KU-Thaigrid Research

By Dr. Arnon Rungsawang
Dept. Computer Engineering/KU

TGCC 2006
30 August 2006

MIKE™ approved

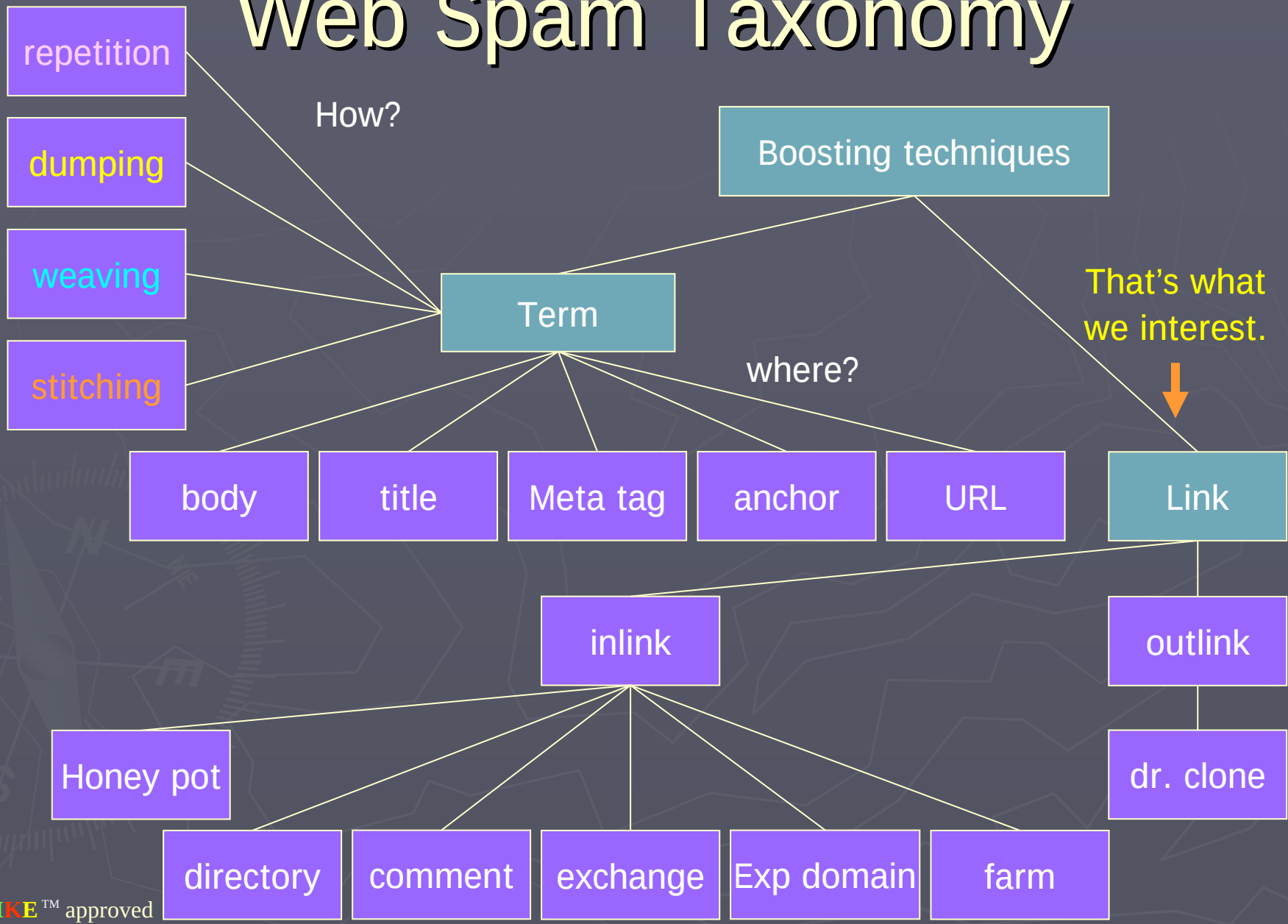
Research Topics

- MIKE Why Google!
  - Combating web spam
  - Age-based PageRank computation
  - Large-scale PageRank calculation
- DAKDL
  - DAKDL structural & functional proteomics on the grid
- Other project from HPCNC, Mechanical Dept., Chemical Dept. (Sci.),...

Combating web spam

- Web Spamming
  - Action intended to mislead search engine into ranking some pages higher than they deserve.
- Effect of Web spamming
  - Quality of search results decrease.
  - Search engine indexes are inflated with useless pages, increasing the cost of each processed query.

Web Spam Taxonomy



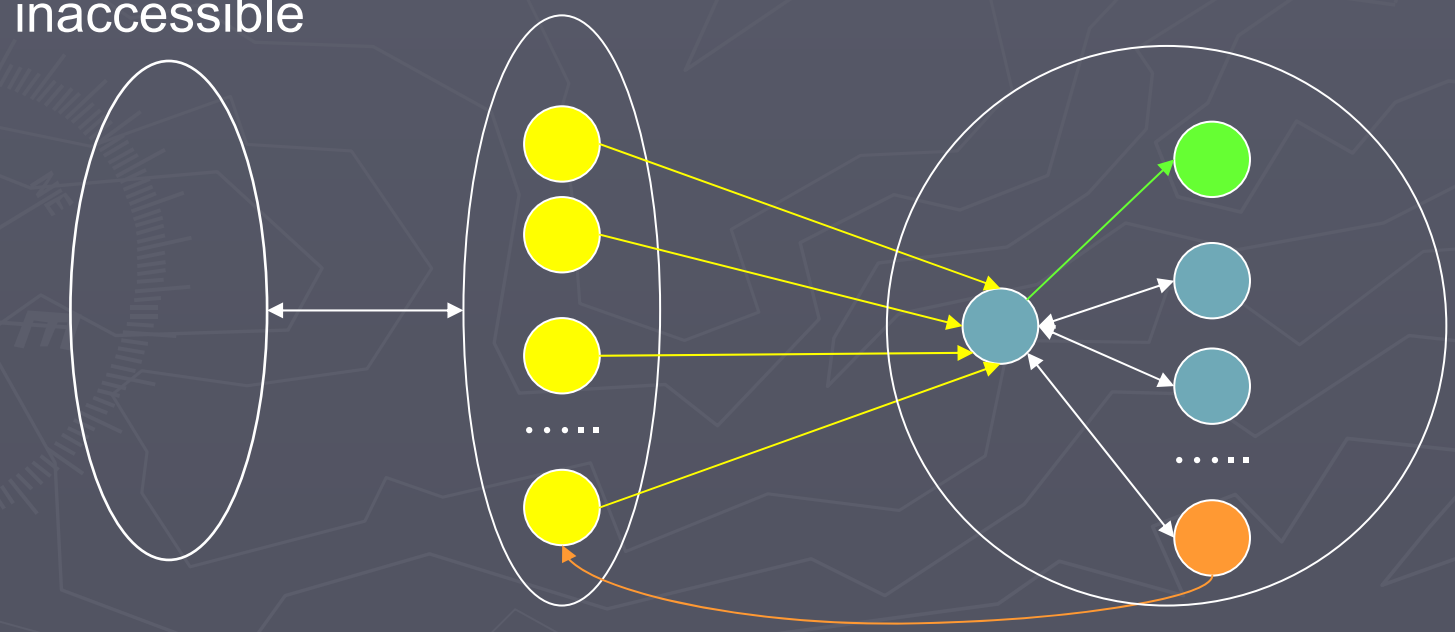
How to spam PageRank

- PageRanK
  - $PR(T) = PR_{static}(T) + PR_{in}(T) - PR_{out}(T) - PR_{sink}(T)$

accessible and partially modifiable

Owner (set of link-farm T)

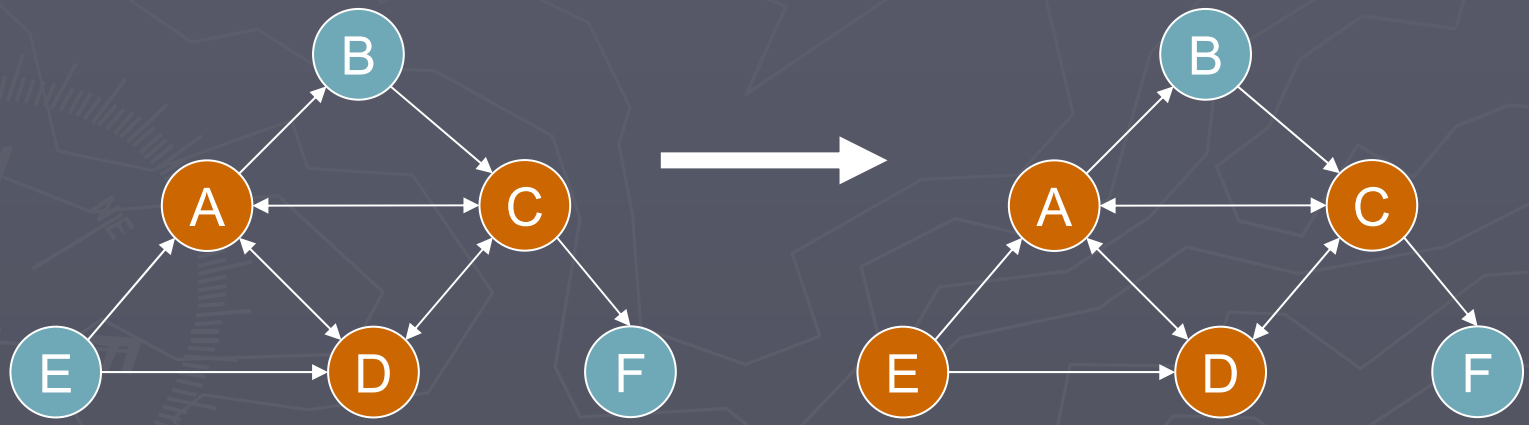
inaccessible



Eg., blog, wiki

Identify Boost Farm

- Set Threshold $T_{io} = T_{pp} = 2$

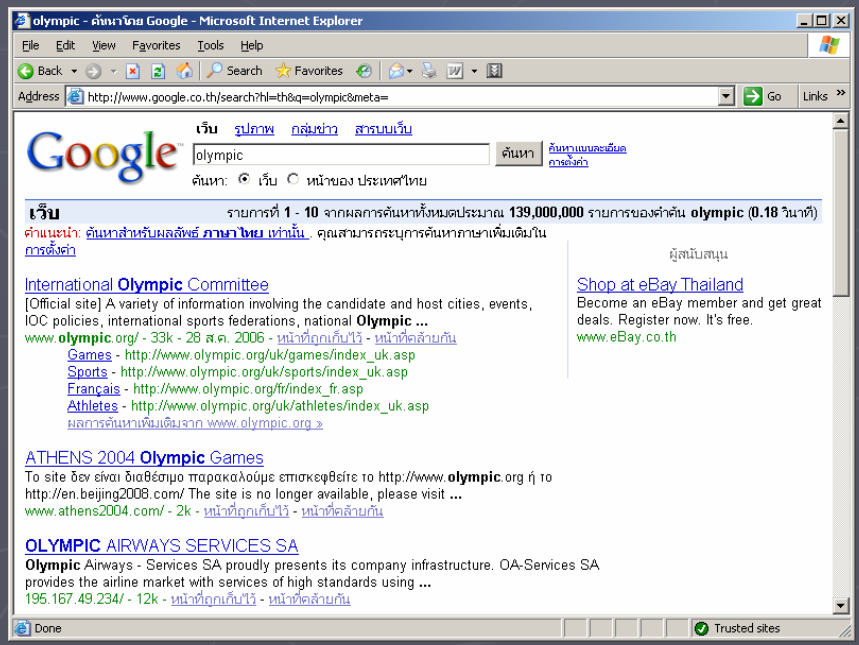


How to un-bias Boost Farm

- Identify boost farm from web graph
- Generate Virtual-link for all boost farm
- Compute Average Change rate of probability of Boost farm (ACB)
- Re-weight Boost Farm link and Virtual-link with computed ACB value
- Compute PageRank with new transition matrix (M)

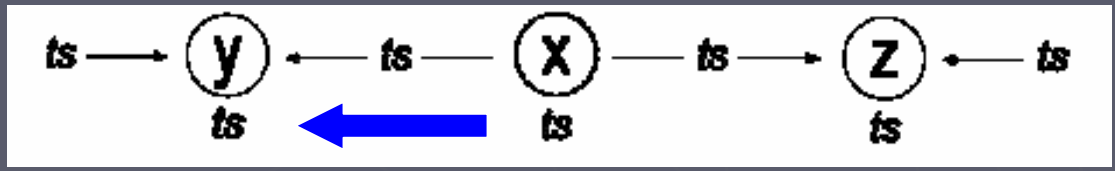
Age-based PageRank Computation

- Rapid change of the Web
- Old Pages receive over PageRank scores.
- New pages receive very few PageRank scores.



Methodology - Aging Factor (AF)

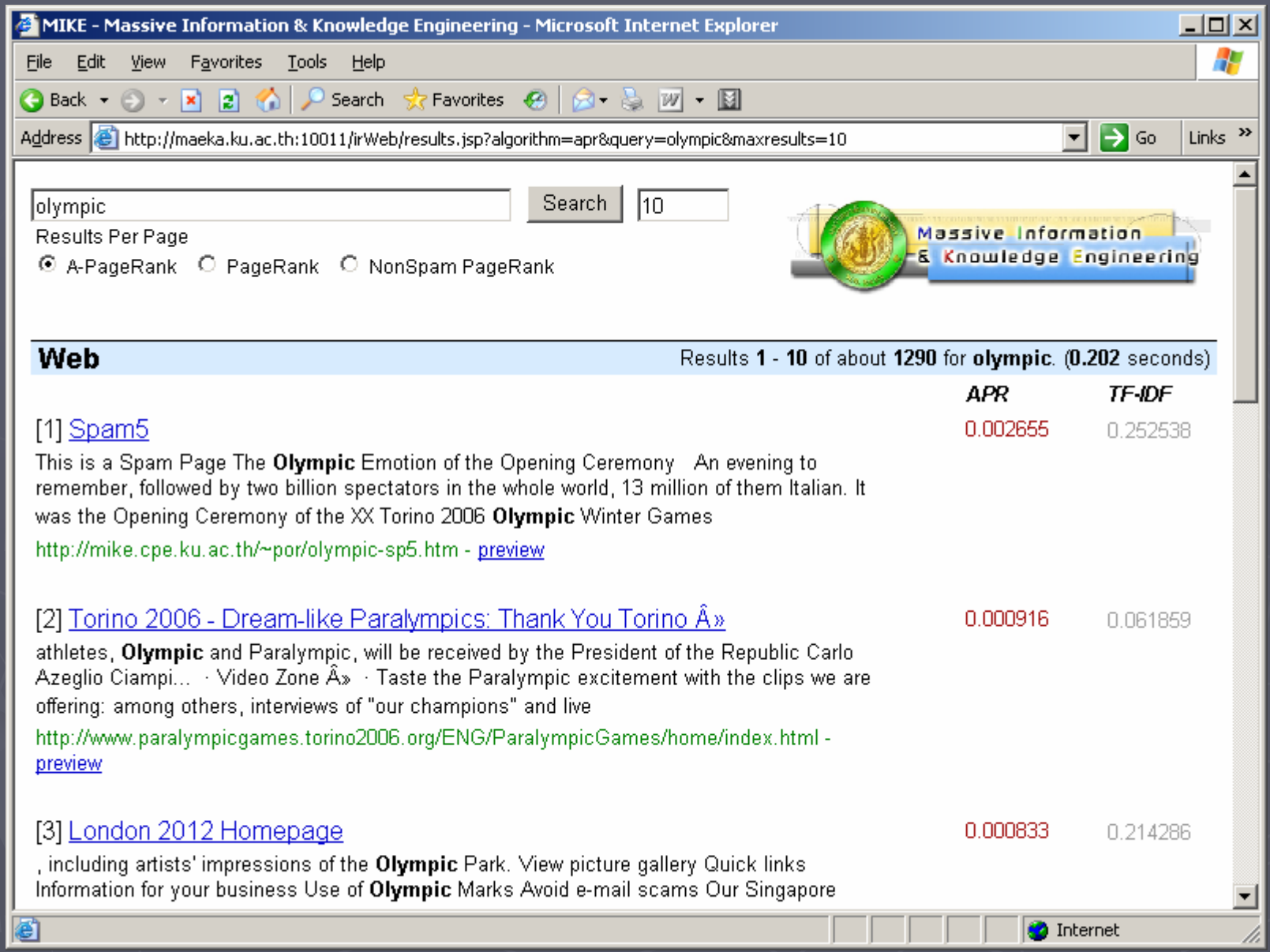
- At web page X



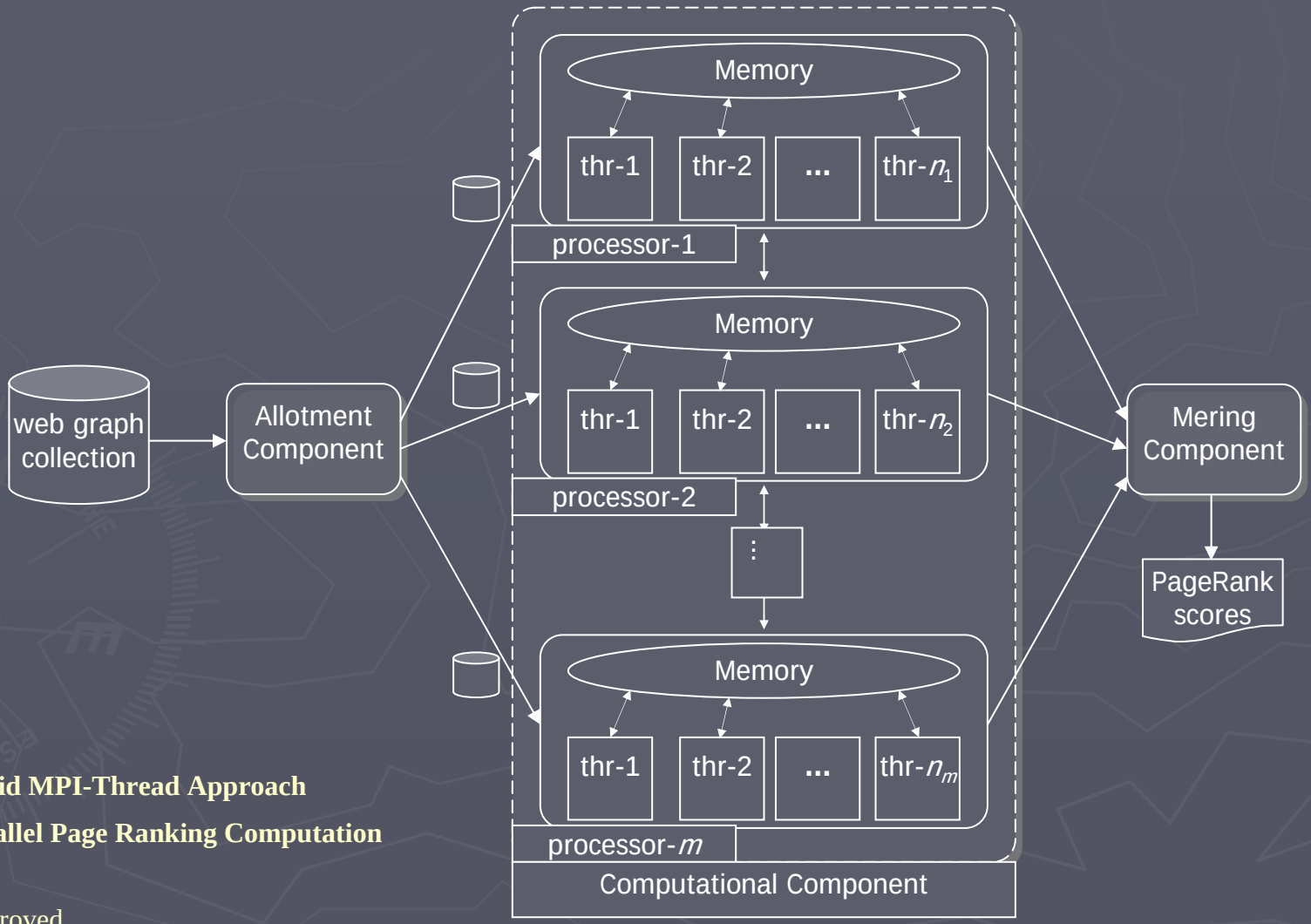
$$AF(x) = \begin{cases} 1 & Max\{ts_i \mid i \in \upsilon\} - ts_x < 3\,\text{months} \\ \dfrac{ts_x - Min\{ts_i \mid i \in \upsilon\} + e}{Max\{ts_i \mid i \in \upsilon\} - Min\{ts_i \mid i \in \upsilon\} + e} & Otherwise \end{cases} \quad (1)$$

- At link between webpage X to Y

$$AF(x,y) = \begin{cases} 1 & Max\{ts_{i,j} \mid (i,j) \in \varepsilon\} - ts_{x,y} < 3\,\text{months} \\ \dfrac{ts_{x,y} - Min\{ts_{i,j} \mid (i,j) \in \varepsilon\} + e}{Max\{ts_{i,j} \mid (i,j) \in \varepsilon\} - Min\{ts_{i,j} \mid (i,j) \in \varepsilon\} + e} & Otherwise \end{cases} \quad (2)$$


MIKE - Massive Information & Knowledge Engineering - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://maeka.ku.ac.th:10011/irWeb/results.jsp?algorithm=apr&query=olympic&maxresults=10
Go Links
olympic
Search
10
Results Per Page
A-PageRank PageRank NonSpam PageRank
Massive Information & Knowledge Engineering
Web
Results 1 - 10 of about 1290 for olympic. (0.202 seconds)
APR
TF-IDF
[1] Spam5
0.002655
0.252538
This is a Spam Page The Olympic Emotion of the Opening Ceremony An evening to remember, followed by two billion spectators in the whole world, 13 million of them Italian. It was the Opening Ceremony of the XX Torino 2006 Olympic Winter Games
http://mike.cpe.ku.ac.th/~por/olympic-sp5.htm - preview
[2] Torino 2006 - Dream-like Paralympics: Thank You Torino Â»
0.000916
0.061859
athletes, Olympic and Paralympic, will be received by the President of the Republic Carlo Azeglio Ciampi... · Video Zone Â» · Taste the Paralympic excitement with the clips we are offering: among others, interviews of "our champions" and live
http://www.paralympicgames.torino2006.org/ENG/ParalympicGames/home/index.html - preview
[3] London 2012 Homepage
0.000833
0.214286
, including artists' impressions of the Olympic Park. View picture gallery Quick links Information for your business Use of Olympic Marks Avoid e-mail scams Our Singapore
Internet

Large-scale PageRank computation



A Hybrid MPI-Thread Approach

for Parallel Page Ranking Computation

Some technical investigations

- We need fully control all our computing resource in local grid.
  - CPU/Process
  - Memory
  - Disk space
- We need high network bandwidth.
- And, we need a lot of disk space.

DAKDL structural & functional proteomics on the grid

Data Analysis and Knowledge Discovery Lab.
Room 803 Building15 Department of Computer Engineering Faculty of Engineering
Kasetsart University Phaholyothin Rd.,Chatuchak Bangkok 10900, Thailand
(662)942-8555 Ext. 1444 Fax (662)575-6245
E-mail : fengknw@ku.ac.th

เทคโนโลยีกริด & ชีวสารสนเทศด้านโปรตีน

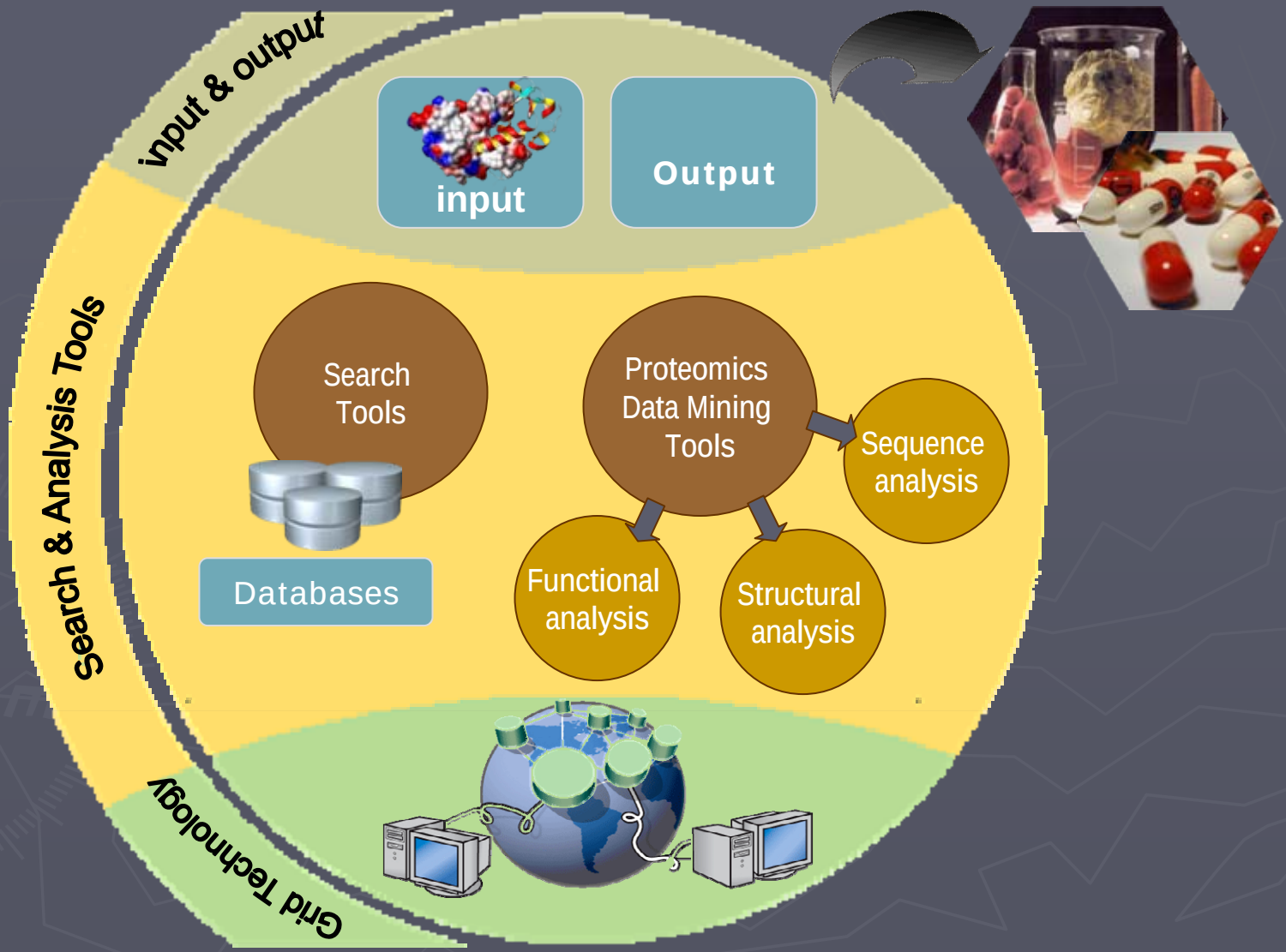
เทคโนโลยีชีวสารสนเทศด้านโปรตีน

- มีข้อมูลเพิ่มขึ้นอย่างรวดเร็ว
- วิธีการในการวิเคราะห์ในรูปแบบใหม่

เทคโนโลยีกริด (Grid)

- ประมวลผลได้รวดเร็ว

สถาปัตยกรรมของระบบ



KU-Thaigrid Research

TGCC 2006
30 August 2006

Thank you for your attention...