POLSKO-JAPOŃSKA AKADEMIA TECHNIK KOMPUTEROWYCH

ポーランド日本情報工科大学

Warszawa, 29.01.2015

Dr hab. Adam Wierzbicki
Profesor Polsko-Japońskiej Wyższej Szkoły
Technik Komputerowych

OPINIA
O ROZPRAWIE DOKTORSKIEJ

MGR INŻ. ROBERTA KŁOPOTKA

"INVASIVE ANALYSIS OF SOCIAL NETWORKS"

1. **Jakie zagadnienie naukowe jest rozpatrzone w pracy i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?**

Praca dotyczy zagadnienia identyfikacji modelu wzrostu dwumodalnej sieci społecznej na podstawie bieżącego stanu tej sieci. Przewidywanie przyszłych własności sieci społecznych o strukturze dwumodalnej, takich jak np. sieci klient-kupiony produkt, artykuł-autor, ma liczne zastosowania w badaniach naukowych (np. w socjologii) oraz przemyśle (np. marketing). Dlatego wielu badaczy proponowało modele wzrostu takich sieci, starając się włączyć możliwości modelowania szeregu zjawisk, obserwowanych w rzeczywistych sieciach. W pracy rozpatrywany jest model CSUIM (Cold Start User-Item Model), uwzględniający efekt wpływu rekomendacji na rozwój sieci w warunkach tzw. zimnego startu (niewielkiej ilości produktów nabywanych przez poszczególnych klientów).

Tytuł pracy („Invasive Analysis of Social Networks") jest tłumaczony przez autora faktem, że jego model dotyczy sytuacji, w której sieć jest tworzona pod wpływem rekomendacji, które są jednocześnie obliczane na podstawie struktury tworzonej sieci. Rozpatrywany model CSUIM wykorzystuje efekt „preferential attachment", który jest zjawiskiem obserwowanym w realnych sieciach, ale który może być wzmocniony przez proste rekomendacje w warunkach „cold start".

Z praktycznego punktu widzenia cenną własnością modelu wzrostu byłaby możliwość odtworzenia jego parametrów z rzeczywistej obserwowanej sieci społecznej tak, aby móc przewidywać jej przyszłe własności. Niestety stopień złożoności modelu CSUIM powodował, iż dotychczas nie udało się uzyskać metody jego identyfikacji na podstawie sieci, co ograniczało stosowanie tego modelu jedynie do jakościowego opisu dynamiki sieci społecznej. W niniejszej rozprawie podjęto wyzwanie identyfikacji modelu CSUI z sieci. Celem było w pierwszej kolejności określenie z wystarczającą precyzją parametrów wzrostu dwudzielnej sieci społecznej w tym modelu, stosując mieszaninę symulacji, regresji i metod analitycznych, a następnie jakościowe uchwycenie faktu, że w pewnym momencie parametry modelu zostały zmienione. Osiągnięcie tych celów implikuje, że CSUIM jest przydatny do analizy ilościowej dwudzielnych grafów sieci społecznych.

regresji i metod analitycznych, a następnie jakościowe uchwycenie faktu, że w pewnym momencie parametry modelu zostały zmienione. Osiągnięcie tych celów implikuje, że CSUIM jest przydatny do analizy ilościowej dwudzielnych grafów sieci społecznych.

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**
Praca zawiera bibliografię, liczącą około stu pozycji. Jednak nie ma w pracy wyraźnie wydzielonego rozdziału poświęconego analizie literatury. Rozdział 2.1 zawiera krótki przegląd literatury dotyczącej modeli generujących sieci społeczne. Przegląd ten nie jest wyczerpujący i ma istotne braki. Głównym pominiętym wątkiem w literaturze są badania wskazujące na istnienie sieci, które nie mogą zostać wygenerowane przez model „preferential attachment" stosowany w pracy (sieci nie posiadające własności „scale free"). Przykładem są sieci współpracy naukowców, które wykazują własność odcięcia rozkładów stopnia (występowanie maksymalnej wartości stopnia):
   - M.E. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad.Sci. USA 98, 404-409, 2001

   Dalszy przegląd literatury na temat sieci społecznych o charakterystykach odmiennych od sieci "scale free" można znaleźć w książce:

   - M. Newman, A. Barabasi, D. Watts, The structure and dynamics of networks, Princeton University Press, 2006

3. **Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?**
Zagadnienie postawione w pracy polega na dopasowaniu parametrów modelu generującego sieć społeczną do zadanej sieci (dla której znany jest jej stan w wybranej chwili czasu). Autor nazywa to zagadnienie różnie, niekiedy „odwróceniem modelu", „odzyskiwaniem modelu" lub „identyfikacją modelu". W literaturze często spotyka się określenie „model fitting", czyli „dopasowanie modelu". Konkretniej, zagadnieniem rozpatrywanym w pracy jest dopasowanie modelu określonego jako CSUIM do zadanych sieci.

   Tak zdefiniowane zagadnienie badawcze zostało rozwiązane przy pomocy metody badawczej polegającej na symulacyjnym badaniu populacji sieci generowanych z tego samego modelu CSUIM i opracowaniu metody dopasowania modelu przez odwrócenie relacji między oryginalnymi parametrami modelu i obserwowalnymi metrykami sieci. Koncentrowano się na wykrywaniu parametrów ze stanu sieci w zadanej chwili czasu. Badano kilka różnych podejść:

   1. analityczne – wyprowadzenia wzorów na postać parametrów z matematycznego opisu dynamiki rozwoju grafu dwudzielnego w modelu CSUIM,

   2. uczenie maszynowe – prognoza parametrów z wybranych własności grafów dla różnych modeli uczenia maszynowego: regresja liniowa, maszyna wektorowa,

proces Gaussa i perceptron wielowarstwowy, przy czym danymi uczącymi były parametry użyte w generatorze CSUIM oraz globalne własności grafów wygenerowanych na bazie tych parametrów,

3. regresyjno-symulacyjne – polegające na odejściu od typowego dla uczenia maszynowego "czarnoskrzynkowego" ujęcia odwracania modelu i wniknięcia metodą analityczno-eksperymentalną w zależności pomiędzy parametrami modelu CSUIM a własnościami grafu oraz interferencji między tymi zależnościami, co prowadziło do lokalnej symulacji i lokalnego uczenia dla znanych własności grafu oraz kolejno identyfikowanych parametrów modelu.

W pracy zakłada się, że w analizowanej sieci społecznej, będącej przedmiotem procesu identyfikacji parametrów, spełnione są warunki modelu CSUIM, wobec czego nie weryfikuje się rzeczywistego spełnienia tych warunków w danych.

Ponadto zakłada się, że rozpatrywany graf sieci społecznej jest "wystarczająco duży", wobec czego nie analizuje się wpływu warunków początkowych (początkowej liczby węzłów i krawędzi), oraz zakłada się, że fluktuacje stochastyczne podczas generowania grafu mogą być pominięte. Z literatury (cytowanej także w pracy) wiadomo, że wpływ ten jest zauważalny dla małych grafów i jest on dość złożony.

Wreszcie zakłada się, że wpływ poszczególnych parametrów generatora na charakterystykę własności generowanych grafów da się od siebie odseparować, ale łatwo zauważyć, że parametry te wpływają w sposób matematycznie dość uwikłany, stąd zależności wykorzystywane w procesie identyfikacji modelu mają charakter przybliżony, a także ograniczony zakres obowiązywania.

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?**

Głównym wynikiem pracy jest dopasowanie modelu CSUIM – identyfikacja parametrów modelu z obserwowanego grafu. Dotychczas tak złożony model generatora nie został dopasowany do obserwowanej sieci społecznej.

- Wynik ten jest rozwiązaniem ciekawego i nietrywialnego zagadnienia teoretycznego z uwagi na złożony i dotychczas nie wyrażony w formie jawnej łączny rozkład prawdopodobieństwa struktury sieci generowanej z modelu CSUIM.
- Dopasowanie modelu pozwala na przewidywanie przyszłych bądź pośrednich stanów rozwoju sieci. Ponieważ CSUIM jakościowo dobrze opisuje wiele zjawisk w dwu-modalnych sieciach społecznych, więc otwiera to nowe możliwości analizy sieci społecznych.
- Dopasowanie modelu pozwala na wykrywanie ewentualnych zmian parametrów w trakcie rozwoju sieci.
- Dopasowanie modelu otwiera możliwości aplikacyjne - badania skuteczności oddziaływań zewnętrznych na sieć społeczną w aspekcie zmian parametrów rozwoju sieci.

Na ten rezultat złożyło się pogłębienie teoretycznego rozumienia modelu CSUI przez odkrycie przybliżonych matematycznych zależności między podstawowymi parametrami tego modelu na drodze eksperymentalno-badawczej.

- Udało się m.in. uzyskać wgląd w relację między parametrem procesu odbijania (ang. *bouncing*) a obserwowanymi miarami grafu (modularnością)
- Udało się pokazać w przybliżeniu liniową zależność własności rozkładu stopni wierzchołków od proporcji przełączania między preferencyjnym a równomiernym losowaniem końców krawędzi. W literaturze znane są analogiczne rezultaty dla mieszanin rozkładów wykładniczego i potęgowego, jednakże nie mogły być one bezpośrednio przeniesione na rozważany przypadek z uwagi na: (1) niespełnienie założenia o mieszaninie rozkładów, (2) dwu-modalność rozkładu, (3) potencjalny wpływ parametru mechanizmu odbijania. Dlatego empiryczne pokazanie tych własności jest nowym wynikiem.
- Udało się pokazać przybliżoną niezależność brzegowych rozkładów stopni wierzchołków obu modalności, co nie jest oczywiste z uwagi na wpływ parametru mechanizmu odbijania.

Stworzono analityczną procedurę identyfikacji parametrów modelu CSUIM oraz badano praktyczną przydatność uzyskanej metody. W literaturze wyprowadza się często modele zakładając pewne własności "w granicy" oraz stosując uciąglenia, np. ciągłość stopni wierzchołków. Niniejsze badanie pokazało, że takie założenia mogą być zbyt silne w konkretnym zastosowaniu identyfikacji. Ten negatywny wynik jest ważnym przyczynkiem do postulowania empirycznej weryfikacji takiej procedury wyprowadzania wzorów, co obecnie nie jest w literaturze powszechne.

Przeprowadzono studium klasycznej identyfikacji modelu metodami uczenia maszynowego. Choć wyniki były dokładniejsze, poddano je krytycznej analizie. Uzyskane wyniki wskazują, że metody uczenia maszynowego mogą nie być w stanie odkryć relacji liniowych wtedy, gdy są one uwikłane, wobec czego praca wskazuje, iż techniki uczenia maszynowego mogą być użyteczne w zastosowaniu lokalnym, jeśli nie sprawdzają się przy konstrukcji modelu globalnego.

5. **Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienie uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?**
   Rozprawa jest zwięzła i jasna, a jej redakcja jest poprawna.

6. **Jakie są słabe strony rozprawy i jej główne wady?**
   Praca ogranicza się do identyfikacji jednego tylko modelu sieci dwudzielnych, t.j. modelu CSUIM. Model ten, choć dość złożony i uwzględniający różne zjawiska rzeczywistych sieci społecznych, jest w wielu przypadkach dość sztywny. Główna jego wada to rozważanie wyłącznie grafów spójnych, podczas gdy dostępne i analizowane dane rzeczywiste charakteryzowały się istnieniem wielu niespójnych składowych. Dalej, założenie o dodawaniu tylko stałej liczby krawędzi w jednym kroku wydaje się nie być potwierdzane przez dane empiryczne, gdyż węzły

niskich stopni stanowią dużą część grafów, co by wskazywało, iż dodawana jest zmienna liczba krawędzi w każdym kroku, oscylując wokół pewnych wartości średnich. Ponadto należałoby zbadać, na ile założenie o dodawaniu krawędzi tylko z nowych węzłów jest słuszne.

Należy tu zaznaczyć, że model CSUIM wydaje się być nieadekwatny do wielu zastosowań opisywanych w pracy, a także do wielu realnych, dwudzielnych sieci społecznych. Rozważmy na przykład problem rekomendacji produktów dla użytkowników. Rozważany w pracy graf dwudzielny (trudno tu nawet posłużyć się terminem „sieci społecznej") składa się zatem z użytkowników i produktów. Jakie mogłyby być własności takiego grafu?

- Jeśli rozważamy dobra podstawowe, to każdy użytkownik musiałby je kupować w podobnej ilości. Różnica może dotyczyć wyłącznie marki tych dóbr. Jednak model generujący graf musiałby zakładać, że każdy użytkownik musi wybrać przynajmniej jedną ze zbioru marek produkujących to samo dobro podstawowe. Model CSUIM nie robi takiego założenia.
- Jeśli rozważymy dobra luksusowe, sytuacja jest odwrotna. Wyłącznie nieliczna grupa użytkowników kupowałaby takie dobra. Znów, model CSUIM nie bierze pod uwagę takiego ograniczenia, sprowadzającego się w zasadzie do ograniczenia ilości użytkowników, którzy mogą być połączeni z określonym towarem luksusowym.
- Podobne ograniczenie ilości użytkowników może dotyczyć także bardziej powszechnych towarów. Dobrym przykładem są wycieczki lub podróże, a także spektakle teatralne, kinowe itd. Dla każdego takiego towaru liczba użytkowników, którzy mogą go kupić, jest ograniczona.
- Model CSUIM nie bierze pod uwagę relacji rynkowych. Jeśli jakiś towar stawałby się bardzo popularny, co przewiduje model „preferential attachment", powinna rosnąć jego cena, a to z kolei powinno ograniczyć jego popularność.

Powyższa, pobieżna analiza problemu wskazuje, że dla zastosowania, jakim jest rekomendacja towarów, dóbr lub usług warto rozważyć zastosowanie zupełnie innych modeli sieci. Może to prowadzić autora w przyszłości do interesujących wyników naukowych.

Wreszcie ograniczono się wyłącznie do odzyskiwania parametrów grafu zadanego tylko w jednym punkcie czasu. Być może analiza wielu momentów czasu poprawiłaby dokładność odzyskiwania parametrów, choć niewątpliwie skutkowałaby większą złożonością obliczeń parametrów oraz o wiele trudniejszym uzasadnieniem matematycznym zakładanych własności i powiązań pomiędzy parametrami modelu a miarami pozyskanymi z takiego ciągu grafów.

7. **Jaka jest przydatność rozprawy dla nauk technicznych?**

Doświadczenia uzyskane w wyniku zastosowania wyżej opisanych metod badawczych posłużyły do opracowania autorskiego podejścia do odkrywania parametrów modelu z danych. Parametry są rekonstruowane etapowo, posługując się generatorem CSUIM do przeszukiwania przestrzeni kolejnego

parametru i budowania w niej lokalnego modelu analitycznego, posługując się analizą regresji. Podejście to można postrzegać szerzej jako przykład realizacji od dawna stawianego postulatu, aby z metodami uczenia maszynowego integrować wiedzę dziedzinową. W tym przypadku taką wiedzą było wykrycie pewnych prawie liniowych relacji między własnościami grafu a parametrami generatora, obowiązującymi jednakże w ograniczonym zakresie i tylko przy ustaleniu innych parametrów generatora, wpływających na relację w sposób nieliniowy, co przesłaniało tę zależność w wypadku "globalnego" podejścia metodą uczenia maszynowego. Ponadto studium pokazało, iż skuteczność uczenia maszynowego można podnieść poprzez integrację w szerszym strukturalizowanym procesie odkrywania wiedzy.

To ostatnie podejście stało się podstawą do eksploracji wykrywalności zmian parametrów w trakcie rozwoju sieci. Okazało się, że zmiany te da się zauważyć po wystarczająco długim okresie obserwacji.

Praca zawiera materiał doświadczalny weryfikujący skuteczność opracowanych metod identyfikacji. Dyskutuje się ich ograniczenia oraz wskazuje na kierunki dalszych badań.

Uzyskane wyniki pokazują, że możliwa jest wystarczająco dokładna identyfikacja parametrów modelu wzrostu sieci jako grafu dwudzielnego dla sieci zachowującej się podobnie jak model CSUIM poprzez użycie metod symulacyjno-regresyjnej i analitycznej oraz możliwe jest ilościowe zauważenie zmiany parametrów modelu. Ponieważ parametry generatora zdają się dobrze charakteryzować niektóre grafy dwudzielne, więc same parametry modelu CSUIM mogą być miarami stosowanymi do opisu grafów dwudzielnych np. dla poszukiwania grafów podobnych lub też dla odróżniania ich od siebie.

Możliwość wykrywania zmian parametrów modelu może w praktyce posłużyć jako np. sygnalizator niepokojących zmian w zachowaniu sieci, wskazując w przybliżeniu moment, w którym zmiana nastąpiła, co pozwoli na działania diagnostyczne. Może być też użyta do np. oceny skutków oddziaływania z zewnątrz na sieć, np. poprzez kampanię marketingową.

WNIOSKI

Podsumowując stwierdzam, że **rozprawa spełnia wymagania stawiające rozprawom doktorskim przez obowiązujące przepisy.**