論文の内容の要旨

Efficient Sequence Data Analysis with Hidden Markov Models
（隠れマルコフモデルによる高速な系列データの解析手法）

氏名　藤原　靖宏

Sequential data analysis, a relatively young and interdisciplinary field of computer science, is the process of extracting patterns from large sequence data sets by combining methods from statistics and artificial intelligence with database management.

With recent tremendous technical advances in processing power, storage capacity, and Internet, sequential data analysis is seen as an increasingly important approach by modern business. This is because it can give an informational advantage by transforming unprecedented quantities of sequence data into business intelligence. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. Since sequential data analysis can bring real value for real applications, demand for novel technologies in sequential data analysis is growing these days.

The Hidden Markov Model (HMM) is a ubiquitous tool for representing probability distributions over sequences of observations. The basic theory of HMM was developed in the late 1960s. Since HMMs, which assess sequential data as sequences of state transitions, are robust against noise, significant applications that use HMMs have emerged, including sequence labeling, speech recognition, mental task classification, biological analysis, traffic monitoring, and anomaly detection. This thesis mainly handles three research problems; The first problem is exact and efficient states sequence detection for single HMM and static sequence of arbitrary length, and the second problem is efficient identification of the model whose state sequence has the highest likelihood for the given query sequence, exactly (i.e., an HMM that actually has a high-probability path for the given sequence is never missed by the algorithm.). The third problem is efficient monitoring of streaming data sequences to find the best model without exception. And, to show the generality of the proposed approach for other data structure, I applied the proposed approach for the problem to monitoring best centrality nodes of time-evolving graphs.

I propose Staggered decoding for the first problem, SPIRAL for the second and third problem, and Sniper for the fourth problem. The proposed approach is based on two ideas; approximation and pruning. Approximation is an idea that aggregates several

state to discard unlikely states or models. And pruning is an idea that computes exact likelihood of viable states/models by pruning unlikely state transition. The proposed approach has the following attractive characteristics:

-High-speed: Solutions based on previous approaches are prohibitively expensive for large HMM data. The proposed approach uses carefully designed approximations to efficiently identify the most likely model.

-Exact: The proposed approach does not sacrifice accuracy; it returns the highest likelihood model without any omission.

-Applicability: The proposed approach can be applied not only for HMM but other data structures such as time-evolving graphs.

In order to achieve high performance and to find the exact answer, the proposed approach first prunes many states/models with approximate likelihoods at low computation cost. The exact likelihood computations are limited to the minimum necessary, which yields a dramatic reduction in the total search cost. Experiments compared the proposed method with the method based on the previous approaches. As expected, the experiments demonstrate the superiority of the proposed approach.