TASS – Techniki Analizy Sieci Społecznych

Koncepcja

Analiza porównawcza sieci współpublikowania   
z siecią autorów patentów

Aleksander Łosiewicz

Daniel Petrykowski

15.12.2019

1. Wstęp

Założeniem projektu było zbadane podobieństwo sieci badaczy wygenerowanej na zasadzie wspólnego patentowania z siecią badaczy wygenerowaną na zasadzie współpublikowania.   
W raporcie tym chcielibyśmy skupić się na przedstawieniu zarówno różnic jak i podobieństw pomiędzy obiema sieciami.

1. Dane
   1. Sieć badaczy wygenerowana na zasadzie wspólnego patentowania

Przez prowadzących zostały dostarczone dane dla pewnej grupy patentów w postaci plików xml. Z nich została odczytany numer id patentu za pomocą stworzonego skryptu w języku python. Kolejnym krokiem było na podstawie odczytanych numerów id pozyskanie danych o właścicielach patentów, w tym celu wykorzystaliśmy biblioteke google\_patent\_scraper i pobraliśmy potrzebne dane ze strony Google Patents (<https://patents.google.com/>).

Próbowaliśmy pobrać dane na temat 10101 patentów, z czego z sukcesem uzyskaliśmy informacje dla 10000 patentów. Co daje skuteczność powyższej operacji na poziomie 99%.

Ostatecznie w ten sposób uzyskaliśmy informacje ok. 16 000 osobach które są autorami lub współautorami patenów.

* 1. Sieć badaczy wygenerowana na zasadzie współpublikowania

Dla porównania sieci wynikającej z wspólnego patentowania i publikowania również potrzebna była informacja o autorach publikacji. Dane na ten zamierzaliśmy pobrać ze strony Google Scholar (<https://scholar.google.com/>). Jednak, że w tym przypadku natrafiliśmy na znaczne utrudnienia, ponieważ gogle nie udostępnia żadnych mechanizmów pozwalający sprawnie pobierać informacje na temat publikacji, a w szczególności ich autorów. Ze znalezionych przez nas dostępnych rozwiązań, które starliśmy się wykorzystać można wyróżnić dwa:

1. scholar (<http://www.icir.org/christian/scholar.html>) – parser dla Google Scholar, napisany w Python. Pozwalał on znaleźć publikacje napisane przez danego naukowca, jednakże nie było już możliwości wyświetlenia informacji na temat autorów poszczególnych publikacji, a same linki do publikacji prowadziły do różnych stron z różnym formatowanie co uniemożliwiło wykorzystanie techniki web scrapingu.
2. scholarly 0.2.5 (<https://pypi.org/project/scholarly/>) – biblioteka Python umożliwiająca ściągnięcie informacji na temat autorów publikacji z Google Scholar (ilości cytowań, przynależności do danej organizacji, id konta w Google Scholar, obszarze badań). Wadą tego rozwiązania podobnie jak poprzednio była nie możliwość wyświetlenia informacji na temat autorów z jakim dana publikacja była współpublikowana. Dodatkowo rozwiązanie to ogranicza się jedynie do osób, które założyły konto na Google Scholar, co sprawia, że nie jesteśmy w stanie przy wykorzystaniu tego narzędzia uzyskać informacje o innych naukowcach.

Ostatecznie z racji braku dostępnych lepszych rozwiązań, zdecydowano się na wykorzystanie biblioteki scholarly. Przy jej wykorzystaniu pobrano id konta naukowca, aby w kolejnym kroku przy jego wykorzystaniu stworzono url do strony profilowej naukowca w Google Scholar na której są wyświetlone wszystkie jego publikacje. Strona to została pobrana i poddana scrapingowi w celu uzyskania informacji z kim współpracował dany naukowca przy swoich publikacjach.

Jednak technika ta umożliwiła na wysłanie jedynie ok. 300 zapytań do Google Scholar, gdyż przy większej ilości zapytań ruch sieciowy był blokowany.

Ostatecznie z przewidzianych do ściągnięcia informacji o 16000 osobach udało nam się uzyskać informacje o zaledwie 469 z czego przy wysłaniu 2294 zapytań do serwera, aż 1825 zakończyło się niepowodzeniem z powodu braku konta na portalu Google Scholar.

1. Czyszczenie danych

Dane pobrane na temat autorów patentów z racji, że mają przewidziany z góry odpowiedni format nie wymagały żadnych specjalnych zabiegów.

Natomiast dane na temat współpublikowania z racji tego, że były uzyskiwane z portalu Google Scholar to charakteryzowały się dużym zanieczyszczeniem (przykład poniżej).

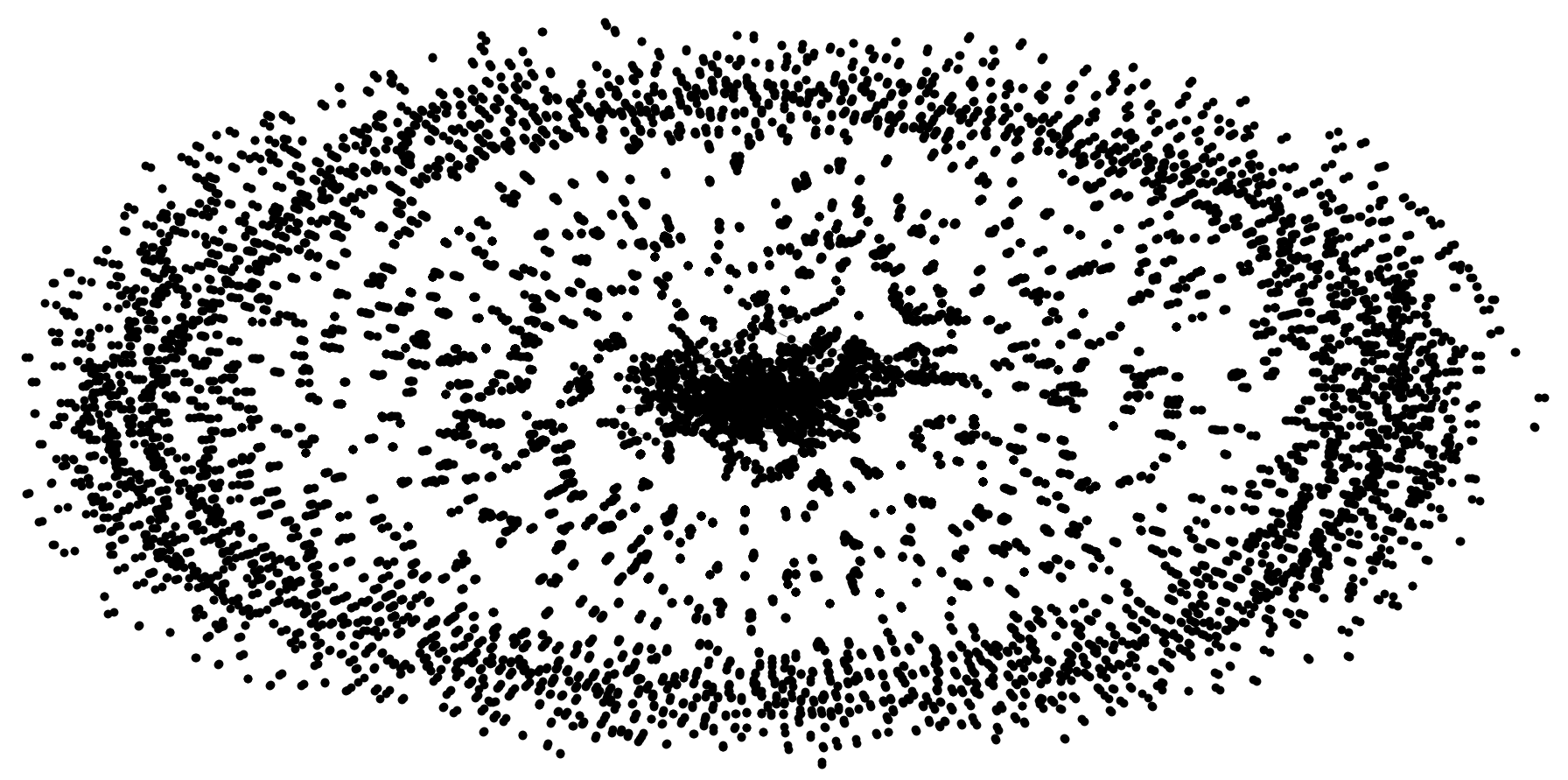
Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

W celu zminimalizowaniu zanieczyszczenia w danych został stworzony skrypt który wskazywał podejrzane miejsca, a następnie dane były czyszczone ręcznie. Co było możliwe za sprawą ich niewielkiej ilości.

1. Analiza danych

**Sieć badaczy wygenerowana na zasadzie wspólnego patentowania:**



Rysunek 1 Wizualizacja sieci badaczy wygenerowana na zasadzie wspólnego patentowania

Jak widać z powyższej wizualizacji już na pierwszy rzut oka można wyróżnić osoby, które są słabo powiązani pomiędzy sobą i współpatentowali w małych dwu osobowych grupach. Takich osób jest 2342. Jak również osoby które są znaczenie bardziej skłonne do współpracy i na powyższej wizualizacji znajdują się w centrum.



Rysunek 2 Wizualizacja centralnej części sieci badaczy wygenerowana na zasadzie wspólnego patentowania

Podstawowe właściwości badanej sieci:

|  |  |
| --- | --- |
| Rząd sieci | 14165 |
| Rozmiar sieci | 34170 |
| Przechodniość | 0.7199811495221332 |
| Gęstość | 0.00034062182972237977 |
| Największy możliwy rząd rdzenia | 28 (w grafie występuje takich 29) |
| Uśredniony współczynnik gronowania | 0.756153873397604 |
| Rząd największej składowej spójnej | 2143 |
| Rozmiar największej składowe spójnej | 8523 |
| Średnia najkrótsza ścieżka dla składowej spójnej | 8.749668976316611 |

Liczba skrajnych stopnie wierzchołków:

|  |  |
| --- | --- |
| Stopień wierzchołka | Liczba wierzchołków o danym stopniu |
| 1 | 2342 |
| 2 | 2584 |
| 3 | 2046 |
| 4 | 1690 |
| 5 | 1367 |
| 6 | 959 |
| ……. | |
| 57 | 1 (Andrzej Krueger) |
| 59 | 1 (Zbigniew Śmieszek) |
| 68 | 1 (Wojciech Balcerowiak) |
| 88 | 1 (Juliusz Pernak) |
| 129 | 1 (Renata Fiszer) |

Histogram przedstawiający rozkład stopni wierzchołków:

Obraz zawierający zrzut ekranu

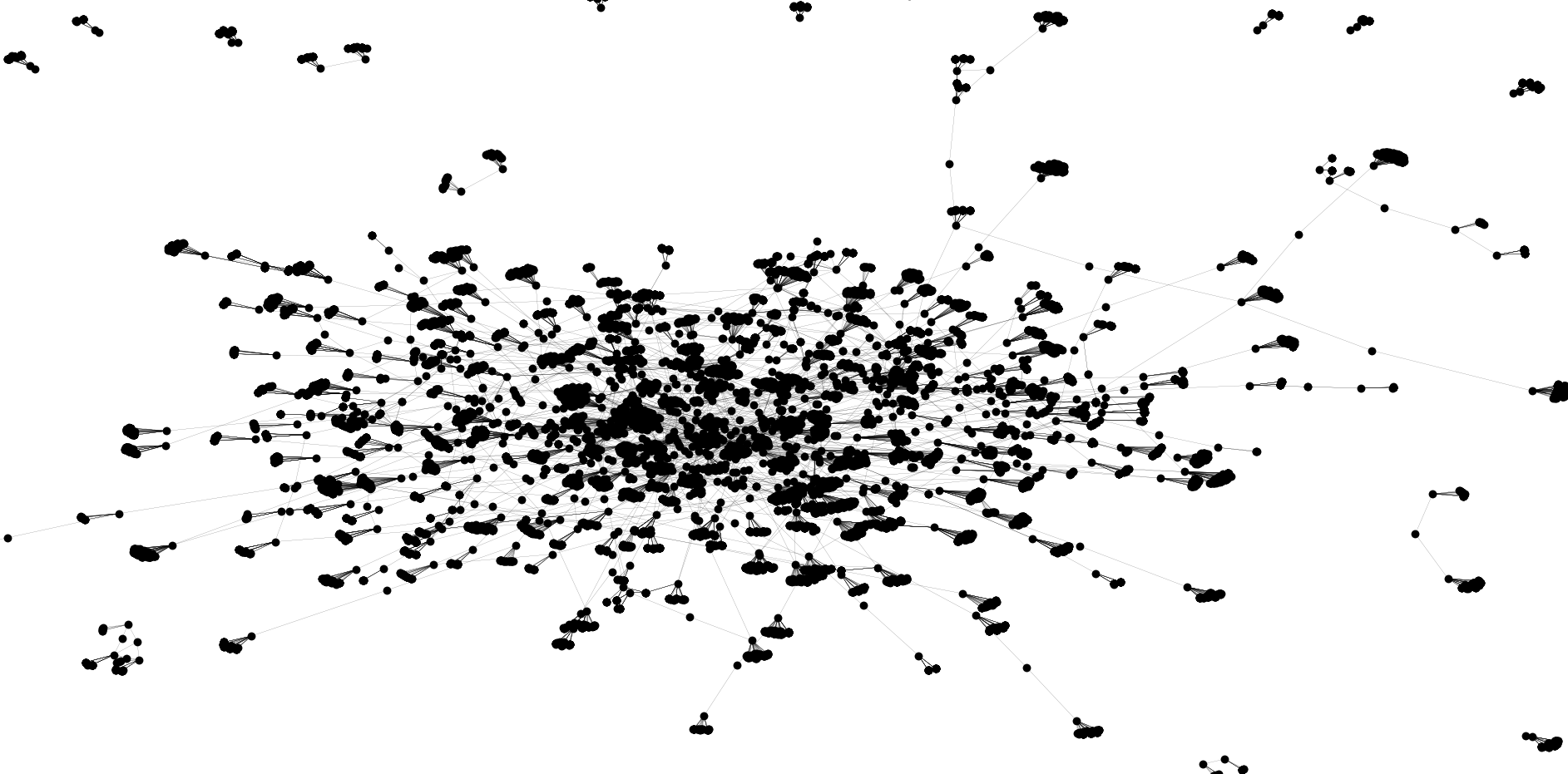
Opis wygenerowany automatycznie

**Sieć badaczy wygenerowana na zasadzie współpublikowania:**

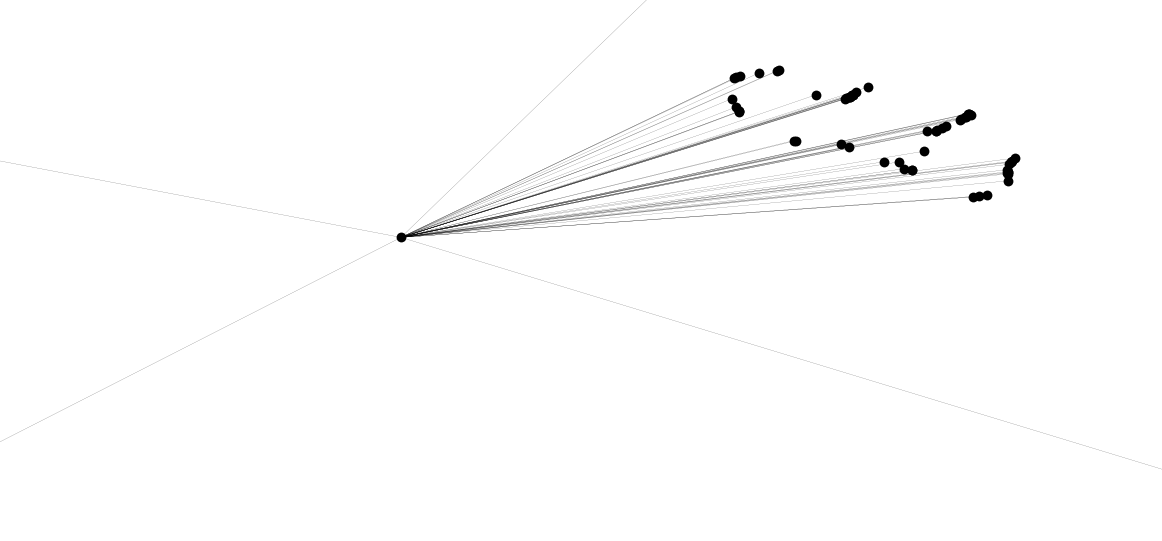


Rysunek 3 Wizualizacja sieci badaczy wygenerowana na zasadzie współpublikowania

Podobnie jak w przypadku wizualizacji dla sieci wygenerowanej na zasadzie wspólnego patentowania także i w przypadku sieci wygenerowanej na zasadzie współpublikowania można zauważyć, że istnieje grupka osób która współpracuje jedynie w swoim wąskim gronie. Warto jednak zauważyć, że jest takich osób znacznie więcej oraz, że mimo wszystko grupki są znacznie większe, co sugeruje, że przy większej ilości danych mogło by się okazać, że osoby te znacznie bardziej współpracują z pozostałymi badaczami niż jest widoczne w tym wypadku.



Rysunek 4 Wizualizacja centralne części sieci badaczy wygenerowana na zasadzie współpublikowania



Rysunek 5 Wizualizacjia grafu - pojedynczy autor publikacji wraz z osobami współpublikującymi z nim

Podstawowe właściwości badanej sieci:

|  |  |
| --- | --- |
| Rząd sieci | 13275 |
| Rozmiar sieci | 14318 |
| Przechodniość | 0.009135189371808876 |
| Gęstość | 0.00016250847344127697 |
| Największy możliwy rząd rdzenia | 7 (w grafie występuje takich 9) |
| Uśredniony współczynnik gronowania | 0.037232854182817215 |
| Rząd największej składowej spójnej | 10702 |
| Rozmiar największej składowe spójnej | 11828 |
| Średnia najkrótsza ścieżka dla składowej spójnej | 9.042866275716804 |

Liczba skrajnych stopni wierzchołków:

|  |  |  |
| --- | --- | --- |
| Stopień wierzchołka | Liczba wierzchołków o danym stopniu | Stopień wierzchołka dla wygenerowanej na zasadzie współpatentowania |
| 0 | 3 | Nie dotyczy |
| 1 | 11887 |
| 2 | 711 |
| 3 | 165 |
| 4 | 38 |
| 5 | 21 |
| ………… | | |
| 86 | 2 | Nie dotyczy |
| 87 | 1 (M Li) | 15 |
| 88 | 1 (Y Liu) | 4 |
| 89 | 1 (J Debski) | 4 |
| 95 | 1 (J Wu) | 15 |
| 96 | 1 (G Giannini) | 6 |

Histogram przedstawiający rozkład stopni wierzchołków:

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

1. Wnioski

Z danych uzyskanych podczas fazy analizy można wysnuć wnioski, że współpraca jeżeli chodzi o współpublikowanie jest znacznie silniejsza niż współpatentowanie. Jest to zapewne spowodowane tym, że znacznie częściej jest tworzona nowa publikacja niż publikowane nowe rozwiązanie, które wymaga znacznie większego wysiłku i większej ilości czasu. Fakt ten potwierdza porównanie rzędu i rozmiaru największej składowej spójnej, gdyż jest ona znacznie większa dla sieci współpublikowania.

Bardzo niski jest natomiast współczynnik gęstości dla obu sieci co oznacza, że bardzo rzadko tworzą się kliki w której każdy współpatentuje lub publikuje z każdym. Zastanawiać może natomiast zaskakująco niski współczynnik gronowania dla sieci wspólpublikowania, wynika to zapewne z faktu, że w tym przypadku istnieje bardzo dużo wierzchołków stopnia pierwszego (jak pokazuje rysunek 5). Jest to spowodowane faktem, że podczas fazy zbierania danych na temat osób współpublikujących z daną wyjściową osobą, dla której takich informacji szukaliśmy. Dane te nie były filtrowane, czy znalezione osoby znajdują się również w zbiorze osób współpatentujących. Taki filtr mógłby pozwolić na dokładniejsze zbadanie czy istnie korelacja mówiąca o tym, że osoby wpółpatentujące również razem współpublikują.