Sieci złożone

Sprawozdanie

Przemysław Barcicki

13 stycznia 2023

1 Wstęp

Celem mojego zadania było zebranie danych na temat pewnej sieci złożonej oraz zbadanie jej pod względem różnych właściwości ze świata teorii grafów. Różnych sieci które charakteryzują się tą złożonością, jest wiele, przykładowo są to sieci współpracy, znajomości, kontaktów seksualnych czy nawet Internet.

Do swoich badań wybrałem połączenia między artykułami na internetowej encyklopedii Wikipedia, gdzie, wierzchołkami w grafie są różne artykuły, a krawędziami (jednokieurnkowymi) są odnośniki i nawiązania między artykułami. Dzięki historii zmian przypisanego do każdej podstrony można dodatkowo zbadać dane artykułu pod względem rozwoju danej dziedziny, czy też rozwoju samego artykułu, w domenie czasowej. Aby nie dochodziło do dużego wybuchu ilości odnośników w swoich badaniach ograniczyłem się tylko do jednej konkretnej kategorii (artykuł *Informatyka* ma 535 odnośników do innych artykułów z czego tylko 9 dotyczy surowej kategorii Informatyka[1])

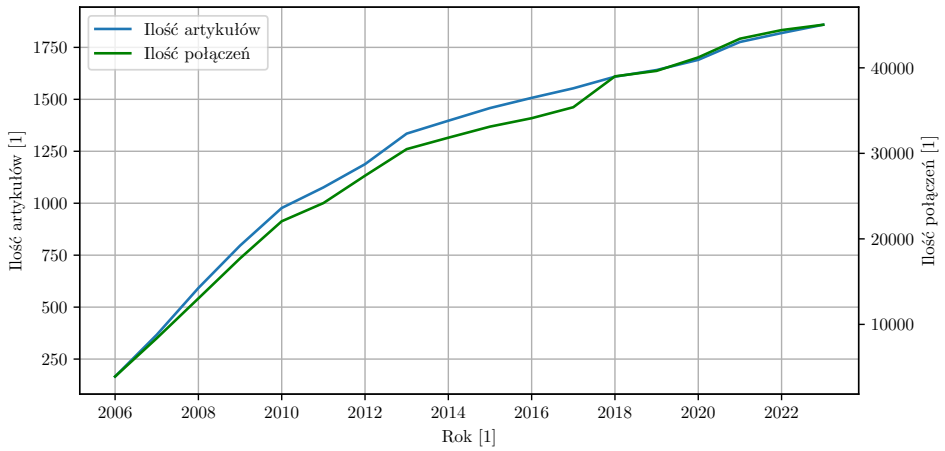
2 Zbieranie danych

Zbieranie danych bazowało na wysyłaniu zapytań do ogólnodostępnego interfejsu programowego wikipedii, gdzie odpowiednio dobierając argumenty można było pytać o konkretne zasoby zawarte na stronie, bez ściągania jej całej, razem z odpowiednim filtrowaniem danych. Tak jak wspomniałem wcześniej, dane filtrowane były na podstawie kategorii. W grafach wyjściowych znajdują się wszystkie artykuły z kategorii oraz dodatkowo artykuł znajdujący się na obrzeżach który jako pierwszy wychodzi z danej kategorii (najczęściej będą to wierchołki ze stopniem 1).

[1]Stan na 03.01.2023

3 Analiza danych

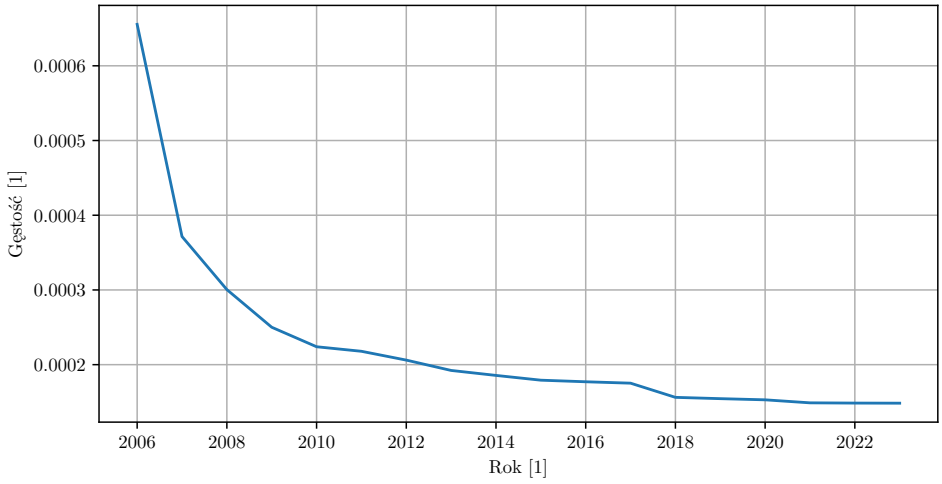
Analizę będę przedstawiał na zbiorze danych z polskiej kategorii informatyka. Najprostszym elementem do analizy jest przegląd jak zmieniała się wielkość kategorii w czasie. Wykres (3) obrazuje tą zmianę w czasie.



Rysunek 1: Wykres przedstawiający wielkość kategorii Informatyka w czasie. Niebieskim kolorem przedstawiono artykuły, zielonym ilość połączeń między artykułami.

Można zauważyć, że ilość połączeń między akrtykułami jest powiązana liniowo z ilością artykułów, sugeruje to że, artykuły dodawane do kategorii, nie nawiązują do dużej ilości innych artykułów w tej samej kategorii.

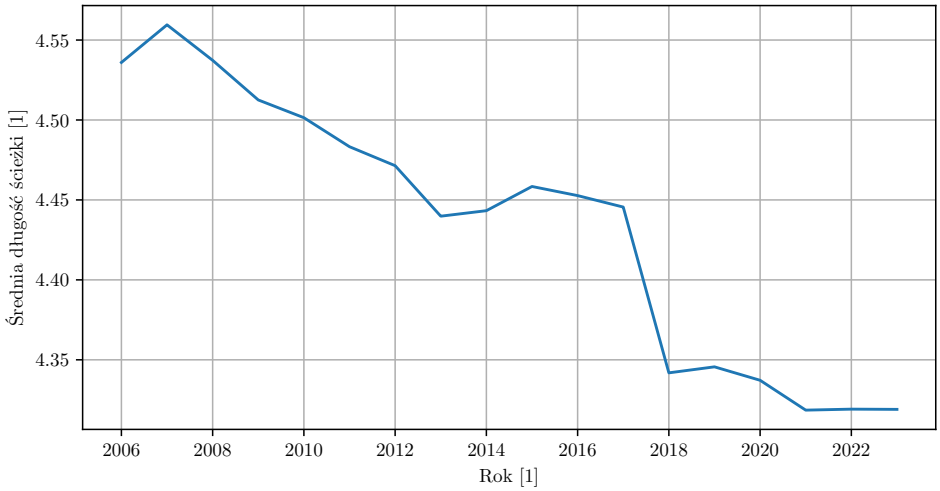
Kolejno można przedstawić, jak zmieniała się gęstość grafu na przestrzeni czasu. Sama wielkość opisuje stosunek między ilością połączeń w grafie a ilością możliwych połączeń.



Rysunek 2: Wykres przedstawiający gęstość grafu opisującego kategorię Informatyka w czasie.

Na wykresie (2) widać, że wraz z rozwojem informatyki na przestrzeni ostatnich 15 lat, graf się rozrzadził. Jest to logiczne, wraz z powstawaniem nowych podkategorii, zasięg informatyki zwiększa się, lecz niekoniecznie są one mocno powiązane z istniejącymi już kategoriami.

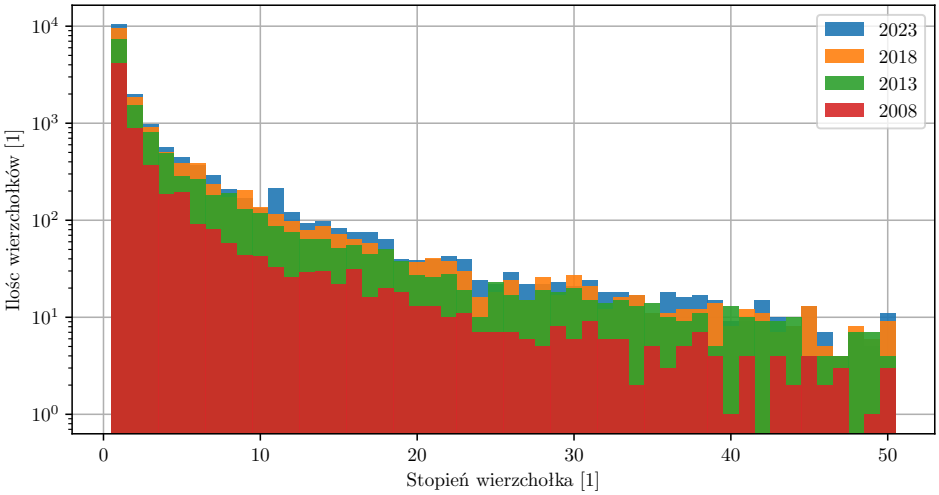
Aby zbadać średnią długość (najkrótszej) ścieżki musimy skorzystać z pewnych uproszczeń, w których dla każdego grafu wybieramy największy słabo spójny komponent. Tym procesem nie tracimy dużo informacji z wykluczonych wierzchołków, ponieważ w każdym roku, wielkość największego komponentu stanowi przynajmniej 99.5% wielkości całego grafu.



Rysunek 3: Wykres przedstawiający średnią długość najkrótszych ścieżek między wierzchołkami.

Spadająca wartość średniej długości ścieżki na wykresie (3) może sugerować konsolidację poprzez powstanie tematów, kategorii czy artykułów które spinają ze sobą dalsze witryny, ale także, co widać po wykresie ilości wszystkich połączeń, że istniejące już tematy, zostały uzupełnione o odpowiednie odnośniki do istniejących wierzchołków.

Kolejną istotną charakterystyką sieci połączeń jest rozkład stopni wierzchołków. Wielkość ta, może nam sugerować jak dobrze połączone są ze sobą konkretne tematy w konkretnej kategorii. Histogram tego rozkładu przedstawia rysunek (4), a tabela (1) zawiera informację o maksymalnym i średnim wierzchołku.



Rysunek 4: Wykres przedstawiający rozkład stopni wierzchołków w grafie dla różnych lat.

Tabela 1: Tabela przedstawiająca zmianę maksymalnej i średniej wartości stopnia wierzchołka na przestrzeni lat.

Rok	Max	Średnia	Artykuł
2008	301	3.943	Pulpit
2011	311	4.551	Język angielski
2014	390	4.81	Język angielski
2017	423	4.931	Język angielski
2020	847	4.972	Skróty używane w informatyce
2023	847	5.115	Skróty używane w informatyce

Pewną dodatkową charakterystyką opisującą ważność danego wierzchołka jest jak wiele najkrótszych ścieżek przechodzi przez dany element w grafie, ta charakterystyka opisuje nam pewien najważniejszy centraly przegub którego usunięcie zmieniłoby strukturę całej sieci. W tabeli (2) zawarto informację o ilości ścieżek przechodzących przez jaki wierzchołek.

Tabela 2: Tabela przedstawiająca wierzchołki z największą ilością ścieżek na przestrzeni lat.

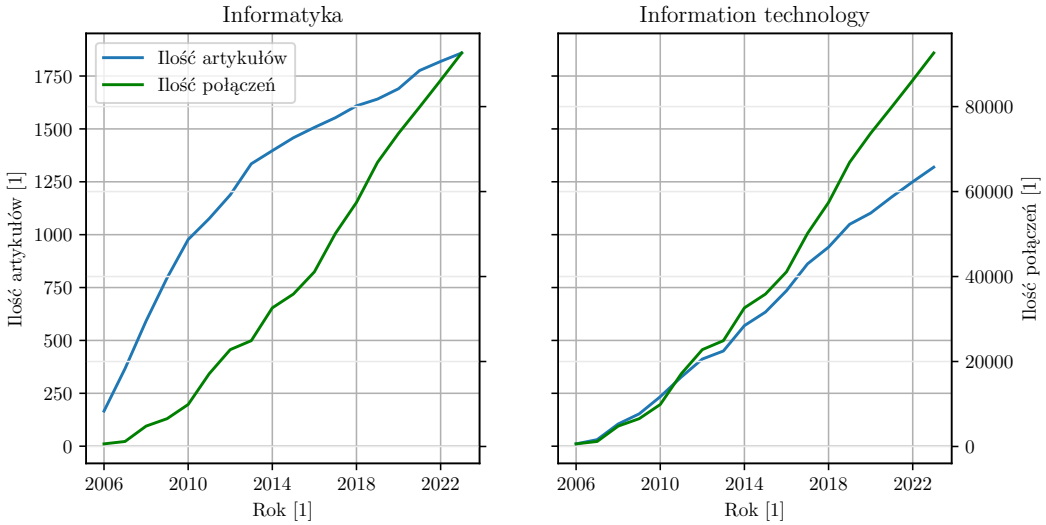
Rok	Artykuł	Ścieżki
2008	Moc obliczeniowa	6.12%
2011	Program komputerowy	2.8%
2014	Program komputerowy	2.08%
2017	Program komputerowy	1.71%
2020	Informatyka	2.78%
2023	Informatyka	2.6%

4 Wnioski

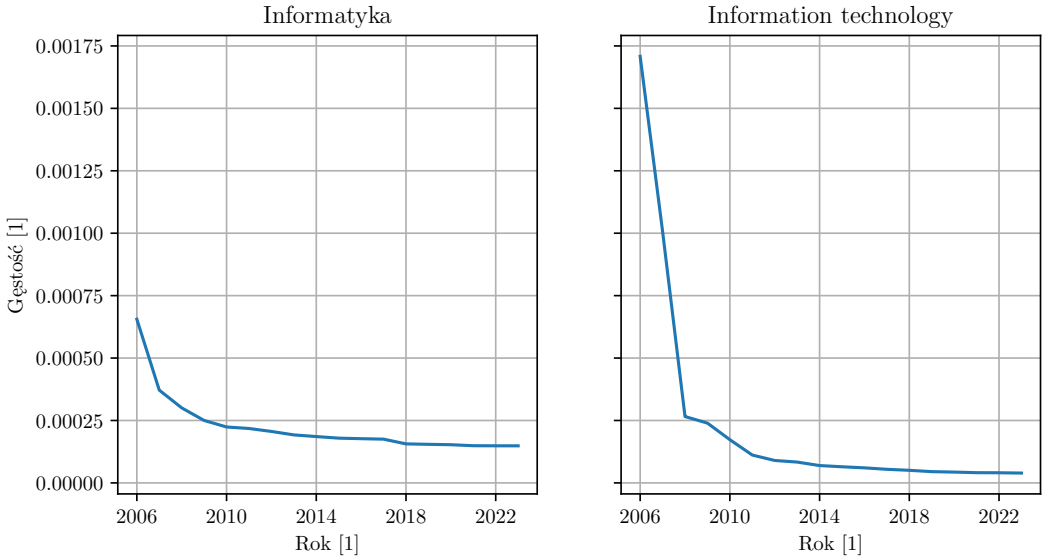
Na przedstawionych wykresach, widać rozwój kategorii Informatyka, razem z rozwojem samej Wikipedii. Wraz z czasem pojawiło się tam kilkukrotnie więcej artykułów i połączeń między nimi, co można powiązać w dużej mierze z rozwojem informatyki, ale także większą dostępnością internetu dla ludzi w całym kraju. Im więcej ludzi z niej korzysta, tym więcej osób się znajdzie, którzy (za darmo) poświęcą swój czas i wiedzę dla rozwoju ogółu.

Dodatek A Angoljęzyczne porównanie

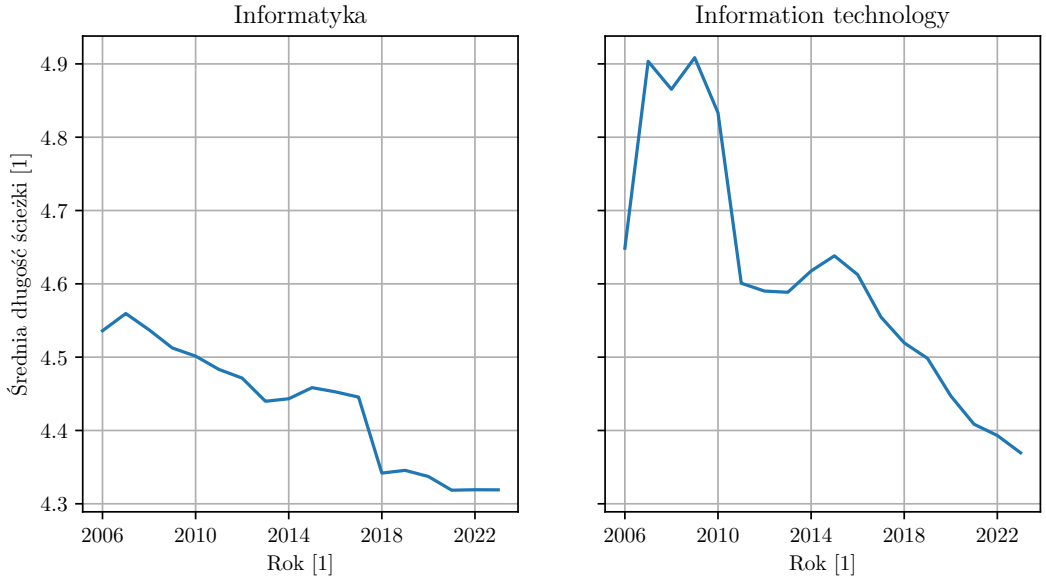
Poniższe rysunki (5, 6, 7, 8) oraz tabele (4, 3) przedstawiają porównanie polskiej oraz anglojęzycznej wersji kategorii Informatyka na stronie Wikipedia.



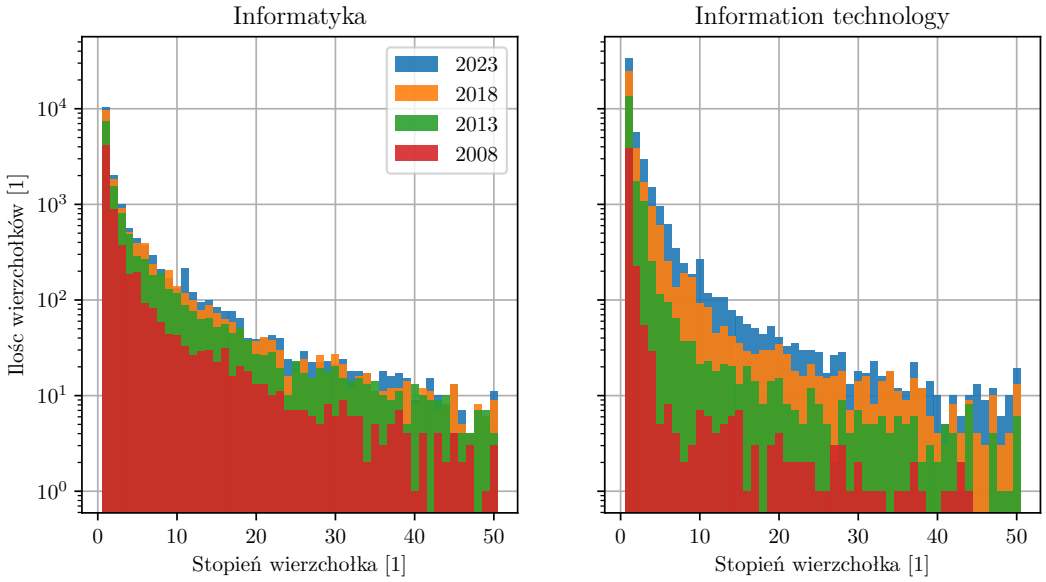
Rysunek 5: Wykres przedstawiający wielkość kategorii Informatyka w czasie. Niebieskim kolorem przedstawiono artykuły, zielonym ilość połączeń między artykułami.



Rysunek 6: Wykres przedstawiający gęstość grafu opisującego kategorię Informatyka w czasie.



Rysunek 7: Wykres przedstawiający średnią długość najkrótszych ścieżek między wierzchołkami.



Rysunek 8: Wykres przedstawiający rozkład stopni wierzchołków w grafie dla różnych lat.

Tabela 3: Tabela przedstawiająca zmianę maksymalnej i średniej wartości stopnia wierzchołka na przestrzeni lat.

Rok	Max	Średnia	Artykuł
2008	626	2.255	Information industry
2011	626	2.761	Information industry
2014	627	3.003	Information industry
2017	741	3.29	Linus Torvalds
2020	899	3.534	Digital media use and mental health
2023	998	3.781	Silicon Valley

Tabela 4: Tabela przedstawiająca wierzchołki z największą ilością ścieżek na przestrzeni lat.

Rok	Artykuł	Ścieżki
2008	Information industry	5.75%
2011	Real-time business intelligence	1.94%
2014	Information society	6.85%
2017	Metadata	7.85%
2020	Business intelligence	4.52%
2023	Information technology	3.16%

Dodatek B Kod źródłowy

Cały kod źródłowy, tego sprawozdania, jak i kodu użytego do zbierania i obróki danych znajduje się w repozytorium Github.
`https://github.com/mlodybercik/sieci-zlozone-wikipedia`