Wrocław, 20.04.2022

D | Streszczenie

Metoda naukowego dociekania prawdy jest jednym z najważniejszych osiągnięć ludzkości. Jest to jasno zdefiniowana ścieżka postępowania która prowadzi do prawd o wszechświecie przez powtarzalne, weryfikowalne kroki. Na przestrzeni dekad stosowanie metody naukowej pozwoliło naukowcom znaleźć rozwiązania coraz to bardziej skomplikowanych problemów, z każdym krokiem otwierając jednak nowe pytania. Z biegiem czasu tak kumulujące się odkrycia wygenerowały również ogromne zbiory publikacji naukowych. Był to zaczątek dziedziny scientometrii, meta-nauki skupionej na analizie i zrozumieniu nauki jako procesu, a także ludzkiej z nim interakcji. Dziedzina ta z biegiem lat opierała się coraz bardziej na zgromadzonych danych i zaadoptowała wiele technik z innych dziedzin naukowych, takich jak informatyka czy nauka o sieciach. Wraz z przyrostem danych o nauce, więzi między wspomnianymi dyscyplinami stają się coraz mocniejsze i dają początek nowym, bardziej zaawansowanym podejściom wymaganym do rozwiązania problemów z jakimi boryka się obecnie analiza nauki.

D.1 Ogólna Charakterystyka Dziedziny

W ogólnym ujęciu scientometria jest interdyscyplinarną dziedziną skupioną na analizie procesów zachodzących w nauce. Głównymi jej zadaniami jest odnajdywanie i rozumienie zjawisk oraz relacji zachodzących w ramach szeroko pojętej działalności naukowej całej ludzkości. O ile intuicyjnie nauka może być rozumiana jako proces złożony głównie z badań i eksperymentów, tak abstrakcyjne rozumienie zjawiska nie jest użyteczne w analizach.

Dla potrzeb problemowego podejścia do nauki, musi ona być rozumiana w dobrze zdefiniowanej i policzalnej formie. Odwołać się tutaj można do tego jak

badania nad procesem ewolucji prowadzone są na bazie obserwowania jej efektów – ta sama idea zastosowana może zostać tutaj. Nauka o nauce musi zatem opierać się na produktach nauki: zweryfikowanych w procesie recenzji publikacjach naukowych.

W tym ujęciu, publikacja naukowa jest podstawowym elementem składowym policzalnego wyniku działania nauki. Publikacja jest tekstowym dokumentem, który wprowadza do nauki między innymi nowe modele, teorie i podejścia do problemów. Z praktycznego punktu widzenia, operowanie na obiektach jakimi są publikacje jest znacznie prostsze. Dodatkowo, praktyka cytowania innych prac naukowych zapewnia publikacjom dobrze zdefiniowany kontekst w czasie, a także w przestrzeni badań naukowych. Struktura cytowań nadaje publikacyjnej reprezentacji nauki porządku i zapewnia szkielet, który może zostać poszerzany przez dodatkowe dane, jak np. o autorach publikacji czy miejscach ich opublikowania, a także o dodatkowe powiązania które lepiej reprezentują skomplikowaną strukturę procesu naukowego.

Tak uzyskana reprezentacja jest jednym z najważniejszych obiektów badań w scientometrii, a także jednym z głównych czynników kształtujących ją jako naukę interdyscyplinarną. Metody analizy nauki czerpią z teorii grafów, analizy sieci, uczenia maszynowego, przetwarzania języka naturalnego i innych dziedzin, tworząc mocny i zróżnicowany zestaw narzędzi do zrozumienia wielu oblicz badań naukowych w skali globalnej. Waga tych metod rośnie z czasem oraz ilością dostępnych danych, znajdując coraz szersze zastosowania we wspomaganiu podejmowania decyzji przez ludzi i jednostki powiązane z nauką.

Poniżej omówione zostaną wybrane koncepty i problemy scientometrii.

D | Streszczenie

Metoda naukowego dociekania prawdy jest jednym z najważniejszych osiągnięć ludzkości. Jest to jasno zdefiniowana ścieżka postępowania która prowadzi do prawd o wszechświecie przez powtarzalne, weryfikowalne kroki. Na przestrzeni dekad stosowanie metody naukowej pozwoliło naukowcom znaleźć rozwiązania coraz to bardziej skomplikowanych problemów, z każdym krokiem otwierając jednak nowe pytania. Z biegiem czasu tak kumulujące się odkrycia wygenerowały również ogromne zbiory publikacji naukowych. Był to zaczątek dziedziny scientometrii, meta-nauki skupionej na analizie i zrozumieniu nauki jako procesu, a także ludzkiej z nim interakcji. Dziedzina ta z biegiem lat opierała się coraz bardziej na zgromadzonych danych i zaadoptowała wiele technik z innych dziedzin naukowych, takich jak informatyka czy nauka o sieciach. Wraz z przyrostem danych o nauce, więzi między wspomnianymi dyscyplinami stają się coraz mocniejsze i dają początek nowym, bardziej zaawansowanym podejściom wymaganym do rozwiązania problemów z jakimi boryka się obecnie analiza nauki.

D.1 Ogólna Charakterystyka Dziedziny

W ogólnym ujęciu scientometria jest interdyscyplinarną dziedziną skupioną na analizie procesów zachodzących w nauce. Głównymi jej zadaniami jest odnajdywanie i rozumienie zjawisk oraz relacji zachodzących w ramach szeroko pojętej działalności naukowej całej ludzkości. O ile intuicyjnie nauka może być rozumiana jako proces złożony głównie z badań i eksperymentów, tak abstrakcyjne rozumienie zjawiska nie jest użyteczne w analizach.

Dla potrzeb problemowego podejścia do nauki, musi ona być rozumiana w dobrze zdefiniowanej i policzalnej formie. Odwołać się tutaj można do tego jak

badania nad procesem ewolucji prowadzone są na bazie obserwowania jej efektów – ta sama idea zastosowana może zostać tutaj. Nauka o nauce musi zatem opierać się na produktach nauki: zweryfikowanych w procesie recenzji publikacjach naukowych.

W tym ujęciu, publikacja naukowa jest podstawowym elementem składowym policzalnego wyniku działania nauki. Publikacja jest tekstowym dokumentem, który wprowadza do nauki między innymi nowe modele, teorie i podejścia do problemów. Z praktycznego punktu widzenia, operowanie na obiektach jakimi są publikacje jest znacznie prostsze. Dodatkowo, praktyka cytowania innych prac naukowych zapewnia publikacjom dobrze zdefiniowany kontekst w czasie, a także w przestrzeni badań naukowych. Struktura cytowań nadaje publikacyjnej reprezentacji nauki porządku i zapewnia szkielet, który może zostać poszerzany przez dodatkowe dane, jak np. o autorach publikacji czy miejscach ich opublikowania, a także o dodatkowe powiązania które lepiej reprezentują skomplikowaną strukturę procesu naukowego.

Tak uzyskana reprezentacja jest jednym z najważniejszych obiektów badań w scientometrii, a także jednym z głównych czynników kształtujących ją jako naukę interdyscyplinarną. Metody analizy nauki czerpią z teorii grafów, analizy sieci, uczenia maszynowego, przetwarzania języka naturalnego i innych dziedzin, tworząc mocny i zróżnicowany zestaw narzędzi do zrozumienia wielu oblicz badań naukowych w skali globalnej. Waga tych metod rośnie z czasem oraz ilością dostępnych danych, znajdując coraz szersze zastosowania we wspomaganiu podejmowania decyzji przez ludzi i jednostki powiązane z nauką.

Poniżej omówione zostaną wybrane koncepty i problemy scientometrii.

D.1.1 Dane Bibliograficzne

Pojęcie danych bibliograficznych ukształtowało się w ramach długiej historii dokumentowania dzieł naukowych (Clarke, 2015). Jest ono zazwyczaj rozumiane jako dane odnoszące się do publikacji naukowych, w szczególności danych na poziomie pojedynczych publikacji.

Taka jednostka publikacyjna opisana być może atrybutami takimi jak tytuł, abstrakt, słowa kluczowe, autorzy, czy też data i miejsce publikacji. Od danych tego typu dodatkowo oczekuje się wyczerpującej informacji na temat publikacji cytowanych przez dany dokument. Ta ostatnie cecha jest powodem, dla którego dane bibliograficzne naturalnie kojarzyć się mogą z grafami.

Część źródeł oferuje dane bibliograficzne w postaci grafowej, co często wiąże się z zastosowaniem przez dane źródło nowoczesnej myśli wedle której dane bibliograficzne reprezentowane są jako zbiory obiektów opisanych atrybutami. Takie rozwiązanie pozwala bytom innym niż same dokumenty (np. autorom czy czasopismom) być reprezentowanymi zarówno jako obiekt jak i atrybut. W tej postaci są one zdefiniowane przez swoje powiązania z innymi elementami zbioru danych. Ta współczesna forma przedstawienia danych nie posługuje się zatem bardziej tradycyjnym rozumieniem publikacji jako pojedynczego rekordu w tabeli czy też bazie danych.

Korzyścią wynikającą z tej formy reprezentowania danych jest fakt, iż otrzymana struktura grafowa może być rozwinięta do postaci obejmującej całą przestrzeń publikowania. Sieci heterogeniczne skonstruowane na bazie danych bibliograficznych zawierają wiele typów relacji, reprezentując zatem złożone zależności między wieloma obiektami. Otwiera to dodatkowe możliwośći analizy i wnioskowania, tym samym oferując szerszą perspektywę na naukę jako zjawisko widziane przez

pryzmat danych o publikacjach.

D.1.2 Sieci

Ponieważ problemy scientometrii koncentrują się na danych powiązanych relacjami, teoria grafów i analiza sieci znajdują szerokie zastosowanie w tej dziedzinie. Analiza sieci jest zastosowaniem teorii grafów do grup obiektów połączonych wiązaniami. Teoria grafów pierwotnie używana była głównie w fizyce, jednakże znalazła późniejsze zastosowanie w informatyce do szerokiej gamy problemów dzięki coraz wydajniejszym algorytmom oraz urządzeniom o większej zdolności obliczeniowej. Analiza sieci stała się z czasem szeroko wykorzystywanym, interdyscyplinarnym zbiorem narzędzi znajdującym zastosowanie w zróżnicowanych problemach. Przykładami tych zastosowań jest analiza sieci społecznościowych (Butts, 2008), zarządzanie (Zheng et al., 2016), psychopatologia (McNally, 2016) oraz główny temat zainteresowania tej pracy: analiza nauki.

Analiza sieci zajmuje się grafami: matematycznymi obiektami które reprezentują obiekty powiązane ze sobą za pomocą wiązań relacji. Najbardziej podstawowy graf składa się zatem ze zbioru obiektów (nazywanych wierzchołkami) oraz ze zbioru relacji (krawędzi). Każda krawędź zdefiniowana jest przez dwa wierzchołki które łączy.

Definition 19 *Graf $\mathcal{G}$ jest zbiorem wierzchołków i krawędzi $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ takich, że $e_{ij} = (v_i, v_j)$ gdzie $v_i, v_j \in \mathcal{V}, v_i \neq v_j$ oraz $e_i \in \mathcal{E}$.*

Prostym czynnikiem rozróżniającym dwa typy grafów jest skierowanie: graf może być skierowany lub nie. W grafie skierowanym, poprzednio zdefiniowana krawędź $e_{ij} = (v_i, v_j)$ oznaczałaby jedynie połączenie wychodzące z v_i i wchodzące do v_j. Grafy skierowane są często używane do reprezentacji obiektów takich jak pub-

likacje naukowe – w ich przypadku, relacja cytowania jest precyzyjnie określona jako zaczynająca się w jednym punkcie, a kończąca w drugim. Należy zwrócić uwagę, że cytowania nie mogą się zapętlać, a zatem $v_i \neq v_j$.

Definition 20 *Graf skierowany $\mathcal{G}$ to graf w którym $e_{ij} \neq e_{ji}$.*

Definition 21 *Graf skierowany $\mathcal{G}$ to graf w którym zdefiniowane są dodatkowo dwa mapowania: $init : E \rightarrow V$ oraz $ter : E \rightarrow V$. Mapowania te przypisują wierzchołek początkowy $init(e)$ oraz wierzchołek końcowy $ter(e)$ do każdej krawędzi e. Krawędź taka jest zatem skierowana z $init(e)$ do $ter(e)$ (**Diestel, 2017**).*

Kolejną własnością rozróżniającą grafy jest waga. Wcześniej zdefiniowaliśmy graf bez wag, a zatem taki, w którym wszystkie krawędzie mają taką samą wartość. Przypisując tym krawędziom inne wagi (ważąc je) możemy reprezentować bardziej zróżnicowane relacje w grafie. Przykładem wykorzystania grafów ważonych w analizie nauki jest tworzenie cieci współautorstwa. W takiej sieci, autorzy publikacji naukowych powiązani są krawędzią, jeżeli pracowali razem nad przynajmniej jedną publikacją, a wagą takiej krawędzi jest równa liczbie prac napisanych wspólnie.

Definition 22 *Ważony graf $\mathcal{G}$ to taki graf, w którym każda krawędź ma przypisaną wagę $\omega(e) \in \mathbb{R}$ (**Fletcher et al., 1991**).*

Dla zachowania spójności oznaczeń, możemy myśleć o grafie nieważonym jako o grafie w którym $\omega(e) = 1 \; \forall \; e \in \mathcal{E}$.

Powyższe definicje są wystarczające dla wprowadzenia pojęcia grafu, ale nie wyczerpują wszystkich możliwych ich typów. Przykładowo, zarówno wierzchołki jak i krawędzie mogą być dodatkowo opisane dodatkowymi cechami. Bardziej za-

awansowane typy sieci to np. sieci heterogenicznie, w których występuje wiele typów wierzchołków i krawędzi, a także sieci wielodzielne – sieci, które podzielone zostać mogą na wiele całkowicie niezależnych podsieci.

D.1.3 Wybrane Problemy w Analizie Nauki

Ponieważ rozprawa ta zajmuje się metodami przetwarzania danych pod kątem zastosowania w procesie wspomagania decyzji podejmowanych przez naukowców, konieczne jest także spojrzenie na różnorodne problemy występujące w analizie nauki i scientometrii.

Każdy element wkładu zaprezentowanego w tej pracy naukowej odnosi się do co najmniej jednego z problemów omawianych poniżej. Może to być bezpośrednie rozwiązanie danego problemu lub jedynie wykorzystanie jednego z podejść do tych zagadnień w ramach prezentowania bardziej skomplikowanych rozwiązań. Poniższe sekcje wprowadzą te zadania w postaci formalnej, lecz częściowo zgeneralizowanej, tym samym oferując dostateczne zrozumienie zagadnień, gdy zostaną one wspomniane w dalszej części rozprawy.

D.1.3.1 Automatyczna Ocena Jakości

Ocena jakości to problem ewaluacji obiektu w celu ustalenia stopnia w jakim obiekt ten odpowiada zdefiniowanym normom i kryteriom jakości.

Ocena jakości jest zazwyczaj rozumiana przez pryzmat zastosowań komercyjnych, w których jest ona częścią procesu zapewniania (kontroli) jakości. Problem ten może być jednak rozpatrywany z innych perspektyw, w tym naukowej. Dobrym przykładem jest ocena jakości obrazów która znajduje zastosowanie np. w systemach biometrycznych, a na temat której prowadzi się wiele badań i publikuje liczne prace

Z formalnego punktu widzenia, ocena jakości to proces mapowania zbioru cech wejściowych

$$B = \begin{pmatrix} b_1 & b_2 & \dots & b_n \end{pmatrix} \tag{D.1}$$

na zbiór ocen zgodności z normami

$$B \to S, S = \begin{pmatrix} s_1 & s_2 & \dots & s_m \end{pmatrix} \tag{D.2}$$

O ile nie ma ściśle zdefiniowanej zależności pomiędzy oceną zgodności a kontrolą jakości, to w literaturze powiązanie między nimi jest wskazywane (Liepiņa et al., 2014). Dla naszych potrzeb rozumieć będziemy oceny zgodności jako odpowiadające pewnym aspektom jakości danego obiektu. Możemy w związku z tym postulować, że oceny zgodności mogą być agregowane do pojedynczej oceny jakości q, przy czym metoda agregacji jest w pełni zależna od samego zadania oceny jakości. Biorąc pod uwagę, że jakość jest definiowana w zależności od stopnia spełniania kryteriów jakości, sama ocena jakości q może być normalizowana:

$$\tilde{q} \in [0, 1] \tag{D.3}$$

Ocena jakości może być prowadzona przez człowieka lub przez zautomatyzowany proces przystosowany do tego zadania. Głównym powodem zastosowania ręcznej (ludzkiej) oceny jakości jest brak umiejętności metod automatycznych do postrzegania obiektów w ten sam sposób co ludzie. np. w przypadku ręcznie pisanego tekstu (M. Zhang, 2013). Różnice te prowadzą do sytuacji w której zadaniem automatycznych procesów ewaluacji jest emulacja ludzkich zachowań, czego przykład

znaleźć można np. w ocenie jakości obrazów (Esses et al., 2018)(Yu et al., 2017).

Alternatywą jest proces uczenia nadzorowanego, w którym zautomatyzowana procedura oceny jakości uczy się przybliżania oceny jaką przyznałby człowiek bez próby replikowania ludzkiego spojrzenia na obiekt. Takie podejście ma zastosowanie np. w rozumieniu kontekstu i znaczenia tekstu na potrzeby tłumaczenia między językami (Chatzikoumi, 2020).

D.1.3.2 Predykcja cytowań

Predykcja cytowań to problem polegający na wykorzystaniu zbiory cech charakteryzujących publikację naukową w celu ustalenia liczby cytowań jakie publikacja ta otrzyma do ustalonego momentu w czasie.

W większości badań nad tym problemem, miara cytowań przybiera jedną z dwóch postaci: pojedynczej, skumulowanej liczby cytowań po czasie τ od opublikowania, lub serii czasowej cytowań.

W toku serii momentów czasu $\tau_0, \tau_1, ..., \tau_n$ publikacja naukowa otrzyma $\lambda_0, \lambda_1, ..., \lambda_n$ cytowań. W przypadku predykcji cytowań jako serii czasowej, zadanie jest zatem zdefiniowane poprzez założenie wartości $k < n$ do oznaczenia przez jak wiele lat cytowania były obserwowane. Zadaniem jest wykorzystanie zaobserwowanej serii cytowań $\lambda_0, \lambda_1, ..., \lambda_k$ aby przewidzieć przyszłą serię $\lambda_k, \lambda_{k+1}, ..., \lambda_n$ (Abrishami & Aliakbary, 2019).

Jeżeli zadaniem predykcji jest zwrócenie jednej liczby, tj. sumy cytowań po danym okresie, cel oznaczyć można jako

$$\Lambda_{k+1,n} = \sum_{i=k+1}^{n} \lambda_i \tag{D.4}$$

Parametry wejściowe w takim przypadku mogą być zarówno serią czasową jak i następującym agregatem:

$$\Lambda_{0,k} = \sum_{i=0}^{k} \lambda_i \tag{D.5}$$

Poza danymi o przeszłych cytowaniach, dane wejściowe zawierać mogą wiele cech które zidentyfikowano lub zakłada się jako związane z liczbą cytowań, np. czasopismo w którym publikacja została zamieszczona (Bornmann et al., 2014) czy cechy charakteryzujące współautorstwo publikacji (Sarigöl et al., 2014). Przy wykorzystaniu takich dodatkowych cech, problem przybiera następującą postać:

$$\begin{pmatrix} b_1 & b_2 & \dots & b_n & \lambda_0 & \lambda_1 & \dots & \lambda_k \end{pmatrix} \rightarrow \Lambda_{k,j} \tag{D.6}$$

Problem przewidywania cytowań może zostać również przedstawiony w postaci problemu klasyfikacji w którym kolejne klasy użyte są do zaznaczenia, że publikacja osiągnęła nie mniej niż graniczną dla klasy liczbę cytowań. Takie ujęcie problemu nie ma zasadniczego wpływu na postać danych wejściowych, ale stanowczo ogranicza przestrzeń wyjściową.

D.1.3.3 Identyfikacja tematów

Identyfikacja tematów to zadanie automatycznego odnalezienia zbioru tematów poruszanych w jednym lub więcej obiektów tekstowych, a także przypisania każdemu z tych obiektów tematów które są w nim poruszane.

Intuicyjnie, temat rozumieć można jako zbiór charakterystycznych słów – dostatecznie precyzyjny zbiór w takiej formie jest rozpoznawalny dla człowieka jako temat. Najprostszym podejściem do identyfikacji tematów jest zatem analiza

wysokiej frekwencji występowania słów. Te najpopularniejsze słowa mogą być wystarczające do określenia tematów w danej dziedzinie.

Bardziej zaawansowanym rozwiązaniem jest analizy współwystępowania słów. Dla dostatecznie dużego zbioru obiektów tekstowych możliwe jest wyliczenie miary współwystępowania dowolnej pary słów.

Współwystępowanie κ_{ij} słów i oraz j to liczba wystąpień dokumentów które zawierały lub były opisane zarówno słowem i jak i j.

Ponieważ sama liczba współwystąpień nie jest porównywalna w skali danego zbioru oraz pomiędzy zbiorami, racjonalną decyzją jest wprowadzenie współczynnika współwystępowania (**Callon et al., 1991**):

$$\varepsilon_{ij} = \frac{\kappa_{ij}^2}{\kappa_i \times \kappa_j} \tag{D.7}$$

gdzie κ_i oraz κ_j wyrażają liczbę dokumentów które zawierały lub były opisane odpowiednio terminami i oraz j.

Z pomocą tego proporcjonalnego współczynnika zbudować możemy macierz współwystępowania wypełnioną wartościami dla każdej pary słów. Taka struktura może zostać również przedstawiona w formie ważonego grafu, co umożliwia zastosowanie procesu gronowania (klastrowania) dla znalezienia struktur odzwierciedlających tematy.

Gronowanie lub klastrowanie zbioru obiektów jest procesem podziału tych obiektów na grupy (grona, klastry) tak, aby wszystkie obiekty w danej grupie były bardziej podobne do siebie – wedle zdefiniowanego kryterium – niż do dowolnego elementu spoza ich zbioru.

Proces gronowania jest tu zatem uwarunkowany dwoma głównymi czynnikami.

Pierwszym z nich jest zastosowany algorytm gronowania, np. k-średnich, gronowania hierarchicznego czy dwumodowego. Wybrana metoda gronowania będzie miała wpływ na sposób w jaki obiekty przypisywane są do zbiorów zgodnie z drugim głównym czynnikiem – miarą podobieństwa. Zastosowana miara definiuje praktyczne rozumienie tego, co oznacza że dane obiekty znajdują się w tym samym zbiorze. Przykładowo, jeżeli obiektem jest dokument tekstowy a miarą podobieństwa jest podobieństwo semantyczne, to grona skupiać będą dokumenty o podobnym słownictwie i poziomie językowym.

Wracając natomiast to współwystępowania, każde grono znalezione przez tą procedurę zawierać będzie zbiór bardzo często występujących ze sobą słów. Jeżeli powołamy się na poprzednio przytoczoną intuicję, zrozumiemy takie grono jako obiekt reprezentujący temat w tekście. Warte zauważanie jest tutaj, że zadanie to może zmienić się w zależności od danych – zbiór bogaty w interdyscyplinarne publikacje może wymagać dodatkowego pochylenia się nad słabo współwystępującymi terminami które nadal mogą mieć znaczenie (K. Dong et al., 2018).

W danych bibliograficznych powiązania między dokumentami mogą zostać dodatkowo zaobserwowane na bazie cytowań – zarówno bezpośrednich, jak i w postaci współcytowania oraz sprzężenia bibliograficznego. Te trzy relacje są znanymi metodami rozpinania sieci cytowań na zbiorze publikacji. Gronowanie przeprowadzone może zostać na dowolnej z tych struktur, nawet bez użycia innych metryk podobieństwa dokumentów. Takie gronowanie również prowadzić może do struktur odzwierciedlających tematy naukowe. Zadanie zmienia się wtedy w problem zidentyfikowania jaki temat dana struktura sobą reprezentuje.

Zastosowanie tutaj ma zupełnie inne podejście do identyfikacji tematów. Probabilistyczne modele tematów można ogólnie scharakteryzować na przykładzie LDA (Blei et al., 2002), który jest klasycznym modelem tego typu. Warto za-

uważyć, że pomimo zupełnie innego podejścia tu zastosowanego, intuicyjne rozumienie tematu jako zbioru słów jest nadal w użyciu.

Generatywne modele tematów zakładają, że zbiór dokumentów D jest rozkładem wielomianowym na zbiorze tematów T. Każdy temat $t \in T$ jest z kolei rozkładem wielomianowym na zbiorze słów W. Rozkłady te nie są znane, a zadaniem modelowania tematów jest ich inferencja.

Proces LDA jest warunkowany na rozkładzie Dirichleta oraz trzech założeniach a priori: właściwych dla rozkładu Dirichleta α i β, a także założonej liczbie tematów k występujacych w zbiorze. Tematy reprezentuje się zazwyczaj poprzez wymienienie n słów w o najwyższym prawdopodobieństwie wystąpienia w temacie t:

$$P(w|t) = \frac{N_{wt} + \beta}{\sum_w N_{wt} + W\beta} \tag{D.8}$$

Gdzie N_{wt} oznacza sumę liczby przypisań słowa w do tematu t. Model prowadzi zatem wnioskowanie od końca, rekonstruując rozkłady które wygenerowały zbiór dokumentów na bazie wyniku ich działania. Po zakończeniu tego procesu, nauczony model może być wykorzystany do wnioskowania o tematach poruszanych przez dokument który nie znajdował się w pierwotnym zbiorze dokumentów.

D.1.3.4 Ewolucja i Trendy Tematów

Rozszerzeniem zadania identyfikacji tematów jest śledzenie ich rozwoju w czasie. Śledzenie ewolucji tematu to zadanie stworzenia reprezentacji dyskretnych stanów tego tematu dla każdego zdefiniowanego momentu w czasie, z uwzględnieniem relacji między tymi stanami które wskazywałyby na kierunek i rodzaj zmian.

Jako przykład śledzenia ewolucji tematów posłużą badania opublikowane przez He *et al.*(Q. He et al., 2009). W sytuacji w której dysponujemy dwoma tematami w formie rozkładów słów $\mathcal{W}_1$ i $\mathcal{W}_2$, możemy je porównać za pomocą podobieństwa cosinusowego.

$$sim(\mathcal{W}_1, \mathcal{W}_2) = \frac{\mathcal{W}_1 \cdot \mathcal{W}_2}{||\mathcal{W}_1||||\mathcal{W}_2||} \tag{D.9}$$

Na to prawdopodobieństwo nałożona może zostać wartość progowa, za pomocą której odróżnimy zaktualizowaną wersję już istniejącego tematu, temat-dziecko wynikłe z tematu macierzystego, oraz temat niepowiązany. Dla modeli generatywnych, dokładniejszym sposobem porównania dwóch tematów jest policzenie prawdopodobieństwa warunkowego dla $T(\tau_i)$ – zbioru tematów zidentyfikowanych dla zbioru dokumentów D_i w momencie czasu τ_i – warunkowanego $T(\tau_{i-1})$:

$$P(T(\tau_i)|T(\tau_{i-1})) \propto sim(T(\tau_i), T(\tau_{i-1})) \tag{D.10}$$

Opis ten nie jest wyczerpujący, ale reprezentuje podstawową ideę podejścia do śledzenia rozwoju ewolucji tematów w czasie. Bardziej zaawansowane metody są cytowane w przeglądzie literatury. Idea śledzenia zmian w tematach jest natomiast pośrednio związana z ideą śledzenia ich trendów. Trend tematu to seria czasowa wartości reprezentujących jedną lub więcej zmienną opisującą policzalne miary procesu publikacji na tenże temat.

Do zbioru tych miar należeć mogą informacje takie jak liczba dokumentów opublikowanych na dany temat w ciągu roku, czy też liczba autorów którzy te publikacje napisali. Miary te mogą być znormalizowane i przeliczone w odniesieniu do poprzedniego momentu czasu. Przewidywanie rozwoju trendów tematu jest zatem zadaniem zbliżonym do przewidywania liczby cytowań – jest to mapowanie

zbioru zmiennych wejściowych do wyjściowej serii czasowej. Zagadnienie może zatem zostać opisane na zbiorze B o m zmiennych wejściowych b oraz zbiorze U zawierającym n miar wyjściowych u jako:

$$B = \begin{pmatrix} b_1 & \dots & b_m \end{pmatrix} \rightarrow \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix} = U \tag{D.11}$$

O ile podobne do śledzenia rozwoju tematu, trendy skupione są przede wszystkim na ocenie i śledzeniu tematu w sensie popularności i wpływu na sferę publikowanych dokumentów z założeniem, że jesteśmy w stanie identyfikować dany temat w każdej chwili czasu. Miary którymi trendy są mierzone pomóc mogą w zrozumieniu niektórych zjawisk mających wpływ na to, czemu pewne zaszły pewne zmiany w samych tematach. Takiego powiązania doszukiwać się możemy w dziedzinie analizy frontów naukowych (*research fronts*).

Definition 23 *Front naukowy to grono publikacji naukowych reprezentujący pewien obszar nauki. Front naukowy złożony jest z gęsto współcytowanego rdzenia najważniejszych publikacji (core papers), a także z publikacji źródłowych które cytują przynajmniej jedną publikację ze rdzenia* [1].

Trendy frontów naukowych mogą być analizowane poprzez cechy które, w przypadku serii czasowych, definiowałyby popularność i trendy tematu. Nowe fronty pojawiają się z biegiem lat w wyniku procesów takich jak podziały i scalenia istniejących frontów, a nawet pełnoprawne powstanie frontu bez poprzednika – a wraz z tymi procesami zmieniać się będą miary je opisujące. Analiza tych charakterystyk wspomaga naukowców analizujących zmiany w strukturze frontów naukowych, oferując dodatkowy kontekst zmianom, które zaszły na poziomie

[1] https://clarivate.com/webofsciencegroup/essays/research-fronts/

sieciowym.

D.2 Motywacja

Problemy z którymi naukowcy borykają się na co dzień mają naturę praktyczną. Gdzie jest warto publikować? Jakie dziedziny są obiecujące z punktu widzenia przyszłego wzrostu? W których dynamicznie następują odkrycia? Jakie ważne publikacje ukazały się ostatnio w mojej dziedzinie? Te pytania mogą wydawać się proste, ale odpowiedź na nie staje się trudniejsza z każdym rokiem wzrostu jaki obserwujemy w światowej nauce. Odpowiedzi na cytowane pytania wymagają zrozumienia procesów i stanu nauki, która obecnie jest zbyt rozległym zjawiskiem, aby pojedyncza osoba mogła sobie poradzić z tym zadaniem.

Ten wzrost skali światowej nauki jest powodem, dla którego metody i narzędzia zdolne wspomóc naukowców w rozumieniu obecnej sytuacji są konieczne. Codziennie publikowane jest wiele artykułów w dziesiątkach dziedzin i czasopism, każde o zróżnicowanej jakości i wpływie na całokształt nauki. Pojedynczy naukowiec może być w stanie śledzić część z nich, ale bez wsparcia maszyny nie może on analizować zmian w makro-skali, ani zrozumieć dynamiki powstawania i wzrostu tematyk naukowych.

Dodatkowo zaobserwować można trend postępującej kwantyfikacji nauki. Na każdym poziomie, od lokalnego do międzynarodowego, instytucje wprowadzają numeryczne miary i wskaźniki wydajności oraz naukowego sukcesu, których prostym przykładem jest ilość uzyskanych cytowań. Ta sytuacja powoduje konieczność dodatkowego wsparcia przy podejmowaniu decyzji mających wpływ na rozwój ich kariery naukowej.

O ile wiele istniejących metod może zostać zaadoptowanych do rozwiązywania

tego typu problemów, rozwijane są one jako podejścia ogólne. Przez brak dostosowania do konkretnej dziedziny sprawia, że ich skuteczność nie jest tak wysoka jak w przypadku wyspecjalizowanych narzędzi. Dla problemów o złożonych danych, takich jak analiza nauki, konieczne jest kreowanie dedykowanych metod.

W tej pracy nacisk położony został na wykorzystanie głębi danych o nauce oraz ich struktury. Celem tego podejścia jest osiągnięcie lepszych wyników oraz metod dedykowanych do problemów analizy nauki. Osiągnięte jest to poprzez eksplorację metod transformacji danych do formy kompatybilnej z innymi technikami uczenia maszynowego, jednakże bez utraty specyficznych właściwości danych bibliograficznych. O ile zatem metody te prezentowane będą na konkretnych problemach, np. przewidywania cytowań publikacji naukowej, to uwaga skupiona jest na tym jakie dane i w jaki sposób są przygotowane i wykorzystane do tego celu.

Główna motywacja tej pracy – stworzenie metod predykcji na danych bibliograficznych – jest zatem realizowana poprzez zaproponowanie metod transformacji, dekompozycji, analizy oraz inferencji na tychże danych. Narzędzia predykcji i analizy stworzone w ten sposób wykorzystane zostać mogą do wsparcia procesów decyzyjnych naukowców i organizacji naukowych.

D.3 Cele

Głównym celem przedstawionych badań naukowych było stworzenie metod przewidywania na sieciowych danych bibliograficznych. Cel ten został rozłożony na pomniejsze cele związane z metodami transformacji, dekompozycji i predykcji na danych empirycznych. Każda z zaproponowanych metod miała za zadanie zachowanie informacji zawartej w postaci pierwotnej, lecz być łatwo przetwarzalna metodami oraz modelami do analizy i predykcji. Cel ten podyktowany był unikalnymi

własnościami danych naukowych oraz chęcią lepszego ich wykorzystania w transformacji i dekompozycji empirycznych danych naukowych do formatu łatwo przetwarzalnego maszynowo. Założeniem stawianym danym wyjściowym tych procesów jest zachowanie informacji charakterystycznej dla danych bibliograficznych bez utraty na kompatybilności z istniejącymi metodami oraz modelami analizy i predykcji. Cel ten wynika z unikalnej struktury i własności danych bibliograficznych.

Dodatkowym celem rozprawy było zaproponowanie modeli wspierania decyzji w karierach naukowych w ramach klasycznych metod uczenia maszynowego. Motywowane jest to znaczną ilością istniejących danych bibliograficznych. W szczególności, poruszane są następujące problemy:

1. Opracowanie metody do transformacji sieciowych, heterogenicznych danych o publikacjach naukowych do postaci zbioru wewnętrznie spójnych sieci homogenicznych
2. Stworzenie metody automatycznego wykrywania tematów naukowych z sieciowych danych o publikacjach
3. Zaproponowanie metody automatycznej oceny jakości dokumentów zorganizowanych w strukturę sieciową
4. Stworzenie metody przewidywania przyszłej liczby cytowań dla nieopublikowanej jeszcze publikacji naukowej w oparciu o dane sieciowe
5. Opracowanie metody przewidywania przyszłego zainteresowania społeczności naukowej danymi tematami
6. Stworzenie analizy oraz metody mapowania danych o publikacjach do ich sukcesu publikacyjnego wyrażonego w liczbie cytowań

Praca nad wspomnianymi zadaniami i metodami oparta była o empiryczne dane bibliograficzne otrzymane z odpowiednich źródeł. Procesy ewaluacji również

opierały się na danych rzeczywistych. Dodatkowo, część metod ewaluowana była z pomocą ekspertów w dziedzinie w celu lepszej oceny ich działania.

D.4 Wkład w Dziedzinę

Rezultatem dokonanej pracy naukowej opisanej w przedstawionej pracy jest zróżnicowany wkład w dziedzinę, ze szczególnym naciskiem położonym na operowanie danymi bibliometrycznymi. Zaproponowane zostały:

1. Rozszerzenie algorytmu *Loopy Belief Propagation* dla danych w postaci dyskretnych dystrybucji
2. Iteracyjna metod oceny jakości dokumentów naukowych powiązanych w sieć
3. Szczegółowa analiza wpływu różnych relacji cytowania w sieciach publikacji naukowych na identyfikację tematów
4. Metoda dekompozycji i transformacji heterogenicznych sieci danych o publikacjach naukowych na zbiór sieci homogenicznych
5. Metoda przewidywania przyszłych cytowań publikacji w heterogenicznej sieci danych bibliograficznych
6. Prototyp metody rekomendacji miejsca publikacji pracy naukowej na bazie potencjalnych przyszłych cytowań zamiast kierowania się uśrednioną jakością miejsca publikacji