Large Scale Data Analysis and Modelling in Online Services and Advertising

[Extended Abstract]

Thore Graepel
Online Services and Advertising Group
Microsoft Research Ltd.
7 J J Thomson Avenue
CB3 0FB Cambridge, UK
thoreg@microsoft.com

Ralf Herbrich
Online Services and Advertising Group
Microsoft Research Ltd.
7 J J Thomson Avenue
CB3 0FB Cambridge, UK
rherb@microsoft.com

The last five years have seen a tremendous growth in online search, advertising and gaming services. Today, it is extremely important to analyse large collections of user interaction data as a first step in building predictive models for these services. These collections pose challenges not only in their size—often exceeding tera-bytes (TB) of data—but also in their heterogeneity. We report on two applications of large scale data analysis performed at Microsoft Research and how data analysis guided model development:

1. We present the unique challenges involved in building a new advertisement ranking algorithm for Paid Search. To guide the modelling task, we developed a tool-chain which allows the storage and querying of weeks of click-through data from raw logs. In this task, we managed more than 3 billion advertisement impressions occupying more than 1 TB of information. Ultimately, we were able to answer within minutes questions about the empirical dependency of the click-through rate on user features such as *client IP* and advertisement features such as *match type*. Central to this task was the development of a very fast object-relational mapping for converting data between the F# [3] type systems and the SQL type-system in order to build a data store of click-through meta-information about users and advertisements. This tool-chain guided the development of features for training a Bayesian click-through estimation algorithm based on *expectation propagation* [2]. We discuss how systems issues such as memory consumption and algorithmic performance influenced the modelling process. We also discuss the issue of scientific programming languages capable of dealing with CPU intensive task while allowing rapid prototyping and give a quick overview of F#, a functional programming language ideally suited for this task.

2. We share some insights gained in the context of the data analysis and modelling tasks that went into the development of Halo 3's online ranking and matchmaking algorithm. Halo is a series of first person shooter games and constitutes one of the most successful series in the history of gaming. At its core, Halo 3 uses the well-known Bayesian TrueSkill ranking and matchmaking system [1]. Before its launch, we performed thousands of simulations of ranking behaviour on over 3 million players varying two parameters:

   (a) Speed of convergence. The convergence of the original TrueSkill system is close to optimal but this fast initial convergence is counter-intuitive to gamers and can lead to a rejection of this system.

   (b) Skill-level display. During matchmaking, waiting time and tightness of the match are in opposition. By re-mapping the skill level display, matchmaking can be greatly helped to minimise waiting time while constructing seemingly tight matches.

   We also discuss the limitations of this simulation and present results on how the running online part of the game today compares with the results of the simulations. As of today, over 800,000 unique players play over 2 million Halo 3 matches every 24 hours.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
ACM 978-1-60558-193-4/08/08.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: General; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Algorithms, Economics, Experimentation, Performance

1. REFERENCES

[1] R. Herbrich, T. Minka, and T. Graepel. TrueSkill(TM): A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576, 2007.

[2] T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

[3] D. Syme, A. Granicz, and A. Cisternino. *Expert F#*. APress, 2007.