

# Research Statement

Kuan-ming Lin

January 22, 2007

Since I took the Database course taught by Prof. Yang, I have been attracted by the excitement of data mining. Broadly speaking, I am interested in finding useful patterns in large stores of data which can be handled only by computers. In earlier years, such data of large size were collected by business or organizations and kept in proprietary data warehouses. Therefore, many data mining techniques were specialized to tackle individual cases, and their generalizability has not been deeply studied. Nowadays, as the Internet has become the largest worldwide information storage, it is possible for the public to perform data analysis without access to proprietary databases. So, it is high time I participated in data mining research. In particular, I am curious in mining three categories of public-accessible electronic data, namely biological databases[5][3], Web documents[1], and the Semantic Web.

In recent years, a number of repositories of biological data has been published on the Internet, free and public to the global research community. Particularly, the emerging of high throughput methods contributes to plentiful public databases representing different aspects of biological processes. The data collectively are believed to provide more information than what they were explained in individual studies. For example, we could assign function to unknown genes by integrative analysis on multiple microarrays, genome sequences, NMR and X-ray data, etc. As a result, how to extract complementary information across heterogeneous datasets is an appealing data mining problem for discovering unknown biological relations. Since there is usually no underlying theoretical model available for biological datasets, this type of data mining is very challenging and worth studying.

As the rapid growing of textual information on the Web, text mining on Web documents is receiving much attention. That is, retrieving interesting information from Web documents could have a high commercial potential value. In academic research, Web text mining is a unique challenge over traditional information retrieval in that Web documents are by nature highly unstructured yet interconnected through hyperlinks. Therefore, designing new techniques for mining Web document could lead to great applications, and I will be happy to be involved in related research.

The Semantic Web is a solution to extend the current Web content to provide intelligent mining and analysis on the Web. Unlike the currently prevalent HTML and XML, Semantic Web languages explicitly define associations between semantics, or meaning, with the Web document content. The semantics are formally specified in well-designed ontologies; therefore, they can be either shared via the Internet or refined for local needs. The current standard for the Semantic Web exists (OWL, a W3C Recommendation), but integration from heterogeneous Semantic Web groups and ontologies are still in its infancy. Besides, the current semantic reasoning and ontology updating tools are still far from satisfactory. Therefore, I would like to explore this new field using my already cumulated knowledge on data mining and database integration.

Throughout the last two years, I started to develop skills in the data mining subfields, ranging from theoretical studies like approximation algorithms for integrative data mining, to machine learning issues like feature selection and discovery and SVM[4][2], to applications in bioinformatics and Web mining. My courses and publications focused on bioinformatics and statistical learning, and I gained experiences on Web text mining through my research internship and a summer school lecture series in Taiwan. I also published two papers at Duke: one based on the CPS 260 final project[5], and the other on the RIP final report[3]. In this year, I hope to continue my research in data mining and join a research group in Duke who can support me throughout my PhD study.

## References

- [1] C.-C. Huang, K.-M. Lin, and L.-F. Chien. Automatic Training Corpora Acquisition through Web Mining. In *Proceedings of The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 193–199, 2005.
- [2] K.-M. Lin. Reduction Techniques for Support Vector Machines. Master's thesis, National Taiwan University, Taipei, Taiwan, 2002. Adviser: Chih-Jen Lin.
- [3] K.-M. Lin and J. Kang. Exploiting Inter-gene Information for Microarray Data Integration. In *Proceedings of The ACM Symposium on Applied Computing (SAC)*, March 2007. To appear.
- [4] K.-M. Lin and C.-J. Lin. A Study on Reduced Support Vector Machines. *IEEE Transactions on Neural Networks*, 14(6):1449–1559, 2003.
- [5] K.-M. Lin, J. Zheng, and J. Zhang. Automatic Structure Prediction of HIV Coreceptor CCR5. In *Proceedings of Bioinformatics in Taiwan Symposium (BIT)*, Taichung, Taiwan, September 2006.