



Master's thesis in
Information- and Communication Technology

Title:
Incremental Web Crawling as a Competitive Game of Learning Automata

Candidates:
Svein Arild Myrer
Morten Goodwin Olsen

Supervisor:
Ole-Christoffer Granmo, AUC



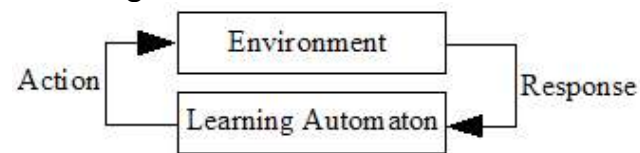
Problem

An increasing amount of versatile services keep finding their way onto the Web. Especially the possibility of producing dynamic content has been an accelerant factor and is the reason why we now conveniently can see the latest development of our favorite stocks in near real-time from our own living rooms.

However, for automated data mining applications the highly dynamic nature of these new services is not convenient at all. As a matter of fact, a complete new set of challenges emerge where traditional crawling strategies are shown to be sub-optimal.

In our work we have addressed this new problem area of monitoring highly dynamic data sources of different importance. We have used the concept of an incremental web crawler that works by selectively updating parts of its local copy in an incremental fashion. In our novel approach we have considered the incremental crawling task as a continuous learning problem where scheduling of monitoring tasks is combined with parameter estimation in an on-line manner. By mapping the problem to two variants of the so called knapsack problem we have proposed two solutions based on a machine learning technique known as learning automata.

Learning Automata



The principles of learning automata have attracted considerable interest in the last decade because they can learn the optimal action when operating in (or interacting with) unknown stochastic environments. Furthermore, they combine rapid and accurate convergence with low computational complexity. The above figure

shows in essence how a learning automaton interacts with an environment in order to improve its performance by a learning process.

The web crawling - knapsack problem connection.

The so called knapsack problem is a popular combinatorial optimization problem which arises whenever we meet a resource allocation problem imposed by some constraint. We have mapped the incremental crawling task to two variants of the knapsack problem; namely the binary and the fractional knapsack problem. In both cases we have considered the item values to be of unknown distribution and the item weights to be equal. The table below summarizes the relations between the fractional knapsack problem and the incremental crawling task as we mapped it. The binary case is similar only that we do not consider fractions of an item, only binary values.

Knapsack size	Crawler capacity
Fraction of item	Polling rate of a web page
Value of item	Update rate of a web page
Weight of item	Cost of polling a web page
Number of items	Number of web pages to monitor

Proposed Algorithms

Competitive Game of Learning Automata

A solution to the fractional knapsack variant were approached by extending a learning algorithm used to solve the parameter optimization problem. By connecting independent automata into a competitive game and govern when rewards and penalties should be regarded, we designed a scheme that were not only able to adapt to a set knapsack size / crawler capacity, but also create a stochastic competition between the automata that adaptively improved performance.

Fixed Partitioning Automaton

As a solution to the binary knapsack variant we presented a learning automaton which was an extension of the Object Migration Automaton previously presented as a solution to the equi-partitioning problem. Our proposed solution is designed to partition items into two partitions of fixed, but possibly unequal, sizes where one partition contained the most valuable items.

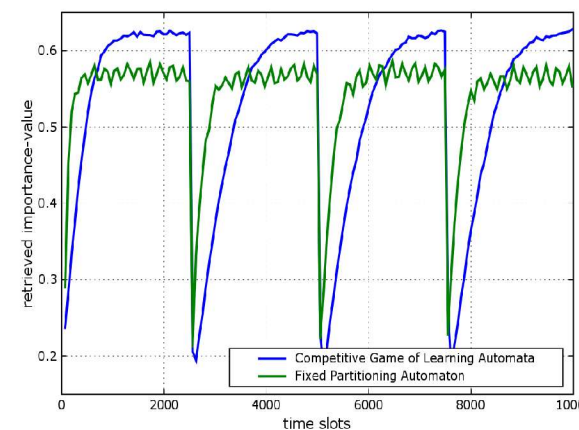
Experimental results

We have empirically evaluated the performance of our proposed automata solutions using two different environment models; one where we considered updates to disappear at once and another where we considered updates to remain, but just until the next update occurred.

As our performance metric we have used a normalized value that reflected the importance-value of the information which the crawler were able to retrieve during the simulation runs.

Adaptability

Both solutions were designed to cope with a dynamic or non-stationary environment. The figure below shows the automata operating in an environment where changes overwrite information and that switches every 2500th time slot. We can observe their adaptive capability.



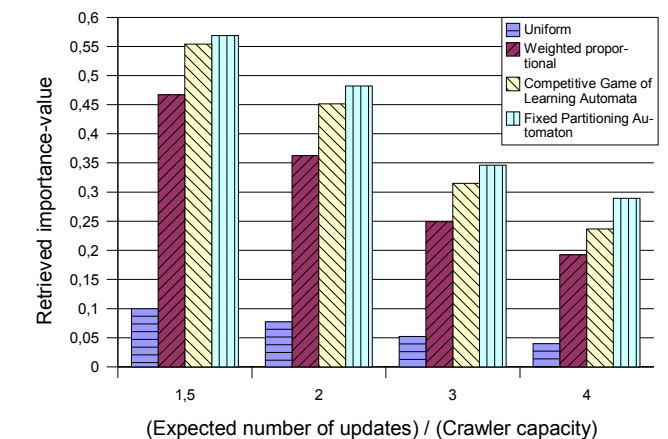
Comparison to alternative algorithms

The performance of both solutions were compared with relevant algorithms. These policies are:

Uniform: Resources are distributed uniformly among the items, resulting in a round robin scheme for the re-crawling process.

Weighted proportional: Resources are allocated in a proportional manner. This algorithm assumes that the values of the items are known.

The figure below shows our solutions outperform alternative algorithms in a stationary environment where updated information disappear at once.



Conclusion

Both of our proposed solutions were shown to outdo the uniform scheme and in most cases the weighted proportional scheme. Most notably we outperformed the round robin scheme by factors up to 550% in certain situations. Our solutions were also shown to successfully being able to operate in non-stationary environments where the Fixed Partitioning Automaton showed a faster adaptive behavior compared to the Competitive Game of Learning Automata in all investigated situations.