Using Grid to Facilitate Diseasome Analysis from Taiwan National Health Insurance Research Database

Yu-Chuan (Jack) Li and Ming-Chin Lin, Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan
Outline

- Introduction of NHIRD
- Frequency Distribution of Diseasesome
- Comorbidity Analysis
- Conclusion
The National Health Insurance Research Database (NHIRD)

- 10 years of data
- Coverage: about 99% residents in Taiwan
  (23 million people from 530 hospitals and 17,000 clinics)
- 360 million outpatient visits / year
- 25 million inpatient-day / year
The NHIRD is opened for research by application.

The NHIRD consists of claim records with numbers and text.

Demographics, Diagnoses (ICD 9-CM 2001 version), Medications, Procedures, Exams and Costs data.

Raw data size: 200GB / year.
Frequency of Visits

- Analyze database by patient visits
  - Frequency data over time (X-axis) and Age (Y-axis)
  - Heatmap visualization

Dermatophytosis of foot
Frequency of Visits (cont.)

- Analyze database by patient visits
  - Bottleneck --> Disk I/O Speed
  - Using 12 Apple Mac mini with external Firewire Hard Drive (400 Mbps)
  - Collective bandwidth on I/O: 4.8 Gbps
Frequency of Visits (cont.)

WWW

Grid (Globus)

Result DB

Send grid commend

Jan Feb Mar Apr May June Jul Aug Sep Oct Nov Dec
## Frequency of Visits (cont.)

### Big Vs. mini

<table>
<thead>
<tr>
<th></th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big</td>
<td>Strong CPU</td>
<td>Expensive</td>
</tr>
<tr>
<td></td>
<td>Strong I/O speed</td>
<td>Hard to upgrade</td>
</tr>
<tr>
<td>mini</td>
<td>Cheap</td>
<td>Mild CPU</td>
</tr>
<tr>
<td></td>
<td>Low maintain fee</td>
<td>Low I/O speed</td>
</tr>
</tbody>
</table>
Frequency of Visits (cont.)

- Difficulty on doing job on single machine
  - Limitation of database size
    - Take very long time to generate index table
  - Limitation of scaling up
    - Hard to improve the performance
    - Performance vs Price curve --> not linear
Disease Frequency HeatMap (NHIRD 2000)
Taiwan NHIRD 2000-2002

Influenza

Erythema multiforme

Lung Cancer
Hepatitis B with coma

3-year seasonal change of “Cough”

male

female
Influenza
Hand foot and mouth disease
GIS distribution of “Cough”
Cough
Cough
Retrospective study - Comorbidity analysis

The limitation

- Grouping all visit records by unique ID
- Software memory limitation - 2GB memory

<table>
<thead>
<tr>
<th>Essential HYPERTENSION</th>
<th>Jan</th>
<th>Feb</th>
<th>Total transaction record number (2000-2002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>571,099</td>
<td>525,646</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>644,650</td>
<td>645,846</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>752,353</td>
<td>655,867</td>
<td>25,015,172</td>
</tr>
</tbody>
</table>
Disease Comorbidity analysis

For Comorbidity analysis
- ID1{dis1,dis2,dis3,dis4,...}

For example
- 192305,M,HS10710973,01340,2001-04-11,4919|4659|4019|3534|4011|38022|4640|3804|4785|3004|7291|78059|01340|460|4660
- 192505,F,KT71864585,01340,2002-07-10,01100|01340|29532|0113|0119
Solution-
Sorting and segmenting database for grid architecture
Our experience

- Divide NHIDB by month and year of Birthdates
- Divide NHIDB into 1,212 small databases
  - 12 months * 101 years (from 1900 to 2000) = 1,212 segments
- Easily scale up - Linear acceleration
- Low machine specification requirement
Comorbidity

- About 10 diagnoses per person in 3 years
- Clusters of comorbidity are being identified and pre-calculated
- 1TB of comorbidity data processed for 7 days under a 100-PC grid
Endometriosis

| Neoplasm of unspecified nature of other genitourinary organs | 55 | 7275 | 22158 | 11233 | 14.0 |
| Neoplasm of unspecified nature of other specified sites | 17 | 24635 | 22158 | 11233 | 1.0 |
| Neoplasm of unspecified nature of breast | 37 | 39736 | 22158 | 11233 | 1.0 |
| Neoplasm of uncertain behavior of uterus | 1 | 1123 | 22158 | 11233 | 1.0 |
| Neoplasm of uncertain behavior of other and unspecified female genital organs | 4 | 394 | 22158 | 11233 | 20.0 |
| Neoplasm of uncertain behavior of ovary | 82 | 4625 | 22158 | 11233 | 34.0 |
| Neoplasm of uncertain behavior of breast | 19 | 11003 | 22158 | 11233 | 3.0 |
| Neoplasm of unspecified nature of brain | 5 | 5670 | 22158 | 11233 | 1.0 |
| Neoplasm of unspecified nature site unspecified | 13 | 17132 | 22158 | 11233 | 1.0 |
| Neoplasm of unspecified nature of endocrine glands and other parts of n... | 2 | 2541 | 22158 | 11233 | 1.0 |
| Neoplasm of unspecified nature of bone soft tissue and skin | 12 | 39968 | 22158 | 11233 | 0.0 |
| Neoplasm of uncertain behavior of other specified sites | 30 | 1275 | 22158 | 11233 | 46.0 |
| Neoplasm of unspecified nature of digestive system | 6 | 31112 | 22158 | 11233 | 0.0 |
| Neoplasm of uncertain behavior of liver and biliary passage | 5 | 5386 | 22158 | 11233 | 1.0 |
| Neoplasm of uncertain behavior of neurofibromatosis | 1 | 1123 | 22158 | 11233 | 16.0 |
Conclusion

- Linear improvement of performance is achievable if the data are properly segmented.
- A heatmap for visualization of frequency distribution over season and patient age is useful for huge data sets.
- A geographical relationship of frequency distribution can also be visualized.
Conclusion (cont.)

- Comorbidity is one area that has great potential but very computation-intensive.
- Complete comorbidity data can be crossed with genome, haplome and bibliome data to achieve greater utility.
Thank you