20 marca 2022

Dr hab. Paweł Sobkowicz
Centrum Doskonałości NOMATEN
Narodowe Centrum Badań Jądrowych
Ul A. Sołtana 7, 04-500 Otwock Świerk

Recenzja

rozprawy doktorskiej mgr inż. Jana Chołoniewskiego, pt. "Modelling dynamics of news media", przedstawionej Radzie Naukowej Dyscypliny Nauki Fizyczne Politechniki Warszawskiej.

1. Wstęp

Recenzowana praca dotyczy dogłebnej analizy i modelowania światowych serwisów informacyjnych. Zgodnie z intencjami Autora, praca łączy aspekty socjologiczne, medioznawcze i analizę danych – ze szczególnym uwzględnieniem twórczego zastosowania metod fizyki statystycznej w celu wydobycia nietrywialnych prawidłowości charakteryzujących współczesne dziennikarstwo i procesy związane z publikowaniem informacji przez źródła profesjonalne. Tematyka jest istotnym uzupełnieniem wcześniejszych prac dotyczących tworzenia i rozchodzenia informacji w sieciach społecznościowych.

Rozprawa, przygotowana w języku angielskim ma 107 stron i podzielona jest na pięć rozdziałów, z czego pierwszy wprowadza tematykę i przedstawia zagadnienie badawcze, trzy kolejne omawiają szczegółowo uzyskane przez Autora wyniki a ostatni podsumowuje rezultaty pracy. Bibliografia rozprawy liczy 147 pozycji, a sama rozprawa oparta jest na wynikach uzyskanych z trzech publikacji w których Autor rozprawy jest pierwszym autorem. W dorobku publikacyjnym Autora znajdują się także dwie publikacje częściowo związane z tematyką rozprawy.

2. Najważniejsze elementy i osiągnięcia rozprawy

Przedstawiona do recenzji rozprawa ma nie tylko multidyscyplinarny charakter ale także umiejętnie łączy różne metodyki badawcze:

- Analizę złożonego zbioru danych o nietrywialnych charakterystykach (w tym niezbędne „czyszczenie" zbioru danych. Oznacza to silne osadzenie dalszych analiz na podstawach empirycznych – a jest to cecha, którą stosowanie metodologii nauk fizycznych wnosi jako istotną wartość do nauk społecznych.
- Zastosowanie zaawansowanych metod zaczerpniętych z fizyki statystycznej do wydobycia ze zbioru danych ukrytych prawidłowości, wykazania zarówno zgodności ze znanymi prawidłowościami (jak prawo skalowania fluktuacji) jak też analiza indywidualnych odchyleń od tych prawidłowości.
- Zaproponowanie modelu, mającego na celu odtworzenie zaobserwowanych prawidłowości, a zatem próba wskazania ich możliwych źródeł.

Co istotne, badania prowadzone przez Autora zaowocowały także powstaniem dedykowanego oprogramowania *NewsMapper* (https://newsmapper.sta.si/) pozwalającego na wykrywanie duplikatów wiadomości i analizę zjawiska takiego kopiowania.

Przedstawiona rozprawa wykorzystuje szereg zróżnicowanych narzędzi: NLP (Natural Language Processing), analizę szeregów czasowych, analizę statystyczną (w szczególności analizę zależności fluktuacji czasowych, ale także analizę ochyleń (*residuals*) od prawa skalowania Taylora, agentowy model oparty o sieciowa wersje modelu niezależnych kaskad, analizę PCA czy wreszcie analizę korelacji pomiędzy konceptami stanowiącymi podstawę klasyfikacji zbioru danych. Tak szeroka paleta narzędzi potwierdza opanowanie przez Autora zakresu umiejętności niezbędnych dla uzyskania stopnia naukowego doktora nauk fizycznych.

Najważniejsze osiągnięcia pracy

- Potwierdzenie, że podobnie jak ma to miejsce w przypadku sieci społecznych, także w przypadku publikacji o charakterze profesjonalnym (dziennikarskich, informacyjnych, portali agregujących itp.) w odniesieniu do kluczowych charakterystyk aktywności (liczby artykułów, liczby artykułów odnoszących się do pojedynczego wydarzenia (event) czy liczby podmiotów (*publishers*) uczestniczących w opisach takiego wydarzenia występują rozkłady typu heavy-tail (rys 2.3). Podczas gdy obserwacje takie były znane wcześniej dla zachowań użytkowników sieci społecznych, ich występowanie w omawianym środowisku nie jest wcale oczywiste.
- Potwierdzenie zgodności aktywności z prawem skalowania Taylora (Temporal Fluctuations Scaling, TFS). Podobnie jak w poprzednim przypadku, nie było to oczywiste (co, w pewnym sensie potwierdza analiza odchyleń od TFS przedstawiona w rozdziale 4). Prawo TFS wydaje się obowiązywać dla szerokiego zakresu okien czasowych i rozważanej tematyki (*concepts*), mimo, że jak pokazuje rysunek 2.2 poszczególne przebiegi czasowe dla danego tematu mogą być radykalnie różne dla różnych źródeł.
- Zdefiniowanie modelu agentowego mającego odtworzyć charakterystyki obserwowane w analizowanym zbiorze danych, opartego na założeniu niezależnych kaskad. W szczególności ciekawe jest wprowadzenie do modelu parametru odpowiadającego zwiększonej częstości kaskad dla określonych tematyk (*hype parameter*), koniecznego, by dopasować model do obserwacji.

Podsumowując, praca zawiera szereg różnorodnych i interesujących wyników. Dotyczących stosunkowo jeszcze nieadekwatnie zbadanego aspektu roli profesjonalnych źródeł tworzenie i rozprzestrzeniania informacji we współczesnej infosferze.

3. Uwagi szczegółowe

W tak rozbudowanej metodologicznie i treściowo rozprawie trudno uniknąć niedociągnięć i błędów redakcyjnych. Poniżej przedstawiam listę elementów, uszeregowanych w kategorii istotności.

- Rozdział 4, poświęcony odchyleniom od prawa TFS (*residuals*) zawiera znaczną liczbę analiz o charakterze technicznym. Jednak zdecydowanie brak jest wyjaśnienia, dlaczego indywidualne odchylenia obserwowanych aktywności od konkretnego modelu teoretycznego mają mieć istotne znaczenie. Sam model TFS wiąże średnie wartości aktywności o odchyleniami standardowymi. Co jednak mierzą odchylenie od tego prawa? Jaką informację wnoszą do zrozumienia opisywanego środowiska?
- W analizie wykładników prawa TFS, Autor wyróżnił trzy skale okna czasowego i dopasował trzy zależności liniowe wykładnika od wielkości okna (rys 2.5, tabela 2.5). Rozprawa nie podaje wystarczającego wyjaśnienia merytorycznego dla takiego podziału, a wykresy z rysunków 2.7 i 2.8 nie sugerują występowania dobrze zdefiniowanych reżimów czasowych.

  Wręcz przeciwnie, dane ze wspomnianych rysunków sugerują stosunkowo prostą zależność wykładnika α od wielkości okna czasowego Δ : $\alpha = \alpha_0 + A(log\ \Delta)^2$. Funkcja ta dla każdego ze słów kluczowych ma tylko dwa parametry, a wspomniane rysunki 2.7 i 2.8i sugerują stosunkowo zbliżone ich wartości dla różnych słów kluczowych (z nielicznymi wyjątkami). Unika się w ten sposób dość sztucznego podziału na trzy reżimy dopasowani liniowych, a jednocześnie powstaje ciekawe pytanie o przyczynę takiej zależności – której odtworzenie mogłoby być dodatkowym zadaniem dla modelu agentowego.
- Kluczowa dla zbioru danych definicja zdarzenia (*event*) realizowana przez algorytmy Event Registry nie jest dostatecznie jasno zdefiniowana (co zapewne wynika z natury algorytmu). Nie jest więc możliwe oszacowanie jakości procesu kategoryzacji poszczególnych artykułów – na przykład identyfikacji artykułów (*stories*) przypisanych do konkretnego zdarzenia, w tym ich ilości czy określenie występowania zarówno statystycznych jak systematycznych błędów w charakterystyce zbioru danych. Uwaga ta nie odnosi się do wkładu Autora, jednak w mojej opinii powinna zostać wspomniana przy wprowadzaniu zbioru danych Event Registry.
- W modelu agentowym użyta jest uproszczona struktura sieciowa źródeł informacji (*giant component of pruned graph*). Podejście takie niejako automatycznie pomija procesy cykliczne w sieci obejmującej wydawców

reprezentujących przeciwne (spolaryzowane) poglądy polityczne. Zarówno obserwacja rynku wydawniczego i dziennikarskiego, jak też dyskusji w sieciach społecznych wskazują, że polemiki (kłótnie) mogą być istotnym elementem napędzającym długoterminowe kaskady. Zatem potencjalnie model o bardziej realistycznej strukturze sieciowej i obecności polemik między spolaryzowanymi źródłami mógłby „zastąpić" model z siecią prostą i parametrem *hype*. Istotnym elementem w porównaniu tych dwóch modeli byłoby występowanie w modelu z polemikami różnic między tematami (słowami kluczowymi) neutralnymi i polaryzującymi. Brak dyskusji takiego podejścia jest o tyle zaskakujący, że rola negatywnych emocji była badana w grupie prof. Hołysta.

- Część rysunków w pracy przedstawiona jest w sposób nieczytelny lub utrudniający czytelnikowi analizę. Przykładowo: zmiana skali w dolnych panelach rys 4.12; nieczytelne schematy kolorów na rysunkach 2.7 i 2.8; błędny opis osi pionowej na rys 4.3 i brak konkluzji dotyczących przedstawianych tam danych; brak informacji o strukturze klastrowej tematów na wykresach pokazujących korelacje.

Zaznaczyć jednak trzeba, że opisane powyżej niedociągnięcia nie deprecjonują przedstawionych w rozprawie ciekawych wyników.

4. Wnioski końcowe

W mojej opinii rozprawa mgr inż. Jana Chołoniewskiego zawiera bogaty zestaw analiz dotyczących bogatego i złożonego zbioru danych. Część wyników uzyskanych przez autora rozprawy została przedstawiona w trzech międzynarodowych publikacjach. Praca w zupełności spełnia warunki stawianym rozprawom doktorskim. Stawiam wniosek o dopuszczenie pracy do publicznej obrony.

Paweł Sobkowicz

Dr hab. Paweł Sobkowicz