prof. dr hab. inż. Tadeusz Czachórski
Instytut Informatyki, Wydz. Automatyki Elektroniki i Informatyki
Politechniki Śląskiej, ul. Akademicka 16, 44-100 Gliwice

Gliwice, 10 październik 2018 r.

Recenzja pracy doktorskiej mgr inż. Macieja Rafała Buraka
p.t. *"Application of inhomogeneous continuous time Markov chains for call center modeling"*
(„Modelowanie Call Center z wykorzystaniem niejednorodnych łańcuchów Markowa z czasem ciągłym")

Kontekst rozprawy

Praca dotyczy modelowania i oceny efektywności pracy centrów obsługi klientów (*call centers*), tworzonych przez przedsiębiorstwa dla zapewnienia zdalnego (telefonicznego) kontaktu z klientami, np. w biurach podróży, liniach lotniczych, firmach ubezpieczeniowych czy bankowości. Rozwijane od lat osiemdziesiątych ubiegłego wieku i nabierające znaczenia wraz z rozprzestrzenieniem się telefonii komórkowej i Internetu, pracujące całą dobę centra tego typu, uwzględniające w swojej pracy dodatkowo zlecania mniejszych firm, stały się znaczącym elementem rynku usług, zatrudniającym w Polsce dziesiątki tysięcy, a w Unii Europejskiej miliony ludzi. Właściwy dobór rozmiaru (liczby czynnego w danym momencie wykwalifikowanego personelu odpowiadającego na pytania klientów) takiego centrum w funkcji zmiennego w czasie i mającego losowy charakter zapotrzebowania jest poważnym problemem ekonomicznym, bo większość kosztów funkcjonowania *call centers* to koszta osobowe.

Call center jest obiektem, którego pracę można w naturalny sposób odwzorować w modelu teorii obsługi, w którym strumień nadchodzących w losowych odstępach czasu klientów (zapytań) jest obsługiwany przez pewną liczbę równoległych stanowisk obsługi ze wspólną kolejką, w regulaminie której można zawrzeć reguły dostępu do stanowisk. Problem jest od około dwudziestu lat studiowany, lecz wciąż otwarty, a prezentowane modele różnią się przyjętymi założeniami i stopniem złożoności. Wiele istniejących modeli opiera się na założeniu stacjonarności: wyniki stacjonarnego modelu kolejkowego są wyznaczane niezależnie dla poszczególnych okresów planowania i dla uśrednionych wartości parametrów. Nie oddaje to dynamiki zmian obciążenia i wynikającego stąd stanu systemu, co zmniejsza dokładność modelu.

Analiza stanów nieustalonych jest w teorii kolejek znacznie bardziej złożona niż stanów ustalonych, co komplikuje modelowanie *call centers* i ogranicza rozmiar badanych systemów. Omawiana praca poszerza praktyczne możliwości modelowania i analizy tych systemów w przypadku stanów nieustalonych, wykorzystując niejednorodne w czasie łańcuchy Markowa i wprowadzając modyfikacje algorytmów numerycznych dla ich rozwiązywania, zwiększające ich efektywność obliczeniową i dokładność.

Struktura rozprawy

Praca jest napisana w języku agielskim, składa się 6 rozdziałów i liczy ok. 100 stron.

Rozdział 1 przedstawia problem, opisuje działanie centrum obsługi klientów i zasady oceny jego wydajności. Teoria kolejek ma swoje korzenie w liczących sto lat markowowskich modelach Erlanga dotyczących centrali telefonicznej z nieskończonym źródłem klientów i analogicznych modelach Engseta dla skończonego źródła klientów. Modele te można odnieść wprost do centrum obsługi klientów. Dodatkowe szczegóły mogą dotyczyć zachowania się klientów: czy rezygnują oni z obsługi w momencie nadejścia, gdy kolejka jest zbyt długa, albo zniecierpliwieni zbyt długim czasem czekania w pewnej chwili odchodzą, czy w przypadku niepowodzenia zgłaszają sie ponownie. W przypadku, gdy pojemność kolejki jest ograniczona i wszystkie w niej miejsca są zajęte, klienci nie są przyjmowani. Klienci moga być obsługiwani według kolejności nadejścia lub zgodnie z innymi regułami. Model powinien określić prawdopodobieństwo, że zgłoszenie będzie kolejkowane, określić rozkład czasu czekania i rozkład długości kolejki lub wartości średnie tych wielkości, prawdopodobieństwo, że klient zrezygnuje natychmiast lub zanim zostanie obsłużony, itp. Są to wielkości zależne od liczby równoległych stanowisk obsługi – im jest ich więcej, tym lepsza jakość obsługi, lecz wyższe koszty własne; model można wykorzystać do wyboru kompromisowego punktu pracy.

Rozdział 2 wprowadza podstawowe pojęcia dotyczące łańcuchów Markowa, służących do konstrukcji dyskutowanych modeli, omawiając łańcuchy z ciągłym (CTMC) i dyskretnym czasem (DTMC), o stałych (łańcuchy jednorodne) i zależnych od czasu (łańcuchy niejednorodne) parametrach. Dla określenia prawdopodobieństw stanów trzeba rozwiązać układy równań, których liczba odpowiada liczbie stanów. W przypadku analizy stanów ustalonych są to równania algebraiczne, w przypadku analizy stanów nieustalonych – różniczkowe. Brak pamięci w procesach Markowa powoduje, że w przypadku CTMC czas pobytu w stanie ma rozkład wykładniczy, czyli w klasycznych modelach czasy pomiędzy nadejściam klientów i czasy obsługi mają rozkłady wykładnicze. Po przedstawieniu procesu urodzin i śmierci oraz modelu $M/M/S/S/$, który pozwala obliczyć prawdopodobieństwo odrzucenia klienta (formuła B Erlanga) oraz modelu $M/M/S$ i związanej z nim formuły C Erlanga określającej prawdopodobueństwo kolejkowania, Autor przedstawia krótki przegląd modeli, w których odchodzi się od nierealistycznego założenia dotyczącego wykładniczego rozkładu obsługi (wprowadzając rozkłady typu fazowego) oraz wprowadza rozkład czasu zniecierpliwienia, wykładniczy w modelu $M/M/S/K+M$ lub ogólny w modelu $M/M/S/K+G$. Omawia też modele, w których Poissonowski strumień wejściowy ma intensywnośc zależną od liczby klientów w systemie typu $M(n)/M/S/K$, $M(n)/M/S/K+M$, $M(n)/M/S/K+G$, co pozwala oddać zniechęcenie klientów widzacych długą kolejkę. Są to wszystko modele odnoszące się do stanu ustalonego. Następnie Autor omawia możliwości rozwiązywania równań Chapmana-Kołmogorowa w przypadku zmiennych w czasie parametrów, a w szczególności algorytm uniformizacji, który jest alternatywą dla typowych numerycznych rozwiązań typu ODE. CTMC jest tutaj zastąpiony przez proces z dyskretnym czasem, w którym w jednym przedziale czasu jest możliwa tylko jedna tranzycja, a rozwiązania można poszukiwać w sposób typowy dla DTMC, a więc w sposób bardziej efektywny numerycznie. Dodatkowo można tutaj zastąpić niejednorodny CTMC serią jednorodnych DTMC.

Rozdział 3 dotyczy zaproponowanej modyfikacji metody uniformizacji i przedstawia główne

wyniki pracy. Przedstawia podstawowy algorytm uniformizacji i jego rozbudowę dla wykrywania stanów ustalonych, skupiając się na złożoności obliczeniowej i ograniczeniach błędów. Wprowadza modyfikacje dotyczące wykrywania stanów ustalonych oparte na zbieżności przyporządkowanego procesu DTMC i jego zastosowanie do niejednorodnego w czasie CTMC. Poprawa efektywności obliczeniowej wynika bezpośrednio ze zmniejszenia liczby iteracji podporządkowanego wektora DTMC (operacji mnożenia przez macierz): liczba iteracji potrzebnych do podjęcia decyzji o aproksymacji rozwiązania danego kroku wartością stanu ustalonego bywa znaczaco mniejsza od liczby iteracji w algorytmie klasycznym.

Dodatkowe usprawnienia wynikają z własności CTMC przedstawiających proces narodzin i śmierci.

Rozdział 4 przedstawia przeprowadzone eksperymenty numeryczne sprawdzające wpływ modyfikacji na wydajność obliczeń na podstawie dwu przykładów numerycznych. Jeden z modeli zakłada dyskretne zmiany obsady osobowej (liczby serwerów), drugi ciągłe zmiany natężenia strumienia zapytań w funkcji czasu. Dla zmian dyskretnych liczne wykresy przedstawiają zachowanie się systemu (średnia liczba zadań w systemie, prawdopodobieństwo natychmiastowej obsługi, prawdopodobieństwo odrzucenia) jak również charakterystykę modelu (błędy aproksymacji stanu ustalonego, liczba iteracji) w funkcji czasu w modelach $M/M/S_t/K$ oraz $M(n)/M/S_t/K$, a następnie w modelu $M(n)/M/S_t/K+M$. Podobne wyniki są następnie podane dla zmiennego w sposób płynny w czasie strumienia zgłoszeń i przy dużej liczbie serwerów, np. $S = 300$ i dodatkowej kolejce o maksymalnej pojemności 300 zgłoszeń, w modelach $M_t/M/S/K+M$ i $M_t(n)/M/S/K+M$. Analizowana jest zbieżność obliczeń w systemach z różnym obciążeniem, a zyski z modyfikacji są badane dla różnego rozmiaru systemów.

Rozdział 5 testuje zaproponowane podejście z wykorzystaniem dostępnych rzeczywistych danych dotyczących bankowego centrum usługowego. Uwzględniono rzeczywiste, monitorowane zmiany strumienia zgłoszeń, zmiany liczby czynnych doradców, liczbę oczekujących i obsługiwanych klientów, zanotowane prawdopodobieństwo natychmiastowej obsługi i analizowano wyniki modelu. Dla danych dotyczących tego centrum przy średniej liczbie 200 lub 300 i maksymalnej liczbie 500 agentów modyfikacje poprawiające efektywność operacji mnożenia wektora i macierzy pozwoliły na ok. 60-krotne zmniejszenie czasu obliczeń w porównaniu z ortyginalnym algorytmem uniformizacji, zastosowanie modyfikacji wykrywania stacjonarności pozwoliło dodatkowo dwukrotnie zwiększyć efektywność obliczeniową. Ma to duże znaczenie przy planowaniu obsady, wymagającym wielokrotnego przeliczania modelu przy różnych parametrach.

Rozdział 6 przedstawia wnioski.

Obszerna bibliografia zawiera ok. 120 pozycji i wskazuje na bardzo dobrą znajomość literatury. Również przedstawione na początku pracy rozpoznanie literaturowe jest bez zarzutu.

Praca napisana jest starannie, bardzo dobrym językiem. Dobrze dokumentowane są wykonane prace i eksperymenty numeryczne.

Rezultaty pracy

Pod względem teoretycznym praca proponuje modyfikację szeroko stosowanego algorytmu uniformizacji z wykrywaniem stacjonarności zaproponowanego przez Trivediego i in. w 1992 roku. Pozwala ona uniknąć błędu przedwczesnego wykrywania stacjonarności, analizując numerycznie funkcję konwergencji podporządkowanego łańcucha Markowa czasu dyskretnego w algorytmie uniformizacji, by określić przewidywany błąd zastąpienia rozwiązania nieustalonego

jego stacjonarnym przybliżeniem. Modyfikacja dodatkowo zwiększa efektywność modeli, w których rozkład stacjonarny może być obliczony z wydajnością lepszą niż metoda potęgowa. Zaproponowana heurystyka redukcji stanów pozwala zmniejszyć złożoność obliczeniową w stopniu zależnym od wielkości systemu, przy jednoczesnym ograniczeniu wzrostu błędu rozwiązania do z góry założonej wartości. Modyfikacja jest szczególnie korzystna dla niejednorodnych w czasie łańcuchów Markowa z czasem ciągłym, których stan jest bliski ustalonemu przez dużą część analizowanego okresu.

Pod względem praktycznym, w zastosowaniu do *call centers* zaproponowane modyfikacje znacząco (np. o dwa rzędy) poprawiają wydajność obliczeniową, co pozwala na modelowanie i optymalizację obsady dla praktycznie każdego spotykanego w rzeczywistości centrum obsługi klientów. Zastosowane podejście umożliwia łatwą implementację dowolnych Markowowskich modeli narodzin i śmierci, w tym również modeli zależnych od stanu, pozwalających na odwzorowanie złożonych rozkładów czasu obsługi i czasu cierpliwości do opuszczenia kolejki przed obsługą, czy rezygnację z obsługi przez klienta w momencie nadejścia z powodu zbyt długiej kolejki. Podkreślmy, że dotychczas spotykane modele oparte na niejednorodnych w czasie CTMC były ograniczone do $M/M/S$.

Wyniki opublikowane w pięciu artykułach w czasopismach, w tym w INFORMS Journal on Computing z impact factorem oraz przedstawione na trzech konferencjach.

Podsumowanie:

Uważam, że rozprawa doktorska mgr inż. Macieja Rafała Buraka spełnia warunki stawiane rozprawom doktorskim przez ustawę o stopniu i tytułach naukowych. Autor umiejętnie stawia i rozwiązuje oryginalny i trudny problem o dużym znaczeniu praktycznym. Wnioskuję o przyjęcie tej pracy jako rozprawy doktorskiej i dopuszczenie jej do publicznej obrony. Jeżeli są spełnione warunki formalne ustalone przez Radę, to przedstawiam także tę pracę do wyróżnienia ze względu na bardzo dobry poziom matematyczny, wzorową metodykę przeprowadzonych badań i użyteczność wyników.