----------------------------------------------------------------------------------------------

STRESZCZENIE

AUTOR /nazwisko, imię /:

**Chołoniewski, Jan**

TYTUŁ :

**Modelling dynamics of news media**

STRONY:

**107**

PROMOTOR /imię, nazwisko, tytuł nauk./:

**Janusz, Hołyst, prof. dr hab.**

PROMOTOR POMOCNICZY /imię, nazwisko, tytuł nauk./:

**Julian, Sienkiewicz, dr inż.**

----------------------------------------------------------------------------------------------

STRESZCZENIE PRACY /14-16 wierszy/ Rodzaj pracy: doktorska

~~habilitacyjna~~

W pracy analizujemy i modelujemy działalność internetowych serwisów informacyjnych, łącząc ilościowe podejście medioznawcze z metodami fizyki statystycznej. Głównymi celami prezentowanych badań są: (i) charakterystyka dynamiki takich serwisów poprzez analizę statystyczną ich aktywności publikacyjnej z punktu widzenia prawa skalowania fluktuacji (prawa Taylora), (ii) znalezienie modelu agentowego odtwarzającego obserwowane własności statystyczne, oraz (iii) analiza rezyduów prawa skalowania fluktuacji w kontekście geografii i politycznej stronniczości internetowych serwisów informacyjnych.

Prezentowana rozprawa ma charakter interdyscyplinarny: łączy podejścia i techniki znane z fizyki statystycznej z najnowszymi osiągnięciami w obliczeniowych badaniach mediów, które leżą na przecięciu nauk społecznych i informatyki. Jednym z rezultatów przedstawionych badań było opracowanie komercyjnego oprogramowania dedykowanego agencjom prasowym, zaprojektowanego i napisanego przez autora tej rozprawy we współpracy z Słoweńską Agencją Prasową. Opracowana platforma Software-as-a-Service nosi nazwę NewsMapper; jej rozwój został wyróżniony na stronie inicjatywy Innovation Radar prowadzonej przez Komisje Europejską i otrzymał dodatkowe finansowanie ze środków Google Digital News Initiative.

Szybka cyfryzacja naszego życia codziennego stworzyła możliwości analizy wcześniej

niedostępnych obszarów. W szczególności, pojawienie się Web 2.0 i mediów społecznoś-

ciowych (np. Twitter, blogosfera, Facebook) dało podwaliny pod dyscyplinę obliczeniowych

nauk społecznych. W jej rozwój niemal od początku angażują się fizycy. Ich badania

obejmują między innymi kwantyfikację i modelowanie emocji i opinii w społecznościach internetowych, analizy sieci współpracy naukowej, obserwacje dynamiki języków czy

przetwarzanie i dyfuzja informacji. Medioznawstwo jest subdziedziną nauk społecznych.

Mając zatem dostęp do baz setek milionów artykułów i dysponując technikami przetwarzania języka naturalnego, wkraczamy na teren medioznawstwa obliczeniowego. Naszym

celem jest wzbogacenie tego podejścia o metody znane z fizyki statystycznej i fizyki

układów złożonych.

Media informacyjne stanowią ważną część współczesnego społeczeństwa, a Internet

stał się nie tylko doskonałą platformą do masowej wymiany informacji ale umożliwił także

łatwe obserwacje i modelowanie tego procesu. Wydaje się, że panuje konsensus co do tego,

że wymianę informacji między internetowymi serwisami informacyjnymi można opisać

za pomocą sieci. Metody modelowania tematycznego umożliwiły analizę długofalowych

trendów w zawartości publikowanej w informacyjnych mediach online oraz ich reakcje

na nagłe i znaczące wydarzenia. Cechy serwisów informacyjnych, takie jak stronniczość mediów czy struktura władzy/własności, mogą być kwantyfikowane. Rozprzestrzenianie się wiadomości może być modelowane za pomocą agentowych modeli epidemicznych,

jednak większość badań dotyczy rozprzestrzeniania fake newsów w mediach społecznoś-

ciowych.

Nasze badania dotyczą modelowania aktywności internetowych serwisów informacyjnych rozumianej jako liczba opublikowanych artykułów na dany temat w pewnym

oknie czasowym. Na początku pokazujemy, że internetowe serwisy informacyjne tworzą

specyficzny system złożony, w którym aktywność poszczególnych jednostek jest zgodna

z prawem skalowania fluktuacji Taylora. Następnie szczegółowo omawiamy zawartość

zbiorów danych i pokazujemy także jak wykładniki skalowania zależą od skali czasowej i

wielkości danej jednostki oraz opisujemy trzy charakterystyczne skale czasowe dla aktywności mediów informacyjnych.

Na podstawie tych wyników stawiamy hipotezę, że zaobserwowane skalowanie fluktuacji jest wynikiem nietrywialnej dynamiki kolektywnej, którą można modelować za

pomocą dość prostego podejścia agentowego. Bazując na modelu niezależnych kaskad

(ang. Independent Cascades Model) tworzymy model dynamiki serwisów informacyjnych,

który wykazuje skalowanie fluktuacji zbliżone do tego, które zaobserwowano w danych.

Na początku opisujemy kolejne kroki prowadzące do otrzymania danych wejściowych

do modelu. Po pierwsze, pokazujemy jak wykrywać pary podobnych artykułów używając metod przetwarzania języka naturalnego i jak wykorzystać tę metodę do śledzenia

przepływu informacji pomiędzy redakcjami. Po drugie, agregując korelacje publikowanych

treści tworzymy sieć serwisów informacyjnych i badamy jej podstawowe cechy. Następnie badamy skalowanie fluktuacji czasowych w modelu niezależnych kaskad i opisujemy

modyfikację konieczną do uzyskania nietrywianlej zależności wykładników skalowania od

rozmiaru okna czasowego.

Na koniec, badamy rezydua prawa skalowania fluktuacji dla poszczególnych jednostek układu. Określają one względną zmienność sygnałów generowanych przez te jednostki. Bazując na tej obserwacji, wprowadzamy nowatorską metodę badania układów

wykazujących skalowanie fluktuacji czasowych, opartą na agregacji odchyleń od prostej

dopasowania; następnie wykorzystujemy tę metodę do dalszej charakterystyki aktywności

internetowych serwisów informacyjnych w kontekście ich kraju pochodzenia, poruszanych

tematów, jak i stronniczości politycznej.

Podczas gdy prawa potęgowe i sieci są obecne w poprzednich analizach informacyjnych mediów online, prawo skalowania fluktuacji czasowych w danych dotyczących

aktywności serwisów informacyjnych zostało opisane po raz pierwszy. Co więcej, analiza zaproponowanego modelu nie tylko pokazuje jego jakościową zbieżność z danymi empirycznymi, ale wskazuje także na powszechną obecność skalowania fluktuacji w modelach

epidemicznych. Zbadaliśmy także jaką informację niosą rezydua prawa skalowania fluktuacji czasowych i zinterpretowaliśmy je dla przypadku aktywności internetowych serwisów

informacyjnych.

Prezentowana rozprawa ma charakter interdyscyplinarny: łączy podejścia i techniki znane z fizyki statystycznej z najnowszymi osiągnięciami w obliczeniowych badaniach mediów, które leżą na przecięciu nauk społecznych i informatyki. Wierzymy, że

prezentowana tu praca jest nowym przykładem data science wykonywanej z pozycji fizyka.

Otrzymane przez autora teoretyczne i praktyczne wykształcenie w zakresie analizy systemów złożonych pozwaliło na wykorzystanie metod opartych o sieci (jak ICM) czy analizy fluktuacji (jak prawo TFS i reszty do niego). W efekcie stworzyliśmy nową wiedzę

na temat statystycznych własności produkcji newsów online, modelowania aktywności

serwisów informacyjnych online oraz zastosowań prawa skalownia fluktuacji czasowych.

Jednym z rezultatów przedstawionych badań było opracowanie komercyjnego oprogramowania dedykowanego agencjom prasowym, zaprojektowanego i napisanego przez

autora tej rozprawy we współpracy z Słoweńską Agencją Prasową. Opracowana platforma Software-as-a-Service nosi nazwę NewsMapper; jej rozwój został wyróżniony na

stronie inicjatywy Innovation Radar prowadzonej przez Komisje Europejską1 i otrzymał

dodatkowe finansowanie ze środków Google Digital News Initiative. W momencie pisania

tego streszczenia, platforma NewsMapper była subskrybowana przez trzy krajowe agencje

prasowe (słoweńską, polską, chorwacką), kilka słoweńskich portali informacyjnych oraz

jednostkę rządową. Podstawową funkcjonalnością platformy jest wykrywanie duplikatów

treści klientów na innych stronach internetowych i agregowanie danych o duplikatach w

celu uzyskania różnych statystyk wykorzystania treści przez osoby trzecie (w czasie, ale

także według kategorii, tematu, autora, kanału, pory dnia, dnia tygodnia itp.) Każdego

dnia program znajduje około 2 000 duplikatów w około 10 000 artykułów opublikowanych

na ponad 600 domenach w 4 językach.