

Grid Enabled Data/Text Mining for Systems Biology Knowledge Base Development



Eric Bremer, PhD
Pediatric Brain Tumor Research Program
Children's Memorial Hospital
Chicago, Illinois
egbremer@northwestern.edu



Text Mining

- Aid in the understanding of genomic and gene expression data with the goal of therapeutic intervention.
- **Initial Project** –
 - Articles from 5 years of 20 journals
 - About 150,000 articles
- **Problem** - More than 24 hours to process about 5,000 articles on a single desk top computer
- **Time Sensitive** – only as good as the last update



Initiation of the Grid Project

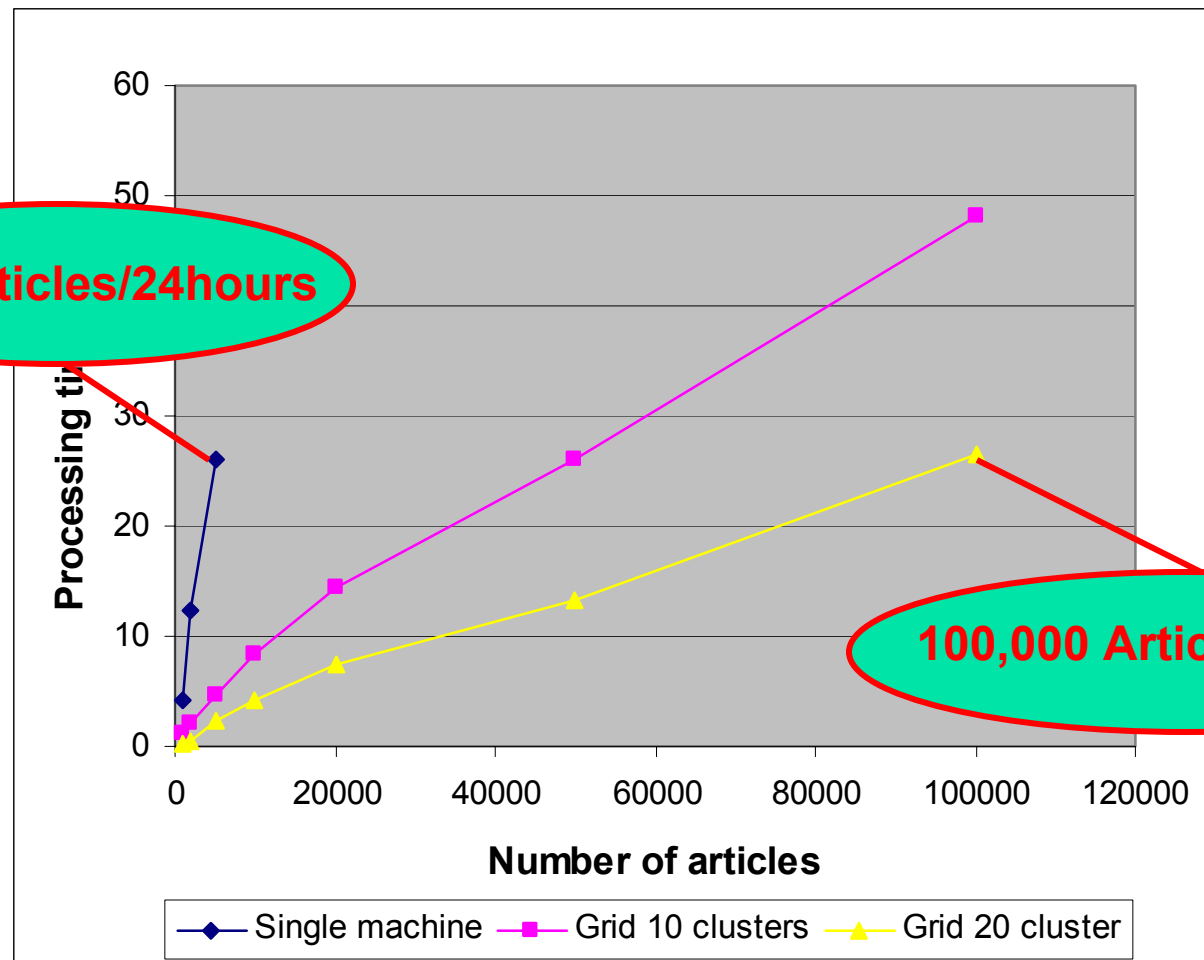
- **Pros of a Grid Solution**
 - Need for increased computational power
 - Minimize the infrastructure required
- **Cons of a Grid Solution**
 - Few applications are grid enabled
 - Institutional psychology and sociology



Grid Validation

- Timing
 - Expected improvement
 - Efficiency
- Fidelity
 - Are the same results returned?
 - Single computer vs. grid
 - Repeated runs on the grid

Decreased analysis time in the grid environment



5,000 Articles/24hours

100,000 Articles/24hours



Fidelity

- 1,000 abstracts run on the grid
 - Divided in to 10 “chunks” (100 articles)
 - Repeated 3 times
- Results
 - Each run returned 430 records
 - >98% concordance
 - Variability seen when more than one pattern per sentence



Lessons Learned

- **Institutional Psychology**
 - Administration
 - Excited to be part of research
 - Only use their machines at night
 - Information Technology
 - Easier maintenance and updating
 - Use of untapped resources
 - Still wonder at times if we are the cause of network slowdown
- **Use of Professional Services**
 - Set up grid
 - Grid enable your applications
 - LexiQuestMine(SPSS)
 - GetItRight(CTH Technologies)
- **Initial Project**
 - Go ahead with the project
 - “Pilot Project” has limited usefulness
 - Other ways to use the grid



What is Next?

- Expansion of text corpus to 10-15 years of 100 journals
- Automation of process – update on a regular schedule
 - Downloading of journal articles
 - Concept extraction and processing



Additional Grid Projects

- Analysis of large data sets:
 - Moving programs to the data for analysis



Acknowledgements

- **Brain Tumor Research Lab**

- David George
- Yonghong Zhang
- Jason Monroe
- Tadaki Tomita
- Chris McCabe
- David Paul
- Kimberly Anderson
- Krissy Dulek
- Yuichi Tange
- Shekhar Mayanil
- Beth Stahl
- Hiromichi Nakazaki

- **Univ. Ulster**

- Werner Dubitzky
- J'Kumar Natarajan
- Daniel Berrar

- **SPSS**

- Fabrice Leroy
- Catherine DeSesa
- Eric Martin
- Olivier Jouve

- **United Devices**

- Niranjana Mulay



Fidelity Test (single vs. grid)

Run Number (1000 articles)	Single machine	Grid 10 clusters	Grid 20 clusters	Run Number (5000 articles)	Single machine	Grid 10 clusters
1	735	743	741	1	4143	4156
2	735	742	732	2	4143	4157
3	735	743	743	3	4143	4148
4	735	742	734	4	4143	4138
5	735	734	745	5	4143	4148