A Grid Infrastructure for Text Mining of Full Text Articles and Creation of a Knowledge Base of Gene Relations

Jeyakumar Natarajan[1], Niranjan Mulay[2], Catherine DeSesa[3],
Catherine J. Hack[1], Werner Dubitzky[1], and Eric G. Bremer[3]

[1] Bioinformatics Research Group, University of Ulster, UK
{j.natarajan,cj.hack,w.dubitzky}@ulster.ac.uk
[2] United Devices Inc, Austin, TX, USA
niranjan@ud.com
[3] Brain Tumor Research Program, Children's Memorial Hospital,
Feinberg School of Medicine, Northwestern University, Chicago, IL USA
egbremer@northwestern.edu, cdesesa@comcast.net

Abstract. We demonstrate the application of a grid infrastructure for conducting text mining over distributed data and computational resources. The approach is based on using LexiQuest Mine, a text mining workbench, in a grid computing environment. We describe our architecture and approach and provide an illustrative example of mining full-text journal articles to create a knowledge base of gene relations. The number of patterns found increased from 0.74 per full-text articles from a corpus of 1000 articles to 0.83 when the corpus contained 5000 articles. However, it was also shown that mining a corpus of 5000 full-text articles took 26 hours on a single computer, whilst the process was completed in less than 2.5 hours on a grid comprising of 20 computers. Thus whilst increasing the size of the corpus improved the efficiency of the text-mining process, a grid infrastructure was required to complete the task in a timely manner.

1 Introduction

Whilst the past decade has witnessed an inexorable rise in the volume, diversity and quality of biological data available to the life-science research community, arguably the more measured rise in natural language documents provides a richer source of knowledge (Figure 1).

With many documents such as research articles, conference proceedings, and abstracts being made available in electronic format they have the potential to be automatically searched for information. However, given the volume of literature available, a semi-manual approach which would require reading a large number of articles or abstracts resulting from a keyword-based search, is infeasible. This has prompted researchers to apply text-mining approaches to life-science documents. Text mining is concerned with identifying non-trivial, implicit, previously unknown, and potentially useful patterns in text [1]. For example identifying references to biological entities [2-5], molecular interactions (protein-protein, gene- protein, drug-protein/gene), interactions of complexes and entities [6-9], sub-cellular localization of proteins [10, 11], pathways and structure of biological molecules [12-14]. Text mining has also been

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 101–108, 2005.

used to complement other techniques to achieve more complex tasks such as assisting in the construction of predictive and explanatory models [15] and assisting the construction of knowledge-bases and ontologies [16-19].

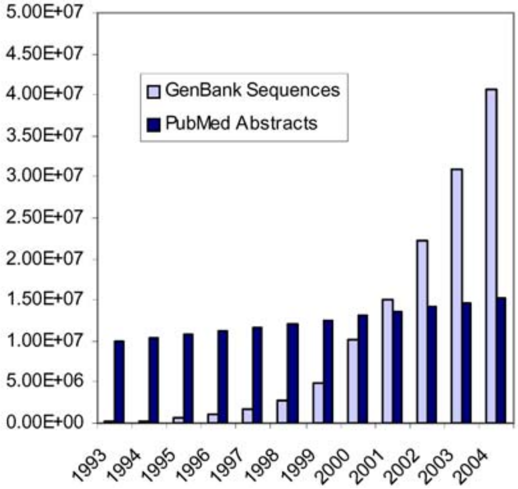

Fig. 1. The past decade has seen a rapid growth in databases of biological data such as GenBank, whilst the growth in literature databases such as PubMed has remained linear

Each of these text mining applications employs one or more of a range of computationally expensive techniques, including natural language processing, artificial intelligence, machine learning, data mining, and pattern recognition. Predominantly these examples have used homogeneous and localized computing environments to analyze small data sets such as collections of subject specific abstracts. However, as the text collections, the domain knowledge, and the programs for processing, analyzing, evaluating and visualizing these data, increase in size and complexity the relevant resources are increasingly likely to reside at geographically distributed sites on heterogeneous infrastructures and platforms. Thus text mining applications will require an approach that is able to directly deal with distributed resources; grid computing is promising to provide the necessary functionality to solve such problems. Grid computing is a generic enabling technology for distributed computing, which can provide the persistent computing environments required to enable software applications to integrate instruments, displays, computational and information resources across diverse organizations in widespread locations [20]. In its basic form, grid technology provides seamless access to theoretically unlimited computing resources such as raw compute power and physical data storage capacity.

In this paper we describe the use of a grid infrastructure for conducting text mining over distributed data and computational resources. Full-text journal from the fields of molecular biology and biomedicine were used to create a knowledge base of gene annotations. It is envisaged that this knowledge base may be used to annotate or interpret high-throughput data-sets such as microarray data. The approach is based on extending LexiQuest Mine, a text mining workbench from SPSS (SPSS Inc, Chicago) [21] in a UD grid computing environment (United Devices Inc, Texas) [22].

2 Text Mining in LexiQuest Mine

LexiQuest Mine [21] employs a combination of dictionary-based linguistic analyses and statistical proximity matching to identify entities and concepts. Pattern matching rules are then used to identify the relationship between the named entities (Figure 2). The LexiQuest Mine natural language processing engine contains modules for linguistic, non-linguistic, named entity, events and relations extractor. Each module can be separately or consequently used, depending on the nature of the problem. User defined dictionaries for specific domains (e.g., gene, protein names in life science) may also be incorporated.

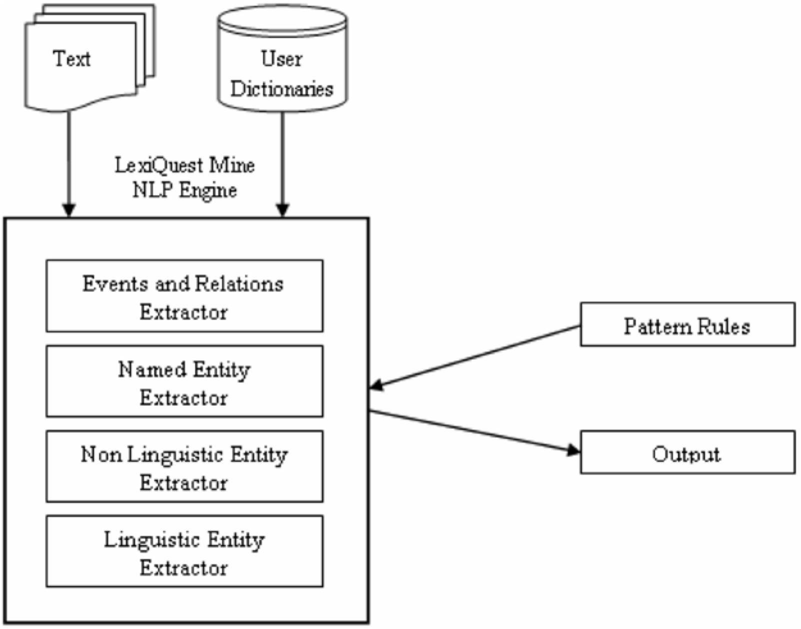
Text
User Dictionaries
LexiQuest Mine NLP Engine
Events and Relations Extractor
Named Entity Extractor
Non Linguistic Entity Extractor
Linguistic Entity Extractor
Pattern Rules
Output

Fig. 2. Text mining in the LexiQuest Mine environment

The Pattern Rules component contains specific patterns or rules to extract relationships. The rules are regular expressions based on the arrangement of named entities (gene, protein names), prepositions and keywords that indicate the type of relationship between them, for example,

gene[a-z](\s)interact\s[a-z]*(\s)+*gene[a-z]*(\s).

To perform the extraction of gene relations using pattern matching on gene names and relations, LexiQuest Mine relies on:

- Dictionaries of gene names
- Dictionaries of synonyms of gene names
- Linguistics tags following protein names which are used to automatically identify unknown genes (e.g. protein, kinase, phosphate)
- Dictionaries of gene relations (e.g., binds, inhibits, activates, phosphorylates)

3 Grid Deployment of Text Mining

Figure 3 illustrates the grid infrastructure Grid MP from United Devices Inc, Texas, USA [22]. Grid MP features a Linux based dedicated Grid server (labeled Grid MP Services here), which acts as a master or hub of the Grid. It balances compute demand by intelligently routing 'job' requests to suitable resources. The Grid MP Agent is software program that runs on each of the compute nodes, identifying the node and its capabilities to the Grid Server. The compute devices may run any Operating System (e.g. Windows, Mac, or Linux). The Grid MP Agent is responsible for processing work and for automatically returning the result files to the Grid MP Services. Users connect to the grid using a web interface or command line application-specific service scripts to deploy and run their applications. Grid MP software transforms the execution of an application from a static IT infrastructure, where the execution of an application is always tied to specific machines, to a dynamic virtual infrastructure where applications execute on different machines based on resource availability, business priority, and required service levels. The common steps involved in grid deployment and running of an application include:

- Grid MP Services and Agents are installed to lay the foundation for the virtual infrastructure. Administrators define and install policies that govern the use of this infrastructure across multiple applications and users.
- Application Services (e.g. LexiQuest Mine) are created and deployed.
- Users may interact with their application which is transparently executed on the virtual infrastructure created by Grid MP
- Results are collected by Grid MP Services and passed back to the end user.

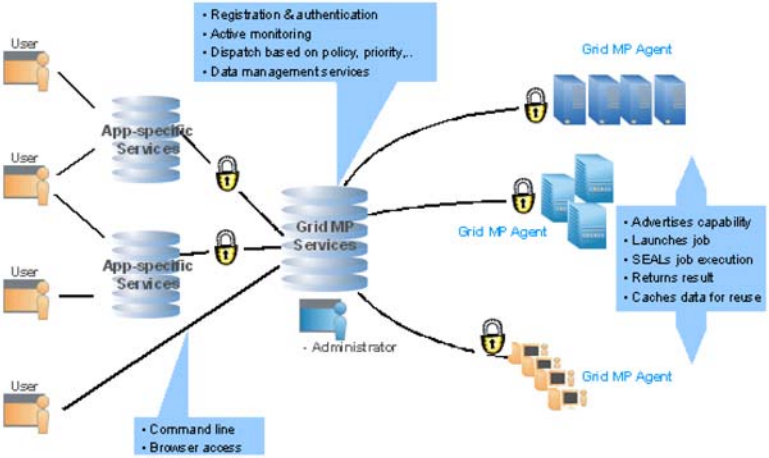

Fig. 3. Grid deployment of an application in UD Grid environment

4 Evaluation of Grid Based Text Mining for the Extraction of Gene Annotations

The aim of this work was to create a knowledge base of gene relationships which can be used to annotate and interpret microarray data. A corpus of approximately 125 000 full-text articles from 20 peer-reviewed journals in the field of molecular biology and biomedicine (1999-2003) (Table 1) was mined. The text mining methodology has been described previously [19] and comprises the following natural language processing (NLP) steps:

- *Sentence tokenization* to separate the text into individual sentences;
- *Word tokenization* to break pieces of text into word-sized chunks;
- *Part-of-speech tagging* to assign part-of-speech information (e.g., adjective, article, noun, proper noun, preposition, verb);
- *Named entity tagging* to find gene names and their synonyms and to replace the gene synonyms with unique gene identifier.
- *Pattern matching* to extract gene relations.

A sample output of final gene relations extracted using our system is illustrated in Table 2. In addition to the gene relations, we also extracted the PubMed ID and section ID corresponding to each gene relation using the pre-inserted XML tags in the corpus. This will help users to identify the source article and section from which the relations were extracted. The advantage here is that users can get corresponding gene annotations from full text articles for their genes of interest from their initial PubMed query results. The section tag helps users to identify the sections other than abstracts, which contains more gene relations for further research.

Table 1. List of downloaded journals and publisher's web sites

Journal Name	URL
Biochemistry	http://pubs.acs.org/journals/bichaw/
BBRC	http://www.sciencedirect.com/science/journal/0006291X
Brain Research	http://www.sciencedirect.com/science/journal/00068993
Cancer	http://www3.interscience.wiley.com/cgi-in/jhome/28741
Cancer Research	http://cancerres.aacrjournals.org
Cell	http://www.cell.com/
EMBO Journal	http://embojournal.npgjournals.com/
FEBS Letters	http://www.sciencedirect.com/science/journal/00145793
Genes and Development	http://www.genesdev.org/
International Journal of Cancer	http://www3.interscience.wiley.com/cgi-in/jhome/29331
Journal of Biological Chemistry	http://www.jbc.org/
Journal of Cell Biology	http://www.jcb.org/
Journal of Neuroscience	http://www.jneurosci.org/
Nature	http://www.nature.com/
Neuron	http://www.neuron.org/
Neurology	http://www.neurology.org/
Nucleic Acid Research	http://nar.oupjournals.org/
Oncogene	http://www.nature.com/onc/
PNAS	http://www.pnas.org/
Science	http://www.sciencemag.org/

Due to the high-through put computational tasks and huge amount of data (125 000 full-text articles from 20 different journals), we are unable to run the above application in a reasonable time frame using single computer. As an eventual solution to above situation, we have ‘grid-enabled’ the above application using UD Grid environment. The execution of this task on a single computer was compared with the grid-enabled application in terms of the computational efficiency and fidelity. The computational efficiency compares the time taken to complete the task, whilst fidelity provides a quantitative assessment of the quality of the results, i.e. whether the same results are returned from a single machine as on the grid, and whether repeated runs on the grid system produce consistent results.

Table 2. Sample output of gene relation extracted from full-text articles

PubMed ID	Gene 1	Gene 2	Relation	Source
12881431	APOBEC2	AICDA	Mediates	Abstract
12101418	NS5ATP13TP2	P53	Inhibits	Introduction
15131130	Map3k7	nf-kappa b	Activates	Methods
12154096	Igf-1	Pkb	Activates	Results

The time taken to complete the task and the CPU time was calculated for a single machine and two grid infrastructures comprising of 10 computers and 20 computers respectively (Figure 4). Whilst the CPU time remains approximately constant for both the grid and the single computer, the advantage of the grid structure becomes apparent when looking at the total time to complete the task. Once the literature database contained more than 5000 full-text articles it was impractical to run on single machine.

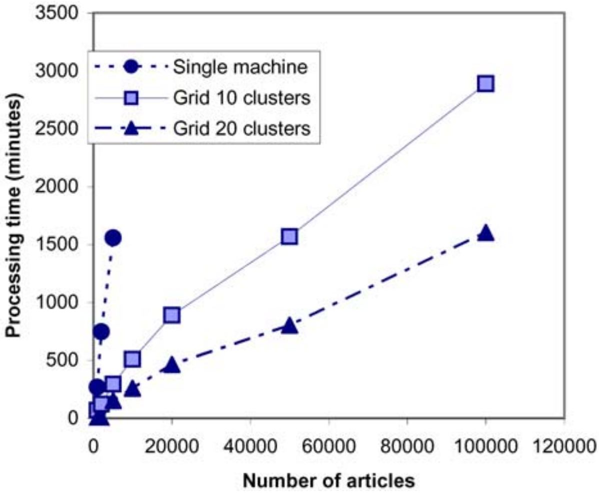

Fig. 4. Processing time to complete text mining of full-text articles using a single computer (✿) and a grid of 10 computers (■) and a grid of 20 computers (▲)

Figure 5 illustrates the average number of patterns returned per article over 5 runs on a single machine and the two grid infrastructures described previously. The error bars show the maximum and minimum number of patterns found, illustrating that there was no significant difference observed on the different architectures. The number of patterns returned in repeated runs also remained constant. However as the size of the corpus increased from 1000 to 5000 articles, the number of patterns found per article increases from an average of 0.74 to 0.83, indicating the clear advantage of using a larger corpus.

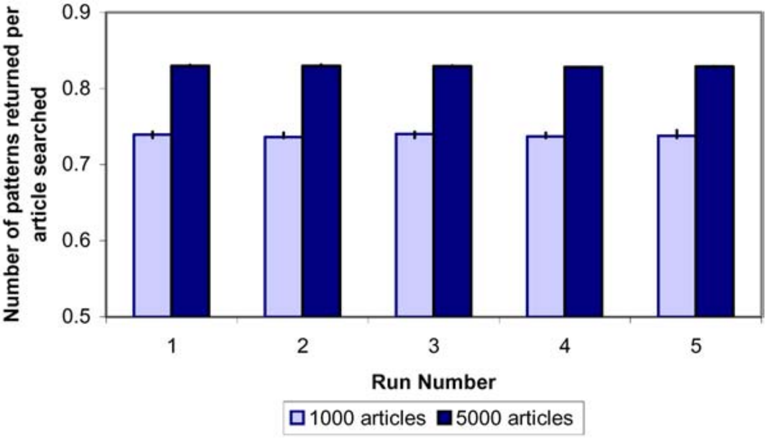

Fig. 5. The average number of patterns returned per article in the corpus using a single machine and two grid structures comprising of 10 and 20 cluster computers. The process was repeated over 5 runs and using a corpus of 1000 articles and a corpus of 5000 articles

5 Conclusion and Future Directions

Biological texts are rich resources of high quality information and the automatic mining of such repositories can provide the key to interpreting the large volumes of life-science data currently available. This work has demonstrated that as the size of the corpus increases, the text-mining process becomes more efficient in terms of extracting information from articles. However mining larger databases has significant implications in terms of processing time and computing resources; for example it was impractical to mine a corpus of more than 5000 full-text articles on a single machine. This paper has demonstrated that grid technology can provide the necessary middle-tier layer to allow the use of a distributed computing environment for such tasks. In addition to those described in this paper, we are also in the process of using the above infrastructure for developing other applications such as finding gene-disease relationship, disease-drug relation etc. using the same full-text articles. We plan to integrate these data with our in-house microarray data that integrates text and data mining applications in grid infrastructure. Incorporating this technique into the data mining pipeline of microarray analysis has the potential to effectively extract information and thus provide a greater understanding of the underlying biology in timely manner.

References

1. Hearst M. A., Untangling text data mining, Proc. Of ACL, 37 (1999)
2. Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T., Towards Information Extraction: identifying protein names from biological papers, Pacific Symposium on Biocomputing, 707-718 (1998)

3. Eriksson, G., Franzen, K., and Olsson, F., Exploiting syntax when detecting protein names in text, Workshop on Natural Language Processing in Biomedical Applications, 2002, at http://www.sics.se/humle/projects/prothalt/
4. Wilbur, W., Hazard G. F. Jr., Divita G., Mork J. G., Aronson A. R., and Browne A. C., Analysis of biomedical text for biochemical names: A comparison of three methods, Proc. of AMIA Symposium, 176-180, (1999)
5. Kazama, J., Makino, T., Ohta, Y., and Tsujii, J., Tuning Support Vector Machines for Biomedical Named Entity Recognition, Proc. of the Natural Language Processing in the Biomedical Domain, Philadelphia, PA, USA (2002)
6. Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T.: Automated extraction of information on protein-protein interactions from the biological literature, Bioinformatics, 17, 155-161 (2001)
7. Wong, L.: A protein interaction extraction system, Pacific Symposium on Biocomputing, 6, 520-531 (2001)
8. Yakushiji, A., Tateisi, Y., Miyao, Y., & Tsujii, J.: Event extraction from biomedical papers using a full parser, Pacific Symposium on Biocomputing, 6, 408-419 (2001)
9. Sekimizu, T., Park, H.S., & Tsujii, J.: Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, Proceedings of the workshop on Genome Informatics, 62-71 (1998)
10. Craven, M., and Kumlien, J., Constructing biological knowledge base by extracting information from text sources, Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology, 77-76 (1999)
11. Stapley, B. J., Kelley, L. A., and Strenberg, M. J. E., Predicting the sub-cellular location of proteins from text using support vector machines, Pacific Symposium on Biocomputing, 7, 374-385 (2002)
12. Gaizauskas, R., Demetriou, G., Artymiuk, P. J, and Willett, P., Protein structure and Information Extraction from Biological Texts: The PASTA system, Bioinformatics, 19:1, 135-143 (2003)
13. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C., GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data, Jr of Biomedical Informatics, 37, 43-53 (2004)
14. Hahn, U., Romacker, M., and Schulz, S., Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system, Pacific Symposium on Biocomputing, 7, 338-349 (2002)
15. Ideker, T., Galitski, T., and Hood, L., A new approach to decoding life: systems biology, Annu Rev Genomics Hum Genet 2:343-372 (2001)
16. Rzhetsky, A., etc.: GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data, Jr of Biomedical Informatics, 37, 43-53 (2004)
17. Pustejovsky, J., etc.: Medstract: Creating large scale information servers for biomedical libraries, ACL-02, Philadelphia (2002)
18. Wong, L.: PIES a protein interaction extraction system, Pacific Symposium on Biocomputing, 6, 520-531 (2001)
19. Bremner,E.G., Natarajan, J. Zhang,Y., DeSesa,C., Hack,C.J. and Dubitzky,W.,: Text mining of full text articles and creation of a knowledge base for analysis of microarray data, LNAI, Knowledge exploration in Life Science Informatics, 84-95 (2004)
20. Foster I, and Kesselman C (eds), The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann (2004)
21. SPSS LexiQuest Mine available at http://www.spss.com
22. United Devices Grid MP Services available at http://www.ud.com