Árpád GELLÉRT

Prediction-Based Modeling and Estimation in Advanced Computing Systems

Modelare și estimare bazată pe predicție în sisteme de calcul avansate

TEZĂ DE ABILITARE

în Calculatoare și Tehnologia Informației

REZUMAT

2023

Această teză de abilitare este o sinteză a celor mai importante realizări de după teza de doctorat în modelare și estimare bazată pe predicție în domenii precum arhitecturi de calcul, fabrici inteligente, clădiri inteligente, procesarea imaginilor și web mining. Toate metodele prezentate au în comun o largă exploatare a tehnicilor avansate de predicție și a modelelor de prognoză.

În domeniul arhitecturilor de calcul, s-au dezvoltat și evaluat diferite scheme de predicție a valorilor având ca scop creșterea paralelismului la nivelul instrucțiunilor și a performanței de procesare și în același timp scăderea consumului de energie în cazul microprocesoarelor superscalare, multifir sau multicore. Predicția valorilor este o tehnică speculativă aplicată pentru anticiparea rezultatelor instrucțiunilor cu latență de execuție ridicată și deblocarea instrucțiunilor dependente prin execuție speculativă. Predicția trebuie realizată cu acuratețe pentru că fiecare predicție greșită trebuie tratată printr-un mecanism de recuperare a contextului corect, ceea ce introduce timp de execuție suplimentar. Această lucrare aduce contribuții originale în dezvoltarea și evaluarea unor metode avansate de predicție a valorilor aplicate selectiv pe instrucțiuni Load critice. Sunt prezentate scheme de predicție bazate pe contoare respectiv perceptroni integrate în simulatoare ca M-SIM sau Sniper. Numărul mare al parametrilor microprocesoarelor simulate generează o complexitate de simulare ridicată în procesul de explorare a spațiului de proiectare. De aceea, optimizarea multi-obiectiv se realizează aplicând căutarea euristică prin algoritmi genetici.

Tema de cercetare în care m-am implicat cel mai recent propune îmbunătățirea ecosistemului de fabrică inteligentă prin contribuții în predicția și modelarea proceselor de asamblare a produselor. Scopul principal este integrarea unui modul de predicție în stații de asamblare capabile să ghideze muncitorii oferind opțiuni cu privire la următorul pas de asamblare. Ca o primă încercare, pentru furnizarea următorului pas de asamblare s-a folosit un predictor contextual pe două nivele compus dintr-un registru de stare (primul nivel) care indexează o tabelă ce stochează perechi de context împreună cu starea următoare asociată (al doilea nivel). O altă variantă a predictorului contextual pe două nivele extinde fiecare stare de asamblare cu un automat care poate fi în substare stabilă sau instabilă. Din păcate, această schemă a adus îmbunătățiri nesemnificative și, având în vedere că folosește informații suplimentare și pași adiționali în procesul de predicție, s-a dovedit mai puțin eficientă decât schema inițială fără automat. Modelul Markov este un alt predictor contextual pe două nivele care poate stoca stări multiple împreună cu frecvențele lor de apariție pentru fiecare context. Dimensiunea contextului folosit stabilește ordinul modelului. Starea cu frecvența de apariție cea

mai ridicată se extrage din intrarea tabelei de predicție selectată cu registrul de stare și este furnizată apoi ca stare predicționată. Un astfel de predictor poate furniza mai multe opțiuni cu probabilități diferite pentru următorul pas de asamblare, dar pentru sistemele în timp real se poate configura să returneze starea cea mai probabilă. A urmat evaluarea algoritmului de predicție bazat pe potrivire parțială pentru furnizarea următorului pas de asamblare. Acesta combină mai multe modele Markov de ordin diferit. Prima încercare de predicție a modelului de ordin R se realizează cu lanțul Markov de ordin R. Dacă acesta poate predicționa, i se returnează predicția. Altfel, dacă lanțul Markov curent nu poate predicționa, ordinul este decrementat până când predicția este realizabilă sau niciun model n-a reușit să predicționeze, caz în care procesul se încheie fără predicție. Deoarece contextul curent nu poate fi întotdeauna găsit în tabela de predicție, am introdus o schemă de predicție îmbunătățită, care explorează caracteristicile vecine ale utilizatorului dacă caracteristicile reale nu se potrivesc exact. Explorarea contextelor vecine implică schimbarea pe rând a câte unei caracteristici a muncitorului urmată de încercarea generării predicției. Starea predicționată majoritar din contexte vecine va fi considerată predicția finală.

În domeniul clădirilor inteligente, principala contribuție constă într-un sistem inteligent de gestiune a energiei electrice destinat caselor echipate cu panouri fotovoltaice și sisteme de stocare a energiei care ia automat decizii pe baza estimării producției și a consumului de energie electrică. În orice moment sistemul poate combina după necesități energia electrică produsă local cu energia electrică stocată respectiv cu energie din rețeaua electrică. Rolul sistemului de gestiune este ajustarea și sincronizarea consumului și a producției de energie electrică, crescând raportul de autoconsum și reducând presiunea pe rețeaua electrică. Astfel, scad costurile anuale și, ca beneficiu suplimentar, scad și pierderile din rețelele de distribuție. Pe baza predicțiilor, sistemul de gestiune a energiei electrice poate activa electrocasnice când este disponibilă electricitate ieftină și poate întârzia activarea lor când e disponibilă doar electricitate scumpă. Au fost dezvoltate modele Markov, predictoare incrementale, rețele neuronale recurente de tip LSTM și modele hibride și toate au fost evaluate pe seturile de date disponibile.

În ceea ce privește procesarea imaginilor, am dezvoltat o metodă contextuală de eliminare a zgomotului de tip impuls din imaginile în nivele de gri afectate. Algoritmul propus folosește lanțuri Markov pentru înlocuirea zgomotului detectat cu intensitatea care a apărut cel mai frecvent în contexte similare. Contextul unui pixel afectat de zgomot constă în intensitățile pixelilor din strânsa vecinătate și este căutat într-o vecinătate mai largă dar limitată. Am analizat diferite metode de căutare și diferite forme de context. Rezultatele experimentale obținute pe imagini de test au arătat că modelul cel mai eficient aplică o căutare în formă de „*” a contextelor în formă de „+”. Pe lângă performanța de filtrare îmbunătățită pe toate nivelele de

zgomot, s-a redus substanțial și durata de procesare față de modelul de filtrare contextuală cu căutare completă a contextelor de formă pătratică. Am comparat acest filtru Markov cu alte tehnici de filtrare existente în literatura de specialitate, majoritatea lor fiind net depășite. O altă contribuție originală constă în metode de reconstrucție contextuală a zonelor de imagine afectate de factori externi (defecte) sau acoperite de obiecte sau text. Într-o primă fază, utilizatorul trebuie să selecteze zona dorită, iar apoi algoritmul dezvoltat înlocuiește intensitatea fiecărui pixel din zona marcată pe baza informațiilor contextuale neafectate din jur. Procesul de restaurare se aplică din exterior spre interior în cadrul zonei selectate. Pentru înlocuirea intensității unui anumit pixel, se explorează o zonă limitată din jur în vederea identificării intensității care apare cel mai frecvent în contexte similare. Prin folosirea informațiilor contextuale, tehnica propusă poate reconstrui foarte bine detaliile din imagini.

În domeniul web mining, am studiat diverse metode de preîncărcare a obiectelor web pe bază de predicție. S-a evaluat algoritmul de predicție bazat pe potrivire parțială, precum și arbori de decizie cu diferite componente Markov. Obiectul web predicționat se preîncarcă în cache pentru a-l face disponibil în caz de accesare. Experimentele efectuate pe setul de date colectat de Universitatea din Boston arată că metoda optimă este cea bazată pe arbore de decizie care folosește ca trăsături link-ul curent, tipul link-ului și predicțiile generate de lanțurile Markov de ordin 1-4. Această metodă optimă de predicție a depășit toate modelele Markov aplicate individual, dar și algoritmul de predicție bazat pe potrivire parțială.

Árpád GELLÉRT

Prediction-Based Modeling and Estimation in Advanced Computing Systems

HABILITATION THESIS

in Computers and Information Technology

SUMMARY

2023

This habilitation thesis is a synopsis of the most important research achievements after the PhD thesis in prediction-based modeling and estimation in topics like computer architecture, smart factories, smart buildings, image processing and web mining. All the presented methods have in common a large exploitation of advanced prediction techniques and forecasting models.

In computer architecture, different value prediction schemes have been developed and evaluated with the goal of increasing instruction-level parallelism and the overall processing performance and decreasing at the same time the energy consumption of superscalar, multithreaded and multicore microprocessors. Value prediction is a speculative technique, which anticipates the results of high-latency instructions and unlocks subsequent dependent instructions by speculatively executing them earlier. The prediction must be accurate, since any misprediction is treated by a recovery of the correct processor context, which consumes additional cycles. This work brings original contributions in developing and evaluating advanced value prediction methods applied selectively on critical Load instructions. We present counter-based and perceptron-based prediction schemes integrated into simulators like M-SIM or Sniper. The high number of parameters of the simulated microprocessors generates a huge simulation complexity in the design space exploration process. Therefore, the multi-objective optimization is realized by heuristic search through genetic algorithms.

My most recent research topic is aiming to improve the smart factory ecosystem with contributions in predicting and modeling assembly processes. The main goal is the integration of a prediction module into assembly assistance systems able to support factory workers in their manufacturing activities by providing choices for the next assembly step. First, a two-level contextual predictor was used to suggest the next assembly steps which consists in a state register (the first level) which indexes a table storing pairs of state-patterns and their associated next states (the second level). Another variant of the two-level contextual predictor extended each assembly state with an automaton which could be in stable or in unstable substate. Unfortunately, this scheme provided insignificant improvement, and because it used supplementary information and additional steps in the prediction process, it was considered less efficient than the scheme without automata. The Markov model is another context-based two-level predictor which can store multiple next states, together with their number of occurrences for each pattern. The length of the context determines the order of the model. The state with the highest number of occurrences is extracted from the prediction table entry selected with the left-shift state register and is then provided as the predicted one. Such a predictor can provide

multiple next assembly choices with different probabilities, but for time-critical decisions it can be configured to return the most probable state. The prediction by partial matching algorithm was also evaluated as an assembly step predictor. It combines different order Markov predictors. The model of order R first tries to predict with the Markov chain of order R. If the Markov chain can predict, its prediction is returned. Otherwise, if the current Markov chain cannot predict, the order is decremented until a prediction can be done or all the models were evaluated without success and, in that case, no prediction can be provided. Because the current context cannot always be found in the prediction table, an enhanced prediction scheme was considered, which explores the neighboring characteristics of the user if the actual characteristics do not have an exact match. Neighboring context exploration involves changing one characteristic of the worker at a time followed by prediction trial. The step that was predicted the most from the neighboring states will be considered the next assembly step.

In the smart buildings research area, a main contribution consists in smart energy management systems designed for households equipped with photovoltaics and energy storage systems making automated decisions based on forecasted electricity production and consumption. At any time, the system can combine the own produced electricity with the stored electricity and with electricity from the grid. The interest is to adjust and synchronize through prediction the electricity consumption and production increasing self-consumption and reducing the intake from the power grid. Thus, the total annual operating cost is lower and, as additional benefit, the losses in the distribution networks are reduced. Based on the predictions, the energy management system may decide to activate some household appliances when cheap electricity is available and to delay their activation when only high-cost electricity is available. Markov chains, stride predictors, a Long Short-Term Memory and hybrid models were developed and evaluated on the available datasets.

In the image processing topic, a context-based denoising method was developed for grayscale images affected by impulse noise. The proposed algorithm is using Markov chains to replace the detected noise with the intensity having the highest number of occurrences in similar contexts. The context of a noisy pixel consists in its neighbor pixels and is searched in a larger but limited surrounding area. We have analyzed different search methods and different context shapes. The experimental results obtained on the test images have shown that the most efficient model applies the search in form of “*” of contexts having the form of “+”. Beside the better denoising performance obtained on all the noise levels, the computational time has been also significantly improved with respect to the context-based filter which applies full search of full context. We have also compared this Markov filter with other denoising techniques existing in the literature, most of them being significantly outperformed. Another original contribution is a

context-based inpainting method which is using Markov chains to repair pixel colors from images affected by external factors (defects) or to replace pixel colors belonging to image areas covered by objects or texts. First, the user must select the target area and then the developed inpainting algorithm is replacing each pixel intensity from the target area based on the surrounding unaffected context information. The restoration process is applied from the exterior to the interior within the selected target area. For the replacement of a certain pixel intensity, we explored a limited surrounding image area to identify the intensity occurring with the highest probability in similar contexts. Since we use context information, the proposed inpainting technique can very well rebuild the image details.

In the web mining research area, several prediction-based web-prefetchers have been studied. The prediction by partial matching algorithm was evaluated, as well as a dynamic decision tree with different order Markov predictors as components. The predicted web object is prefetched into the cache to make it available for possible next accesses. The experiments performed on a dataset from the Boston University show that the optimal method is the dynamic decision tree which uses as features the current link, the link type and the predictions provided by the Markov chains of orders 1-4. This optimal predictor outperformed all the Markov models applied separately, but also the prediction by partial matching.