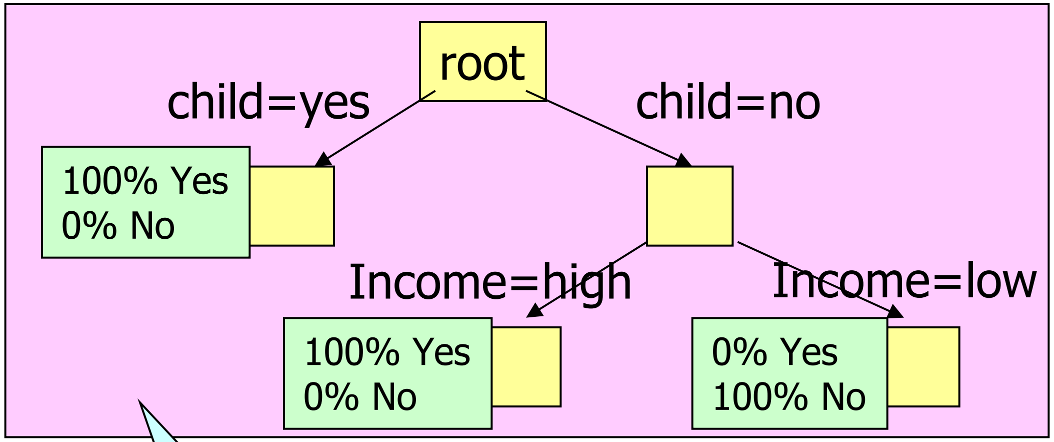
Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

3. Classification

Suppose there is a person.

Race	Income	Child	Insurance
white	high	no	?

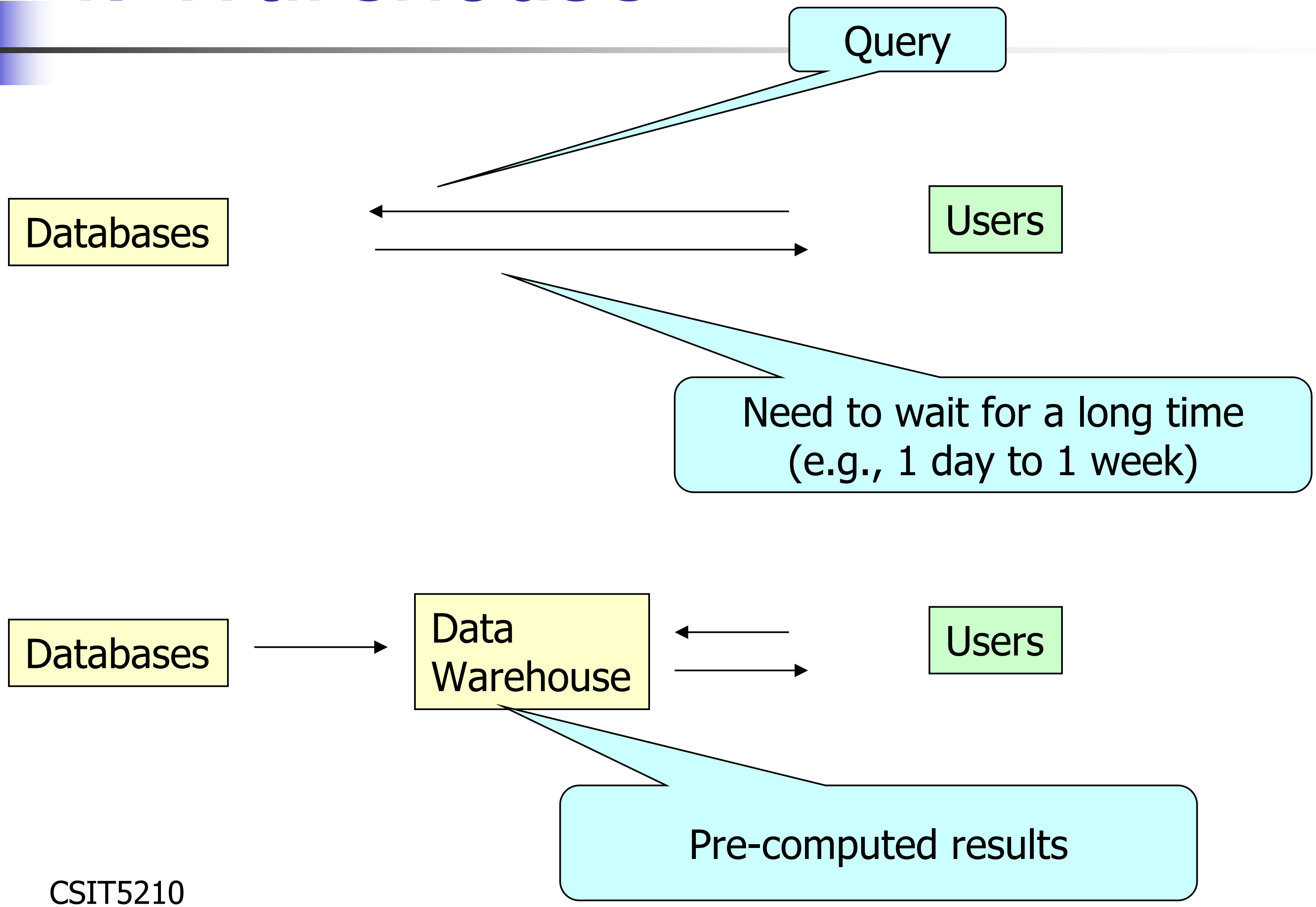


Decision tree

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

4. Warehouse

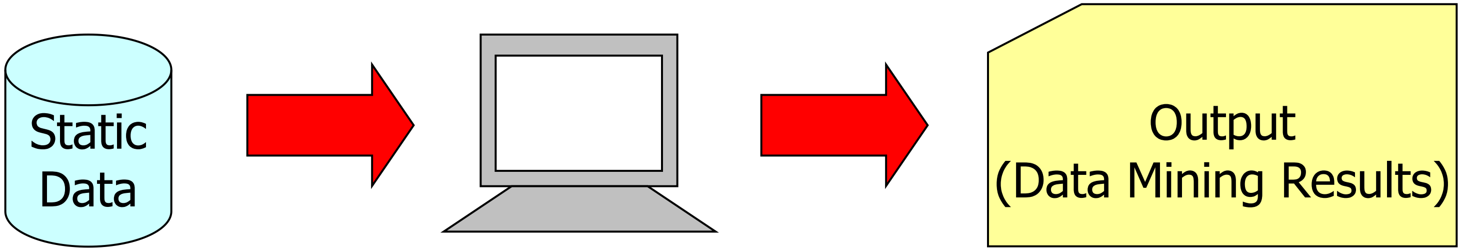

Query
Databases
Users
Need to wait for a long time (e.g., 1 day to 1 week)
Databases
Data Warehouse
Users
Pre-computed results

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

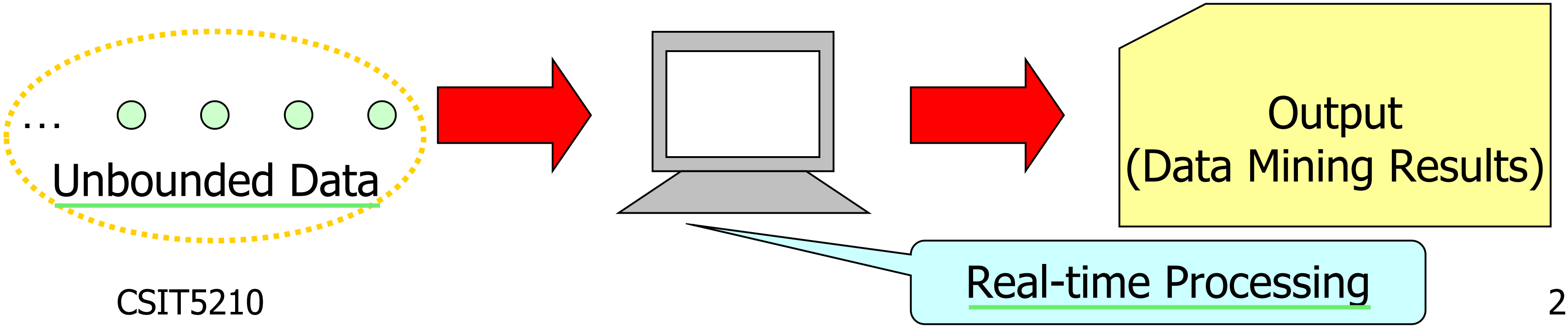
5. Data Mining over Static Data

1. Association
2. Clustering
3. Classification



5. Data Mining over **Data Streams**

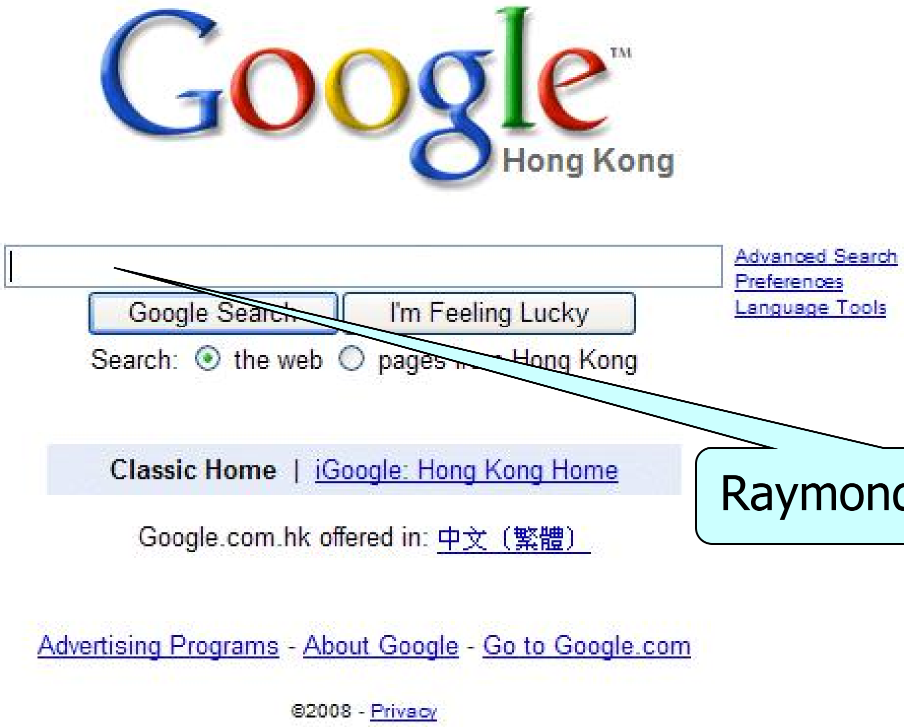
1. Association
2. Clustering
3. Classification



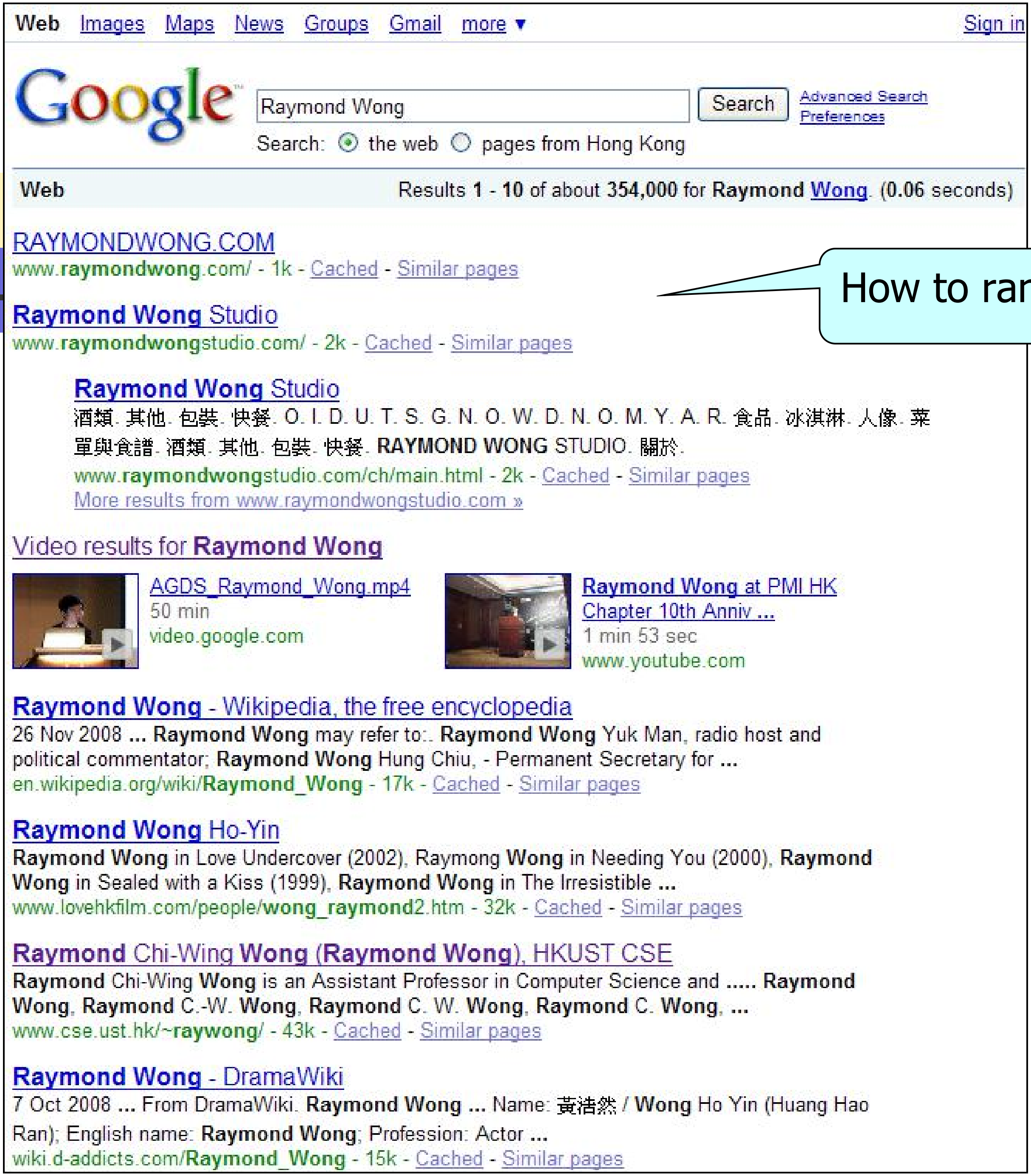
Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Data Mining over Data Streams
6. Web Databases

6. Web Databases


Google™
Hong Kong
Advanced Search
Preferences
Language Tools
Google Search
I'm Feeling Lucky
Search: ◉ the web ○ pages from Hong Kong
Classic Home | iGoogle: Hong Kong Home
Google.com.hk offered in: 中文（繁體）
Advertising Programs - About Google - Go to Google.com
©2008 - Privacy

Raymond Wong



Raymond Wong [Search] Advanced Search Preferences

Search: ● the web ○ pages from Hong Kong

Web Results **1** - **10** of about **354,000** for **Raymond Wong**. (**0.06** seconds)

RAYMONDWONG.COM
www.**raymondwong**.com/ - 1k - Cached - Similar pages

Raymond Wong Studio
www.**raymondwong**studio.com/ - 2k - Cached - Similar pages

Raymond Wong Studio
酒類. 其他. 包裝. 快餐. O. I. D. U. T. S. G. N. O. W. D. N. O. M. Y. A. R. 食品. 冰淇淋. 人像. 菜單與食譜. 酒類. 其他. 包裝. 快餐. **RAYMOND WONG** STUDIO. 關於.
www.**raymondwong**studio.com/ch/main.html - 2k - Cached - Similar pages
More results from www.raymondwongstudio.com »

Video results for **Raymond Wong**

AGDS_Raymond_Wong.mp4
50 min
video.google.com

Raymond Wong at PMI HK Chapter 10th Anniv ...
1 min 53 sec
www.youtube.com

Raymond Wong - Wikipedia, the free encyclopedia
26 Nov 2008 ... **Raymond Wong** may refer to:. **Raymond Wong** Yuk Man, radio host and political commentator; **Raymond Wong** Hung Chiu, - Permanent Secretary for ...
en.wikipedia.org/wiki/**Raymond_Wong** - 17k - Cached - Similar pages

Raymond Wong Ho-Yin
Raymond Wong in Love Undercover (2002), Raymong **Wong** in Needing You (2000), **Raymond Wong** in Sealed with a Kiss (1999), **Raymond Wong** in The Irresistible ...
www.lovehkfilm.com/people/**wong_raymond**2.htm - 32k - Cached - Similar pages

Raymond Chi-Wing **Wong** (**Raymond Wong**), HKUST CSE
Raymond Chi-Wing **Wong** is an Assistant Professor in Computer Science and ..... **Raymond Wong**, **Raymond** C.-W. **Wong**, **Raymond** C. W. **Wong**, **Raymond** C. **Wong**, ...
www.cse.ust.hk/~**raywong**/ - 43k - Cached - Similar pages

Raymond Wong - DramaWiki
7 Oct 2008 ... From DramaWiki. **Raymond Wong** ... Name: 黃浩然 / **Wong** Ho Yin (Huang Hao Ran); English name: **Raymond Wong**; Profession: Actor ...
wiki.d-addicts.com/**Raymond_Wong** - 15k - Cached - Similar pages

How to rank the webpages?